

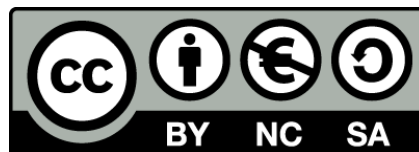


UNIVERSITAT DE  
BARCELONA

# Exactitud de los criterios simplificados para el diagnóstico de la hepatitis autoinmune en población pediátrica

## Una nueva propuesta basada en la clasificación ESPGHAN/NASPGHAN 2009

José Vicente Arcos Machancoses



Aquesta tesi doctoral està subjecta a la llicència **Reconeixement- NoComercial – CompartirIgual 4.0. Espanya de Creative Commons.**

Esta tesis doctoral está sujeta a la licencia **Reconocimiento - NoComercial – CompartirIgual 4.0. España de Creative Commons.**

This doctoral thesis is licensed under the **Creative Commons Attribution-NonCommercial-ShareAlike 4.0. Spain License.**

**UNIVERSITAT DE BARCELONA**

PROGRAMA DE DOCTORADO: *MEDICINA I RECERCA TRANSLACIONAL*

Exactitud de los criterios simplificados  
para el diagnóstico de la hepatitis  
autoinmune en población pediátrica

---

Una nueva propuesta basada en la  
clasificación ESPGHAN/NASPGHAN 2009

TESIS DOCTORAL

**José Vicente Arcos Machancoses**

Barcelona, 2018



UNIVERSITAT DE BARCELONA



**EXACTITUD DE LOS CRITERIOS SIMPLIFICADOS  
PARA EL DIAGNÓSTICO DE LA HEPATITIS  
AUTOINMUNE EN POBLACIÓN PEDIÁTRICA**

---

**UNA NUEVA PROPUESTA BASADA EN LA  
CLASIFICACIÓN ESPGHAN/NASPGHAN 2009**

LÍNEA DE INVESTIGACIÓN: *FISIOPATOLOGIA DE LES MALATIES PEDIÀTRIQUES*

Tutor y director  
**Javier Martín de Carpi**

Sección de Gastroenterología, Hepatología y Nutrición. Hospital Sant Joan de Déu, Barcelona

Codirector  
**Vicent Modesto i Alapont**

Servicio de Pediatría. Hospital Universitari i Politècnic La Fe, València



Unidad Integral de Hepatología Compleja y Trasplante Hepático  
Hospital Sant Joan de Déu – Hospital Vall d'Hebron



*“The greatest challenge to any thinker is stating the problem in a way that will allow a solution”.*

**Bertrand Russell**

*“We can only see a short distance ahead, but we can see plenty there that needs to be done”.*

**Alan Turing**



## Agradecimientos

---

A Cristina Molera y Javier Martin de Carpi, por haber creído desde el principio en este proyecto y por su ayuda durante el desarrollo del trabajo, especialmente por sus consejos, estímulo y la revisión crítica de los sucesivos borradores y manuscritos.

A Vicent Modesto, por la dirección de la tesis en el sentido más amplio posible: por inspirar con su ejemplo el despliegue de un trabajo riguroso desde la humildad y la honestidad que, en esencia, es la base del Método. Sin dejar de lado el reconocimiento a su entusiasmo indestructible y, también, por sus enseñanzas en mis inicios en la Pediatría.

A Ecaterina Julio, Victoria Bovo y Vanessa Crujeiras, por su fundamental e imprescindible colaboración en la recogida de datos.

A los colegas de Hepatología Pediátrica del Hospital Vall d'Hebron, por haber permitido y alentado que me aprovechara de su excelente trabajo en la asistencia a los problemas del hígado de tantos niños y jóvenes.

A todos los compañeros de la Sección de Gastroenterología, Hepatología y Nutrición del Hospital Sant Joan de Déu, a los que nunca estaré suficientemente agradecido por integrarme, de una forma fabulosamente natural, en un equipo de tanta calidad científica y humana. Sin la oportunidad que me brindaron, esta tesis no hubiera podido ser posible.

A Rebeca, mi mujer, por su apoyo incondicional, comprensión y paciencia durante todo el tiempo que le he robado a ella y al resto de mi familia para la realización de esta tesis.

A los pacientes y sus familias, por hacerme ver que la alegría y la determinación por seguir adelante no debe abandonarse nunca, ni tan siquiera en los momentos más adversos de la vida.





# índice



<b>1. Abreviaturas y siglas empleadas</b>	<b>19</b>
<b>2. Listado de figuras</b>	<b>25</b>
<b>3. Listado de tablas</b>	<b>35</b>
<b>4. Introducción</b>	<b>43</b>
<b>4.1. El diagnóstico en Medicina basado en criterios de clasificación</b>	<b>45</b>
4.1.1. La esencia real y la esencia nominal de las enfermedades	46
4.1.2. Controversia en la distinción entre criterios diagnósticos y de clasificación	48
4.1.3. Desarrollo de criterios de clasificación por modelización a través de técnicas de regresión con intención predictiva	51
4.1.3.1. Transformación en un sistema de puntos de un modelo de regresión para predecir un diagnóstico	54
4.1.3.1.1. Estimación de los parámetros del modelo multivariable	55
4.1.3.1.2. Organización de los factores de riesgo en categorías y establecimiento de los valores de referencia	55
4.1.3.1.3. Definición de las características de la referencia para cada variable pronóstica o factor de riesgo	56
4.1.3.1.4. Establecimiento de la distancia de cada categoría respecto a la de referencia en unidades de regresión	56
4.1.3.1.5. Establecimiento del multiplicador fijo o constante B	57
4.1.3.1.6. Determinar el número de puntos de cada categoría de los factores de riesgo o variables pronósticas	57
4.1.3.1.7. Estimar los riesgos asociados a la puntuación total	57
<b>4.2. La enfermedad hepática autoinmune en pediatría</b>	<b>59</b>
4.2.1. Definiciones	59

4.2.2.	Hepatitis autoinmune	59
4.2.2.1.	Epidemiología	60
4.2.2.2.	Presentación clínica e historia natural	63
4.2.2.2.1.	Exploración clínica	65
4.2.2.2.2.	Hepatitis autoinmune y embarazo	67
4.2.2.2.3.	Enfermedades autoinmunes asociadas	67
4.2.2.2.4.	Desarrollo de carcinoma hepatocelular	67
4.2.2.3.	Hallazgos de laboratorio	68
4.2.2.4.	Autoanticuerpos diagnósticos	69
4.2.2.4.1.	Anticuerpos anti-nucleares	70
4.2.2.4.2.	Anticuerpos anti-músculo liso	72
4.2.2.4.3.	Anticuerpos anti-microsomales de hígado/riñón de tipo 1	74
4.2.2.4.4.	Variantes de los anticuerpos frente a microsoma hepático	76
4.2.2.4.5.	Anticuerpos anti-citosol hepático de tipo 1	76
4.2.2.4.6.	Anticuerpos anti-antígeno soluble hepático/anti-hígado-páncreas	77
4.2.2.4.7.	Anticuerpos anti-citoplasma de los neutrófilos	78
4.2.2.4.8.	Anticuerpos anti-receptor de asialoglicoproteínas	79
4.2.2.5.	Histopatología	79
4.2.2.6.	Genética y mecanismos inmunológicos	85
4.2.2.7.	Tratamiento	92
4.2.2.7.1.	Definición de remisión y recaída	92
4.2.2.7.2.	Indicaciones del tratamiento	93
4.2.2.7.3.	Posibilidades terapéuticas	93
4.2.2.7.3.1.	Tratamiento estándar	93
4.2.2.7.3.2.	Tratamientos de segundo escalón	96
4.2.2.7.3.3.	Tratamiento de los casos refractarios	96
4.2.3.	Colangitis esclerosante autoinmune	97
4.2.3.1.	Definición y presentación clínica	97
4.2.3.2.	Diferencias con la hepatitis autoinmune	98
4.2.3.3.	Diferencias con la colangitis esclerosante primaria	99

<b>4.3. Diagnóstico de la hepatitis autoinmune por criterios de clasificación</b>	<b>101</b>
4.3.1. Los criterios clásicos revisados de 1999	101
4.3.2. Los criterios simplificados de 2008	104
4.3.2.1. Desarrollo y validación inicial del sistema simplificado	106
4.3.3. Aplicabilidad de los criterios de clasificación por puntos para la hepatitis autoinmune en población pediátrica	109
4.3.3.1. El caso de las hepatopatías colestásicas	112
4.3.3.2. El caso del fallo hepático fulminante	112
4.3.3.3. El ítem de los autoanticuerpos	112
4.3.3.4. El ítem de la anatomía patológica	113
4.3.4. Los criterios diagnósticos pediátricos propuestos por la ESPGHAN y la NASPGHAN (2009)	115
<b>5. Hipótesis de trabajo</b>	<b>119</b>
<b>6. Objetivos</b>	<b>123</b>
6.1. Objetivo general	125
6.2. Objetivos específicos	126
<b>7. Justificación de la unidad temática</b>	<b>127</b>
<b>8. Diseño</b>	<b>131</b>
8.1. Diseño general, pacientes y material	133
8.1.1. Tipo de estudio	133
8.1.2. Ámbito institucional	133
8.1.3. Muestreo	133
8.1.4. Población	134
8.1.4.1. Criterios de inclusión	134
8.1.4.2. Criterios de exclusión	135
8.1.5. Cálculo del tamaño muestral	135

8.1.5.1.	Estimación del número de pacientes para la elaboración del modelo predictivo de regresión logística _____	135
8.1.5.1.1.	Fase de elaboración o desarrollo _____	136
8.1.5.1.2.	Fase de validación _____	136
8.1.5.2.	Estimación del número de pacientes para el estudio de la validez de los criterios diagnósticos _____	136
8.1.5.3.	Decisión final sobre el tamaño muestral _____	138
8.1.6.	Aplicación de las pruebas a validar _____	139
8.1.7.	Aplicación del patrón de referencia _____	140
8.1.8.	Recogida de otras variables _____	142
8.1.9.	Técnicas de medición y extracción de los datos _____	143
<b>8.2.</b>	<b>Aspectos éticos _____</b>	<b>144</b>
<b>9.</b>	<b><i>Metodología y resultados</i> _____</b>	<b>145</b>
<b>9.1.</b>	<b>Capítulo 1: Estimación de la prevalencia de la HAI y evaluación de la validez de los criterios simplificados de 2008 _____</b>	<b>147</b>
9.1.1.	Metodología específica _____	147
9.1.2.	Análisis estadístico _____	151
9.1.3.	Resultados _____	153
9.1.3.1.	Flujo de pacientes y descripción de las categorías diagnósticas _____	153
9.1.3.2.	Indicadores de validez interna _____	164
9.1.3.2.1.	Indicadores independientes de la prevalencia de la enfermedad _____	164
9.1.3.2.2.	Indicadores sensibles a cambios en la prevalencia de la enfermedad y naturaleza de su relación _____	166
9.1.3.2.3.	Número necesario de pacientes para diagnosticar correcta e incorrectamente a uno _____	171
9.1.3.3.	Cálculo del punto de corte óptimo y del poder discriminante global _____	173
<b>9.2.</b>	<b>Capítulo 2: Revisión sistemática y meta-análisis de estudios sobre la validez de los criterios simplificados de 2008 _____</b>	<b>177</b>

9.2.1.	Metodología específica _____	177
9.2.1.1.	Definición de la pregunta diagnóstica de interés _____	177
9.2.1.2.	Búsqueda en fuentes bibliográficas y selección de estudios _____	178
9.2.1.2.1.	Medline _____	180
9.2.1.2.2.	Embase _____	181
9.2.1.2.3.	Trip Database _____	181
9.2.1.2.4.	Web of Science _____	181
9.2.1.2.5.	Biblioteca Virtual en Salud _____	182
9.2.1.3.	Extracción y presentación de los datos de los estudios primarios _____	182
9.2.1.4.	Riesgo de errores sistemáticos _____	182
9.2.2.	Análisis estadístico _____	183
9.2.3.	Resultados _____	186
9.2.3.1.	Características de los estudios recuperados e incluidos _____	186
9.2.3.2.	Evaluación de la calidad de los estudios primarios _____	188
9.2.3.3.	Exactitud diagnóstica y exploración de la heterogeneidad _____	192
9.2.3.4.	Estudio de la presencia de sesgo de publicación _____	199
<b>9.3.</b>	<b>Capítulo 3: Fiabilidad de los criterios simplificados de 2008 _____</b>	<b>203</b>
9.3.1.	Metodología específica _____	203
9.3.2.	Análisis estadístico _____	204
9.3.3.	Resultados _____	206
<b>9.4.</b>	<b>Capítulo 4: Modelización y validación de un nuevo sistema diagnóstico por puntos a partir de los criterios ESPGHAN/NASPGHAN 2009 _____</b>	<b>213</b>
9.4.1.	Metodología específica _____	213
9.4.1.1.	Selección de las submuestras de elaboración y validación _____	216
9.4.2.	Análisis estadístico _____	217
9.4.2.1.	Desarrollo de un modelo predictivo de regresión logística para el diagnóstico de la HAI a partir de los criterios de 2009 _____	217
9.4.2.1.1.	Manejo de los valores perdidos _____	217



9.4.2.1.2.	Selección del mejor modelo a partir de todas las ecuaciones	219
9.4.2.1.3.	Diagnósticos del mejor modelo seleccionado	221
9.4.2.1.3.1.	Detección de valores alejados que afecten a las estimaciones	221
9.4.2.1.3.2.	Comprobación del supuesto de equidispersión	221
9.4.2.1.3.3.	Bondad de ajuste	223
9.4.2.1.3.4.	Significación global del modelo	223
9.4.2.1.3.5.	Calibración del modelo	223
9.4.2.1.4.	Construcción de una tabla de índices pronósticos	224
9.4.2.1.5.	Punto de corte óptimo del mejor modelo seleccionado	225
9.4.2.2.	Transformación del modelo en un sistema de puntos	225
9.4.2.3.	Validación del nuevo sistema de puntos con los criterios 2009	226
9.4.3.	Resultados	227
9.4.3.1.	Diagnósticos del modelo elegido y tabla de índices pronósticos	230
9.4.3.2.	Sistema de puntos basado en los criterios de 2009	233
9.4.3.3.	Validación interna del nuevo sistema diagnóstico	236
9.4.3.4.	Validación externa del nuevo sistema diagnóstico	242
<b>9.5.</b>	<b>Capítulo 5: Evaluación de la concordancia entre las clasificaciones de los criterios diagnósticos</b>	<b>253</b>
9.5.1.	Metodología específica	253
9.5.2.	Análisis estadístico	253
9.5.3.	Resultados	254
<b>9.6.</b>	<b>Capítulo 6: Evaluación de la utilidad clínica de los criterios diagnósticos</b>	<b>261</b>
9.6.1.	Metodología específica	261
9.6.1.1.	Descripción del contexto clínico para el análisis de los criterios diagnósticos simplificados como índices de predicción	261
9.6.1.1.1.	Probabilidad preprueba de hepatitis autoinmune	261
9.6.1.1.2.	La prueba diagnóstica: Criterios simplificados para la hepatitis autoinmune pediátrica (propuestas de 2008 y 2009)	263
9.6.1.1.2.1.	Riesgo neto de los criterios simplificados	263
9.6.1.1.2.2.	Razones de verosimilitud de resultados positivo y negativo	264

9.6.1.1.2.3. Umbral de acción para los criterios simplificados _____	265
9.6.1.1.3. Efecto de la decisión clínica basada en los criterios simplificados: ¿Hay que tratar a este paciente? _____	267
9.6.1.1.3.1. Umbral terapéutico según el modelo de Pauker-Kassirer modificado por Latour _____	268
9.6.1.2. Simulaciones de la toma de decisiones en varios contextos clínicos ____	268
9.6.1.3. Estudio de la estabilidad diagnóstica de los criterios reducidos _____	269
9.6.1.3.1. Índice de reclasificación neta ( <i>net reclassification improvement</i> )_	269
9.6.1.3.2. Análisis por curvas de decisión ( <i>decision curve analysis</i> ) _____	271
9.6.2. Recursos para los análisis _____	276
9.6.3. Resultados _____	277
<b>10. Discusión _____</b>	<b>291</b>
<b>10.1. De la validez _____</b>	<b>293</b>
<b>10.2. De la fiabilidad _____</b>	<b>304</b>
<b>10.3. De la utilidad clínica _____</b>	<b>307</b>
<b>11. Conclusiones _____</b>	<b>311</b>
<b>12. Bibliografía _____</b>	<b>315</b>
<b>13. Anexos _____</b>	<b>355</b>
<b>13.1. Los estudios de evaluación de sistemas diagnósticos _____</b>	<b>357</b>
13.1.1. El proceso diagnóstico _____	357
13.1.2. Evaluación de la validez de un sistema diagnóstico _____	364
13.1.2.1. Sensibilidad _____	365
13.1.2.2. Especificidad _____	366
13.1.2.3. Valor predictivo positivo _____	366
13.1.2.4. Valor predictivo negativo _____	367
13.1.2.5. Razones de verosimilitud _____	368

13.1.2.5.1. Interpretación de las razones de verosimilitud _____	370
13.1.2.5.1.1. La contribución de Turing _____	374
13.1.2.6. Valor predictivo global _____	380
13.1.2.7. Odds ratio diagnóstica _____	382
13.1.2.8. Efectividad de una prueba _____	382
13.1.2.9. El índice de Youden _____	383
13.1.2.10. Ganancia diagnóstica _____	383
13.1.2.11. Curva de características operativas del receptor _____	386
13.1.2.12. Curva de Lorenz _____	391
13.1.2.12.1. Índices de Gini y Pietra _____	393
13.1.3. Cálculo del tamaño muestral para estudios de evaluación de pruebas o sistemas diagnósticos _____	394
13.1.4. Diseño de estudios para valorar la validez de sistemas diagnósticos _____	399
13.1.4.1. Consideraciones iniciales para la gestión de la arquitectura de un estudio de validación de unos criterios de clasificación _____	399
13.1.4.2. Fases en el estudio de un sistema diagnóstico _____	401
13.1.4.3. Descripción de los diseños genéricos para estudios sobre pruebas o sistemas diagnósticos _____	402
13.1.4.3.1. Estudio de cohortes _____	402
13.1.4.3.2. Estudio transversal _____	403
13.1.4.3.3. Estudio de casos y controles _____	404
13.1.4.4. Sesgos en la validez de sistemas o tests diagnósticos _____	405
13.1.4.4.1. Sesgo de selección _____	405
13.1.4.4.2. Sesgo de información _____	408
13.1.4.4.3. Otros sesgos _____	411
13.1.4.4.4. Errores más comunes en el razonamiento clínico probabilístico _____	412
13.1.5. Evaluación de la fiabilidad de un sistema diagnóstico _____	413
13.1.5.1. Variables categóricas nominales _____	413
13.1.5.2. Variables categóricas ordinales _____	415

13.1.5.3. Variables cuantitativas continuas _____	417
13.1.5.3.1. Desviación estándar intrasujetos _____	417
13.1.5.3.2. Coeficiente de correlación intraclase _____	418
13.1.5.3.3. Método de Bland-Altman _____	419
13.1.6. Lectura crítica de estudios para pruebas diagnósticas _____	420
13.1.6.1. Validez de los resultados _____	420
13.1.6.2. Exposición de los resultados _____	421
13.1.6.3. Aplicabilidad de los resultados _____	421
13.1.7. Índices de predicción clínica como fundamento de la toma de decisiones en Medicina asistencial _____	423
13.1.8. Análisis de decisiones clínicas _____	427
13.1.8.1. Valor de un procedimiento terapéutico o diagnóstico _____	429
13.1.8.1.1. Beneficio neto de un tratamiento apropiado _____	429
13.1.8.1.2. Riesgo neto de un tratamiento inapropiado _____	430
13.1.8.1.3. Riesgo neto de una prueba diagnóstica _____	430
13.1.8.2. Umbrales de acción para una prueba diagnóstica _____	431
13.1.8.3. Umbral terapéutico _____	433
<b>13.2. Criterios originales revisados para el diagnóstico de la HAI (1999) _____</b>	<b>437</b>
<b>13.3. Criterios simplificados para el diagnóstico de la HAI (2008) _____</b>	<b>439</b>
<b>13.4. Criterios diagnósticos de la HAI pediátrica propuestos por la ESPGHAN y la NASPGHAN (2009) _____</b>	<b>441</b>
<b>13.5. Lista STARD (<i>Standards for Reporting of Diagnostic Accuracy</i>) para la comunicación de estudios de validez de pruebas diagnósticas _____</b>	<b>443</b>
<b>13.6. Fundamentos de la revisión sistemática y el meta-análisis de estudios de sistemas diagnósticos _____</b>	<b>447</b>
13.6.1. Definición de la pregunta diagnóstica de interés _____	448
13.6.2. Búsqueda en diversas fuentes de todos los estudios confiables _____	448
13.6.3. Selección de estudios por medio de criterios de inclusión y de exclusión _____	450

13.6.4. Extracción y presentación de los datos de cada estudio _____	451
13.6.5. Evaluación de la homogeneidad entre los estudios primarios _____	452
13.6.6. Métodos de combinación de resultados sobre sistemas diagnósticos ____	454
13.6.6.1. <i>Combinación de sensibilidades y especificidades</i> _____	454
13.6.6.2. <i>Combinación de razones de verosimilitud</i> _____	455
13.6.6.3. <i>Combinación de odds ratio diagnósticas</i> _____	457
13.6.6.4. <i>Scores de efectividad diagnóstica</i> _____	458
13.6.6.5. <i>Curva resumen de características operativas del receptor</i> _____	459
13.6.7. Comentarios sobre el formato de presentación del meta-análisis _____	463
13.7. Lista QUADAS (Quality Assessment Diagnostic Accuracy Studies) para la comprobación de estudios sobre pruebas o sistemas diagnósticos incluidos en revisiones sistemáticas y meta-análisis _____	465
13.8. Formulario de entrada a la base de datos _____	471
13.9. Hoja informativa y de consentimiento informado _____	473

# 1

## Abreviaturas y siglas empleadas



## Abreviaturas y siglas empleadas

---

<b>ALT</b>	Alanina aminotransferasa
<b>ANA</b>	Anticuerpo antinuclear
<b>ANCA</b>	Anticuerpo anti-citoplasma del neutrófilo
<b>ANOVA</b>	<i>Analysis of variance</i> (análisis de la varianza)
<b>Anti-ASGPR</b>	Anticuerpo anti-receptor de asialoglicoproteína
<b>Anti-LC1</b>	Anticuerpo anti-citosol hepático de tipo 1
<b>Anti-LKM1</b>	Anticuerpo anti-microsomal de hígado/riñón de tipo 1
<b>Anti-SLA/LP</b>	Anticuerpos anti-antígeno soluble hepático/anti-hígado-páncreas
<b>Anti-Sm</b>	Anticuerpo anti-músculo liso
<b>APECED</b>	<i>Autoimmune polyendocrinopathy – candidiasis – ectodermal dystrophy</i> (poliendocrinopatía autoinmune tipo 1)
<b>AST</b>	Aspartato aminotransferasa
<b>CASP</b>	<i>Critical appraisal skills programme</i> (programa de habilidades en lectura crítica)
<b>CBP (o PBC)</b>	Cirrosis biliar primaria ( <i>primary biliary cirrhosis</i> )
<b>CCI</b>	Coeficiente de correlación intraclase
<b>CEAI</b>	Colangitis esclerosante autoinmune
<b>CEP (o PSC)</b>	Colangitis esclerosante primaria ( <i>primary sclerosing cholangitis</i> )
<b>CGRE</b>	Colangiografía retrógrada endoscópica
<b>CHC</b>	Carcinoma hepatocelular
<b>CIE</b>	Clasificación internacional de enfermedades
<b>CIEF</b>	Contra-inmunolectroforesis
<b>CMr</b>	Cuadrados medios de los residuos



<b>CMV</b>	Citomegalovirus
<b>CONSORT</b>	<i>Consolidated standards of reporting trials</i>
<b>COR</b>	Características operativas del receptor
<b>CTLA-4</b>	Proteína 4 asociada a linfocitos T citotóxicos
<b>EHNA (o NASH)</b>	Esteatohepatitis no alcohólica ( <i>non-alcoholic steatohepatitis</i> )
<b>ELISA</b>	<i>Enzyme-linked immunosorbent assay</i> (prueba de inmunoabsorción enzimática)
<b>ESPGHAN</b>	<i>European Society for Paediatric Gastroenterology, Hepatology and Nutrition</i> (Sociedad Europea de Gastroenterología, Hepatología y Nutrición Pediátrica)
<b>Et al</b>	Y colaboradores
<b>FA</b>	Fosfatasa alcalina
<b>FHF</b>	Fallo hepático fulminante
<b>FTCD</b>	Forminino-transferasa ciclodeaminasa
<b>GGT</b>	Gamma-glutamil transpeptidasa
<b>HAI (o AIH)</b>	Hepatitis autoinmune ( <i>autoimmune hepatitis</i> )
<b>HLA</b>	<i>Human leukocyte antigen</i> (antígeno leucocitario humano)
<b>HSROC</b>	<i>Hierarchical summary receiver operating characteristic</i>
<b>IAIHG</b>	<i>International Autoimmune Hepatitis Group</i> (Grupo Internacional para el Estudio de la Hepatitis Autoinmune)
<b>IB</b>	Inmunoblot
<b>IC95%</b>	Intervalo de confianza al 95%
<b>IDBD</b>	Inmunodifusión bidimensional
<b>IF</b>	Inmunofluorescencia
<b>IgA</b>	Inmunoglobulina A
<b>IgG</b>	Inmunoglobulina G

<b>IgM</b>	Inmunoglobulina M
<b>IL</b>	Interleuquina
<b>INR</b>	<i>International normalized ratio</i> (relación normalizada internacional)
<b>JPGN</b>	<i>Journal of Pediatric Gastroenterology and Nutrition</i>
<b>LASPGHAN</b>	<i>Latin American Society for Pediatric Gastroenterology, Hepatology and Nutrition</i> (Sociedad Latinoamericana de Gastroenterología, Hepatología y Nutrición Pediátrica)
<b>LES</b>	Lupus eritematoso sistémico
<b>LIA</b>	<i>Line-immuno-assay</i> (inmunoensayo lineal)
<b>LSP</b>	<i>Liver specific protein</i> (proteína específica hepática)
<b>MHC</b>	<i>Major histocompatibility complex</i> (nó complejo mayor de histocompatibilidad)
<b>NASPGHAN</b>	<i>North American Society for Pediatric Gastroenterology, Hepatology and Nutrition</i> (Sociedad Norteamericana de Gastroenterología, Hepatología y Nutrición Pediátrica)
<b>NND</b>	Número de pacientes necesarios para diagnosticar
<b>NNDM</b>	Número de pacientes necesarios para diagnosticar mal
<b>NRI</b>	<i>Net reclassification improvement</i> (índice de reclasificación neta)
<b>ORD</b>	<i>Odds ratio</i> diagnóstica
<b>PALFSG</b>	<i>Pediatric Acute Liver Failure Study Group</i> (Grupo de Estudio sobre el Fallo Hepático Agudo Pediátrico)
<b>pANCA</b>	Anticuerpo anti-citoplasma del neutrófilo con patrón perinuclear
<b>pANNA</b>	Anticuerpo periférico anti-núcleo del neutrófilo
<b>QUADAS</b>	<i>Quality assessment diagnostic accuracy studies</i>
<b>QUORUM</b>	<i>Quality of reporting of meta-analysis</i>
<b>RIA</b>	<i>Radio-immune-precipitation assay</i> (prueba de radio-inmuno-precipitación)

<b>RIC</b>	Rango intercuartílico
<b>RM</b>	Resonancia magnética
<b>RV</b>	Razón de verosimilitud
<b>Se o S</b>	Sensibilidad
<b>Sp o E</b>	Especificidad
<b>STARD</b>	<i>Standards for reporting diagnostic accuracy</i>
<b>TGF</b>	<i>Transforming growth factor</i> (factor de crecimiento transformante)
<b>UD</b>	Umbral diagnóstico
<b>UDT</b>	Umbral diagnóstico-terapéutico
<b>VEB</b>	Virus de Epstein-Barr
<b>VHC</b>	Virus de la hepatitis C
<b>VPN</b>	Valor predictivo negativo
<b>VPP</b>	Valor predictivo positivo
<b>WoE</b>	<i>Weight of evidence</i> (peso de la evidencia)

# 2

## Listado de figuras



## Listado de figuras

---

Figura 1: Tasa de incidencia según edad y sexo de la hepatitis autoinmune en Dinamarca (1994-2012) .....	62
Figura 2: Tasa de incidencia estandarizada por edad y sexo en Dinamarca (1994-2012).....	63
Figura 3: Imagen de inmunofluorescencia indirecta en tejido de roedor de los autoanticuerpos diagnósticos de la hepatitis autoinmune .....	73
Figura 4: Inmunofluorescencia de los anticuerpos anti-mitocondriales y de los anti-microsoma de hígado/riñón .....	75
Figura 5: Intensa plasmocitosis portal en la hepatitis autoinmune, señalada con flechas en el panel de la derecha (tinción con hematoxilina-eosina) .....	81
Figura 6: Hepatitis de interfase con cuerpos apoptóticos señalados con las flechas en un caso de hepatitis autoinmune tipo 1 (teñido con hematoxilina-eosina) .....	81
Figura 7: Varios ejemplos de rosetas de hepatocitos señalados por las flechas. Tinción de hematoxilina-eosina en todos los paneles salvo en el superior derecho, que pone en evidencia el colapso lobulillar a través de una tinción para fibras de reticulina.....	82
Figura 8: Se pueden observar eosinófilos en mayor o menor grado formando parte de las poblaciones constituyentes del infiltrado leucocitario (panel A). En el panel B se señalan cambios hidrópicos de los hepatocitos en una sección de hepatitis de interfase con necrosis multiacinar.....	84
Figura 9: Emperipolesis a la izquierda, con un linfocito dentro de un hepatocito dañado. A la derecha una roseta de hepatocitos en el área de la interfase .....	85
Figura 10: Esquema de la patogenia de la hepatitis autoinmune. ....	89
Figura 11: Esquema de manejo en forma de diagrama de flujo para la enfermedad hepática autoinmune en pediatría .....	95

Figura 12: Puntuación según los criterios simplificados de 2008 de los pacientes con hepatitis autoinmune y sus controles en la muestra de derivación. Misma comparación en la muestra de validación .....	107
Figura 13: Curva de características operativas del receptor de los criterios simplificados de 2008 en la muestra de derivación .....	108
Figura 14: Criterios para el diagnóstico de la hepatitis autoinmune en población pediátrica.....	115
Figura 15: Árbol de decisiones para la elección de aplicar una prueba diagnóstica, en función de la sensibilidad y especificidad propias de la prueba y en base a las probabilidades preprueba reales y heurísticas .....	149
Figura 16: Ejemplo de representación de la ganancia neta en certeza diagnóstica con y sin hacer radiografía de tórax para el diagnóstico de neumonía .....	150
Figura 17: Diagrama de flujo STARD con los resultados de la prueba índice y la asignación diagnóstica final.....	154
Figura 18: Número de pacientes con hepatitis autoinmune en los que se realizó colangio-resonancia magnética en función del año de diagnóstico. ....	155
Figura 19: Histograma con la distribución de frecuencias por edad al diagnóstico en los pacientes con hepatitis autoinmune .....	157
Figura 20: Distribuciones de los parámetros bioquímicos con patrón significativamente diferente entre los dos grupos diagnósticos.....	159
Figura 21: Distribuciones de los títulos de autoanticuerpos entre los dos grupos diagnósticos. ....	161
Figura 22: Diagrama de caja de la puntuación obtenida en los criterios clásicos revisados de 1999 por los pacientes con o sin hepatitis autoinmune.....	162
Figura 23: Diagrama de caja de la puntuación obtenida en los criterios simplificados de 2008 por los pacientes con o sin hepatitis autoinmune. ....	163
Figura 24: Diagrama de caja de las puntuaciones obtenidas en los criterios clásicos revisados de 1999 por los pacientes con los principales diagnósticos diferenciales .....	163

Figura 25: Diagrama de caja de las puntuaciones obtenidas en los criterios simplificados de 2008 por los pacientes con los principales diagnósticos diferenciales.....	164
Figura 26: Relación entre la prevalencia de hepatitis autoinmune y la probabilidad postprueba con puntuaciones iguales o superiores a 6 (resultado positivo: línea azul) e inferiores a 6 (resultado negativo: línea roja) en los criterios simplificados de 2008 .....	167
Figura 27: Relación entre la prevalencia de hepatitis autoinmune y la probabilidad postprueba con puntuaciones de 7 u 8 (resultado positivo: línea azul) e inferiores a 7 (resultado negativo: línea roja) en los criterios simplificados de 2008 .....	168
Figura 28: Comportamiento de la ganancia neta en certeza diagnóstica en función de la prevalencia de la enfermedad para los puntos de corte de 6 y 7 de los criterios simplificados de 2008. ....	169
Figura 29: Curva de características operativas del receptor de los criterios simplificados de 2008 para el diagnóstico de hepatitis autoinmune.....	173
Figura 30: Curva de características operativas del receptor del modelo diagnóstico basado los criterios de 2008 para el diagnóstico de hepatitis autoinmune con el punto de corte en $\geq 6$ . ....	175
Figura 31: Curva de Lorenz de los criterios simplificados para el diagnóstico de hepatitis autoinmune (HAI). La región sombreada gris representa el intervalo de confianza del 95% de la estimación. ....	175
Figura 32: Algoritmo para la toma de decisiones en meta-análisis de sistemas diagnósticos.....	184
Figura 33: Diagrama de flujo de la selección de estudios primarios para la revisión sistemática y el meta-análisis de la validez de los criterios simplificados de 2008 para el diagnóstico de la hepatitis autoinmune.....	187
Figura 34: Principales indicadores de validez de los criterios de 2008 comunicados por los estudios primarios de la revisión sistemática.....	192
Figura 35: Sensibilidad y especificidad de los trabajos originales de la revisión sistemática representados en el plano de características operativas del receptor con sus intervalos de confianza al 95%.....	193
Figura 36: Meta-análisis univariante de la sensibilidad y la especificidad de los criterios simplificados para el diagnóstico de la hepatitis autoinmune. ....	194



Figura 37: Meta-análisis univariante de las razones de verosimilitud de los criterios simplificados para el diagnóstico de la hepatitis autoinmune .....	195
Figura 38: Meta-análisis univariante de la <i>odds ratio</i> diagnóstica de los criterios simplificados para el diagnóstico de la hepatitis autoinmune.....	195
Figura 39: Espacio de características operativas del receptor con la intersección entre la sensibilidad y la especificidad combinadas y sus intervalos de confianza al 95%. .....	196
Figura 40: Plano de características operativas del receptor con el punto resumen y su intervalo de confianza al 95%.....	197
Figura 41: Curva resumen simétrica de características operativas del receptor siguiendo el modelo de Moses-Shapiro-Littenberg.....	198
Figura 42: Representación del punto resumen y la curva COR resumen basada en un modelo jerárquico .....	199
Figura 43: Diagrama en embudo de Deeks que relaciona el logaritmo de la <i>odds ratio</i> diagnóstica con su error estándar para cada estudio primario del meta-análisis .....	200
Figura 44: Representación gráfica del test de asimetría sobre el diagrama en embudo de Deeks .....	201
Figura 45: Gráfico Q-Q normal sin tendencia para la diferencia entre las puntuaciones obtenidas por los criterios de 2008 aplicados por dos observadores distintos. ....	210
Figura 46: Diagrama de Bland-Altman para las diferencias entre las puntuaciones de los criterios simplificados aplicados por dos observadores independientes.....	211
Figura 47: Sensibilidad y especificidad obtenidas para cada punto de corte en la probabilidad de HAI predicha por el modelo máximo basado en los criterios de 2009 con el ítem de los autoanticuerpos categorizado.....	228
Figura 48: Detección de casos influyentes con la distancia de Cook .....	230
Figura 49: Riesgos relativos de cada posible combinación de criterios ESPGHAN/NASPGHAN 2009 respecto al riesgo de un paciente con hipertransaminasemia, sin marcadores virales ni enfermedad de Wilson, colangiografía normal y resto de variables negativas .....	232

Figura 50: Curva de características operativas del receptor. En azul, para el modelo predictivo de regresión logística, basado en todas las variables de los criterios ESPGHAN/NASPGHAN 2009 .....	233
Figura 51: Diagrama de caja de la puntuación obtenida por los nuevos criterios diagnósticos ESPGHAN/NASPGHAN 2009 para la hepatitis autoinmune pediátrica en la muestra de derivación ...	237
Figura 52: Curva de características operativas del receptor del nuevo sistema diagnóstico por puntos basado en los criterios ESPGHAN/NASPGHAN 2009 para la hepatitis autoinmune pediátrica en la muestra de derivación. ....	238
Figura 53: Relación entre la prevalencia de hepatitis autoinmune y la probabilidad postprueba con un resultado positivo y con un resultado negativo en el nuevo sistema diagnóstico por puntos basado en los criterios ESPGHAN/NASPGHAN 2009 .....	240
Figura 54: Curva de Lorenz del nuevo sistema diagnóstico basado en los criterios ESPGHAN/NASPGHAN 2009 para la hepatitis autoinmune pediátrica, sobre la muestra de derivación .....	241
Figura 55: Diagrama de caja de la puntuación obtenida en los nuevos criterios ESPGHAN/NASPGHAN 2009 por los pacientes con y sin hepatitis autoinmune en la muestra de validación.....	242
Figura 56: Diagrama de caja de las puntuaciones obtenidas en el nuevo sistema diagnóstico basado en los criterios ESPGHAN/NASPGHAN 2009 por los pacientes con los principales diagnósticos diferenciales .....	243
Figura 57: Relación entre la prevalencia de hepatitis autoinmune y la probabilidad postprueba con puntuaciones iguales o superiores a 6, e inferiores a 6, en el nuevo sistema diagnóstico por puntos basado en los criterios ESPGHAN/NASPGHAN 2009.....	246
Figura 58: Curva de características operativas del receptor del nuevo sistema diagnóstico por puntos basado en los criterios ESPGHAN/NASPGHAN 2009 para la hepatitis autoinmune pediátrica en la muestra de validación .....	249
Figura 59: Curva de características operativas del receptor del modelo diagnóstico basado en los criterios ESPGHAN/NASPGHAN 2009 para la hepatitis autoinmune pediátrica en la muestra de validación, con el punto de corte en $\geq 6$ .....	249

Figura 60: Curva de Lorenz del nuevo sistema diagnóstico basado en los criterios ESPGHAN/NASPGHAN 2009 para la hepatitis autoinmune pediátrica, sobre la muestra de validación .....	251
Figura 61: Cambios en las categorías diagnósticas basadas en los dos criterios del Grupo Internacional para el Estudio de la Hepatitis Autoinmune.....	256
Figura 62: Cambios en las categorías diagnósticas basadas en los criterios clásicos, revisados en 1999, del Grupo Internacional para el Estudio de la Hepatitis Autoinmune, y en el nuevo sistema de puntos basado en los criterios pediátricos propuestos por la ESPGHAN/NASPGHAN en 2009 .....	258
Figura 63: Complicaciones menores y mayores de la biopsia hepática y su correspondiente incidencia .....	263
Figura 64: Tabla de resultados para el ejemplo del cálculo del índice de reclasificación neta .....	270
Figura 65: Curva de decisión de ejemplo de un modelo de predicción de la invasión de vesículas seminales en el cáncer de próstata.....	274
Figura 66: Curvas de decisión para una distribución teórica con una prevalencia del 20%.....	276
Figura 67: Representación gráfica del cambio de las utilidades de los criterios simplificados de 2008 en función de la probabilidad teórica de HAI.....	278
Figura 68: Nomograma de Fagan para los criterios simplificados de 2008.....	280
Figura 69: Simulaciones con el nomograma de Fagan para los criterios simplificados de 2008 con las razones de verosimilitud positiva y negativa de 13,72 y 0,23.....	281
Figura 70: Gráfico de modificación probabilística ( <i>probability-modifying plot</i> ) de la enfermedad construido con las razones de verosimilitud del meta-análisis de la validez de los criterios simplificados de 2008.....	282
Figura 71: Representación gráfica del cambio de las utilidades del nuevo sistema diagnóstico por puntos basado en los criterios ESPGHAN/NASPGHAN 2009 en función de la probabilidad teórica de HAI.....	283
Figura 72: Nomograma de Fagan para los nuevos criterios diagnósticos ESPGHAN/NASPGHAN 2009. Las razones de verosimilitud positiva y negativa empleadas son 24,9 y 0,04.....	284

Figura 73: Simulaciones con el nomograma de Fagan para los nuevos criterios ESPGHAN/NASPGHAN 2009, transformados en un sistema diagnóstico por puntos, con las razones de verosimilitud positiva y negativa de 24,9 y 0,04.....	285
Figura 74: Gráfico de modificación probabilística ( <i>probability-modifying plot</i> ) de la enfermedad construido con las razones de verosimilitud obtenidas en la validación externa del nuevo sistema diagnóstico por puntos basado en los criterios pediátricos 2009 de hepatitis autoinmune .....	286
Figura 75: Curvas de decisión de los distintos criterios diagnósticos para la hepatitis autoinmune ...	289
Figura 76: Propuesta de criterios diagnósticos por puntos para la hepatitis autoinmune pediátrica ( <i>ESPGHAN Hepatology Committee Position Statement</i> de 2018) .....	303
Figura 77: Componentes del proceso diagnóstico .....	359
Figura 78: Ejemplo de representación de la validez de un sistema diagnóstico en una gráfica Sensibilidad / 1-Especificidad para establecer comparaciones con otros sistemas o pruebas diagnósticas.....	371
Figura 79: Nomograma de Fagan .....	373
Figura 80: Esquema del razonamiento bayesiano para la solución del contraste de la hipótesis Y en base a la información X en un universo $t$ .....	375
Figura 81: Valores predictivos en función de la razón de verosimilitud y la prevalencia .....	381
Figura 82: Ganancia diagnóstica obtenida por la respuesta de una prueba en función de la probabilidad preprueba o prevalencia.....	385
Figura 83: Habitualmente existe un solapamiento en los resultados de las pruebas diagnósticas cuantitativas entre enfermos y sanos que hace imposible encontrar un punto de corte perfecto.....	388
Figura 84: Curva COR paramétrica para datos discretos.....	390
Figura 85: Curva de Lorenz.....	392
Figura 86: Elementos de los índices de Gini y Pietra sobre una curva de Lorenz .....	393

Figura 87: Nomograma de Malhotra e Indrayan (basado en la ecuación de Buderer) para el tamaño muestral de un estudio de validación de pruebas diagnósticas a partir de la prevalencia y la sensibilidad/especificidad esperadas, y el margen de error deseado .....	397
Figura 88: Nomograma de Carley para la sensibilidad con $\alpha = 0,05$ .....	398
Figura 89: Nomograma de Carley para la especificidad con $\alpha = 0,05$ .....	399
Figura 90: Ejemplo de diagrama de Bland-Altman que representa las diferencias entre dos métodos distintos de estimar la osmolaridad de las soluciones de nutrición parenteral en función de la magnitud del valor real .....	419
Figura 91: Diagrama prototípico de la declaración STARD para comunicar el flujo de pacientes a lo largo del estudio.....	422
Figura 92: Análisis de umbrales de probabilidad y decisiones diagnósticas y terapéuticas acordes a la probabilidad preprueba de padecer una enfermedad concreta.....	431
Figura 93: Prueba diagnóstica cuyo resultado se traduce en un cambio de actitud terapéutica a favor del tratamiento .....	435
Figura 94: Prueba diagnóstica cuyo resultado se traduce en un cambio de actitud terapéutica en contra del tratamiento .....	435
Figura 95: Esquema del proceso de búsqueda y selección de artículos para una revisión sistemática de estudios de sistemas diagnósticos .....	449
Figura 96: Gráfico de dispersión. Cada punto representa un estudio .....	450
Figura 97: Ejemplo de plano de la curva resumen de características operativas del receptor. Incluye las regiones de confianza de la curva y el par sensibilidad/especificidad resumen.....	462

# 3

Listado de tablas



## Listado de tablas

---

Tabla 1: Diferencias entre criterios de clasificación y criterios diagnósticos .....	48
Tabla 2: Comparación epidemiológico-analítica entre la hepatitis autoinmune de presentación aguda y las hepatitis víricas agudas .....	64
Tabla 3: Características clínicas y analíticas de la hepatitis autoinmune según la cronopatología: aguda, crónica y asintomática.....	66
Tabla 4: Autoanticuerpos y sus antígenos en la enfermedad hepática autoinmune.....	71
Tabla 5: El antígeno leucocitario humano del complejo mayor de histocompatibilidad II y su papel y asociaciones con la hepatitis autoinmune .....	87
Tabla 6: Datos bioquímicos al diagnóstico de los niños con enfermedad hepática autoinmune .....	98
Tabla 7: Presentación clínica al diagnóstico de las enfermedades hepáticas autoinmunes en la infancia .....	99
Tabla 8: Datos clínicos y analíticos comparados entre la colangitis esclerosante autoinmune y la primaria.....	100
Tabla 9: Guía de uso de los criterios descriptivos revisados para el diagnóstico de hepatitis autoinmune según el sistema clásico de 1999.....	102
Tabla 10: Sensibilidad y especificidad de las diferentes variaciones de los criterios simplificados de 2008 sobre la muestra de derivación.....	109
Tabla 11: Sensibilidad y especificidad de las diferentes variaciones de los criterios simplificados de 2008 sobre la muestra de validación. ....	109
Tabla 12: Resumen de los estudios llevados a cabo en población pediátrica sobre la exactitud de los criterios diagnósticos para la hepatitis autoinmune .....	111
Tabla 13: Frecuencia de las asociaciones entre hepatitis autoinmune y otras enfermedades autoinmunes .....	156



Tabla 14: Enfermedades y frecuencias que integran el grupo de no casos .....	157
Tabla 15: Marcadores bioquímicos relevantes para el diagnóstico acorde a la clasificación final. Para los datos cuantitativos se emplea la mediana y el rango intercuartílico como marcador de tendencia central .....	158
Tabla 16: Proporción de pacientes con títulos de autoanticuerpos superiores a los dinteles establecidos por los criterios simplificados de 2008 en los grupos de casos y no casos de hepatitis autoinmune.....	160
Tabla 17: Distribución descrita a través de la mediana y el rango intercuartílico de los títulos de autoanticuerpos entre los casos de hepatitis autoinmune y el grupo de no casos. ....	160
Tabla 18: Distribución descrita a través de la mediana y el rango intercuartílico de los puntos obtenidos en los criterios diagnósticos para la hepatitis autoinmune .....	162
Tabla 19: Valores predictivos positivos de los criterios simplificados de 2008 según la prevalencia de hepatitis autoinmune en la población .....	170
Tabla 20: Valores predictivos negativos de los criterios simplificados de 2008 según la prevalencia de hepatitis autoinmune en la población. ....	171
Tabla 21: Resumen de los principales indicadores de validez de los criterios simplificados de 2008 para el diagnóstico de hepatitis autoinmune en niños. Entre paréntesis, intervalo de confianza al 95% .....	172
Tabla 22: Indicadores de validez para cada uno de los posibles puntos de corte de los criterios simplificados de 2008 para el diagnóstico de hepatitis autoinmune.....	173
Tabla 23: Simulación de los puntos de corte óptimos para distintas prevalencias y distintas asunciones respecto a la razón de costes idónea. Encuadrado el contexto más parecido al de la población del estudio .....	174
Tabla 24: Características de los estudios incluidos en el meta-análisis .....	186
Tabla 25: Respuestas a las preguntas del listado QUADAS-2 sobre el riesgo de sesgo en cada uno de los estudios primarios de la revisión sistemática .....	191

Tabla 26: Resumen de la calidad metodológica de los estudios incluidos en la revisión sistemática según la metodología QUADAS-2.....	191
Tabla 27: Parámetros del ajuste de la recta de regresión en el test de asimetría de Deeks para el estudio de la presencia de sesgo de publicación .....	201
Tabla 28: Relación de resultados de la aplicación de los criterios simplificados por dos observadores independientes .....	207
Tabla 29: Tabla de contingencia de las clasificaciones de los dos observadores empleando los criterios simplificados de 2008.....	209
Tabla 30: Estudio de la concordancia mediante el estadístico <i>kappa</i> ponderado para la clasificación en no hepatitis autoinmune, enfermedad probable o enfermedad definitiva según los criterios simplificados de la IAIGH.....	209
Tabla 31: Criterios diagnósticos pediátricos para la hepatitis autoinmune propuestos por Mieli-Vegani et al en 2009.....	214
Tabla 32: Definición del criterio de los autoanticuerpos en la propuesta ESPGHAN/NASPGHAN 2009 y su versión modificada para la modelización .....	227
Tabla 33: Calidad del ajuste de los modelos máximos basados en los criterios ESPGHAN/NASPGHAN 2009 para el diagnóstico de hepatitis autoinmune.....	228
Tabla 34: Características de los mejores submodelos restringidos en variables .....	229
Tabla 35: Modelo provisional de regresión logística con intención diagnóstica (predictiva) para hepatitis autoinmune a partir de los criterios ESPGHAN/NASPGHAN de 2009 con los autoanticuerpos categorizados .....	230
Tabla 36: Proceso de transformación de los coeficientes $\beta$ de cada variable del modelo de regresión seleccionado en los puntos de los nuevos criterios diagnósticos de hepatitis autoinmune, basados en la propuesta pediátrica de 2009 .....	234
Tabla 37: Nuevos criterios diagnósticos de hepatitis autoinmune pediátrica basada en la propuesta ESPGHAN / NASPGHAN de 2009 .....	236

Tabla 38: Indicadores de validez para cada uno de los posibles cortes de los nuevos criterios ESPGHAN/NASPGHAN 2009, concebidos como un sistema de puntos, para el diagnóstico de la hepatitis autoinmune pediátrica (muestra de derivación) .....	239
Tabla 39: Simulación de los puntos de corte óptimos para distintas prevalencias y distintas asunciones respecto a la razón de costes idónea .....	239
Tabla 40: Distribución descrita a través de la mediana y el rango intercuartílico de los puntos obtenidos por el nuevo sistema de puntos ESPGHAN/NASPGHAN 2009 para la hepatitis autoinmune .....	242
Tabla 41: Valores predictivos de los nuevos criterios ESPGHAN/NASPGHAN según la prevalencia de hepatitis autoinmune en la población .....	247
Tabla 42: Resumen de los principales indicadores de validez de los nuevos criterios ESPGHAN/NASPGHAN 2009 para el diagnóstico de hepatitis autoinmune en niños. Calculados en la submuestra de validación .....	248
Tabla 43: Indicadores de validez para cada uno de los posibles puntos de corte de los nuevos criterios ESPGHAN/NASPGHAN 2009 para el diagnóstico de hepatitis autoinmune (muestra de validación). .	250
Tabla 44: Simulación de los puntos de corte óptimos para distintas prevalencias y distintas asunciones respecto a la razón de costes idónea .....	250
Tabla 45: Análisis de la utilidad de los criterios simplificados de 2008.....	277
Tabla 46: Análisis de la utilidad de los nuevos criterios pediátricos de 2009 .....	283
Tabla 47: Evaluación a través del índice de reclasificación neta de la contribución de la variable sexo (masculino o femenino) al modelo con los criterios simplificados de 2008 .....	287
Tabla 48: Evaluación a través del índice de reclasificación neta de la contribución de la variable alcohol (<25 o >60 g/día) al modelo con los criterios simplificados de 2008 .....	287
Tabla 49: Evaluación a través del índice de reclasificación neta de la contribución de la variable AMA (sí o no) al modelo con los criterios simplificados de 2008.....	287
Tabla 50: Evaluación a través del índice de reclasificación neta de la contribución de la variable FA/AST (<1,5, de 1,5 a 3 o >3) al modelo con los criterios simplificados de 2008. ....	287

Tabla 51: Evaluación a través del índice de reclasificación neta de la contribución de la variable descarte de enfermedad de Wilson (sí o no) al modelo con los criterios simplificados de 2008 .....	288
Tabla 52: Evaluación a través del índice de reclasificación neta de la contribución de la variable antecedentes personales o familiares de autoinmunidad (sí o no) al modelo con los criterios simplificados de 2008.....	288
Tabla 53: Distribución de casos para valorar la validez de una prueba diagnóstica. ....	365
Tabla 54: Tabla de contingencia donde la evidencia es la prueba a validar y la hipótesis es el criterio de verdad. ....	378
Tabla 55: Relación entre la razón de verosimilitud y el peso de la evidencia .....	379
Tabla 56: Sesgos en estudios de evaluación de sistemas diagnósticos.....	411
Tabla 57: Tabla de contingencia entre los resultados dados por dos observadores diferentes sobre una misma muestra de casos .....	414
Tabla 58: Interpretación de los valores del índice <i>kappa</i> .....	415
Tabla 59: Nomenclatura para los elementos de cada estudio individual sobre un sistema diagnóstico .....	451



# 4

## Introducción



#### **4.1. El diagnóstico en Medicina basado en criterios de clasificación**

---

El diagnóstico en Medicina es un proceso dinámico en el que se intenta tomar decisiones idóneas en presencia de incertidumbre. A lo largo de la historia se han desarrollado diversos instrumentos cuya intervención se traduce en un aumento del grado de certeza con el que se emiten los juicios diagnósticos. Las herramientas clásicas empleadas para dicho fin son la sistemática de la anamnesis y la exploración clínica pero también se disponen de estudios complementarios que, utilizados según las indicaciones óptimas, permiten mejorar el proceso diagnóstico.

Desde un punto de vista funcional, se considera que una prueba diagnóstica es cualquier procedimiento realizado para confirmar o descartar un diagnóstico o incrementar o disminuir su verosimilitud. La utilidad de una prueba diagnóstica depende fundamentalmente de su validez y de su fiabilidad, pero también de su rendimiento clínico y su coste.

La complejidad de determinadas situaciones clínicas y el grado incompleto del conocimiento de las causas y mecanismos de algunas enfermedades han llevado a buscar combinaciones de observaciones, tanto clínicas como de resultados de exploraciones complementarias, para efectuar el diagnóstico. La formalización y estandarización de una mezcla de datos médicos lleva a la aparición de criterios de clasificación, que se evalúan y presentan las mismas propiedades que cualquier otra prueba diagnóstica [1,2].

De este modo, los criterios de clasificación para una entidad nosológica pueden incluir síntomas, un clúster de síntomas (síndrome), datos de la exploración clínica, anormalidades morfológicas, disfunciones fisiológicas, defectos bioquímicos, anormalidades genéticas o ultraestructurales, o evidencia de exposición a agentes etiológicos [3].

En determinadas enfermedades el diagnóstico se realiza mediante criterios de clasificación, que las definen tanto en la práctica clínica como en investigación. De esta manera, se consigue incluir varias condiciones similares en un único



concepto inteligible. Sin embargo, no deja de ser un mero intento de aproximación a la realidad a través de un marco teórico susceptible de ampliarse y detallarse a medida que avanza el conocimiento.

#### **4.1.1. La esencia real y la esencia nominal de las enfermedades**

En el debate sobre el concepto de la enfermedad que tiene lugar en la literatura de la filosofía de la Medicina es frecuente encontrar categorías conceptuales como esencialista, objetivista, naturalista, realista o realista taxonómico. Se aplican ampliamente con el fin de establecer la posición del término “enfermedad” dentro de los fenómenos corrientes de la naturaleza. Según la teoría de los tipos naturales, las enfermedades individuales se pueden agrupar sobre la base de una esencia natural compartida. Varias características de eventos particulares distintos los hacen pertenecer a un grupo natural común y por lo tanto se pueden denominar de la misma forma (semántica de los tipos naturales) [4]. De este modo, bajo la concepción naturalista de la enfermedad, se encuentran definiciones basadas en las respuestas que ofrecen las ciencias biológicas a las preguntas sobre su naturaleza.

Por el contrario, la concepción nominalista rechaza que aquello que hace que un proceso pueda considerarse “enfermedad” sea algo natural o real: existe algo común a lo que llamamos enfermedad, pero no se encuentra en la naturaleza. Bajo este prisma, las entidades morbosas serían más fácilmente caracterizables teniendo en cuenta que se trata de constructos humanos, visión motivada parcialmente por eventos de la historia de la Medicina que han conducido a cambios en los modelos explicativos hegemónicos de cada época. La Medicina ha estudiado la enfermedad en términos de desequilibrio de la homeostasis del cuerpo o sus humores (Hipócrates, Galeno), cambios morfológicos en sus órganos internos (Morgagni), en sus tejidos (Bichat) o en sus células (Virchow), la irritación de los órganos y su reacción (Brown), la invasión de un agente externos (Koch) y alteraciones genéticas.

Actualmente, la definición de las enfermedades incluye como elementos signos y síntomas clínicos, hallazgos anormales en pruebas complementarias, quejas del paciente e incluso circunstancias sociales, tal como se recoge en la Clasificación Internacional de Enfermedades (CIE) [5].

Claude Bernard apuntó en la línea de la concepción nominal de las enfermedades: *“Ni los fisiólogos ni los médicos necesitan pensar en la causa de la vida o la esencia de la enfermedad para llevar a cabo su tarea. Sería como perder el tiempo en la búsqueda de un fantasma”* [6].

Otro pensador e historiador de la Medicina, Lester King, expresó que *“la enfermedad es toda aquella condición que, a la luz de la cultura dominante, se considere dolorosa o inhabilitante y que, al mismo tiempo, se desvíe de la normalidad estadística o de una suerte de estado ideal”* [7].

Por lo tanto, las enfermedades no se “descubren”, sino que, sobre una base natural real, se “inventan” en su acepción global. Se trata de las enfermedades concebidas desde el punto de vista social, subjetivo, atributivo o descriptivo [7]. Lo que las categorías anteriormente mencionadas tienen en común es que implican un rechazo a que los casos considerados “enfermedad” se agreguen en base a fenómenos naturales. Por el contrario, parece afirmarse que lo que comparten dichos casos es que encajan con una descripción o definición dada. Así, estas situaciones consideradas comúnmente anormales, se clasifican como enfermedad si satisfacen determinados criterios de clasificación, atributos o características definitorias [8,9]. Esta visión se denomina *semántica descriptiva* y aunque se trata de un concepto ampliamente estudiado por la filosofía del lenguaje, entronca directamente con la forma a través de la cual se llevan a cabo los diagnósticos en la Medicina contemporánea. De hecho, desde los años 80 y 90 del siglo XX, ha constituido un elemento clave del desarrollo de la Medicina basada en la evidencia y la recuperación de información médica a través de directorios [4,10,11].

Una característica emergente de esta concepción nominal de las enfermedades es que permite estudiar el grado de encaje de determinadas

situaciones compatibles con un diagnóstico con la definición de ese diagnóstico. Como se expondrá en el apartado correspondiente, el objetivo de esta tesis es llevar a cabo una aproximación formal a esta problemática en el caso de las hepatitis autoinmunes en población pediátrica, a través de los métodos de evaluación de pruebas diagnósticas.

#### 4.1.2. Controversia en la distinción entre criterios diagnósticos y de clasificación

El estudio de las enfermedades desde la perspectiva de la semántica descriptiva obliga a definir las según unos criterios. Tradicionalmente se ha distinguido entre criterios de clasificación y criterios diagnósticos atendiendo a los supuestos de particularidades específicas reflejados en la tabla 1.

Tabla 1: Diferencias entre criterios de clasificación y criterios diagnósticos. Adaptado de Belmonte-Serrano. *Reumatol Clin.* 2015;11:188. Con permiso de Elsevier.

	Criterios de clasificación	Criterios o procesos diagnósticos
Objetivo	Seleccionar apropiadamente pacientes para ensayos clínicos	Diagnosticar pacientes con una enfermedad determinada
Número de ítems	Pocos, los imprescindibles para seleccionar bien a los candidatos para estudios clínicos	Todos los datos diagnósticos disponibles que permitan el diagnóstico del paciente
Selección de ítems	Estudio estadístico/epidemiológico elaborado	A criterio del médico a cargo del paciente
Énfasis	Especificidad (evitar falsos positivos)	Sensibilidad (evitar falsos negativos)
Umbral de criterio	Fijo, bien establecido (cualitativo o ponderado)	Indeterminado/arbitrario
Cohorte resultante de su aplicación	Homogénea	Heterogénea

La argumentación es que los criterios de clasificación se han hecho para seleccionar pacientes que van a ser incluidos en algún ensayo clínico y, por tanto, son criterios donde se busca la mayor certeza y homogeneidad de criterio a fin de obtener poblaciones estables y comparables de un estudio a otro. Los criterios

diagnósticos, por otro lado, serían aquellos que permiten establecer un diagnóstico en pacientes individuales y de uso en la práctica clínica diaria, de modo que estos últimos son generalmente amplios y pretenden reflejar las diferentes características fenotípicas de la enfermedad [12].

De este modo, los criterios de clasificación, a pesar de facilitar la comparación de los resultados de los trabajos, presentan la capacidad de restringir su validez externa. El comportamiento general y la respuesta a las intervenciones de la enfermedad a estudio pueden diferir entre los sujetos que cumplen los criterios de clasificación y los que no los cumplen pero también son diagnosticados del mismo proceso. Esto ocurre porque en general los pacientes que cumplen criterios de clasificación pueden ser diagnosticados de esa enfermedad, de modo que frecuentemente dichos criterios son la base para confirmar el diagnóstico de sospecha. Lo contrario no siempre es cierto: algunos pacientes que no llegan a cumplir criterios pueden ser también diagnosticados usando datos adicionales a los incluidos en los criterios de clasificación [13]. Sin embargo, existe la corriente de pensamiento de que diferenciar entre ambos tipos de criterios supone, a efectos prácticos, una falacia debido al uso generalizado de los criterios de clasificación como criterios diagnósticos [12]. De hecho, aunque sean el producto de la necesidad de alcanzar objetivos diferentes, representan dos extremos de un mismo *continuum* [14]. La distancia entre los criterios diagnósticos y de clasificación depende de diversos factores que incluyen la prevalencia de la enfermedad, el área geográfica en la que se presenta y la prevalencia de los diagnósticos diferenciales lógicos. En las situaciones en las que la etiología de la enfermedad es bien conocida, como en el caso de la hepatitis crónica por virus B, los criterios diagnósticos y de clasificación pueden ser tan similares que se utilicen indistintamente. Del mismo modo, cuando unos criterios de clasificación presentan una sensibilidad y una especificidad perfectas (del 100% en ambos casos), el concepto de criterio de clasificación y el de criterio diagnóstico es sinónimo e identificaría correctamente cada caso. Este es el consenso entre los investigadores y clínicos del campo de las enfermedades

reumatológicas, paradigma de las enfermedades diagnosticables por criterios clínicos [15]. Sin embargo, como se ha referido anteriormente, la realidad es que el fenotipo de una enfermedad no es idéntico en todos los casos de una enfermedad dada. En consecuencia, los criterios de clasificación nunca son perfectos y universales, y dejan una proporción variable de pacientes por diagnosticar. Cumplir unos criterios diagnósticos (o no cumplirlos) no es garantía de presentar (o de poder descartar) dicho diagnóstico. Solo el médico, a la luz de la información recogida del proceso de un paciente individual, y después de considerar otros factores (como la prevalencia en su medio de trabajo de las condiciones que entran en el diagnóstico diferencial), puede emitir un juicio diagnóstico. Es un proceso cognitivo de alta complejidad que requiere la síntesis de abundantes datos de todo tipo, que en el caso de la hepatitis autoinmune son fundamentalmente epidemiológicos, clínicos, y de exploraciones complementarias [16]. Esta dificultad intrínseca conlleva que sea imposible establecer unos criterios diagnósticos que funcionen como un algoritmo sencillo. En base a esto se propone renunciar a sugerir criterios diagnósticos y considerar que la mejor aproximación que se puede dar es a través de criterios de clasificación [17]. Sería como considerar el proceso de diagnóstico médico en sí mismo como un proceso de clasificación en el cual se parte de un conjunto de datos para realizar un constructo teórico al que damos un nombre de enfermedad. Los criterios corresponderían a un subconjunto limitado de las manifestaciones que pueden presentarse en una enfermedad, ya que en los criterios de clasificación suelen eliminarse los elementos que son redundantes o que presentan colinealidad (correlación muy estrecha entre sí), así como las manifestaciones tardías o infrecuentes de enfermedad.

En resumen, si en una población dada, el diagnóstico de una enfermedad se puede llevar a cabo con una validez interna y externa suficiente, los criterios de clasificación pueden llegar a ser diagnósticos [14]. Ello dependerá de la epidemiología de la enfermedad y de la bondad de los criterios. En el caso de la hepatitis autoinmune pediátrica es posible que, si los criterios de clasificación

propuestos por las sociedades científicas funcionan con un adecuado rendimiento, puedan funcionar como criterios diagnósticos.

Por todo lo expuesto previamente, en el presente trabajo no se reconocerá el matiz diferencial entre los dos términos y se utilizarán indistintamente. Ambos harán referencia al concepto de “criterio de clasificación” recogido tradicionalmente en la bibliografía.

#### **4.1.3. Desarrollo de criterios de clasificación por modelización a través de técnicas de regresión con intención predictiva**

Los modelos de regresión multivariantes se utilizan ampliamente en la investigación de ciencias de la salud. Con frecuencia, el objetivo en la recolección de datos obedece al afán de explicar las interrelaciones que existen entre ciertas variables o a determinar los factores que afectan a la presencia o ausencia de un episodio adverso determinado. Es ahí donde los modelos de regresión multivariantes pasan a ser un instrumento útil, al suministrar una explicación matemática simplificada de dicha relación [18].

Los modelos de regresión tienen en general una estructura común que sigue el siguiente patrón:

$$\text{Respuesta} = \text{Término de error normal} + \text{Ponderación}_1 \times \text{Predictor}_1 + \dots + \text{Ponderación}_i \times \text{Predictor}_i$$

La variable *respuesta* es la variable dependiente y el resto de variables (*predictores* en la expresión anterior) son las variables independientes. Frecuentemente la función de las variables independientes es servir de covariables a una de ellas considerada de mayor peso (o el objeto del estudio) para proporcionar un ajuste estadístico que minimice el desequilibrio entre la variable explicativa central y el resto de factores de influencia o pronósticos. Es el caso de los modelos de regresión con intención explicativa que intentan medir la magnitud de un efecto. Sin embargo, a veces, la identificación de la relación entre varios posibles factores

predictores constituye el objetivo principal del estudio (servir de modelo pronóstico o predictivo), en cuyo caso cada variable independiente pasaría a ocupar la función de variable de interés.

Un caso particular del estudio de esta interrelación lo constituye la predicción de la probabilidad de diagnóstico de una enfermedad concreta en base a la presencia, ausencia o valor de determinadas variables. La modelización a través de estrategias de regresión logística se ha empleado con este objetivo en numerosas situaciones, incluyendo el diagnóstico de la hepatitis autoinmune [19]. El motivo por el que se emplea un análisis logístico es que la variable dependiente en este caso es una variable con un comportamiento binario (sí/no): el diagnóstico se puede o no se puede establecer. Para poder emplear esta variante de regresión multivariable se necesita conocer si el evento que se intenta predecir (la enfermedad) está presente en cada individuo al final del estudio. En principio no sería adecuado utilizarla cuando el tiempo hasta que ocurre el criterio de valoración representa una característica importante del diseño, para lo cual se debería de emplear el método de regresión de Cox. Solo cuando la duración del seguimiento de la cohorte de estudio sea corta o la proporción de observaciones censuradas sea mínima y similar en los dos niveles de una variable explicativa, la regresión logística se puede considerar una alternativa a la regresión de Cox [20].

El objetivo final será obtener un modelo simplificado que tenga sentido desde una perspectiva biológica, se atenga estrechamente a los datos disponibles y aporte predicciones válidas al aplicarlo a datos independientes. Para estos tipos de modelos, el investigador debe establecer un equilibrio entre el grado de complejidad (y exactitud) y su simplicidad; en otras palabras, balancear la exactitud con que el modelo se ajusta matemáticamente a los datos usados para su derivación frente a su capacidad de generalizar las predicciones a poblaciones externas. Modelos complejos (por ejemplo, aquellos con interacciones múltiples, número excesivo de predictores o predictores continuos que muestran un patrón de riesgo no lineal) tienden a reproducirse pobremente en poblaciones diferentes de la usada en su

creación. Se han propuesto varias recomendaciones para la elaboración de este tipo de modelos [18,21–24]. A continuación, se resumen las más importantes, que serán tenidas en cuenta en el bloque correspondiente de esta tesis:

a) Incorporar la mayor cantidad posible de datos exactos, con distribución amplia en los valores de los predictores.

b) Imputar datos si es necesario por una cantidad de valores perdidos en cada variable superior al 5%, ya que mantener un adecuado tamaño de la muestra es de vital importancia. Esto es especialmente relevante para las pérdidas aleatorias, aunque ante el desconocimiento de la naturaleza de la pérdida, también se recomienda la imputación múltiple a través de algoritmos matemáticos complejos. La presencia de una proporción significativa de valores perdidos ocasiona una pérdida de la potencia estadística y un sesgo en el análisis. Si los valores perdidos superan el 50% del tamaño muestral, se debería de renunciar a operar con esa variable.

c) Especificar de antemano la complejidad o el grado de no linealidad que deberá permitirse para cada predictor.

d) Limitar el número de interacciones e incluir solamente las pre-especificadas y basadas en cierta plausibilidad biológica.

e) Seguir la regla de incluir 10 – 15 eventos de la variable respuesta por cada variable independiente para criterios de valoración binarios, con el fin de evitar la sobresaturación del modelo, y si esto no es posible, utilizar técnicas para la simplificación (o reducción) de los datos.

f) Tener presentes los problemas asociados al uso de las estrategias de selección escalonada; en caso de utilizarlas, preferir la eliminación retrógrada y establecer el valor de  $p = 0,157$  que es equivalente a usar el criterio de Akaike como regla de detención; en caso de muestras pequeñas, relajar aún más la regla de detención ( $p = 0,25 - 0,5$ ) con el fin de no ignorar predictores importantes; utilizar el conocimiento previo como guía en la selección de las variables siempre que sea posible.



g) Como alternativa a las estrategias de selección escalonada se pueden utilizar métodos iterativos que calculen indicadores de capacidad predictiva para todos los posibles modelos (como el  $C_p$  de Mallows o el criterio de Akaike en los modelos de regresión con esta finalidad). Existen paquetes estadísticos comerciales o librerías de códigos para entornos de distribución libre que permiten esta posibilidad.

h) Verificar el grado de colinealidad entre los predictores importantes y utilizar la experiencia y la información que se tenga del tema para decidir qué predictores colineales deben ser incluidos en el modelo final.

i) Validar el modelo final con relación a parámetros de calibración y discriminación, preferiblemente utilizando técnicas de remuestreo (*bootstrapping*) cuando no se pueda hacer una validación externa con otra muestra.

j) Utilizar métodos para la simplificación o reducción de datos si la validación interna muestra predicciones excesivamente optimistas.

k) Emplear, en el caso de los modelos de regresión logística, la prueba de Homer-Lemeshow como medida de calibración y el área bajo la curva de características operativas del receptor (COR) para estudiar su capacidad discriminante.

#### **4.1.3.1. Transformación en un sistema de puntos de un modelo de regresión para predecir un diagnóstico**

Un modelo de regresión con intención predictiva tiene una expresión que puede ser transformable en un sistema de puntos con valores enteros. Esto es particularmente fácil si las variables predictoras son categóricas binarias. Esta metodología de transformación de un modelo de regresión en un sistema de puntos sencillo se ha empleado demostrando una buena concordancia con los diagnósticos del modelo multivariable del estudio Framingham [25]. En el caso de una regresión logística se procede según los pasos que se describen a continuación.

#### 4.1.3.1.1. Estimación de los parámetros del modelo multivariable

Considerando el modelo  $f(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$  donde  $Y$  es la variable resultado o dependiente y, en el caso de una regresión logística,  $Y=0$  indica la ausencia de efecto e  $Y=1$  indica su presencia,  $f(Y)$  es una función de  $Y$  que puede ser representada como la combinación de los factores de riesgo  $X_i$ , y  $X_1 \dots X_p$  son los factores de riesgo candidatos y  $\beta_0, \beta_1 \dots \beta_p$  son los estimadores de los coeficientes de regresión basados en el modelo de regresión apropiado.

#### 4.1.3.1.2. Organización de los factores de riesgo en categorías y establecimiento de los valores de referencia

Si un factor de riesgo es una variable continua, hay que establecer clases contiguas y determinar el valor de referencia para cada una de ellas. Por ejemplo, si  $X_1$  es el título de un autoanticuerpo cuyo ancho de valores va de 1:0 a 1:80 (0 a 80), se pueden usar las categorías 0 – 9, 10 – 19, 20 – 29, 30 – 39... 70 – 79. Para determinar los puntos que se asignaran a cada categoría, es necesario especificar un valor de referencia para cada una de ellas. Las medianas de todos los valores posibles entre los límites del intervalo son generalmente valores de referencia aceptables. En el ejemplo previo serían 4,5, 14,5, 24,5, 34,5...74,5 respectivamente. Una excepción no infrecuente se da cuando existen valores extremos. En la variable anterior, si el grueso de los valores de la titulación del autoanticuerpo está entre 0 y 80, pero existe una posible categoría >80, habría que marcar como valor de referencia para la última categoría el punto medio entre 80 y el valor más extremo. Si es 120, se trataría de 100,5. Para otra variable en la que la categoría con límite indeterminado fuera la primera de la serie, el procedimiento sería el mismo. Si el factor de riesgo está modelado como una serie de variables *dummy* en el paquete estadístico (codificadas como 0 = ausente y 1 = presente) que reflejan las distintas categorías del factor de riesgo inicialmente continuo, el valor de referencia es 0 o 1.

Si el factor de riesgo es binario y está incluido en el modelo como una variable indicadora (0 = ausente y 1 = presente), igual que como en las variables

*dummy*, el valor de referencia es cualquiera de estos dos posibles y no hace falta llevar a cabo ningún otro procedimiento.

En adelante  $W_{ij}$  denotará el valor de referencia para la categoría de orden  $j$  del factor de riesgo  $i$  (o variable explicativa  $i$  para los modelos con esta intención), donde  $i = 1, \dots, p$ , y  $j = 1, \dots, c_i$ , donde  $c_i$  es el número total de categorías para la variable o factor de riesgo  $i$ .

#### **4.1.3.1.3. Definición de las características de la referencia para cada variable pronóstica o factor de riesgo**

A continuación, se determina la categoría apropiada para constituirse como categoría basal o de referencia de cada factor de riesgo. A esta categoría basal se le asignarán 0 puntos en el sistema de puntuación. En general, las categorías que reflejen peores estados dentro de la variable predictiva (estados de *menos salud*) obtendrán puntuaciones positivas, tanto más cuanto mejor reflejen un alejamiento de la normalidad. También se pueden asignar puntuaciones negativas en el caso de que reflejen un estado de salud óptimo o un diagnóstico alternativo al de la variable respuesta para la que se quiere construir el sistema de diagnóstico por puntos. En el caso de los criterios clásicos de clasificación para la hepatitis autoinmune, como se verá más adelante, existen ejemplos de ambas situaciones.

El valor de referencia de la categoría basal se expresa como  $W_{iREF}$ , para cada factor de riesgo  $i$ , siendo  $i = 1, \dots, p$ .

#### **4.1.3.1.4. Establecimiento de la distancia de cada categoría respecto a la de referencia en unidades de regresión**

Posteriormente, para cada factor de riesgo o variable pronóstica, se determina la distancia de cada categoría respecto a la categoría de referencia  $W_{iREF}$ , en términos de unidades de regresión. Específicamente se lleva a cabo la siguiente operación sobre cada categoría  $j$  de cada variable  $i$ :  $\beta_1(W_{ij} - W_{iREF})$ , donde  $i = 1, \dots, p$ , y  $j = 1, \dots, c_i$ .

#### 4.1.3.1.5. Establecimiento del multiplicador fijo o constante B

La constante B es el número de unidades de regresión que reflejan 1 punto en sistema de puntos final. Habitualmente se utiliza la constante  $\beta$  de una variable con peso relevante en la ecuación final, que tenga un efecto intuitivo o con plausibilidad reconocida sobre la variable respuesta o la probabilidad del diagnóstico final. En el caso de variables categóricas binarias la constante B es igual a la constante  $\beta$ . Si se trata de una variable cuantitativa,  $B = X_i \cdot \beta_i$ , siendo X un valor de la variable  $i$  que se puede elegir arbitrariamente por el investigador. En el desarrollo del sistema de puntos del estudio Framingham se fijó un multiplicador fijo equivalente al aumento de riesgo asociado a un aumento de 5 años en la edad. En este ejemplo,  $\beta_1 = 0,05$ , por lo que  $B = (5)0,05 = 0,25$  [25].

#### 4.1.3.1.6. Determinar el número de puntos de cada categoría de los factores de riesgo o variables pronósticas

Los puntos para cada categoría de cada factor de riesgo vienen explicados por el siguiente término:

$$\text{Punto}_{Sij} = \beta_1(W_{ij} - W_{iREF}) / B$$

A la categoría basal de cada factor de riesgo (o a  $X=0$  en el caso de las variables categóricas binarias) se le asignan 0 puntos utilizando esta fórmula.

#### 4.1.3.1.7. Estimar los riesgos asociados a la puntuación total

Para cada variable independiente se calcula el total de puntos que admite. El paso último en la confección del sistema de puntuación es llevar a cabo las estimaciones del riesgo de presentar la variable dependiente asociada a cada puntuación ( $\hat{p}$ ). Este paso es específico del tipo de modelo multivariable empleado y para ello requiere de la fórmula concreta de cada modelo. En el caso de los modelos de regresión logística con intención predictiva, que sería el método a emplear para el desarrollo de un sistema de clasificación con objetivo diagnóstico, la fórmula que estima este riesgo es la siguiente:

$$\hat{p} = \frac{1}{1 + e^{-\sum_{i=0}^p \beta_i X_i}}$$

La idea básica del sistema de puntos es aproximar la contribución de cada factor de riesgo o variable pronóstica a la estimación del riesgo, en el caso que nos ocupa, de clasificarse como enfermo de un determinado proceso morboso. Específicamente estimar  $\sum_{i=1}^p \beta_i X_i$ , que es el componente inmediato que se obtiene a partir del modelo de regresión. Cabe destacar que el sistema de puntos no incluye un apartado específico para la constante del modelo de regresión. Si queremos estimar  $\sum_{i=0}^p \beta_i X_i$ , necesitamos hallar el estimador de la constante,  $\beta_0$ .

## 4.2. La enfermedad hepática autoinmune en pediatría

---

### 4.2.1. Definiciones

La enfermedad hepática autoinmune se define serológicamente por niveles aumentados de inmunoglobulina G (IgG) y la presencia de autoanticuerpos, e histológicamente por la presencia de hepatitis de interfase (infiltrado denso mononuclear y linfoplasmocitario del tracto portal con invasión hacia el parénquima), en ausencia de una etiología conocida [26–28].

De acuerdo con esta definición, existen en pediatría dos entidades donde el daño hepático es motivado por una reacción de tipo autoinmune: la hepatitis autoinmune “clásica” (HAI) y el síndrome de solapamiento (*overlap syndrome* en la voz inglesa) entre HAI y la colangitis esclerosante, que también se conoce como colangitis esclerosante autoinmune (CEAI) [29,30]. Dentro de la HAI, asimismo, se reconocen dos perfiles serológicos en función de los autoanticuerpos que se detecten: la HAI tipo 1, en la que se encuentran anticuerpos antinucleares (ANA) y/o anti-músculo liso (anti-Sm); y la HAI tipo 2, que se caracteriza por la presencia de anticuerpos anti-microsomales de hígado/riñón de tipo 1 (anti-LKM1) o anti-citosol hepático de tipo 1 (anti-LC1).

### 4.2.2. Hepatitis autoinmune

Esta entidad fue inicialmente descrita en 1950 por el profesor sueco Jan Waldenström en un grupo de mujeres jóvenes. Desde el inicio y clásicamente se ha concebido como una hepatopatía inflamatoria progresiva y una causa importante de enfermedad hepática terminal. Los pacientes sufrían una forma grave de afectación hepática sin mejoría espontánea caracterizada por una elevación marcada de los niveles séricos de inmunoglobulinas [31]. A raíz de la primera descripción se

comunicaron diversas cohortes de pacientes con la misma presentación morbosa, cuyo estudio y descripción permitió establecer la observación de que una proporción importante de casos presentaban células de lupus eritematoso sistémico (LES) en sangre periférica (célula fagocítica del sistema inmune que ha fagocitado el material nuclear desnaturalizado de algún otro tipo de célula, por lo general un macrófago o un neutrófilo) y ANA. Esto condujo a la hipótesis, por analogía, de que una pérdida de la tolerancia inmunológica constituía el fundamento patogénico de esta condición [32].

#### **4.2.2.1. Epidemiología**

Se desconoce la prevalencia global exacta de la HAI al no disponer de estudios epidemiológicos más allá de en poblaciones concretas ni con un diseño adecuado que defina claramente los casos según el sistema de clasificación clínica adoptado por el Grupo Internacional para el Estudio de la Hepatitis Autoinmune (*International Autoimmune Hepatitis Group, IAIHG*) [33]. De hecho, a los primeros trabajos realizados con el fin de estimarla, se les critica que están sesgados por la posible inclusión de pacientes con hepatitis crónica por virus C [34]. La metodología utilizada consiste en búsquedas retrospectivas de información en las Unidades de Documentación de grandes hospitales terciarios, lo que implica frecuentemente la imposibilidad de comprobar la presencia o ausencia de todos los criterios clásicos del IAIHG, principalmente la respuesta al tratamiento y los datos histopatológicos [35].

En líneas generales, los estudios llevados a cabo en Europa y Norteamérica señalan unas prevalencias medianas dentro del rango de 11 a 35 casos por cada 100.000 habitantes [36–38]. La HAI afecta a todos los grupos étnicos, aunque predomina en caucásicos, y característicamente muestra una fuerte predisposición a afectar a mujeres, con un ratio de casos mujer/hombre de aproximadamente 4/1 [39,40].

La primera publicación en la que se utilizaron los criterios clásicos revisados del IAIHG informó de una prevalencia pico de 42,9(IC95% 31 – 57,7) casos por cada 100.000 habitantes en nativos de Alaska. Se consideraban únicamente pacientes adultos. El periodo de estudio abarcó desde 1984 hasta 2000 y se encontraron 77 pacientes con posible HAI, que en 42 pacientes fue clasificada como definitiva por el sistema de puntuación, y en 7 como probable [37].

En un estudio epidemiológico de 2010 llevado a cabo en Nueva Zelanda con población infantil y adulta, se informó de una incidencia anual de 2 casos nuevos (IC95% 0,8 – 3,3) por cada 100.000 habitantes en un año (2008). La prevalencia máxima fue en diciembre de ese mismo año, con 24,5 casos (IC95% 20,1 – 28,9) por cada 100.000 habitantes. El análisis estandarizado por edad (reparto etario según informes de la Organización Mundial de la Salud, OMS) demostró una incidencia de 1,7 casos nuevos/100.000 habitantes/año y una prevalencia de 18,9 casos/100.000 habitantes [38].

Más recientemente se ha publicado un trabajo en población danesa con el objetivo, además de definir el pronóstico y las causas de mortalidad relacionadas con la HAI, de estudiar su epidemiología. Se basa en un registro nacional de casos recogidos entre 1994 y 2012 que incluye anatomía patológica en todos los casos, a diferencia de trabajos previos. Halla una tasa de incidencia global de 1,68 casos nuevos/100.000 habitantes/año (IC95% 1,60 – 1,76) durante todo el periodo analizado. Sin embargo, comparando el último año registrado con respecto al primero, se observa un aumento de la incidencia de prácticamente el doble: de 1,37 (IC95% 1,06 – 1,76) a 2,33 casos nuevos/100.000 habitantes/año (IC95% 1,95 – 2,77), con una prevalencia máxima puntual en 2012 de 24 casos por cada 100.000 habitantes (35/100.000 considerando solo niñas y mujeres) [41]. La interpretación más obvia para este hallazgo en una enfermedad infrecuente como la HAI es un aumento de la capacidad diagnóstica de la misma en los últimos años. Sin embargo hay algunos datos que señalan que se trata de un aumento de la incidencia real: el diagnóstico por criterios clínicos consolidados desde más de 10 años antes, la



universalidad del acceso a la atención sanitaria del entorno en el que se desarrolló el estudio y, sobre todo, la similitud de la distribución de los estadios histopatológicos encontrada durante todo el periodo de estudio [42]. En esta línea informa de una prevalencia al diagnóstico de cirrosis de 28,3%, sin analizar por separado población adulta y pediátrica. La metodología empleada permite calcular el cociente de riesgo (*hazard ratio*) de mortalidad con respecto a población normal dado que reclutan una cohorte de comparación de pacientes sanos emparejados por edad y sexo. De este modo se observa que en el primer año tras el diagnóstico los pacientes con HAI tienen una mortalidad 4 veces mayor que la de la población general, que se reduce a 2 veces superior posteriormente (con un 3,6% de mortalidad acumulada por carcinoma hepatocelular a los 10 años) [41].

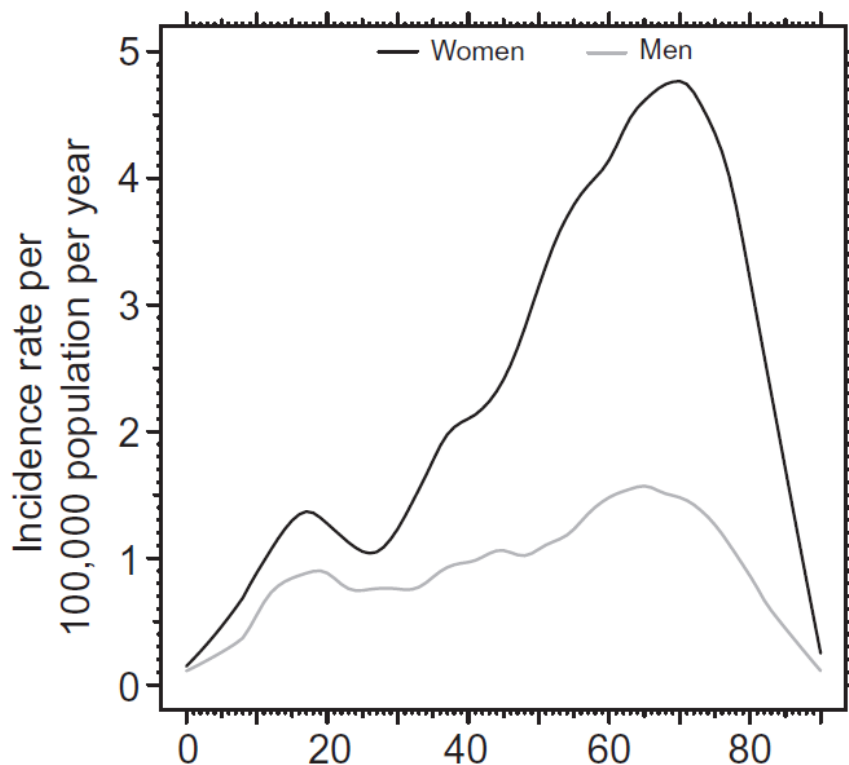


Figura 1: Tasa de incidencia según edad y sexo de la hepatitis autoinmune en Dinamarca (1994-2012). Reproducido de Grønbaek et al. *Journal of Hepatology*. 2014;60:613. Con permiso de Elsevier.

Por lo que respecta a la distribución entre los dos tipos serológicos de HAI, también se desconoce su magnitud real y son varios los trabajos que concluyen que la tipo 2 puede estar infra-representada en las series [43]. Las HAI tipo 2 son precisamente más características del niño y el adulto joven.

El equipo de hepatología pediátrica del King's College Hospital ha descrito un aumento de la incidencia de hasta 7 veces tanto en la HAI tipo 1 como en la 2 en la última década, llegando a suponer un motivo de derivación a dicho centro de referencia en 400 casos anuales, de los cuales un tercio son de tipo 2 y el resto de tipo 1 [44].

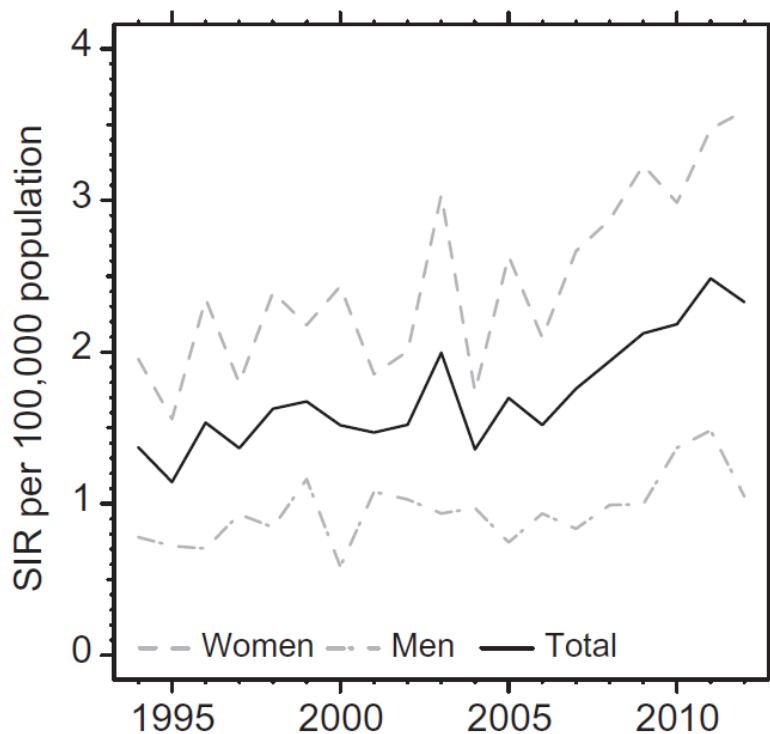


Figura 2: Tasa de incidencia estandarizada por edad y sexo en Dinamarca (1994-2012). Reproducido de Grønbaek et al. *Journal of Hepatology*. 2014;60;614. Con permiso de Elsevier.

#### 4.2.2.2. *Presentación clínica e historia natural*

El espectro de presentaciones clínicas de la HAI abarca una constelación variable de síntomas [45], aunque en adultos la mayoría de los pacientes muestran

un debut clínico insidioso caracterizado por fatiga progresiva, ictericia recurrente, amenorrea, pérdida de peso y, más ocasionalmente, artralgias [46]. Aunque es infrecuente, también pueden ser aparentes las complicaciones de la hipertensión portal secundaria al daño crónico hepático, es decir, sangrado gastrointestinal o hiperesplenismo [29].

Aproximadamente el 25% de los pacientes están asintomáticos y se diagnostican de forma incidental al objetivarse hipertransaminasemia u otros marcadores de citólisis o función hepática alterados. Por otro lado, de un 30 a un 40% de los casos, particularmente niños, adolescentes y adultos jóvenes, se presentan con síntomas y signos que simulan los de cualquier hepatitis aguda, fundamentalmente astenia y anorexia. Los elementos clínico-analíticos más definitorios de HAI en contraste con las hepatitis víricas son el predominio femenino, una hipertransaminasemia con una relación AST/ALT aumentada y la hipergammaglobulinemia (tabla 1) [47].

**Tabla 2: Comparación epidemiológico-analítica entre la hepatitis autoinmune de presentación aguda y las hepatitis víricas agudas. Reproducido de Ferrari et al. QJM. 2004;97:407-12. Con permiso de Oxford University Press.**

	HAI aguda (n = 22)	Hepatitis vírica aguda (n = 41)	Valor p de la diferencia
Edad (años)	39,5 ± 20,2	33 ± 13,1	NS
Mujeres	18 (82%)	10 (24%)	<0,0001*
AST (×LSN)	29,11 ± 16,8	25,9 ± 19,9	NS
ALT (×LSN)	25,33 ± 13,41	40,6 ± 28,3	<0,05**
Relación AST/ALT	1,20 ± 0,55	0,61 ± 0,2	<0,0001**
Bilirrubina (mg/dl)	7,89 ± 6,16	10,1 ± 5,8	NS
Fosfatasa alcalina (×LSN)	1,59 ± 0,79	1,85 ± 0,89	NS
GGT (×LSN)	3,66 ± 3,18	5,1 ± 3,8	NS
γ-Globulina (g/l)	26,9 ± 10,8	13,4 ± 4	<0,0001**
IgG (×LSN)	1,64 ± 0,59	0,76 ± 0,2	<0,0001**
IgA (×LSN)	0,81 ± 0,41	0,63 ± 0,1	<0,05**
IgM (×LSN)	0,81 ± 0,55	0,94 ± 0,5	NS

Valores expresados en media ± desviación estándar y en porcentajes

\*test exacto de Fisher

\*\*test t para muestras independientes

NS: No significativo

La forma de debut como fallo hepático agudo fulminante es muy infrecuente [34]. Un estudio de 2004 revela que el 8,7% de los casos considerando todos los grupos de edad se presentan de forma aguda, siendo el fallo hepático fulminante el tipo de presentación aguda mayoritario. El grupo de niños con una HAI seropositiva para anti-LKM1 es en el que es más frecuente el debut como fallo hepático agudo, que se puede producir hasta en 5% de los mismos [48].

Con independencia del modo de presentación, se encuentra evidencia histológica de cirrosis en al menos un 30% de pacientes, observación que se interpreta como que ha habido enfermedad subclínica durante el tiempo suficiente como para que la fibrosis evolucione hacia dicho estadio. Además, grados avanzados de fibrosis o cirrosis se pueden encontrar también en casos con presentación aguda [49].

Las diferencias entre las formas clínicas mencionadas de la HAI tipo 1 fueron estudiadas por Ferrari *et al.* y publicadas en 2004 [47]. Tal como se puede comprobar en la tabla 2, los rasgos característicos de cada grupo patocronológico son fundamentalmente analíticos, con una marcada elevación de los marcadores de citólisis y colestasis en las formas agudas respecto al resto.

#### **4.2.2.2.1. Exploración clínica**

Los hallazgos en la exploración suelen reflejar la duración y la gravedad de la enfermedad. El hallazgo más común es la hepatomegalia, que comúnmente no es excesivamente llamativa. La esplenomegalia suele acompañar a la cirrosis, pero también se ha observado que puede reflejar el síndrome inflamatorio. La ictericia es un hallazgo usualmente presente en los casos con inicio agudo. Si ya hay degeneración cirrótica pueden ponerse en evidencia los estigmas hepáticos como arañas vasculares, eritema palmar o ascitis. Estos hallazgos potenciales en la exploración clínica, al reflejar el tiempo de evolución y la gravedad, se encuentran en el 75% de los pacientes. El resto, incluyendo los casos infantiles, presentan una

exploración normal, incluyéndose en este grupo de pacientes aquellos en los que la enfermedad se detecta de forma fortuita [50].

**Tabla 3: Características clínicas y analíticas de la hepatitis autoinmune según la cronopatología: aguda, crónica y asintomática. Reproducido de Ferrari et al. QJM. 2004;97:407-12. Con permiso de Oxford University Press.**

	HAI aguda (n=22)	HAI crónica (n=59)	HAI asintomática (n=5)	Valor p de la diferencia
Edad (años)	39,5±20,2	44,2±19,5	48,4 ± 16,7	NS
Mujeres	18 (82%)	52 (88%)	3 (60%)	NS
AST (×LSN)	29,11 ± 16,08	9,09 ± 11,46	2,80 ± 2,77	<0,001 aguda vs crónica, <0,001 aguda vs asintomática**
ALT (×LSN)	25,33 ± 13,41	9,36 ± 9,81	2,99 ± 1,20	<0,001 aguda vs crónica, <0,001 aguda vs asintomática**
Bilirrubina (mg/dl)	7,89 ± 6,16	3,61 ± 5,93	0,72 ± 0,35	<0,001 aguda vs crónica, <0,001 aguda vs asintomática**
Fosfatasa alcalina (×LSN)	1,59 ± 0,79	1,41 ± 0,95	1,17 ± 0,58	NS
GGT (×LSN)	3,66 ± 3,18	2,89 ± 2,93	2,61 ± 3,11	NS
Albúmina (g/dl)	3,37 ± 0,65	3,63 ± 0,60	3,92 ± 0,55	NS
γ-Globulina (g/l)	26,9 ± 10,8	27,4 ± 9,6	25,4 ± 7,9	NS
IgG (×LSN)	1,64 ± 0,59	1,75 ± 0,77	1,59 ± 0,62	NS
IgA (×LSN)	0,81 ± 0,41	0,84 ± 0,51	1,08 ± 0,44	NS
IgM (×LSN)	0,81 ± 0,55	1,17 ± 0,71	0,94 ± 0,87	NS
Autoanticuerpos (>1:40)	ANA y SMA 14 (64%), ANA 2 (9%), SMA 4 (18%), negativos 2 (9%)	ANA y SMA 30 (51%), ANA 6 (10%), SMA 18 (31%), negativos 5 (8%)	ANA y SMA 2 (40%), SMA 3 (60%)	NS
Hallazgos histológicos moderados o graves	10/14 (71%)	27/45 (60%)	3/5 (60%)	NS
Cirrosis*	5 (23%)	22 (37%)	1 (20%)	NS
HLA DR3	6/16 (37,5%)	13/37 (35%)	1/2 (50%)	NS
HLA DR4	2/16 (12,5%)	7/37 (19%)	1/2 (50%)	NS

Valores expresados en media ± desviación estándar y en porcentajes

\*Por anatomía patológica o presencia de ascitis o varices esofágicas.

\*\*test de Kruskal-Wallis.

LSN: Límite superior de la normalidad según el laboratorio.

NS: No significativo.

#### **4.2.2.2. Hepatitis autoinmune y embarazo**

Se sabe también que la HAI puede aparecer durante el embarazo, como se señala en un trabajo sobre el manejo de la HAI en gestantes. Los autores revisan las gestaciones conocidas de 162 mujeres con diagnóstico definitivo por los criterios clásicos revisados de la IAIGH. De ellas, dos pacientes, desarrollaron la HAI de novo durante el embarazo [51]. La aparición post-parto y las exacerbaciones de la una enfermedad preexistente en mujeres cuyo grado de actividad mejoró durante la gestación también se han descrito [52].

#### **4.2.2.3. Enfermedades autoinmunes asociadas**

Aproximadamente el 40% de los pacientes pediátricos con HAI tienen una historia familiar de enfermedades autoinmunes y cerca de un 20% se han diagnosticado de entidades de esta naturaleza concomitantemente o lo hacen durante el seguimiento. No existen diferencias en cuanto a la proporción de antecedentes personales ni familiares de enfermedades autoinmunes entre los tipos 1 y 2 de HAI [29]. Las enfermedades asociadas en los casos pediátricos estudiados por Gregorio *et al.* (en una revisión retrospectiva de 52 pacientes en edad pediátrica) fueron: síndrome nefrótico, tiroiditis autoinmune, enfermedad de Behçet, colitis ulcerosa, diabetes mellitus tipo 1, urticaria pigmentosa, vitiligo y enfermedad de Addison combinada con hipoparatiroidismo (síndrome poliglandular autoinmune tipo 1). Por lo que respecta a los antecedentes familiares, las más prevalentes fueron la diabetes mellitus tipo 1 y enfermedades de la glándula tiroides [29]. El vitiligo y el déficit selectivo de IgA van particularmente unidos a los casos de HAI con celiaquía o diabetes tipo 1 [49].

#### **4.2.2.4. Desarrollo de carcinoma hepatocelular**

Al igual que otras hepatopatías crónicas, la HAI puede progresar a una cirrosis y a un carcinoma hepatocelular (CHC) a pesar del tratamiento inmunosupresor.

En un estudio de 2008 en el que se siguió durante 16 años una cohorte retrospectiva de 243 casos de HAI (todas de tipo 1), se diagnosticaron 15 casos de CHC, lo que representa una incidencia de 1090 casos por cada 100.000 HAI/año. El CHC ocurrió en la misma proporción en hombres y en mujeres y su incidencia máxima fue en el grupo de pacientes con cirrosis al diagnóstico. El tiempo mediano entre el diagnóstico de la cirrosis y el del CHC fue de 102,5 meses (rango intercuartílico –RIC– 12 a 195 meses); y el de supervivencia tras detectarse el carcinoma, de 19 meses (RIC 6 a 36 meses) [53].

Existe otro trabajo más reciente y metodológicamente similar en el que se informa de una incidencia menor, de 459 casos por cada 100.000 pacientes/año (16 CHC en el grupo de 322 HAI).

La conclusión de ambos trabajos, refrendada por otras publicaciones, es que es mandatorio realizar un seguimiento a los pacientes con HAI, fundamentalmente a los que desarrollan cirrosis, con el objetivo de hacer una prevención secundaria del CHC [40].

#### **4.2.2.3. Hallazgos de laboratorio**

Los datos del análisis sanguíneo convencional más frecuentes en las HAI son la elevación de aspartato (AST) y alanina (ALT) aminotransferasas. Puede existir una elevación de menor magnitud de bilirrubina total y fosfatasa alcalina (FA). En una proporción menor de pacientes es posible encontrar signos analíticos de colestasis, que obliga a considerar la posibilidad de obstrucción biliar extra-hepática y formas colestásicas de hepatitis víricas, hepatitis tóxica, CBP, CEP y síndromes de solapamiento [45,54,55].

La HAI también se asocia con elevación de las seroproteínas, particularmente las gammaglobulinas, a expensas de IgG [56]. Los autoanticuerpos más típicos son los ANA, anti-Sm, anti-LKM1 y anti-LC1 [57,58]. Aunque los AMA son más específicos de la CBP, también se pueden encontrar ocasionalmente en la HAI. El hallazgo aislado de un resultado positivo para estos autoanticuerpos no permite el

diagnóstico de HAI porque también se ha descrito se elevación en otras enfermedades con base autoinmune o autoinflamatoria. Ninguno de ellos es patognomónico de la HAI [54].

Otras alteraciones analíticas como los defectos de la hemostasia (descenso del INR –*International Normalized Ratio*– o de los valores de factor V de la coagulación, por ejemplo) o la hipoalbuminemia orientan a insuficiencia hepatocelular, que se puede encontrar en el caso del fallo hepático agudo o en grados avanzados de cirrosis.

#### **4.2.2.4. Autoanticuerpos diagnósticos**

Establecer la sospecha clínica de HAI es motivo para solicitar al laboratorio la detección de los autoanticuerpos incluidos en los criterios diagnósticos, tal como indica el documento de consenso del grupo de trabajo de serología autoinmune del IAIHG [59]. Además de servir como criterio de apoyo diagnóstico, el perfil serológico permite clasificar el caso de HAI en tipo 1 (caracterizado por ANA y/o anti-Sm) y tipo 2 (caracterizado por anti-LKM1 y/o anti-LC1). Los autoanticuerpos que definen cada tipo raramente se encuentran en combinación entre ellos, y cuando lo hacen, el curso clínico y el pronóstico es similar al de las HAI tipo 2 [59]. Los anticuerpos anti-LC1, anti-SLA/LP y anti-citoplasma del neutrófilo (ANCA) se incorporaron a los criterios clásicos en la revisión de 1999 al descubrirse su papel como apoyo diagnóstico en los casos seronegativos para los ANA, anti-Sm y anti-LKM1 [33,60,61].

Estos marcadores serológicos se determinan por técnicas de inmunofluorescencia (IF), observando al microscopio el marcado luminoso sobre muestras de tejido control a las que se les ha aplicado el suero problema del paciente [59,62]. El reconocimiento e interpretación de cada patrón de IF es dependiente del operador y por lo tanto no es infrecuente que se comuniquen errores en su lectura, que se complica más por la baja incidencia de la enfermedad hepática autoinmune en comparación con otras entidades [34]. El consenso del comité de serología autoinmune del IAIHG establece que la primera línea de exploraciones



complementarias a pedir para un estudio de HAI incluye una IF indirecta para determinar la presencia en suero de los autoanticuerpos relevantes para dicha sospecha: ANA, anti-Sm, anti-LKM1, anti-LC1 y AMA (tabla 7). La técnica de IF se lleva a cabo sobre tejido en fresco procedente de secciones de varios órganos de ratón, habitualmente hígado, riñón y estómago, sin fijar y secado al aire. La forma de diluir los agentes reveladores marcados con fluorocromos y el suero del paciente también está descrita en el documento de consenso, así como la preparación de la muestra de tejido, la aplicación del suero y ejemplos de patrones relevantes desde el punto de vista diagnóstico [59].

El proceso de análisis por IF de los autoanticuerpos requiere una cuidadosa preparación y orientación de las secciones de riñón para asegurar que contienen túbulos contorneados proximales y distales. El motivo de ello es que, aunque los anti-LKM1 y los AMA tienen afinidad por los túbulos de la nefrona, el primero tiñe de forma predominante la tercera porción de los túbulos proximales más largos, y el segundo marca principalmente las mitocondrias de los túbulos distales. El empleo de tejidos multi-órgano también contribuye a la distinción entre estos autoanticuerpos porque los AMA, a diferencia de los anti-LKM1, tiñen las células parietales gástricas (figura 5).

#### **4.2.2.4.1. Anticuerpos anti-nucleares**

Los ANA son fácilmente detectables por IF y crean un patrón nuclear en células de hígado, riñón y estómago. Bajo esta técnica, en el caso de la HAI es típica la configuración de un teñido homogéneo en los hepatocitos, aunque también se observa en ocasiones un patrón más granular o punteado [59].

En adultos se consideran positivos títulos de 1:40 o superiores, y en niños el dintel baja a 1:20, mostrando una correlación positiva con la actividad de la enfermedad [63].

Con el objetivo de definir mejor el patrón de los ANA se han utilizado células epiteliales humanas de tipo 2 (línea HEP2) debido a que presentan una estructura

nuclear prominente. Sin embargo utilizarlas como primer escalón diagnóstico en la sospecha de HAI muestra un rendimiento subóptimo debido a un índice de resultados positivos excesivamente elevado en sujetos sanos [34].

**Tabla 4: Autoanticuerpos y sus antígenos en la enfermedad hepática autoinmune. Reproducido de Liberal et al. *Autoimmunity Reviews*. 2014;13;438. Con permiso de Elsevier.**

Autoanticuerpo	Antígeno	Hepatopatía	Valor en HAI	Método de detección convencional	Análisis molecular
ANA	Cromatina	HAI	Diagnóstico en HAI 1	IF indirecta	ELISA, IB, LIA
	Histonas	CBP			
	Centrómeros	CEP			
	Ciclina A	Toxicidad			
	Ribonucleoproteínas	HCVC			
	ADN de doble hebra	HCVB			
Anti-Sm	ADN de una hebra	EHNA	Diagnóstico en HAI 1	IF indirecta	ELISA
	Microfilamentos	Igual que ANA			
Anti-LKM1	Filamentos intermedios	ANA	Diagnóstico en HAI 2	IF indirecta	ELISA, IB, LIA, RIA
	Citocromo P4502D6	HAI 2 HCVC			
Anti-LC1	Forminino-transferasa ciclodeaminasa	HAI 2	Diagnóstico en HAI 2 Mal pronóstico	IF indirecta, IDBD, CIEF	ELISA, LIA, RIA
		HCVC			
Anti-SLA/LP	tRNP(Ser)Sec	HAI HCVC	Diagnóstico en HAI Mal pronóstico	ELISA competitiva	ELISA, IB, RIA
pANCA	Proteínas de la lámina nuclear	HAI CEP, CEAI	Apoya diagnóstico HAI	IF indirecta	No
AMA	Subunidades E2 de los complejos de 2-oxoácido deshidrogenasa, particularmente PDC-E2	CBP	En contra de diagnóstico HAI	IF indirecta	ELISA, IB, RIA

ANA: Anticuerpo antinuclear. Anti-Sm: Anticuerpo anti-músculo liso. Anti-LKM1: Anticuerpo anti-microsomal de hígado/riñón de tipo 1. Anti-LC1: Anticuerpo anti-citosol hepático de tipo 1. Anti-SLA/LP: Anticuerpos anti-antígeno soluble hepático/anti-hígado-páncreas. pANCA: Anticuerpo anti-citoplasma del neutrófilo con patrón perinuclear. AMA: Anticuerpo anti-mitocondrial. HAI: Hepatitis autoinmune. CBP: Cirrosis biliar primaria. CEP: Colangitis esclerosante primaria. CEAI: Colangitis esclerosante autoinmune. HCVC/B: Hepatitis crónica por virus C/B. EHNA: Esteatohepatitis no alcohólica. IF: Inmunofluorescencia. IDBD: Inmunodifusión bidimensional. CIEF: Contra-inmunolectroforesis. ELISA: Prueba de inmunoabsorción enzimática. IB: Inmunoblot. LIA: Inmunoensayo lineal. RIA: Prueba de radio-inmuno-precipitación.

Los ANA pueden llegar a encontrarse hasta en un 52% de pacientes con CBP. En este caso se detectan habitualmente con un patrón diferencial respecto al caso de la HAI, con un punteado nuclear o un patrón en ribete perimembranoso. Esta

configuración se reconoce por IF sobre células HEp2. Otras causas de falsos positivos son una variedad de enfermedades de base autoinmune, tales como el LES, el síndrome de Sjögren y las formas sistémicas de esclerodermia. Además, otras condiciones de naturaleza distinta a la autoinmune también pueden cursar con ANA positivos, lo que demuestra su inespecificidad. Son ejemplos de ello algunos casos de hepatitis víricas, hepatitis tóxica y esteatohepatitis inducida por alcoholismo o no [64].

Los antígenos diana de los ANA en la HAI son heterogéneos y todavía están parcialmente definidos. Hasta el momento actual se han determinado ANA con especificidad para hebras de DNA de cadena simple o doble, ribonucleoproteínas pequeñas nucleares, centrómeros, histonas, cromatina y ciclina A. Un conocimiento completo de las dianas de los ANA podría permitir desarrollar antígenos por metodología recombinante y facilitar su detección por inmunoensayo [57].

Los mecanismos por los cuales se producen ANA en la HAI no son bien conocidos. Se especula con que intervienen la liberación de antígenos nucleares resultante de la destrucción hepatocitaria y/o la pérdida de inmunotolerancia por parte de linfocitos B a estos componentes nucleares [64].

#### **4.2.2.4.2. Anticuerpos anti-músculo liso**

Tal como se ha expuesto, los patrones de IF de los anti-Sm se pueden visualizar tanto en células de hígado como en células de estómago o riñón. Los antígenos frente a los que tiene afinidad están presentes en las paredes arteriales y arteriolas pero también se expresan en la muscularis mucosæ y la lámina propia de los cortes de estómago. Los anti-Sm procedentes de pacientes con HAI muestran como característica particular que se depositan en el músculo liso vascular, los glomérulos y los túbulos renales. Es el denominado patrón VGT (vaso, glomérulo y túbulo). Los patrones VGT y VG de IF se consideran de alta especificidad de HAI, en comparación con el patrón V aislado, que también está descrito en hepatopatías de etiología distinta, enfermedades infecciosas y reumatológicas [59,65].

Los títulos de anti-Sm suelen sobrepasar el valor de 1:80 aunque, como en el resto de autoanticuerpos en pacientes pediátricos, se pueden encontrar valores inferiores, de cerca de 1:20.

Los primeros antígenos en identificarse para este marcador fueron constituyentes de filamentos de actina. Con posterioridad también se ha descubierto que otros integrantes del citoesqueleto como la tubulina, la vimentina, la desmina y la esqueletina muestran afinidad por los anti-Sm [57]. Aunque el antígeno específico de los anti-Sm en los casos de HAI tipo 1 no se ha identificado con seguridad, existen pruebas que señalan como principal candidato a la actina en su forma filamentosa (actina F).

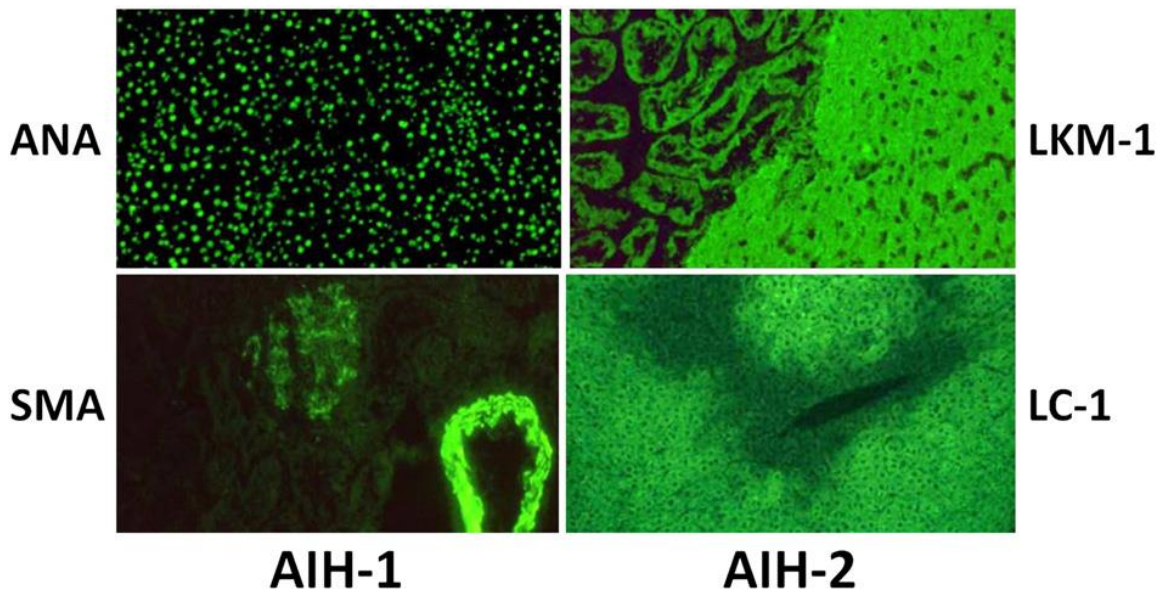


Figura 3: Imagen de inmunofluorescencia indirecta en tejido de roedor de los autoanticuerpos diagnósticos de la hepatitis autoinmune (AIH-1 y AIH-2). ANA: anti-nucleares. SMA: anti-músculo liso (tíñe un vaso, patrón V, y un glomérulo, patrón G). LKM1: anti-microsoma de hígado/riñón de tipo 1. LC-1: anti-citosol hepático de tipo 1. Reproducido de Liberal et al. *Autoimmunity Reviews*. 2014;13;438. Con permiso de Elsevier.

El patrón VGT de la IF es de alta especificidad diagnóstica de HAI tipo 1, pero hasta en un 20% de las formas seropositivas puede estar ausente. Además se sabe que casos de HAI tipo 1, anti-Sm positivos por IF con patrón de afinidad VGT, pueden

darse como falsos negativos por técnicas de biología molecular que empleen actina F purificada. Al igual que los ANA, tampoco son patognomónicos de HAI [58,66,67].

Se ha señalado que los anti-Sm frente a la  $\alpha$ -actina o frente a la actina F pueden ser utilizados como predictores de respuesta al tratamiento. Esta idea se fundamenta en la observación de que los pacientes con respuesta incompleta a los inmunosupresores o con recaídas precoces muestran al diagnóstico valores de dichos autoanticuerpos sensiblemente superiores al de lo que logran remisión mantenida de la actividad de la enfermedad [68].

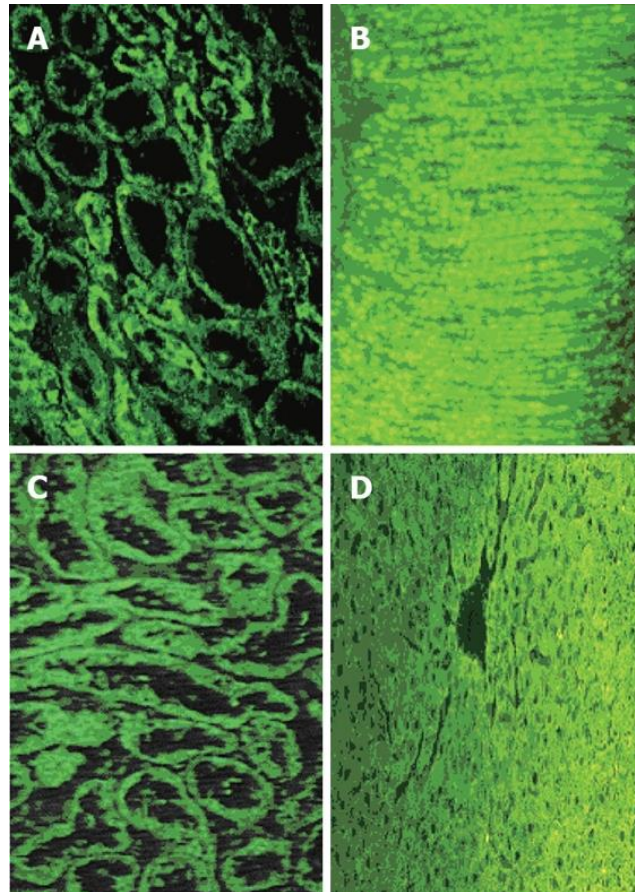
#### **4.2.2.4.3. Anticuerpos anti-microsomales de hígado/riñón de tipo 1**

Por IF, los anti-LKM1 (al autoanticuerpo distintivo de la HAI tipo 2) tiñen el citoplasma hepatocelular y la porción P3 de los túbulos renales. El hecho de presentar afinidad por estos órganos no es específico de los anti-LKM1 porque también puede observarse en los AMA, que a efectos de diagnóstico de HAI tiene una interpretación diferente (de hecho, otorga una puntuación negativa en los criterios clásicos revisados de 1999, apuntando la posibilidad de CBP). El rasgo diferencial entre estos anticuerpos es que el marcado hepático de los AMA es de menor intensidad y el marcado renal es más difuso pero acentuado en los túbulos distales. A esto se añade el que los AMA se detectan por IF sobre tejido de mucosa gástrica [34,69]. Además, se han descrito dos casos de HAI pediátrica en la literatura con hallazgo por IB de AMA de tipo 2 en ausencia de enfermedad biliar y otros autoanticuerpos, con lo que el papel de estos autoanticuerpos puede ser más complejo de lo que se sabe actualmente [70].

Se consideran títulos relevantes de anti-LKM1 aquellos por encima de 1:40 y existe evidencia de su relación con la actividad de la HAI [59]. Al igual que los ANA y anti-Sm, no es patognomónico de esta enfermedad y se puede llegar a detectar en un 5 – 10% de los casos de hepatitis crónica por virus C [34,71].

La diana molecular de los anti-LKM1 es el citocromo P450-2D6 (CYP2D6) y ha sido posible purificarlo de muestras biológicas y desarrollarlo como autoantígeno

comercial por tecnología recombinante. Ello ha permitido que se puedan detectar por inmunoabsorción enzimática (ELISA) con suficiente buen rendimiento para los casos de HAI tipo 2, en los que los títulos suelen ser más altos que otras enfermedades [72]. En los casos de patrón dudoso por IF para los anti-LKM1, por tanto, es posible recurrir a técnicas de biología molecular [34].



**Figura 4:** Inmunofluorescencia de los anticuerpos anti-mitocondriales (AMA) (Paneles A y B) y de los anti-microsoma de hígado/riñón (anti-LKM1) (Paneles C y D). Los AMA se detectan con más intensidad en los túbulos distales de la nefrona (A) y las células parietales gástricas (B) mientras que los anti-LKM1 lo hacen en los túbulos proximales (C) y los hepatocitos de ratón (D). Reproducido de Bogdanos et al. *World J Gastroenterol.* 2008;14;3377. Con permiso de Baishideng Publishing Group.

Se ha comunicado un caso de elevación transitoria combinada de anti-LKM1 y AMA en un caso pediátrico de HAI tipo 2, sin evidencia de lesiones típicas de CBP. Esta última es infrecuente en niños, mientras que la HAI tipo 2 es típica de la edad

infantil. La respuesta al tratamiento en este paciente fue óptima y no se comportó de forma diferente al curso habitual de las HAI [73]. Esta situación poco común es, por consiguiente, posible, y apoya la idea de que la serología de forma aislada no puede constituir un criterio suficiente para el diagnóstico o descarte de la HAI.

#### **4.2.2.4.4. Variantes de los anticuerpos frente a microsoma hepático**

Los anticuerpos anti-microsomales hepáticos están dirigidos principalmente frente a isoformas no-2D6 del citocromo P450, a diferencia de los anti-LKM1.

Los anti-microsoma hepático solo tiñen por IF el citosol del hepatocito, muestran afinidad por el citocromo específico de hígado P4501A2 y aparecen en la hepatitis inducida por dihidralazina y en la hepatitis asociada al síndrome APECED (*autoimmune polyendocrinopathy – candidiasis – ectodermal dystrophy* o poliendocrinopatía autoinmune tipo 1) [74].

También los anticuerpos frente a P4502A6, que se encuentran asimismo en el síndrome APECED y en algunos casos de hepatitis por VHC (virus de la hepatitis C), muestran un patrón de IF similar al de los anti-LKM1.

El término anti-LKM2 fue acuñado para describir los anticuerpos frente a antígenos microsomales dirigidos contra el citocromo P4502C9, que se producían en el contexto de la hepatitis tóxica inducida por el ácido tielínico, un antihipertensivo que actualmente está desaprobadado y retirado de la comercialización [75].

Por último, existen los anti-LKM3, que se producen frente a enzimas de la familia de las 1-UDP glucuronil transferasas. Este autoanticuerpo se describe principalmente en las hepatitis  $\delta$  y solo en una proporción menor de HAI [76,77].

#### **4.2.2.4.5. Anticuerpos anti-citosol hepático de tipo 1**

Los anti-LC1 se describieron originalmente como autoanticuerpos definitorios de HAI tipo 2, aisladamente o en combinación con anti-LKM1. Constituye uno de los criterios clásicos, incorporado después de la revisión de 1999, y continúa siendo indicativo del tipo 2 de las formas de HAI, a pesar de que también se ha encontrado

en pacientes con hepatitis crónica por VHC y en combinación con marcadores serológicos de HAI tipo 1 [61,78].

Los anti-LC1 también se pueden considerar como marcadores del grado de inflamación hepática dado que su título guarda una relación de proporcionalidad directa con la actividad de la enfermedad [57].

El antígeno de los anti-LC1 es una enzima del metabolismo del ácido fólico, la forminino-transferasa ciclodeaminasa (FTCD), que se encuentra a concentraciones altas en tejido hepático. Por lo que respecta a la imagen de IF, los anti-LC1 tiñen el citoplasma del hepatocito respetando parcialmente el área centrolobulillar. La poca intensidad del marcado típico de estos autoanticuerpos hace que cuando se encuentra en coexistencia con los anti-LKM1, el patrón de los primeros quede oscurecido por su interferencia. Para salvar este fenómeno, los anti-LC1 se pueden detectar utilizando citosol hepático por inmunodifusión bidimensional (IDBD) o contra-inmuno-electroforesis (CIEF) empleando un suero control positivo, o por ELISA detectando la reactividad frente al antígeno FTCD [59,72].

#### **4.2.2.4.6. Anticuerpos anti-antígeno soluble hepático/anti-hígado-páncreas**

Los anticuerpos anti-antígeno soluble hepático y los anti-hígado-páncreas se consideraron inicialmente entidades distintas. Los anti-SLA/LP se detectan de rutina por RIA o ELISA aunque se están desarrollando técnicas diagnósticas basadas en las dianas moleculares frente a las que tiene afinidad [34,59].

Las investigaciones iniciales encontraron estos autoanticuerpos en casos de HAI con ausencia de los convencionales o clásicos. Ello llevó a considerar un tercer tipo de HAI que estaría constituido exclusivamente por los casos seropositivos para anti-SLA/LP. Sin embargo los estudios que apoyaron esta propuesta estaban sesgados por punto de corte especialmente elevado para el cribado del resto de autoanticuerpos, motivo por el cual el IAIHG terminó por no aceptar lo que sería la HAI tipo 3 [79].



A pesar de que es posible encontrar anti-SLA/LP en pacientes con hepatitis crónica por VHC con anti-LKM1 positivos, su presencia es altamente específica de HAI y se ha demostrado que su detección en el momento del diagnóstico está relacionada con fenotipos de enfermedad más graves y con peor pronóstico, esto es, con menor supervivencia global y libre de trasplante [80].

La diana de estos autoanticuerpos es un tRNP<sup>(Ser)<sup>Sec</sup></sup>, un complejo de proteínas antigénicas asociadas a una ribonucleoproteína de transferencia, más concretamente O-fosfoseril-tRNA:selenocisteinil-tRNAsintetasa (SepSecS) [81,82].

#### **4.2.2.4.7. Anticuerpos anti-citoplasma de los neutrófilos**

Por IF los ANCA puede presentar un patrón perinuclear (pANCA) o citoplasmático (cANCA). En el caso de la HAI tipo 1, son los pANCA la presentación más frecuente, aunque en el caso particular de esta entidad presenta cierta particularidad: una afinidad añadida por algunos componentes de la membrana nuclear periférica, lo que determina que se le denominen también anticuerpos periféricos anti-núcleo del neutrófilo (pANNA) [83].

Los pANNA son excepcionales en las HAI tipo 2. Sin embargo su detección puede facilitar el diagnóstico del tipo 1 de la enfermedad en situaciones de coexistencia con otros marcadores serológicos [61].

El antígeno propuesto para los pANNA es una proteína nuclear de 50 kDa específica del neutrófilo que pertenece a la macroestructura del poro nuclear, con una alta sospecha de que pueda tratarse de la cadena 5 de la tubulina  $\beta$  [64]. La diana antigénica de los ANCA de la HAI es diferente de la de los ANCA descritos clásicamente. En el caso de los pANCA convencionales existe afinidad por la mieloperoxidasa y se encuentra típicamente en la poliangeítis microscópica. Por otro lado, los cANCA de la granulomatosis de Wegener tienen como diana la proteinasa 3, hecho que demuestra que aunque el patrón de IF sea similar, se trata de autoanticuerpos diferentes [84,85].

#### 4.2.2.4.8. Anticuerpos anti-receptor de asialoglicoproteínas

Los anti-ASGPR (una glicoproteína de membrana tipo 2 también conocida como lectina hepática) es el único anticuerpo dirigido contra un antígeno específicamente hepático descrito hasta el momento actual. La diana es un componente de un extracto crudo de hígado empleado en los intentos de identificar posibles antígenos hepáticos de la HAI. Se ha denominado proteína específica hepática o LSP (*liver specific protein*). Hasta el 90% de los pacientes diagnosticados de HAI son anti-ASGPR seropositivos y se suele encontrar con combinación con ANA, anti-Sm y anti-LKM1.

Del mismo modo que otros autoanticuerpos como los anti-LC1, los niveles en suero de anti-ASGPR también se correlacionan con la actividad inflamatoria y tienen utilidad potencial como herramienta de monitorización de la eficacia del tratamiento.

Sin embargo, su detección requiere del empleo de técnicas de biología molecular con antígenos purificados o recombinantes, que de momento no han conseguido producirse a escala suficiente como para que puedan utilizarse de forma habitual en la práctica clínica.

Por otro lado, tampoco son específicos de HAI. También se han encontrado en hepatitis víricas, tóxicas y en la CBP [57,64].

#### 4.2.2.5. *Histopatología*

Como principio general, en todos los pacientes en los que se sospeche una HAI se debe de practicar una biopsia hepática para estudiarla por anatomía patológica, incluyendo los casos con presentaciones agudas o graves o como fallo hepático fulminante si no hay contraindicaciones para ello [40,86–88]. De hecho el estudio histopatológico del hígado es un prerrequisito para el diagnóstico de la HAI según los criterios clásicos y simplificados propuestos bajo el amparo de la IAIHG y que se han discutido en las secciones previas [19,61,89,90]. Sin embargo no existe un dato histológico de alta especificidad o patognomónico de la HAI [33,91]. Esto ha

llevado a algunos autores a proponer una actitud más conservadora en el procedimiento diagnóstico clásico y evitar realizar biopsia hepática en los casos con datos clínicos y analíticos sugestivos de HAI. Es el caso de la propuesta de Björnsson *et al.* que se ha expuesto en el apartado sobre la aplicabilidad de los criterios diagnósticos de la HAI en población pediátrica [92]. Aunque posteriormente otros estudios han apoyado la actitud de iniciar el manejo de la HAI en dicho escenario clínico, la presencia de datos histopatológicos concordantes previos a la instauración de los inmunosupresores puede facilitar la toma de decisiones terapéuticas [54]. Así lo apoyan las instituciones internacionales más relevantes sobre el estudio de la hepatología, incluyendo la IAIHG [19,40,87,93]. Una variedad de publicaciones concluyen que se necesitan más estudios multicéntricos para validar esta propuesta dado que la toma de muestra por biopsia para estudio histológico no solo tiene interés diagnóstico, sino también para definir el grado de afectación del órgano, el desarrollo de complicaciones y el estadio de la enfermedad, y por consiguiente, el pronóstico de cada caso concreto [94–99].

Un hallazgo considerado como altamente sugestivo de HAI es la presencia de una hepatitis de interfase, también llamada necrosis piecema, que denota inflamación de los hepatocitos adyacentes al tracto portal, en la porción de parénquima del lobulillo yuxta-portal. Generalmente la inflamación ensancha el sistema biliar y separa sus componentes, está conformada por linfocitos y células plasmáticas agrupadas y habitualmente se extiende hacia los lóbulos, fenómeno que se describe como progresión a hepatitis lobular. En algunos pacientes existe destrucción de conductillos biliares sin otros datos de CBP y muestran la misma respuesta al tratamiento que los casos sin biliopatía [100].

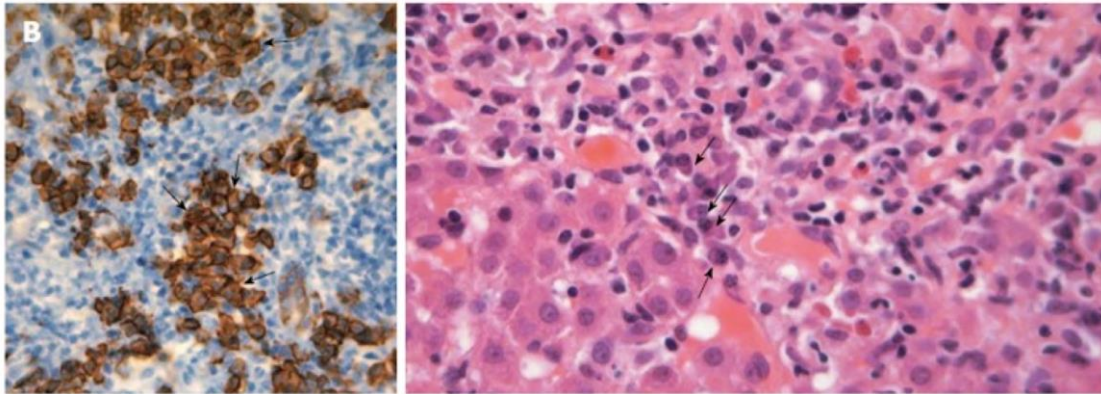


Figura 5: Intensa plasmocitosis portal en la hepatitis autoinmune, señalada con flechas en el panel de la derecha (tinción con hematoxilina-eosina). Las células plasmáticas se encuentran agrupadas y se identifican fácilmente por inmunotinción para CD138 (flechas del panel izquierdo). Reproducido de Gatselis et al. *World J Gastroenterol.* 2015;21;73. Con permiso de Baishideng Publishing Group.

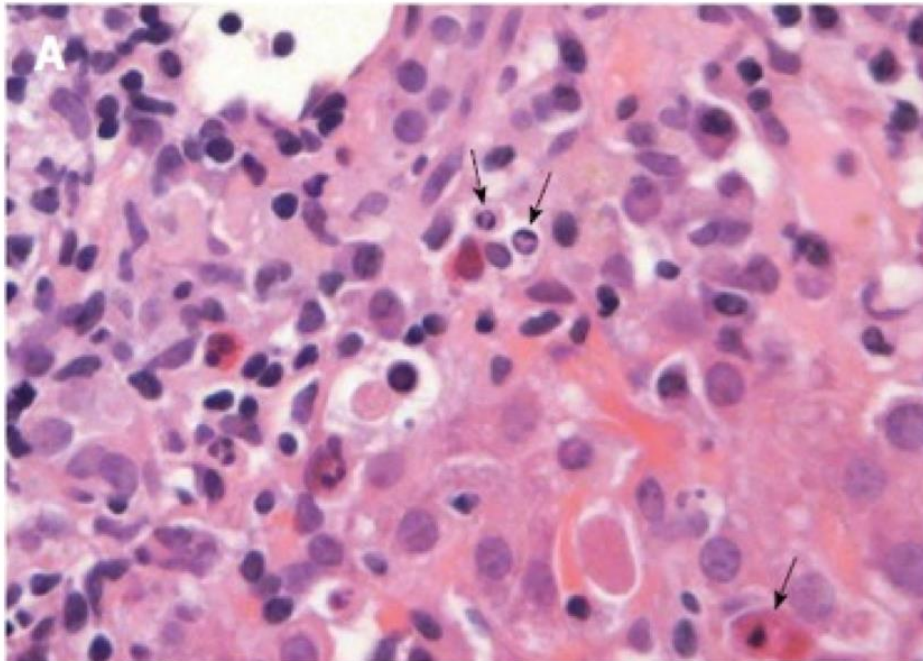


Figura 6: Hepatitis de interfase con cuerpos apoptóticos señalados con las flechas en un caso de hepatitis autoinmune tipo 1 (teñido con hematoxilina-eosina). Reproducido de Gatselis et al. *World J Gastroenterol.* 2015;21;73. Con permiso de Baishideng Publishing Group.

El grado de plasmocitosis es útil para discriminar la HAI de las hepatitis víricas. En raras ocasiones una infección por virus B llega a producir una infiltración por células plasmáticas portal de la intensidad con la que se suele observar en la

HAI. Solo en la hepatitis A se puede encontrar este hallazgo a un nivel que dificulte el diagnóstico diferencial histopatológico [54]. Además del interés diagnóstico, la plasmocitosis es un marcador pronóstico porque traduce un mayor riesgo de recaída si se encuentra durante el tratamiento inmunosupresor. Cerca de una tercera parte de los pacientes con HAI no presentan infiltración plasmocitaria portal por lo que, como se ha comentado anteriormente, su ausencia no descarta el diagnóstico por sí misma [87,101].

Los criterios simplificados de 2008 definen como hallazgos histológicos típicos la emperipolesis y la formación de rosetas hepatocelulares [19,102]. El término emperipolesis describe la penetración activa de una célula en otra más grande.

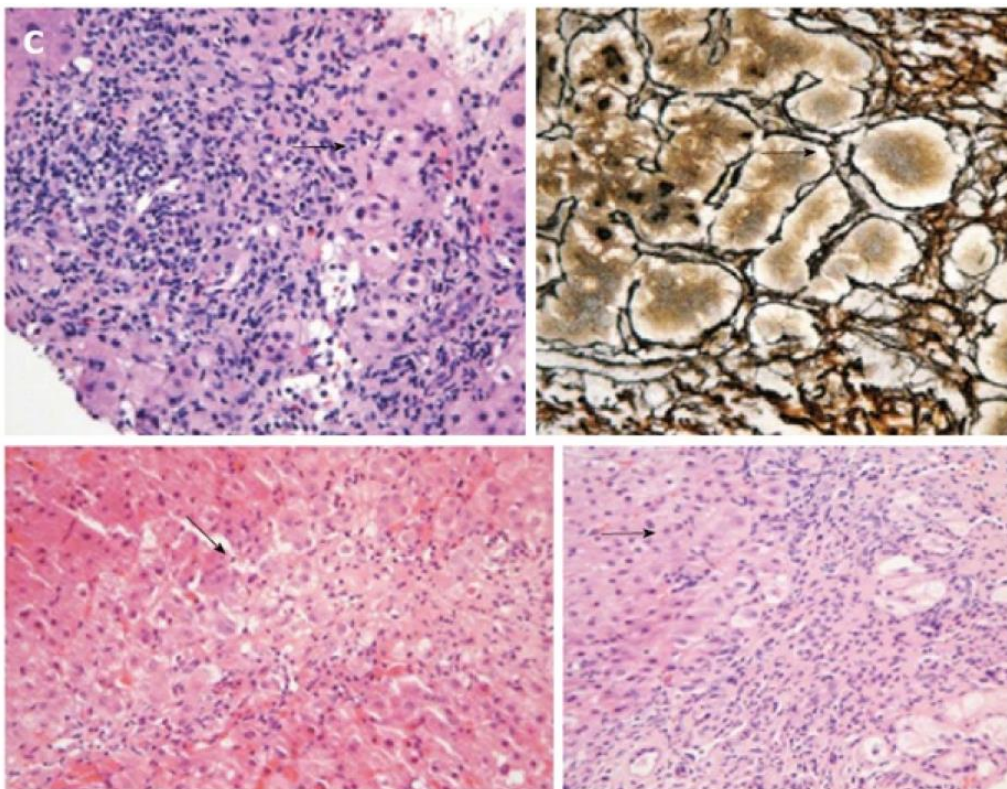


Figura 7: Varios ejemplos de rosetas de hepatocitos señalados por las flechas. Tinción de hematoxilina-eosina en todos los paneles salvo en el superior derecho, que pone en evidencia el colapso lobulillar a través de una tinción para fibras de reticulina. Reproducido de Gatselis et al. *World J Gastroenterol.* 2015;21;73. Con permiso de Baishideng Publishing Group.

La presencia de eosinófilos en tractos portales y dentro del lobillo hepático, así como la necrosis centrolobulillar, también se han descrito en muestras de pacientes con HAI, inicialmente en países asiáticos y en los últimos años en población occidental [103,104]. Se trata de un rasgo que puede dificultar el diagnóstico diferencial con la hepatitis tóxica. El colapso del parénquima, también conocido como necrosis multiacinar, si se describe en un paciente con el perfil clínico adecuado y una serología compatible, también apoya el diagnóstico de HAI [101,105]. En pacientes adultos, con menor proporción de casos leves o de corta evolución, suele encontrarse un grado de fibrosis que puede alcanzar la fase de cirrosis en pacientes no tratados. Se sabe que la actividad necrótica e inflamatoria (la gravedad histológica de la HAI) no guarda relación con los hallazgos bioquímicos ni serológicos de la enfermedad [40,61,87].

Considerando todos los grupos de edad, aproximadamente una tercera parte de los pacientes presentan cirrosis o necrosis en puentes en el momento del debut clínico o el diagnóstico. Dado que se reconoce que estos casos presentan un pronóstico peor que el de los pacientes sin fibrosis hepática, la biopsia hepática es un procedimiento que aporta información muy relevante no solo para el diagnóstico sino también para predecir la evolución de la enfermedad [94–96,98,99].

Adicionalmente, la evaluación de la remisión de la actividad de la enfermedad antes de un ensayo de retirada total o parcial del tratamiento farmacológico debe de hacerse incluyendo información histopatológica. El motivo de esta recomendación es que la presencia de actividad inflamatoria remanente ha demostrado tener un valor predictor de recaída en dicho contexto [40].

La histopatología de los pacientes con HAI que se presentan como fallo hepático puede diferir de la de las formas oligosintomáticas o con clínica insidiosa [106,107]. Esta observación ha llevado a proponer unos criterios diagnósticos específicos para el fallo hepático agudo autoinmune [86]. Igual que en los criterios clásicos y simplificados, la anatomía patológica es uno de los ítems que se deben evaluar. Sin embargo, en él se valora la presencia de dos patrones distintivos de

necrosis hepática masiva. El primero consiste una forma especialmente grave de HAI con necrosis centrolobulillar, que muestra afectación panlobular. El segundo es una necrosis típica de HAI con afectación de la interfase y compromiso ocasional y discreto del área centrolobulillar. En comunicaciones de casos aislados de HAI con fallo hepático también se han descrito folículos linfoides portales, infiltrado masivo de células plasmáticas y perivenulitis central [108,109].

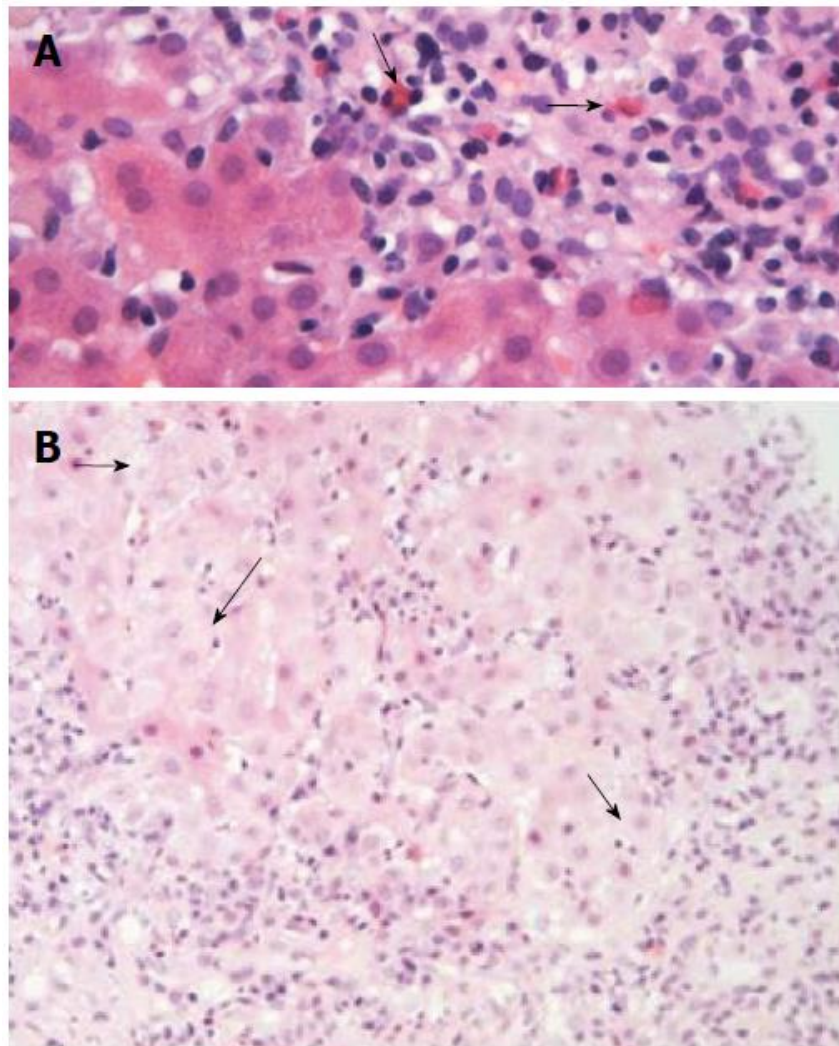


Figura 8: Se pueden observar eosinófilos en mayor o menor grado formando parte de las poblaciones constituyentes del infiltrado leucocitario (panel A). En el panel B se señalan cambios hidrópicos de los hepatocitos en una sección de hepatitis de interfase con necrosis multiacinar. Reproducido de Gatselis et al. *World J Gastroenterol.* 2015;21;74. Con permiso de Baishideng Publishing Group.

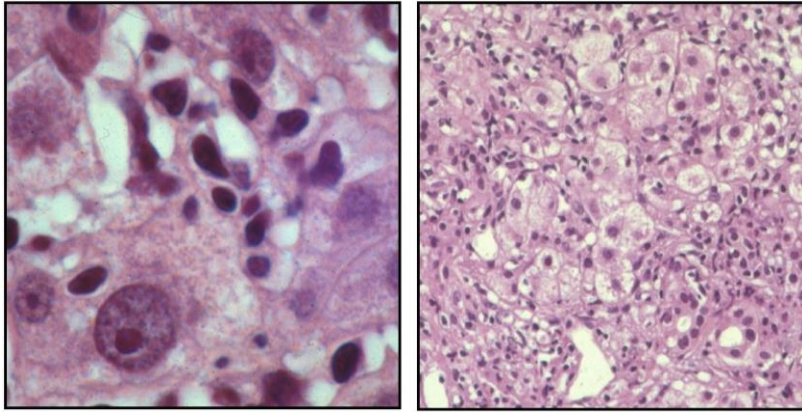


Figura 9: Emperipolesis a la izquierda, con un linfocito dentro de un hepatocito dañado. A la derecha una roseta de hepatocitos en el área de la interfase. Reproducido de Manns et al. *Journal of Hepatology*. 2015;62;S106. Con permiso de Elsevier.

#### 4.2.2.6. *Genética y mecanismos inmunológicos*

En los últimos años se ha descrito la asociación de la HAI con determinados marcadores genéticos y el impacto del inmunofenotipo del paciente en las características clínicas de su proceso concreto [110].

A pesar de que la patogenia de la HAI está todavía siendo objeto de análisis y no se conocen sus detalles con exactitud, ha quedado comprobado que están involucrados los genes del complejo mayor de histocompatibilidad II (MHC II, de *major histocompatibility complex*), y más concreta y directamente con el antígeno leucocitario humano (HLA, de *human leukocyte antigen*) [111]. El vínculo principal es con el HLA-DR3 y con el HLA-DR4 (DRB1\*03 and DRB1\*04) en población europea y norteamericana [112]. En niños, se ha descrito que el HLA-DRB1\*1301 confiere susceptibilidad para el desarrollo de HAI a una edad más precoz y con mayor proporción de casos seropositivos para los anti-Sm y elevaciones superiores de ALT y gammaglobulinas [110,113].

Las conclusiones de los trabajos llevados a cabo hasta el momento sobre este particular se encuentran resumidas en la tabla 8, aunque algunos de los resultados obtenidos son controvertidos. Fortes *et al.* identificaron que la presencia del alelo HLA-DRB1\*1301 imprimía un mayor riesgo de desarrollo de cirrosis [114]. Por su



parte, el grupo de Czaja concluyó que el HLA-DRB1\*03 se encuentra con mayor frecuencia en los casos de debut en niños y jóvenes comparado con el HLA-DRB1\*04 y que, además, otorga un riesgo mayor de obtener una respuesta incompleta a la corticoterapia. Por otro lado los pacientes con HLA-DRB1\*04 son habitualmente mujeres con una mayor proporción de otras enfermedades autoinmunes comórbidas y habitualmente con una buena respuesta a los inmunomoduladores [115].

El MHC II es genéticamente heterogéneo entre los distintos grupos étnicos. Los pacientes HLA-DRB1\*13 y DRB1\*03 positivos tienen una edad al debut más precoz comparado con otros pacientes probablemente porque los grupos étnicos en los que son más prevalentes estos alelos presentan una predisposición a iniciar la HAI a edades más tempranas. De hecho, en sentido contrario, algunas poblaciones, como México y Japón (donde es más frecuente el HLA-DRB1\*04), presentan bajas prevalencias de estos inmunofenotipos y se trata de grupos con una mayor susceptibilidad para un debut tardío de la HAI [114,116,117]. En el caso pediátrico, todavía no hay estudios suficientes como para poder aplicar estos marcadores como indicadores pronósticos o de respuesta al tratamiento [69].

También se ha descrito que El HLA DR7 (DRB1\*0701) y el DR3 (DRB1\*0301) confieren susceptibilidad para desarrollar HAI tipo 2. Las formas con presencia del inmunofenotipo DRB1\*0701 presentan una enfermedad más agresiva con un peor pronóstico global, tanto en términos de supervivencia, como de evolución a cirrosis o necesidad de trasplante hepático [118]. El HLA-DRB1\*0201 se ha encontrado en asociación con HAI tipo 2 aunque este alelo muestra un desequilibrio de ligamiento con DRB1\*0701 y DRB1\*0301, ambos asociados con la misma forma de HAI, lo que implica que su coexistencia en el mismo paciente es más improbable que lo esperable si la transferencia del alelo a las sucesivas generaciones filiales en cada división celular fuera aleatoria [119].

**Tabla 5: El antígeno leucocitario humano del complejo mayor de histocompatibilidad II y su papel y asociaciones con la hepatitis autoinmune. Adaptado de Ferri et al. World J Gastroenterol. 2013;19;4458. Con permiso de Baishideng Publishing Group.**

Referencia	Casos/controles (niños)	HLA estudiado	Conclusiones respecto a la HAI
Donaldson, 1991 [112]	96/100 (0)	DR	Los HLA-DR3 y DR4 son factores de riesgo independientes.
Ota, 1992 [116]	51/0 (0)	DR, DQ	Aumento de la frecuencia para todos los alelos de HLA-DRB1*04, principalmente DRB1*0405. Asociación secundaria con DRB1*15 y DRB1*16.
Czaja, 1993 [115]	86/102 (?)	A, B, C, DR, DQ	HLA-DRB4*0103 se asocia con enfermedades autoinmunes. HLA-DRB1*0301 confiere un mayor riesgo de respuesta pobre al tratamiento y DRB1*0401 se relaciona con menor necesidad de trasplante y mortalidad por causa hepática.
Fainboim, 1994 [120]	52/197 (249)	A, B, C, DR, DQ	No asociación con ningún antígeno de clase I. Aumento de frecuencia de HLA-DR6 (HLA-DRB1), principalmente DRB1*1301. Respecto a los HLA-DQ, asociación con DQB1*0603.
Vázquez- García, 1998 [117]	30/175 (?)	A, B, C, DR, DQ	No asociación con ningún antígeno de clase I. Asociación significativa con DRB1*0404 en pacientes adultos jóvenes. Baja frecuencia de DQB1*0301, que puede jugar un papel protector.
Pando, 1999 [121]	206/208 (122)	DR, DQ	Aumento de la frecuencia de HLA-DRB1*1301, DRB1*0301, DQA1*0103 y DQB1*0603. Asociación de HLA-DRB1*1301 con debut en niños y adolescentes. HLA-DRB1*1302 se comporta como factor protector en población infantil.
Bittencourt, 1999 [122]	139/129 (74)	DRB, DQB1	Aumento de frecuencia de HLA-DRB1*13, DRB1*03, DRB3 and DQB1*06 en la HAI tipo 1. HLA-DRB1*13 es más frecuente en niños que en adultos. HLA-DQB1*0301 puede jugar un papel protector. Aumento de la frecuencia de HAL-DRB1*07, DRB1*03, DRB4 y DQB1*02 en la HAI tipo 2.
Fortes, 2007 [114]	41/111 (13)	A, B, C, DR, DQ	No diferencias entre grupos de la presencia de HLA-A y C. Respecto al HLA I, se observó un aumento en la frecuencia de B*08, B*18, B*45 y B*50. HLA-B*40 puede comportarse como factor protector. Respecto al HLA II, se observó un aumento en la frecuencia de DQB1*02, DQB1*04, DRB1*03, DRB1*13 y DRB3. Los HLA-DRB1*1301 y DRB1*0301 fueron más frecuentes en niños.
Czaja, 2008 [123]	210/498 (0)	DRB1*03, DRB1*04, DRB1*13	La frecuencia de HLA-DRB1*13 es mayor en pacientes sin DRB1*03 ni DRB1*04. Los casos de CEP muestran una proporción de HLA-DRB1*13 similar a la de los casos de HAI.

HLA: Complejo mayor de histocompatibilidad. HAI: Hepatitis autoinmune. CEP: Colangitis esclerosante primaria. ?: No citado en la fuente original.

En 2014 se publicó el primer estudio de asociación del genoma completo sobre la HAI, que ha confirmado las observaciones aisladas descritas previamente y ha arrojado más luz sobre la base genética de la enfermedad. Se ha encontrado que la HAI tipo 1 se asocia, además de con variantes alélicas del MHC II, con variantes de genes implicados por primera vez: SH2B3 y CARD10 [124]. El significado de cada inmunofenotipo en relación con estos loci se desconoce a día de hoy [125].

La revisión de 1999 de los criterios clásicos incorpora la presencia de HLA DR3 y DR4 como parámetros adicionales que otorgan puntuación positiva [61,126].

En la respuesta inflamatoria característica de la HAI intervienen linfocitos T, principalmente colaboradores, linfocitos B, macrófagos y células *natural killer*. Los factores desencadenantes de la respuesta inflamatoria todavía no se conocen [111,127]. Se han propuesto varios mecanismos que explicarían parcialmente los hallazgos inmunológicos involucrados en la patogenia de la HAI.

En base a las conclusiones de diversos estudios llevados a cabo en adultos y niños, se conocen algunas vías potenciales para explicar el daño observado en la HAI, tal como la pérdida del equilibrio de los mecanismos inmunorreguladores. Algunos de estos trabajos describen una reducción en el número y función de las células T CD4+/CD25+, que suele representar entre un 5 y un 10% de los linfocitos T reguladores CD4+ totales en controles sanos [111,127–133]. Estas células suprimen la proliferación y la expresión de citoquinas de las células T CD4+ y CD8+. Entre sus otras funciones está la regulación a la baja de la función de los macrófagos, células dendríticas, células *natural killer* y linfocitos B (ver figura 10) [129].

Todos los hallazgos de tipo inmunológico son más pronunciados en la fase de presentación inicial que después de la inducción de la remisión farmacológica [129,133–135]. Los linfocitos T reguladores, en el marco de su función inmunosupresora, inducen la expresión de citoquinas antiinflamatorias como la interleuquina-4 (IL), la IL-10 y el factor de crecimiento transformante (TGF) beta [136,137]. Los marcadores de superficie que participan en los mecanismos antiinflamatorios son el factor de necrosis tumoral inducido por glucocorticoides

(CD62L), la proteína 4 asociada a linfocitos T citotóxicos (CTLA-4) y el FOXP3 (*fork head/winged helix transcription factor*) [129,138]. Se propone que si se conocieran mejor los mecanismos por los que falla el sistema de inmunotolerancia, se podrían desarrollar nuevos tratamientos basados en la recuperación de la función de los linfocitos T reguladores [139,140].

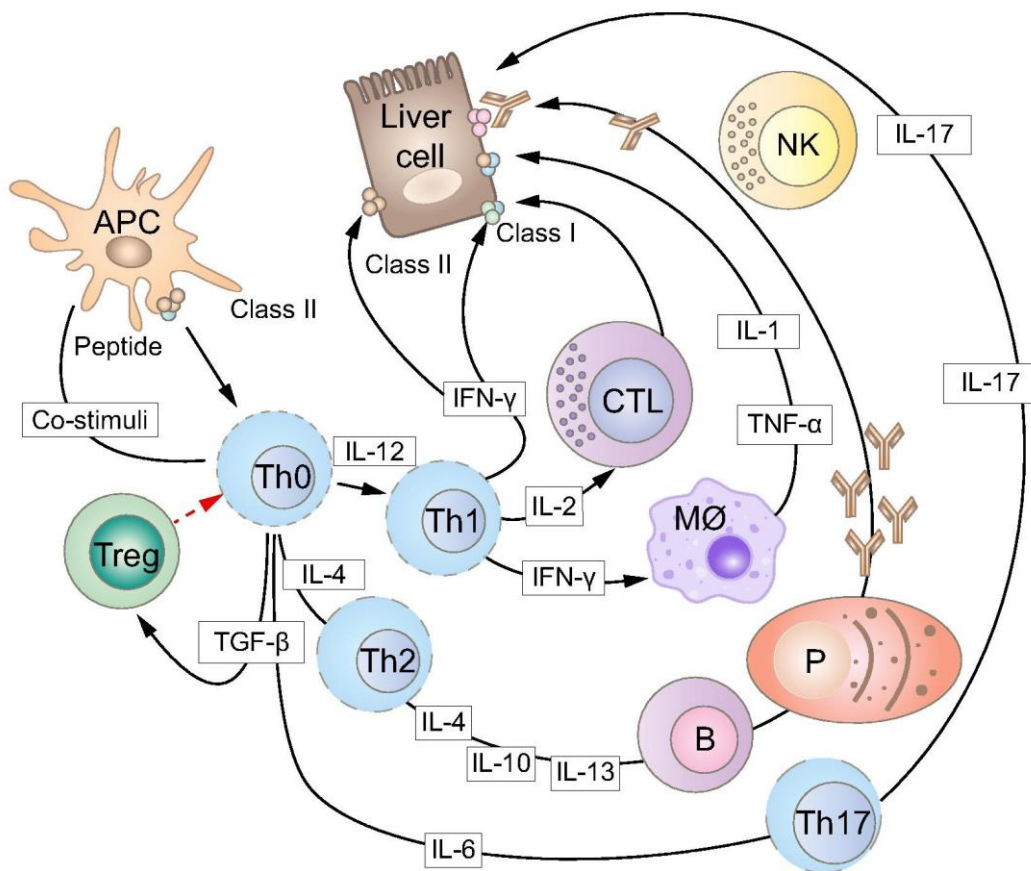


Figura 10: Esquema de la patogénesis de la hepatitis autoinmune. B: Linfocito B. P: Célula plasmática. Th: Linfocito T colaborador. CTL: Linfocito T citotóxico. MØ: Macrófago activado. NK: Célula *natural killer*. Treg: Linfocito T regulador. IL: Interleuquina. TGF: Factor de crecimiento transformante. APC: Célula presentadora de antígeno. Reproducido de Manns et al. *Journal of Hepatology*. 2015;62;S101. Con permiso de Elsevier.

Las células *natural killer* (CD3+ y CD56+) se encuentran en número reducido, produciendo menores cantidades de IL-4 e IL-2 en los pacientes con HAI. Como consecuencia de este fenómeno, existe una expresión a la baja de CTLA-4 en la

membrana plasmática de los linfocitos T CD4+, hecho que juega un papel central en la autoagresión al hepatocito, sobre todo durante la fase activa de la enfermedad [128,140]. Se ha descrito que los niveles de CTLA-4 están reducidos en células inflamatorias de sangre periférica de pacientes con HAI, en comparación con los niveles de CD80+ y CD86+, que se sobreexpresan en leucocitos infiltrantes hepáticos [69]. Otro trabajo ha demostrado que el receptor de la citoquina CCR5 se expresa de forma preferente en células Th1. Esta citoquina interviene en la captación de la citoquina proinflamatoria IFN- $\gamma$ , el interferón gamma, produciendo células T CD4+ en los nichos inflamatorios, como puede ser el tejido hepático, y promoviendo el daño tisular en la HAI [141,142]. Otra posibilidad teórica que se ha planteado involucra la presencia de células CD4+ y/o CD8+ autorreactivas, que serían las responsables de la hepatitis. Se fundamenta en la observación de niveles aumentados de estas células (hasta 10 veces) en pacientes con HAI respecto a controles sanos [129,143].

Otros estudios han sugerido que las mutaciones en otros genes distintos de los del MHC pueden conllevar modificaciones en las proteínas de membrana de las células inmunitarias que podrían explicar parte de la patogenia de las enfermedades autoinmunes. En esta línea, se ha observado que es posible el desarrollo de HAI sin evidencia de polimorfismos infrecuentes o mutaciones en el HLA. Los resultados del reciente trabajo de asociación del genoma completo mencionado previamente lo corroboran [124].

Algunas mutaciones en marcadores de superficie linfocitarios pueden representar indicadores moleculares de autoinmunidad en la HAI. Entre ellos tenemos el CTLA-4 (CD152), en cuyo gen se han descrito mutaciones asociadas a susceptibilidad de padecer HAI [144–149]. Sin embargo estos resultados son controvertidos porque no se han confirmado en un estudio posterior en población no norteamericana [145]. CTLA-4 se expresa en la superficie de membrana de los linfocitos T e induce tolerancia periférica fijando CD80 y CD86 en las células presentadoras de antígenos. De este modo CTLA-4 compite con la molécula

coestimulante CD28, reduciendo la respuesta inmune [110]. Actualmente se considera a CTLA-4 un coordinador principal de la regulación inmune. Existen líneas de investigación centradas en desarrollar fármacos que estimulen el mecanismo de CTLA-4, con la idea de que puedan servir de tratamiento de enfermedades autoinmunes. Recientemente se ha aprobado el uso de una proteína de fusión compuesta por una inmunoglobulina unida al dominio extracelular del antígeno citotóxico de linfocito T CTLA-4 (abatacept). Se trata de un modulador de coestimulación selectivo que se ha autorizado en Estados Unidos para el tratamiento de la artritis reumatoide en casos de respuesta inadecuada a la terapia anti-TNF $\alpha$  [150,151]. Este mecanismo podría ser útil también para la inducción de la remisión clínica en la HAI, pero de momento todavía no han ensayos al respecto.

Otro modelo teórico bajo investigación para explicar el mecanismo por el que se produce la HAI es que involucra polimorfismos en el gen *Fas* (que codifica para CD95), que llevan a una sobreexpresión en la membrana del linfocito. El CD95 forma parte de la familia del factor de necrosis tumoral e induce la apoptosis a través de una fusión con su ligando natural (FasL/CD95L). Así, indirectamente, controla el número de linfocitos activados por antígeno [152].

Se ha demostrado que los pacientes con HAI presentan un mayor número de linfocitos T CD95+ (Fas+), tanto CD4+ como CD8+. En estos casos, se observa una activación constante y mantenida de linfocitos T efectoras, para lo cual se necesita el compromiso de la células reguladoras T CD95+/CD4+ [153]. También se ha observado que los pacientes presentan un incremento de las células mononucleares Fas+ y FasL+, con un aumento in vitro de la producción de TNF-a e IFN- $\gamma$ . Estas citoquinas pueden participar en la aceleración del proceso de muerte celular programada que acontece en la HAI. Otro indicador de apoptosis, el aumento de monocitos CD14+, también se ha descrito en estos pacientes [154,155]. La relación de algunas mutaciones en los genes del receptor Fas con la HAI todavía está siendo objeto de estudio en la actualidad [156–158].

Hoy por hoy todavía no se dispone de evidencia suficiente que relacione polimorfismos en los genes de citoquinas con la HAI. A este respecto, las más estudiadas sin resultados claros son TGF- $\beta$ 1 y TBX21 (reguladora del desarrollo hacia linaje T del linfocito y controladora de la expresión de IFN- $\gamma$ ) [159–165].

#### **4.2.2.7. Tratamiento**

##### **4.2.2.7.1. Definición de remisión y recaída**

Se habla de remisión cuando se llega, después de instaurar un tratamiento, a una desaparición completa de los síntomas, se normalizan los niveles de transaminasas (algunos autores proponen que por debajo de dos veces el límite superior de la normalidad), de IgG y descienden los títulos de anticuerpos, junto con una resolución de la inflamación histológica [50]. Habitualmente la respuesta anatomopatológica acontece después de la mejoría bioquímica, que por su parte ha demostrado no reflejar el grado de resolución histológica [166–168]. Existe evidencia de que después de un seguimiento medio de 4 años tras el inicio de los inmunosupresores, disminuye la inflamación portal en hasta un 95% de los casos, acompañándose también de una mejoría del grado de fibrosis [166].

Las recaídas se caracterizan por una elevación de las transaminasas una vez alcanzado el estado de remisión. Ocurre durante el tratamiento en un 40% de los pacientes y habitualmente obligan a aumentar la dosis de corticoides. El determinante más importante descrito para la aparición de recaídas es la falta de adherencia al tratamiento, que es particularmente relevante en población adolescente [169]. En los casos más agresivos, el riesgo de recaída es mayor en pautas de corticoides a días alternos, que se utiliza principalmente en pediatría por los efectos perjudiciales a largo plazo sobre el crecimiento. Se ha demostrado que dosis más bajas pero diarias son más efectivas en el mantenimiento de la remisión y minimiza el uso de mega-bolos de esteroides durante las recaídas, sin afectar la talla final [170].

#### 4.2.2.7.2. Indicaciones del tratamiento

Dado que la HAI responde exquisitamente al tratamiento inmunosupresor, se debe de iniciar ante la sospecha para evitar la progresión de la enfermedad, después de descartar causas infecciosas y metabólicas [93]. El objetivo del tratamiento es reducir o anular la inflamación hepática, inducir la remisión clínica y prolongar la supervivencia del paciente [171,172]. La rapidez y el grado de la respuesta al tratamiento dependen de la gravedad en el momento del diagnóstico. En el caso pediátrico, la cirrosis al debut se ha descrito entre el 44 y el 80% de los pacientes [30,173,174]. A pesar de ello, la mortalidad por este motivo es baja y la mayoría de los pacientes permanecen clínicamente estables y con una buena calidad de vida con el tratamiento a largo plazo [175,176].

#### 4.2.2.7.3. Posibilidades terapéuticas

Excluyendo las formas de presentación como fallo hepático fulminante con encefalopatía, la HAI responde de forma satisfactoria al tratamiento inmunosupresor con independencia del grado de fibrosis hepática, con una tasa de remisión de cerca del 80% [93].

##### 4.2.2.7.3.1. Tratamiento estándar

El tratamiento convencional para la HAI consiste en prednisolona (o prednisona) a dosis de 2 mg/Kg/día (máximo 40 – 60 mg/día) hasta la consolidación del descenso de las transaminasas, para posteriormente rebajar la dosis de forma paralela a dicho descenso durante 4 – 8 semanas hasta alcanzar la dosis de mantenimiento de 2,5 – 5 mg/día [177–179]. En la mayoría de pacientes se consigue un descenso del 80% en el nivel de transaminasas dentro de los primeros dos meses de tratamiento, pero su normalización completa puede tardar varios meses [49]. Durante las primeras 6 – 8 semanas de tratamiento, las pruebas de función hepática se deben de efectuar semanalmente para vigilar la aparición de efectos adversos graves a los corticoides y permitir el ajuste de dosis si es necesario. Es motivo de diferencias entre los protocolos de los distintos centros el momento de inicio de la



azatioprina, indicada como inmunomodulador y ahorrador de dosis de corticoide. En algunas propuestas de manejo, la azatioprina se inicia solo en presencia de efectos adversos importantes a los corticoides o cuando la disminución de las transaminasas es solo parcial en monoterapia. Se empieza a 0,5 mg/Kg/día y si no se encuentran signos de toxicidad (fundamentalmente hematológica, pero también puede ser pancreática, hepática o neurológica) se aumenta la dosis hasta un máximo de 2 – 2,5 mg/Kg/día [176]. En otros centros, la azatioprina se inicia a dosis de 0,5 – 2 mg/Kg/día después de unas pocas semanas de haber iniciado el corticoide y cuando las transaminasas comienzan a descender. Independientemente de la pauta, el 85% de los pacientes con HAI requieren eventualmente la adición de azatioprina al esquema farmacoterapéutico [93]. Algunos centros emplean una combinación de corticoide y azatioprina desde el diagnóstico, pero esta estrategia obliga a un seguimiento estrecho por la posibilidad de hepatotoxicidad por azatioprina, especialmente en pacientes previamente ictericos en los que puede ser complicado de identificar. En pacientes con HAI tipo 1 en remisión a largo plazo se puede intentar retirar la prednisona y mantener la azatioprina en monoterapia, con lo que se consigue mantener la remisión en hasta un 70% de los niños [29].

Clásicamente, el punto final del tratamiento sería cuando se ha conseguido una función hepática normal mantenida durante dos o tres años y la biopsia hepática no muestra inflamación. Cabe tener en cuenta no se alcanza la curación histológica en un 55% de los pacientes en los que se consigue normalización clínica y analítica mantenida. En estos casos no se debe de retirar el tratamiento, lo que obliga a practicar una biopsia de control previa a la retirada de la farmacoterapia. En una serie del King's College Hospital se pudo retirar el tratamiento en un 19% de los niños con HAI tipo 1 y en ninguno con HAI tipo 2 [29]. En otra serie de 163 pacientes italianos, de los que 28 eran niños, se produjo recaída en todos los casos en que se retiró el tratamiento, por lo que los autores aconsejan mantenerlo de forma indefinida [95].

Es posible determinar la actividad de la tiopurín-metiltransferasa para dosificar la azatioprina en función de la misma [49]. También se ha señalado que la medición de los metabolitos 6-tioguanina y 6-metilmercaptapurina puede ser útil para diagnosticar toxicidad por el fármaco y seguir la adherencia al tratamiento. En esta línea, se disponen de valores objetivo de 6-tioguanina para la enfermedad inflamatoria intestinal pero no se han determinado para el caso concreto de la HAI [180].

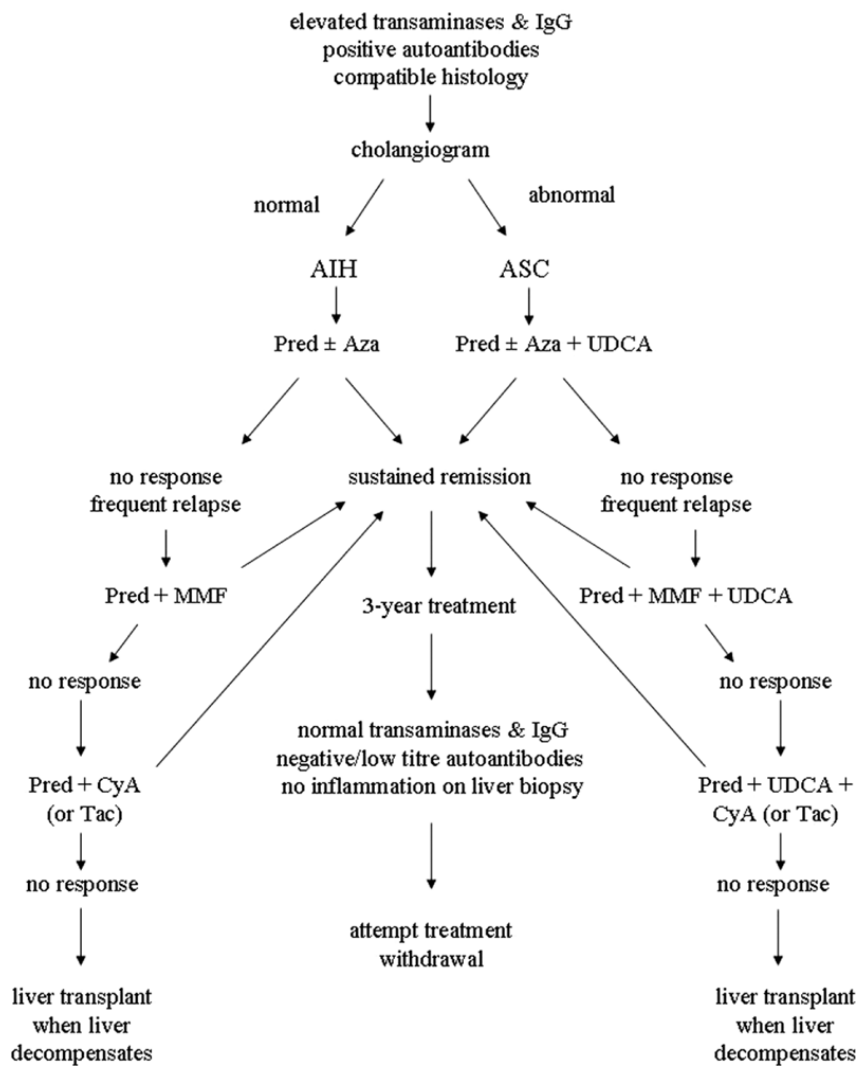


Figura 11: Esquema de manejo en forma de diagrama de flujo para la enfermedad hepática autoinmune en pediatría. AIH: Hepatitis autoinmune. ASC: Colangitis esclerosante autoinmune. Pred: Prednisolona. Aza: Azatioprina. UDCA: Ursodiol. MMF: Micofenolato de mofetil. CyA: Ciclosporina A. Tac: Tacrolimus. IgG: Inmunoglobulina G. Reproducido de Mieli-Vergani et al. J Pediatr Gastroenterol Nutr. 2009;49;159. Con permiso de Lippincott Williams & Wilkins.

#### 4.2.2.7.3.2. *Tratamientos de segundo escalón*

Se ha conseguido inducir la remisión clínica y analítica con ciclosporina A durante 6 meses en niños con HAI *naïve* a tratamientos farmacológicos, con suplementación con prednisona y azatioprina posterior durante un mes, tras el cual se retira la ciclosporina [181,182]. Se emplea a dosis de 4 mg/Kg/día en tres dosis, incrementando la dosis si es necesario, cada 2 o 3 días, para alcanzar una concentración plasmática de 200 – 300 ng/mL durante los primeros tres meses. Si se obtiene respuesta clínica y bioquímica, se reduce la dosis de ciclosporina hasta una concentración de 150 – 250 ng/mL durante los siguientes 3 meses, antes de interrumpirla. Todavía faltan más estudios para evaluar esta propuesta de inducción a la remisión [183].

Por lo que respecta al tacrolimus, se trata de un inmunosupresor más potente que la ciclosporina, pero con potenciales efectos adversos más significativos. Tampoco existe evidencia suficiente que apoye su uso como parte del tratamiento de la HAI dejando de lado comunicaciones de casos aislados en pacientes adultos [184].

Tampoco existe evidencia de la efectividad de la budesonida o del ácido ursodesoxicólico en la HAI pediátrica [93].

#### 4.2.2.7.3.3. *Tratamiento de los casos refractarios*

El micofenolato de mofetil es el profármaco del ácido micofenólico. Su efecto sobre la síntesis de purinas conlleva una proliferación deficiente de linfocitos T y B. En el 10% de los pacientes en los que el tratamiento inmunosupresor estándar no consigue inducir la remisión estable, o en los pacientes intolerantes a la azatioprina, el micofenolato puede ser útil a dosis de 20 mg/Kg dos veces al día en combinación con prednisolona [185]. Si a pesar del tratamiento con micofenolato no se consigue frenar la evolución de la HAI, o si aparece reacciones adversas que obliguen a la retirada del fármaco (dolor de cabeza, diarrea, náusea, mareo, alopecia y neutropenia son las más frecuentes) se debe de considerar el uso de inhibidores de

la calcineurina. En este escenario el tacrolimus puede ser útil en combinación con prednisona [184].

### **4.2.3. Colangitis esclerosante autoinmune**

#### **4.2.3.1. Definición y presentación clínica**

La colangitis esclerosante puede presentarse en niños, adolescentes y adultos jóvenes con rasgos autoinmunes floridos tales como elevación de IgG, títulos altos de autoanticuerpos y hepatitis de interfase en el estudio de la biopsia hepática[19]. A finales de la década de 1980 un estudio retrospectivo demostró que durante el seguimiento de niños con hipergammaglobulinemia y títulos positivos de ANA y/o anti-Sm, en algunos pacientes se ponía de manifiesto la evidencia radiológica de colangitis esclerosante [186]. Las pruebas de la posibilidad de evolución de HAI a CEAI vinieron posteriormente a raíz de publicaciones de series de casos [187–189]. Sin embargo, en ninguno de estos trabajos se informaba de la realización de colangiogramas en el momento de la presentación de la enfermedad.

Esta práctica se ha recomendado recientemente en la revisión de Mieli-Vergani en 2009, a instancias de la ESPGHAN/NASPGHAN (sociedades europea y americana de gastroenterología, hepatología y nutrición pediátrica) [93]. Basado en esta propuesta, se llevó a cabo un trabajo prospectivo de 16 años de duración reclutando niños con evidencia serológica e histológica de enfermedad hepática autoinmune. Se les practicó una prueba de imagen biliar, una sigmoidoscopia y una biopsia rectal al debut y aproximadamente la mitad de los casos presentaron cambios biliares compatibles con colangitis esclerosante. Aunque la afectación biliar no fue tan avanzada como en los casos de colangitis esclerosante aislada o primaria, se les diagnosticó de CEAI [30]. Una cuarta parte de estos pacientes no presentaron evidencia histológica de compromiso de la vía biliar intrahepática a pesar de presentar anomalías macroscópicas en la colangio-resonancia magnética o en la

colangiografía retrógrada endoscópica, lo que traduce que la anatomía patológica tiene un rendimiento inferior que el de las pruebas de imagen. Este mismo estudio demostró que la CEAI fue incluso más frecuente que la CEP sin signos de autoinmunidad, dado que esta se observó solo en 9 niños de toda la muestra [30].

#### 4.2.3.2. Diferencias con la hepatitis autoinmune

El modo de presentación de la CEAI es similar al de la HAI tipo 1 [30]. Se diagnostica de enfermedad inflamatoria intestinal (EII) al 45% de los niños con CEAI, proporción que baja al 20% en los casos de HAI.

Al debut ninguna prueba de laboratorio ha demostrado tener capacidad discriminante entre CEAI y HAI, aunque la AST tiende a ser más elevada en la segunda. Por otro lado, la relación FA/AST tiende a ser superior en la CEAI.

La mayoría de los pacientes, cerca de un 90%, presentan niveles aumentados de IgG en las dos enfermedades y también en ambas son muy infrecuentes los pacientes seronegativos.

Tabla 6: Datos bioquímicos al diagnóstico de los niños con enfermedad hepática autoinmune. Valores expresados en mediana (rango intercuartílico). Adaptado de Gregorio et al. *Hepatology* 2001;33:546. Con permiso de Wiley.

	HAI	CEAI
Bilirrubinemia (normal < 20 mmol/L)	35 (4 - 306)	20 (4 - 179)
Albuminemia (normal > 35 g/L)	35 (25 - 47)	39 (27 - 54)
AST (normal < 50 UI/L)	333 (24 - 4830)	102 (18 - 1,215)
INR (normal <1,2)	1,2 (0,96 - 2,50)	1,1 (0,9 - 1,6)
GGT (normal < 50 UI/L)	76 (29 - 383)	129 (13 - 948)
FA (normal < 350 UI/L)	356 (131 - 878)	303 (104 - 1710)
FA/AST	1,14 (0,05 - 14,75)	3,96 (0,2 - 14,2)

HAI: Hepatitis autoinmune. CEAI: Colangitis esclerosante autoinmune. AST: Aspartato aminotransferasa. INR: International normalized ratio. GGT: Gamma-glutamyl transpeptidasa. FA: Fosfatasa alcalina.

Al respecto de los autoanticuerpos, se sabe que los anti-SLA son más difíciles de encontrar en la CEAI que en la HAI. Solo recientemente se ha descrito que los

pANCA atípicos se encuentran preferentemente en la CEAI (74% de los casos frente al 45% de los pacientes con HAI) [30].

**Tabla 7: Presentación clínica al diagnóstico de las enfermedades hepáticas autoinmunes en la infancia. Adaptado de Gregorio et al. Hepatology 2001;33:546. Con permiso de Wiley.**

Parámetro	HAI-1	HAI-2	CEAI
Mediana de edad (años)	11	7	12
Forma de presentación (%)			
Hepatitis aguda	47	40	37
Fallo hepático agudo	3	25	0
Debut insidioso	38	25	37
Complicación de hepatopatía crónica	12	10	26
Enfermedades inmunes asociadas (%)	22	20	48
Enfermedad inflamatoria intestinal (%)	20	12	44
Colangiograma anormal (%)	0	0	100
ANA o anti-Sm (%)	100	25	96
Anti-LKM1 (%)	0	100	4
pANNA (%)	45	11	74
Anti-SLA (%)	58	58	41
Hepatitis de interfase (%)	92	94	60
Anomalías de la vía biliar (%)	28	6	31
Cirrosis (%)	69	38	15

HAI: Hepatitis autoinmune. CEAI: Colangitis esclerosante autoinmune.

#### **4.2.3.3. Diferencias con la colangitis esclerosante primaria**

A diferencia de lo que ocurre en la CEP, que afecta de forma preferente a varones, la CEAI muestra una distribución por sexos equilibrada [190]. Desde el punto de vista de la bioquímica, la AST suele estar sensiblemente elevada en la CEAI, mientras que en la CAP sin autoinmunidad está normal o discretamente elevada. La FA y la GGT, por su parte, están más elevadas en la CEP que en la CEAI. Por lo que respecta a la anatomía patológica, la presencia de hepatitis de interfase no es un discriminador específico de ninguna de estas dos entidades [30,190]. La coexistencia

de enfermedad inflamatoria intestinal es más frecuente en la CEP. Cuando ocurre en estos dos grupos de pacientes no suele ser fenotípicamente diferente, siendo la pancolitis sin afectación rectal la extensión característica, y leves por lo que respecta a la expresión clínica [191]. Un estudio reciente ha dado como resultado, en pacientes con enfermedad inflamatoria intestinal y hepatopatía autoinmune, que la afectación de intestino delgado es más frecuente en la CEAI que en la CEP. Es todavía objeto de discusión si estas lesiones corresponden a una categoría nueva de enfermedad inflamatoria intestinal [192].

**Tabla 8: Datos clínicos y analíticos comparados entre la colangitis esclerosante autoinmune y la primaria. Adaptado de Gregorio et al. Hepatology 2001;33:546. Con permiso de Wiley.**

	CEAI	CEP
Proporción de mujeres	55%	30%
Elevación de AST	Moderada / Marcada	No / Leve
Elevación de FA	Leve / Moderada	Moderada / Marcada
Niveles de IgG	Elevados en el 89%	Pueden estar elevados
ANA o anti-Sm	96%	Variable (hasta 77%)
pANCA	74%	26 – 94%
Hepatitis de interfase	Normalmente presente	Puede estar presente
Coexistencia con EII	44%	80%

CEAI: Colangitis esclerosante autoinmune. CEP: Colangitis esclerosante primaria. AST: Aspartato aminotransferasa. FA: Fosfatasa alcalina. IgG: Inmunoglobulina G. ANA: Anticuerpos antinucleares. Anti-Sm: anti-músculo liso. pANCA: Anticuerpo anti-citoplasma del neutrófilo con patrón perinuclear.

### **4.3. Diagnóstico de la hepatitis autoinmune por criterios de clasificación**

---

#### **4.3.1. Los criterios clásicos revisados de 1999**

Clásicamente se ha reconocido que el diagnóstico de la HAI se basa en la presencia de una elevación de las transaminasas y de la IgG, seropositividad para autoanticuerpos y evidencia histopatológica de hepatitis de interfase. Este principio se tuvo en cuenta para el desarrollo de los criterios clásicos por parte del IAIHG, que fueron desarrollados con el objetivo de homogeneizar la definición de la enfermedad a efectos de diseño y metodología de trabajos de investigación. En el año 1993 se publica el primer consenso del grupo de estudio internacional, que es revisado y modificado en 1999 con resultado de los criterios clásicos que continúan vigentes actualmente [33,61].

Hasta un total de seis estudios publicados e indexados concluyen que los criterios clásicos muestran una sensibilidad y especificidad elevadas para el diagnóstico de HAI (97 – 100%). Incluso en pacientes con colangitis esclerosante primaria (CSP) y otras anomalías biliares, la especificidad para la exclusión de HAI definitiva alcanza valores entre 96 y 100%. Sin embargo, una proporción de entre 8 y 52% de pacientes alcanzan puntuaciones que los clasifican como HAI probable, reduciendo la especificidad global hasta 45 – 92% [71,193–196].

El motivo de la revisión de los criterios clásicos en 1999 es precisamente mejorar la capacidad del sistema de excluir los pacientes con enfermedad de la vía biliar [89,197–199]. En vista de este defecto inicial, los sucesivos estudios sobre criterios diagnósticos para la HAI intentan comprobar su exactitud también con pacientes con enfermedades que cursan con colestasis y con los síndromes de solapamiento, el principal punto débil de los sistemas de clasificación diagnóstica de esta entidad.



Este sistema de clasificación incorpora una serie de puntuaciones positivas y negativas, lo que permite al clínico o al investigador dar un peso a los rasgos clínicos, analíticos o histopatológicos del paciente en el que se sospecha una HAI (ver tabla con los criterios pormenorizados en *Anexos*). Como se ha comprobado, la sensibilidad y especificidad del sistema diagnóstico es superior al 90% y también ha demostrado ser útil en un escenario clínico con pacientes de todas las edades con características atípicas de la enfermedad [61].

**Tabla 9: Guía de uso de los criterios descriptivos revisados para el diagnóstico de hepatitis autoinmune según el sistema clásico de 1999. Adaptado de Qiu et al. J Hepatol. 2011;54:340-347. Con permiso de Elsevier.**

Característica	Orienta a diagnóstico definitivo	Orienta a diagnóstico probable
Histopatología hepática	Hepatitis de interfase con actividad moderada o grave con/sin hepatitis lobular o necrosis en puentes centro-pontinos. Sin lesiones biliares o granulomas bien definidos u otros cambios prominentes sugestivos de una etiología distinta a la autoinmune.	Igual que en "definitivo".
Bioquímica sérica	Cualquier alteración en los niveles de transaminasas, especialmente (pero no exclusivamente) si la fosfatasa alcalina sérica no está llamativamente elevada. Concentraciones plasmáticas normales de $\alpha$ 1-antitripsina, cobre y ceruloplasmina.	Igual que en "definitivo" pero podrían incluirse pacientes con alteraciones en la cupremia o el nivel de ceruloplasmina, una vez que se haya descartado la enfermedad de Wilson con los métodos diagnósticos apropiados.
Ig sérica	Concentración plasmática de globulina sérica total, gammaglobulina o inmunoglobulina G mayor de 1,5 veces el límite superior de normalidad del laboratorio.	Cualquier elevación de globulina sérica total, gammaglobulina o inmunoglobulina G por encima del límite superior de normalidad del laboratorio.
Anticuerpos	Seropositividad para ANA, SMA o anti-LKM1 a títulos superiores a 1:80. Títulos más bajos (particularmente de anti-LKM1) pueden ser significativos en niños. Seronegatividad para AMA.	Igual que en "definitivo" pero a títulos de 1:40 o superiores. Se incluyen aquí los pacientes que son seronegativos para estos anticuerpos pero positivos para otros especificados en el texto.
Marcadores de infección vírica	Seronegatividad para marcadores de infección activa por virus A, B o C u otros hepatotropos.	Igual que en "definitivo".
Otros factores etiológicos	Consumo medio de alcohol inferior a 25 g/día. No historia reciente de uso de fármacos o drogas hepatotóxicos.	Consumo medio de alcohol inferior a 50 g/día en ausencia de historia reciente de uso de fármacos o drogas hepatotóxicos. Podrían incluirse pacientes que han tomado alcohol o han recibido productos con potencial toxicidad hepática si existe evidencia de progresión del daño hepático después del cese de la exposición a los mismos.

Ig: Inmunoglobulina. ANA: Anticuerpo anti-nuclear. SMA: Anticuerpo anti-músculo liso. Anti-LKM1: Anticuerpo anti-microsomal de hígado/riñón de tipo 1. AMA: Anticuerpo anti-mitocondrial.

Los datos de laboratorio y anatomopatológicos sugestivos de colestasis conllevan una asignación negativa de puntos. En los casos raros en los que no se detectan autoanticuerpos (HAI seronegativa) la presencia de anti-receptor de asialoglicoproteína (anti-ASGPR), anti-antígeno soluble hepático/anti-hígado-páncreas (anti-SLA/PR) o anti-citoplasma del neutrófilo con patrón perinuclear (pANCA) atípico aportan peso a favor de la clasificación como HAI posible (tabla 3) [200]. Los criterios clásicos revisados de 1999 también incorporan información sobre la respuesta a tratamiento con corticoesteroides. Se define HAI definitiva como la suma pre-tratamiento de más de 15 puntos, y post-tratamiento de más de 17 puntos. En esta línea, se considera HAI probable una puntuación pre-tratamiento entre 10 y 15 puntos, y post-tratamiento entre 12 y 17 [78].

La distinción entre un diagnóstico definitivo y otro probable está en relación sobre todo con la magnitud de la elevación de la IgG o la titulación de autoanticuerpos, así como al grado de exposición al alcohol, a infecciones o a fármacos o drogas con potencial hepatotóxico. Se ha demostrado además que la designación diagnóstica probable o definitiva no refleja diferencias en la validez del diagnóstico o en la respuesta al tratamiento, aunque son escasos los estudios llevados a cabo con el fin de comparar estos dos grupos de HAI [201].

Existe un trabajo de validación de los criterios clásicos en una población exclusivamente pediátrica, publicado en 2004. El estudio se lleva a cabo sobre una muestra de 21 HAI, 4 colangitis esclerosantes primarias (CEP) y 3 CEAI. Un 86% (18 de las 21) de las HAI se clasificaron como definitivas y las otras 3, como probables. Los pacientes con CEP aislada obtuvieron puntuaciones bajas y, por tanto, fueron correctamente diagnosticados. Sin embargo, los 3 casos con síndrome de solapamiento puntuaron como HAI. Los autores concluyen que los criterios clásicos pueden tener también el mismo papel en niños que en adultos y sugieren que emplear la GGT en el criterio de la relación FA/AST (o ALT) como alternativa a la FA, puede aumentar la especificidad en este grupo etario, donde los valores de FA son variables en función de la actividad ósea propia del crecimiento normal [202].

### 4.3.2. Los criterios simplificados de 2008

El mismo grupo promotor de los criterios diagnósticos clásicos de 1999, el IAHG, publicó en 2008 unos criterios simplificados para el diagnóstico de la HAI (ver en *Anexos*). La justificación del desarrollo de un sistema alternativo fue que los criterios iniciales resultaban complejos, insuficientemente validados e incluían una serie de criterios de bajo valor desde el punto de vista de la plausibilidad biológica. Además, el hecho de encontrar un sistema más parsimonioso podría favorecer su aplicabilidad en la práctica clínica real. Con este objetivo, Hannes *et al.* diseñan un estudio retrospectivo que incluye 359 casos de HAI y 393 controles [19].

Se trata de un trabajo multicéntrico internacional en el que participan once hospitales de diez países de Norteamérica, Sudamérica, Europa y Asia, todos ellos especializados en enfermedades hepáticas (y ninguno destacado como específicamente pediátrico en el manuscrito publicado en 2008). Los datos se recogieron entre enero de 2005 y septiembre de 2006. En una primera parte se reclutaron casos y controles para configurar una submuestra de análisis o derivación (*training set*), y una vez desarrollado un modelo con finalidad predictiva, se continuó con la incorporación de pacientes para una muestra de validación (*validation set*). Los autores no especifican la técnica de muestreo (si es consecutiva o de otra naturaleza). Al no asignarse de forma aleatoria la pertenencia a estos dos subconjuntos de pacientes, la prevalencia de HAI varió entre los dos: Un 56,4% en la muestra de derivación (de un total de 443 pacientes) y un 27,7% en la muestra de validación (de un total de 393 pacientes).

Al respecto de la distribución de diagnósticos en ambas submuestras destaca la presencia de 10 casos de síndrome de solapamiento en la de derivación, con 8 cirrosis biliares primarias (CBP) con HAI y 2 CEAI, mientras que en la de validación no se incluyó ninguna.

En la descripción de los pacientes y métodos los autores especifican que, para la inclusión como casos, los pacientes debían de cumplir los criterios

diagnósticos clásicos de 1999 (ver en *Anexos*) con una puntuación suficiente como para considerarlos diagnósticos definitivos antes del inicio del tratamiento inmunosupresor y, como criterio adicional, deberían de presentar una respuesta positiva a dicho tratamiento. No establecen explícitamente los criterios que se siguieron para definir la mejoría.

Los únicos criterios de exclusión considerados fueron la presencia de hepatopatía previa, estar bajo tratamiento inmunosupresor o inmunomodulador o haber sido objeto de trasplante hepático.

El modelo máximo incluyó una serie de variables elegidas sobre la base de opinión de expertos: edad, sexo, presencia de autoanticuerpos (anti-Sm, ANA, AMA, anti-LKM1, anti-SLA/LP), nivel de gammaglobulina, de IgG, de inmunoglobulina A (IgA), de inmunoglobulina M (IgM), ausencia de hepatitis vírica e histopatología hepática.

Al respecto de esta última, como diferencia respecto al mismo ítem en los criterios clásicos, se definieron tres categorías para graduar la posibilidad de que los hallazgos con microscopía convencional fueran realmente sugestivos de HAI. La infiltración de los tractos portales por células plasmáticas es una característica reconocida de la HAI, sin embargo no es patognomónica ni altamente específica y además su ausencia no descarta el diagnóstico [101]. Por ello, según la propuesta de los profesores Dienes (Colonia) y Lohse (Hamburgo), la anatomía patológica puede ser atípica, compatible o típica. Las observaciones que definen esta última son la hepatitis de interfase, infiltrado en tracto portal con extensión hacia el lóbulo de células linfocíticas o linfoplasmocíticas, emperipolesis (penetración activa de una célula en otra célula de mayor tamaño) y formación de rosetas hepáticas (transformación microacinar de los hepatocitos) [203]. Para que una anatomía patológica pueda ser considerada típica de HAI deben de estar presentes cada uno de los datos descritos previamente. En caso de que se observe una hepatitis crónica con infiltración linfocítica sin estar presentes todos los datos típicos, se cataloga la anatomía patológica como compatible. De este modo, es atípico el hallazgo de

cualquier otro dato sugestivo de diagnósticos alternativos, como las inclusiones grasas en la esteatohepatitis no alcohólica (EHNA) y la presencia de conductillos biliares escasos en la enfermedad de Alagille. En el estudio de Hennes *et al.*, los patólogos de cada centro no estuvieron cegados respecto a los datos de la historia clínica de los pacientes, es decir, trabajaron en condiciones normales.

#### **4.3.2.1. Desarrollo y validación inicial del sistema simplificado**

Para la formulación del modelo con intención predictiva, se llevó a cabo un análisis univariante en la muestra de derivación. Las variables en las que se detectaron diferencias entre los casos de HAI y los controles con hepatopatías de otras etiologías se seleccionaron para la posterior elaboración del modelo. La comparación se llevó a cabo mediante el *test* U de Mann-Whitney para el caso de las variables cuantitativas y con el *test* exacto de Fisher para el de las categóricas. El valor diagnóstico de cada variable y modelo de combinación de variables se evaluó con el área bajo la curva COR. Para facilitar el uso práctico (“en la cabecera del paciente”) el sistema incorporó de forma categorizada ordinal las variables cuantitativas continuas, asignando valores crecientes de puntos a cada escalón [19].

El estudio de la influencia combinada de cada discriminador potencial se llevó a cabo por técnicas de regresión logística con eliminación o por pasos “hacia atrás” (*stepwise with backward elimination*). Los autores hablan de que se seleccionaron varios modelos, sin especificar ni describir su ecuación resultante, cuya validez fue evaluada en la submuestra reclutada a tal fin con el cálculo de la sensibilidad, la especificidad y sus correspondientes intervalos de confianza al 95% según el método de Clopper-Pearson [204,205]. Se compararon según la técnica de DeLong y el modelo resultante después de trabajar sobre la muestra de derivación fue el que combinaba los niveles de IgG, el título de ANA y anti-Sm y la histopatología hepática, con un área bajo la curva COR de 0,99. Por decisión de los investigadores se incluyó también la variable “exclusión de hepatitis vírica” que se comporta como categórica binaria [19,206].

El modelo con un comportamiento más óptimo en la muestra de validación fue el siguiente con su correspondiente asignación de puntos:

- ANA o anti-Sm con título  $\geq 1:40 \rightarrow 1$  punto.
- ANA o anti-Sm con título  $\geq 1:80$ , o anti-LKM1  $\geq 1:40$ , o anti-SLA positivo  $\rightarrow 2$  puntos (con un máximo restringido a dos puntos sumando los dos resultados de anticuerpos).
- Niveles de IgG por encima del límite superior de la normalidad  $\rightarrow 1$  punto.
- Niveles de IgG 1,1 veces por encima del límite superior de la normalidad ( $>10\%$  superior)  $\rightarrow 2$  puntos.
- Anatomía patológica atípica, compatible o típica  $\rightarrow 0, 1$  y  $2$  puntos, respectivamente.
- Exclusión de hepatitis vírica  $\rightarrow 2$  puntos.

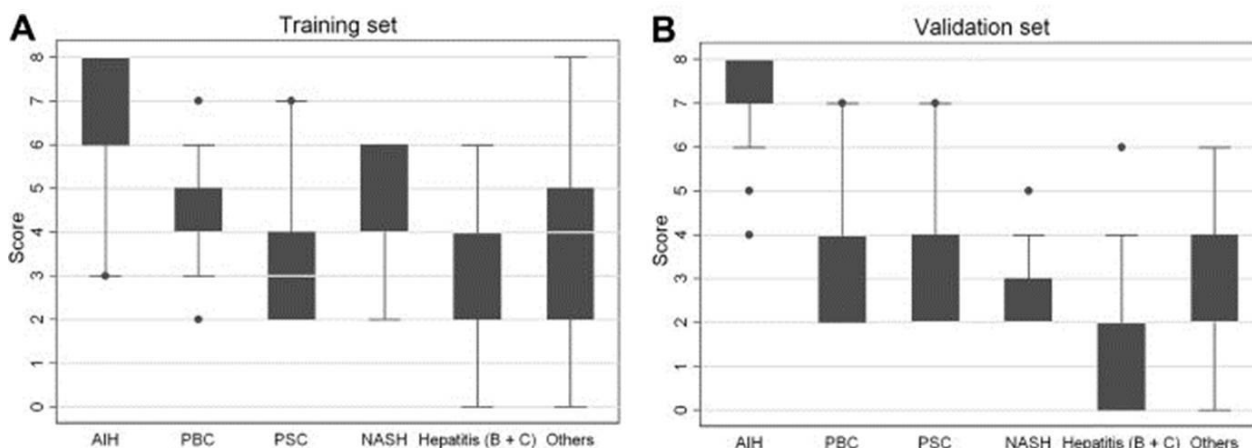


Figura 12: Puntuación según los criterios simplificados de 2008 de los pacientes con hepatitis autoinmune y sus controles en la muestra de derivación (A). Misma comparación en la muestra de validación (B). AIH: Hepatitis autoinmune. PBC: Cirrosis biliar primaria. PSC: Colangitis esclerosante primaria. NASH: Esteatohepatitis no alcohólica. Reproducido de Hennes et al. *Hepatology*. 2008;48:175. Con permiso de Wiley.

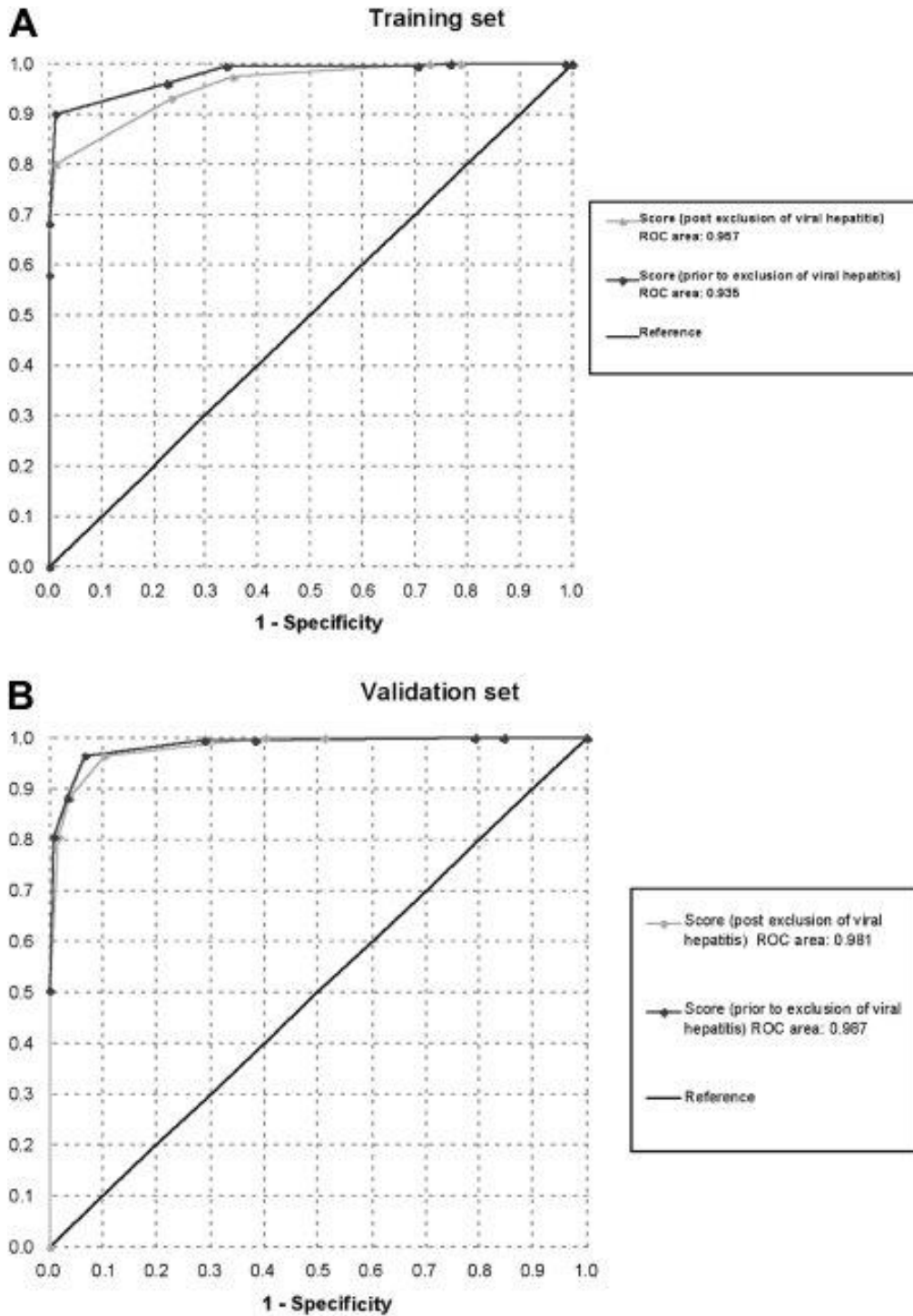


Figura 13: Curva de características operativas del receptor (COR) de los criterios simplificados de 2008 en la muestra de derivación. Comparación antes y después de la exclusión de hepatitis vírica (A). Curva COR con la misma comparación en la muestra de validación (B). Reproducido de Hennes et al. *Hepatology*. 2008;48:172. Con permiso de Wiley.

Considerando todos los pacientes (incluyendo los afectos de hepatitis vírica) se encontraron los indicadores de validez reflejados en las tablas 10 y 11.

**Tabla 10: Sensibilidad y especificidad de las diferentes variaciones de los criterios simplificados de 2008 sobre la muestra de derivación. Reproducido de Hennes et al. Hepatology. 2008:48;173. Con permiso de Wiley.**

Punto de corte	Sensibilidad	Intervalo de confianza al 95%	Especificidad	Intervalo de confianza al 95%
Puntuación $\geq 3$ sin considerar casos con hepatitis víricas	186/191 (97%)	94%-99%	100/154 (65%)	57%-72%
Puntuación $\geq 4$ sin considerar casos con hepatitis víricas	178/191 (93%)	89%-96%	118/154 (77%)	69%-83%
Puntuación $\geq 5$ sin considerar casos con hepatitis víricas	153/191 (80%)	71%-86%	152/154 (99%)	95%-100%
Puntuación $\geq 5$ incluyendo controles con hepatitis vírica	215/224 (96%)	93%-98%	105/159 (66%)	58%-73%
Puntuación $\geq 6$ incluyendo controles con hepatitis vírica	202/224 (90%)	86%-94%	123/159 (77%)	70%-84%
Puntuación $\geq 7$ incluyendo controles con hepatitis vírica	153/224 (68%)	62%-74%	157/159 (99%)	96%-100%

**Tabla 11: Sensibilidad y especificidad de las diferentes variaciones de los criterios simplificados de 2008 sobre la muestra de validación. Reproducido de Hennes et al. Hepatology. 2008:48;173. Con permiso de Wiley.**

Punto de corte	Sensibilidad	Intervalo de confianza al 95%	Especificidad	Intervalo de confianza al 95%
Puntuación $\geq 3$ sin considerar casos con hepatitis víricas	90/93 (97%)	91%-99%	114/127 (90%)	83%-94%
Puntuación $\geq 4$ sin considerar casos con hepatitis víricas	82/93 (88%)	80%-94%	122/127 (96%)	91%-99%
Puntuación $\geq 5$ sin considerar casos con hepatitis víricas	75/93 (81%)	71%-88%	125/127 (98%)	94%-100%
Puntuación $\geq 5$ incluyendo controles con hepatitis vírica	90/93 (97%)	91%-99%	224/240 (93%)	89%-96%
Puntuación $\geq 6$ incluyendo controles con hepatitis vírica	82/93 (88%)	80%-94%	232/240 (97%)	94%-99%
Puntuación $\geq 7$ incluyendo controles con hepatitis vírica	75/93 (81%)	71%-88%	238/240 (99%)	97%-100%

#### 4.3.3. Aplicabilidad de los criterios de clasificación por puntos para la hepatitis autoinmune en población pediátrica

En Medicina, la adopción de unos criterios diagnósticos para una determinada entidad, se considera un intento de estandarizar las observaciones y



los procedimientos clínicos entre distintos ámbitos, entornos asistenciales o centros de atención sanitaria. El objetivo es conseguir un “lenguaje común” o consensuar una definición para poder empezar a discutir sobre ella.

Los criterios diagnósticos deben incluir medidas bien definidas y ser de fácil aplicación para permitir la clasificación de los pacientes en un entorno clínico real y de forma homogénea. Desde el punto de vista pragmático para el que se conciben, idealmente deben de permitir distinguir pacientes candidatos a asignarse tratamientos específicos o diferenciales, incluso aunque el diagnóstico no esté claramente establecido.

En el caso concreto de la pediatría, la situación más común es que los criterios diagnósticos desarrollados en población adulta se adapten posteriormente a las especificidades del grupo etario. Se reconoce, por otro lado, que la actitud óptima sería que los criterios diagnósticos empleados en pediatría se ideen a partir de experiencias clínicas publicadas sobre población exclusivamente infantil [207].

Cumplir una serie de criterios aumenta la posibilidad de la enfermedad pero en ocasiones difícilmente permite excluir otras entidades. Es el caso paradigmático de los criterios originales de Jones para la fiebre reumática, que la definen como la combinación de artritis, fiebre, elevación de la velocidad de sedimentación globular y la presencia de infección por estreptococo del grupo A [208]. Una proporción considerable de niños con artritis idiopática juvenil también muestran las mismas características clínicas [209].

En el caso de la HAI ocurre un caso análogo: los criterios inicialmente fueron establecidos para adultos y posteriormente se adaptaron con cambios para población infantil. De este modo, en su primera versión de 1993, los criterios clásicos basados en el consenso del IAIHG no consideraban que los pacientes pediátricos requirieran un sistema diagnóstico específico o separado. Así se menciona en el trabajo de 2012 publicado por Mileti *et al.*, sobre la validación de los criterios simplificados en población íntegramente constituida por menores de dieciocho años [210]. Otros dos estudios han intentado evaluar los criterios

diagnósticos en niños y han mostrado resultados divergentes en algunos puntos. Se trata del trabajo de Ebbeson *et al.* de 2004, sobre los criterios clásicos revisados de 1999, y el de Hiejima *et al.* de 2011, sobre los criterios simplificados de 2008 [202,211].

Desde un punto de vista amplio los autores de los tres estudios concluyen que los criterios de 1999 y de 2008 son aplicables para la HAI en niños con unos buenos resultados por lo que respecta a la especificidad en la mayoría de los pacientes, pero también muestran algunas limitaciones.

**Tabla 12: Resumen de los estudios llevados a cabo en población pediátrica sobre la exactitud de los criterios diagnósticos para la hepatitis autoinmune. Adaptado de Ferri *et al.* World J Gastroenterol. 2012;18;4472. Con permiso de Baishideng Publishing Group.**

Estudios	Ebbeson <i>et al.</i> [202]	Mileti <i>et al.</i> [210]	Hiejima <i>et al.</i> [211]
Lugar y año de publicación	Canadá 2004	Estados Unidos 2012	Japón 2011
Tamaño muestral	28	68	56
Diagnósticos	21 HAI 4 CEP 3 CEAI	37 HAI según criterios 1999 31 HAI según criterios 2008 40 no HAI	20 HAI 36 no HAI
Criterios evaluados	1999	1999 y 2008	1999 y 2008
Resultados	18 de 21 (86%) con HAI puntuaron como HAI definitiva y 3 de 21 (14%) como probable. Todos los pacientes con CEP puntuaron como no HAI.	Criterios clásicos: 29 de 31 (94%) clasificados como HAI definitiva y 2 de 31 (6%) como probable. Criterios simplificados: 25 de 31 (81%) clasificados como HAI definitiva y 2 de 31 (6%) como probable. Los criterios de 2008 tienen una S de 87% y una E de 89% y no consiguieron identificar 4 pacientes con HAI y FHF	Criterios clásicos: S de 100% y E de 81%. Criterios simplificados: S de 55% y E de 86%. Los 5 pacientes con CEP se clasificaron como HAI por ambos criterios.
Conclusiones	Los criterios clásicos son útiles en pediatría. Emplear la GGT puede mejorar la especificidad en niños.	Los criterios de 2008 tienen una buena S y E. A los pacientes con FHF se les debe de aplicar los criterios de 1999. Globulina e IgG puede ser usadas indistintamente.	La E de los criterios simplificados es elevada. Los criterios simplificados no diferencian entre HAI y CEP y no parecen ser una herramienta útil en pediatría.

HAI: Hepatitis autoinmune. CEP: Colangitis esclerosante primaria. CEAI: Colangitis esclerosante autoinmune. GGT: Gamma-glutamil transpeptidasa. S: Sensibilidad. E: Especificidad. FHF: Fallo hepático fulminante. IgG: Inmunoglobulina G

#### **4.3.3.1. El caso de las hepatopatías colestásicas**

En primer lugar, en Ebbeson *et al.* se encuentra que los niños con diagnóstico de CEP son adecuadamente clasificados como “no HAI” por los criterios clásicos, mientras que en Hiejima *et al.* se demuestra que los cinco pacientes con este diagnóstico de su muestra son catalogados erróneamente como HAI tanto por el sistema de 1999 como por el simplificado [202,211]. Esta observación juntamente con el hecho de que es posible la coexistencia de CEP con HAI (síndrome de solapamiento), conlleva que la consideración de los marcadores bioquímicos de colestasis en los criterios de 1999 puede optimizar su especificidad, pero no así la de los de 2008. En consecuencia se añade la recomendación de practicar una prueba de imagen colangiográfica, en concreto mediante colangio-resonancia magnética (colangio-RM) o colangiografía retrógrada endoscópica (CGRE) [40,178]. Esta propuesta de ampliar el estudio de los casos pediátricos con una colangiografía se ha recogido posteriormente en los criterios de la ESPGHAN/NASPGAHN de 2009 [93].

#### **4.3.3.2. El caso del fallo hepático fulminante**

En segundo lugar tenemos la limitación mencionada por Mileti *et al.* sobre la confiabilidad de los criterios simplificados de 2008 para diagnosticar las HAI que se presentan como fallo hepático fulminante (FHF) [210]. La gravedad y la rapidez de instauración del cuadro hacen particularmente interesante disponer de un recurso para el diagnóstico etiológico preciso. Si se dispusiera de unos criterios diagnósticos de HAI que funcionasen bien en este escenario clínico, justificaría el inicio precoz del tratamiento inmunosupresor con la consiguiente esperable mejoría de la morbimortalidad [207].

#### **4.3.3.3. El ítem de los autoanticuerpos**

Los autoanticuerpos utilizados en los sistemas diagnósticos de la HAI (tanto los de 1999 como los de 2008 los incluyen) se presentan en niños a títulos frecuentemente más bajos que los usuales con los que aparecen en adultos [212].

La reactividad de ANA, anti-Sm y anti-LKM1 es baja en niños lo que conduce a considerar relevante y a dar valor a títulos de 1:20 de ANA y anti-Sm, y de 1:10 de anti-LKM1 [28].

Tanto los criterios clásicos revisados como los simplificados consideran el nivel de significación para los títulos de autoanticuerpos en 1:40. Esta categorización es por convención, pero fundamentada en el hecho de que los laboratorios de análisis clínicos en el pasado no han realizado rutinariamente diluciones por debajo de este umbral. Por consiguiente, un niño con títulos reales de 1:20 podría considerarse falso negativo.

Por este motivo y por la posibilidad de una forma seronegativa, en un paciente pediátrico con una historia clínica compatible pero con autoanticuerpos negativos no debe descartarse la HAI del diagnóstico diferencial [213].

#### **4.3.3.4. El ítem de la anatomía patológica**

En los criterios de 1999, cada dato histopatológico es puntuado por separado en función de su especificidad de HAI, resultando en tres categorías con puntuación positiva (formaciones en roseta, infiltración linfoplasmocitaria y hepatitis de interfase en orden creciente de asignación) y en tres categorías con puntuación negativa (datos de otras etiologías, afectación biliar con datos definidos y nada de lo anterior) [61]. En contraste, los criterios simplificados solo incluyen dos parámetros: histología compatible e histología típica [19].

Lo cierto es que la biopsia hepática puede no ser posible al inicio de la sospecha de la HAI en un niño, o incluso durante el seguimiento ulterior, dado que muchos casos habitualmente cursan con disfunción hepática y coagulopatía secundaria. En estos casos, puede ser necesaria la administración de vitamina K (el componente deficitario por la enfermedad hepática típicamente no mejora con esta medida), la transfusión previa de plasma fresco congelado o recurrir a un acceso alternativo al percutáneo convencional, tal como la vía transyugular [214]. El clínico a veces no dispone de información suficiente y se enfrenta a la necesidad de iniciar

de forma empírica tratamiento con corticoesteroides sin un conocimiento real de la condición histológica del paciente [207].

Björnsson *et al.* han cuestionado la importancia de la anatomía patológica en los casos típicos de HAI con serologías claramente positivas y en ausencia de otras posibilidades diagnósticas alternativas. Han observado que el 95% de los pacientes con HAI (excluyendo fallos hepáticos y síndromes de solapamiento) tienen unos datos compatibles en la biopsia hepática. El estudio retrospectivo que llevan a cabo destaca que el 86% de los casos sospechosos por clínica y parámetros analíticos, pero con histopatología atípica, acaban recibiendo tratamiento inmunosupresor, con lo que esta exploración no contribuye a un cambio de actitud médica. En esta línea, también ponen de relieve que los casos de HAI con una anatomía patológica tanto atípica como compatible no difieren significativamente por lo que respecta a la proporción de casos seronegativos o a los niveles de gammaglobulina [92,97].

Se reconoce el interés de disponer de una muestra hepática obtenida por biopsia para estudiar la histopatología antes de proceder al inicio de un tratamiento que altere la arquitectura y el grado de infiltración linfocitaria y, por lo tanto, dificulte el diagnóstico posterior. No obstante es un proceder apoyado ampliamente el que, si no es posible obtener una muestra para estudio histológico, se inicie el tratamiento empíricamente en pacientes con alta sospecha clínica y datos de laboratorio sugestivos [207]. Con todo, aunque la experiencia de Björnsson demuestra que la HAI se puede diagnosticar con suficiente seguridad a través de una combinación de los datos clínicos y analíticos, aún en ausencia de información anatomopatológica, la guía de práctica clínica de la *American Association for the Study of Liver Diseases* mantiene la recomendación de realizar una biopsia hepática en todos los casos en los que no esté contraindicado [40].

#### 4.3.4. Los criterios diagnósticos pediátricos propuestos por la ESPGHAN y la NASPGHAN (2009)

Tanto los criterios clásicos del IAIHG como los criterios simplificados de 2008 incluyen ítems positivos y negativos y se han desarrollado como sistemas diagnósticos para ser empleados en investigación en población general, sin considerar específicamente las características de los pacientes pediátricos [19,61].

En el año 2009, un grupo de trabajo conjunto de la ESPGHAN y de la NASPGHAN encabezado por Giorgina Mieli-Vergani, Solange Heller y Paloma Jara, después de revisar los principales elementos diferenciales entre las formas de HAI que se presentan en niños y adultos, proponen siete criterios diagnósticos relevantes para ser usados en pediatría [28,93].

Elevated transaminases	ANA and/or SMA
Positive autoantibodies	(titer $\geq$ 1:20) = type 1 AIH
	Anti-LKM1 (titer $\geq$ 1:10)
	= type 2 AIH
	Anti-LC1 = type 2 AIH
	Anti-SLA = present in type 1
	or 2 AIH or in isolation
Elevated immunoglobulin G	
Liver biopsy	Interface hepatitis
	Multilobular collapse
Exclusion of viral hepatitis	
Exclusion of Wilson disease	
Normal cholangiogram	
(nuclear magnetic resonance	
or retrograde cholangiography)	

---

AIH = autoimmune hepatitis, ANA = antinuclear antibody, anti-LKM-1 = anti-liver/kidney microsomal antibody type 1, anti-LC1 = anti-liver cytosol type 1 antibody, anti-SLA = anti-soluble liver antigen antibody, SMA = anti-smooth muscle antibody.

**Figura 14: Criterios para el diagnóstico de la hepatitis autoinmune en población pediátrica. Reproducido de Mieli-Vergani et al. J Pediatr Gastroenterol Nutr. 2009;49;159. Con permiso de Lippincott Williams & Wilkins.**

En primer lugar, se considera el signo analítico más constante, que es la hipertransaminasemia, sin incluirse ningún síntoma clínico debido a su variabilidad y la existencia mayoritaria de formas asintomáticas o paucisintomáticas.

Los autoanticuerpos se consideran positivos a título de 1:20 en el caso de los ANA y los anti-Sm; a título de 1:10, para los anti-LKM1 o simplemente por el hecho de detectarse, con independencia del título, anti-LC1 y anti-SLA/LP. Se reconoce la posibilidad de las HAI seronegativas, pero no se contempla en los criterios diagnósticos debido a que, aunque se sabe que esta presentación es infrecuente en adultos, la prevalencia y características clínicas todavía están por definir completamente en niños [215].

Al igual que en los criterios clásicos y los simplificados, la presencia de hipergammaglobulinemia es un elemento diagnóstico importante debido a su especificidad.

Al respecto de la anatomía patológica, a diferencia de los criterios clásicos revisados de 1999 y los simplificados de 2008, no se describen categorías con pesos de más o menos entidad. Se considera positiva la existencia de hepatitis de interfase (lesión aguda) o colapso multilobular (lesión de tendencia a la cronicidad).

Los dos criterios negativos son la exclusión de hepatitis vírica (al igual que en los criterios simplificados) y, además, se añade la exclusión de la enfermedad de Wilson, cuya forma de presentación infantil suele no incluir síntomas neurológicos y cursa con hipertransaminasemia oligo o asintomática, lo que dificulta el diagnóstico diferencial inicial [216,217].

Finalmente se incluye la presencia de una prueba de imagen de la vía biliar (colangio-RM o CGRE) con resultado normal. En pediatría, la CEP se asocia frecuentemente con rasgos autoinmunes floridos como por ejemplo la elevación de niveles de autoanticuerpos (particularmente anti-Sm), hipergammaglobulinemia o hepatitis de interfase, lo que determina que se hable, en su lugar, de CEAI. Estos hallazgos son compartidos con la HAI. A menudo no se encuentran marcadores bioquímicos de colestasis como elevación de FA o GGT, sobre todo al debut de la enfermedad. Por este motivo se recomienda fundamentar el diagnóstico de la CEAI en estos estudios colangiográficos [93]. Como se ha comentado en apartados previos, la CEAI se considera tan prevalente como la HAI de tipo 1 en la infancia,

pero afecta a niños y niñas por igual. Responde satisfactoriamente al tratamiento inmunosupresor con reducción de la inflamación del parénquima hepático, normalización de las transaminasas y mejoría de los síntomas clínicos si los hay [30]. Los criterios del IAIHG no permiten diferenciar entre HAI y CEAI [93]. La presencia de este ítem, además de la reducción del dintel de positividad de los autoanticuerpos, constituye el elemento más específico de los criterios ESPGHAN/NASPGHAN de 2009 respecto a los previos.

Esta propuesta todavía no ha sido validada en un estudio con un diseño específico. De hecho, se basan en elementos teóricos y no se han planteado para funcionar ni como un sistema de puntos ni como una serie de criterios de presencia obligatoria.





# 5

## Hipótesis de trabajo



## Hipótesis de trabajo

---

Las versiones simplificadas de los criterios diagnósticos clásicos para la hepatitis autoinmune (propuesta de 2008 y propuesta ESPGHAN/NASPGHAN de 2009) son adecuadas y útiles para ser usadas durante la edad pediátrica y originan clasificaciones clínicamente concordantes entre ellas y con el diagnóstico médico experto.



# 6

## Objetivos



## **6.1. Objetivo general**

---

Estudiar la exactitud de los criterios diagnósticos pediátricos simplificados para la hepatitis autoinmune (propuesta de 2008 y propuesta ESPGHAN/NASPGHAN de 2009), la discrepancia de las clasificaciones clínicas basadas en dichos criterios y su utilidad para la toma de decisiones clínicas.



## 6.2. Objetivos específicos

---

1) Conocer la prevalencia de HAI en una población pediátrica con diagnóstico de sospecha.

2) Comprobar la validez de los criterios simplificados de 2008. Además de utilizar los puntos de corte sugeridos en la bibliografía, encontrar otro punto óptimo, penalizando por igual los falsos negativos y los falsos positivos.

3) Revisar sistemáticamente la literatura sobre los criterios simplificados de 2008. Realizar, si es factible, un meta-análisis de pruebas diagnósticas, añadiendo el resultado del segundo objetivo a los comunicados por los autores de los trabajos originales.

4) Estudiar la fiabilidad de los criterios simplificados de 2008.

5) Diseñar un nuevo modelo con finalidad predictiva para las HAI pediátricas, basado en los criterios ESPGHAN/NASPGHAN de 2009, que se pueda aplicar a las situaciones en las que esta posibilidad está incluida en el diagnóstico diferencial. Transformarlo posteriormente en unos criterios de clasificación que se comporten como un sistema de puntos.

6) Evaluar la validez de los nuevos criterios desarrollados en el punto anterior.

7) Establecer el grado de concordancia entre las clasificaciones realizadas por los dos sistemas diagnósticos y el diagnóstico experto.

8) Verificar la utilidad de las clasificaciones realizadas por los dos sistemas diagnósticos sobre la base de la teoría del análisis de decisiones clínicas.

# 7

## Justificación de la unidad temática



## Justificación de la unidad temática

---

La diversidad de los objetivos específicos de la tesis se puede agrupar en tres bloques temáticos, cada uno de los cuales contempla como objeto el análisis de los principales conceptos en torno a la bondad de un sistema diagnóstico.

En primer lugar, tenemos el estudio de la **validez**, que es, en esencia, la verificación de que los criterios de clasificación son una medida adecuada de la probabilidad del diagnóstico.

Existe también el interés de comprobar la **fiabilidad o reproducibilidad**, que consiste en verificar si los criterios producen resultados consistentes cuando se repiten, operados por médicos distintos, en las mismas condiciones y se interpretan sin conocer otra información previa.

Finalmente conviene explorar la **utilidad** de los criterios de clasificación para la toma de decisiones terapéuticas en un entorno y unas condiciones clínicas reales.

La distribución por capítulos de la tesis se hará siguiendo la lógica de la separación entre estos puntos de vista, aplicados a cada uno de los dos criterios simplificados que se han expuesto en la introducción para el diagnóstico de la hepatitis autoinmune en pediatría: la propuesta de 2008 y un potencial sistema de puntos generado a partir de los criterios de la ESPGHAN y la NASPGHAN del 2009.

Los elementos que definen la validez, fiabilidad y utilidad de los sistemas diagnósticos confluyen para determinar cómo de exactos son estos sistemas, lo que sirve de fundamento para justificar que los apartados que integran esta tesis comparten el mismo objetivo general.

En las secciones siguientes se expondrá, primero, el diseño y la metodología general del estudio. Posteriormente y de forma sucesiva se resolverán los objetivos específicos y se detallarán los resultados organizándolos por capítulos siguiendo la estructura siguiente:

**Capítulo 1** – Se abordarán los objetivos 1 y 2, al respecto de la obtención de la prevalencia de la hepatitis autoinmune en una muestra pediátrica que represente la población diana para la aplicación de los criterios simplificados de 2008, y el cálculo de los estimadores de validez en dicha población.

**Capítulo 2** – Se resolverá el objetivo específico 3, a través de una revisión sistemática y un meta-análisis sobre los principales indicadores de validez de los criterios simplificados de 2008.

**Capítulo 3** – Tratará el objetivo específico 4, sobre la fiabilidad de los mismos criterios simplificados de 2008.

**Capítulo 4** – Dará respuesta a los objetivos específico 5 y 6, con la obtención de un modelo predictivo basado en los ítems de los criterios ESPGHAN/NASPGHAN de 2009, que posteriormente se transformará en un sistema de puntos. Se estudiará la validez tanto de la ecuación resultante como de los criterios por puntos.

**Capítulo 5** – Se abordará el objetivo específico 7 a través de una evaluación de la concordancia entre las clasificaciones realizadas por los sistemas diagnósticos estudiados.

**Capítulo 6** – Finalmente, para llegar al objetivo específico 8, se considerarán los dos sistemas diagnósticos simplificados como sendos índices de predicción clínica para establecer su utilidad en la práctica asistencial.



8

Diseño



## **8.1. Diseño general, pacientes y material**

---

### **8.1.1. Tipo de estudio**

Desde un punto de vista amplio, se ha diseñado como un estudio analítico transversal de fase III de valoración de la exactitud de pruebas diagnósticas.

Se añade un apartado con unas características particulares al respecto de la arquitectura del diseño: la revisión sistemática y el meta-análisis de estudios diagnósticos del segundo capítulo.

### **8.1.2. Ámbito institucional**

El estudio ha tenido lugar en los hospitales Sant Joan de Déu y Vall d'Hebron de Barcelona. Concretamente, ha sido promovido por la Unidad Integral de Hepatología Compleja y Trasplante Hepático Pediátrico, que integra las áreas de Hepatología Pediátrica de los dos centros.

Se trata de los dos hospitales terciarios que concentran la atención especializada a la patología hepática infantil de Catalunya. Se atienden casos procedentes tanto de las respectivas áreas sanitarias como del resto del país dado que son hospitales de referencia de las hepatopatías en niños para este territorio. Anualmente se atienden alrededor de 600 niños.

### **8.1.3. Muestreo**

Consecutivo, durante un periodo de estudio que se extiende desde enero de 2005 hasta enero de 2017. Con recogida de datos ambispectiva.

Abarca una primera fase de recopilación retrospectiva de los datos, que se ha completado por revisión manual de las historias clínicas del archivo de ambos hospitales.



Posteriormente, desde enero de 2016, ha continuado con una fase de recogida prospectiva. Se ha utilizado para ampliar la muestra, confirmar los datos de la fase retrospectiva y enmendar posibles clasificaciones diagnósticas erróneas iniciales.

La información se ha volcado a un archivo informático mediante un formulario específico prediseñado (que se describe más adelante).

#### **8.1.4. Población**

Está constituida por los niños y adolescentes en situación clínica y/o analítica compatible con HAI, de modo que esta posibilidad esté incluida en el diagnóstico diferencial. La idea ha sido obtener una muestra que represente a la población en la que tenga sentido aplicar los criterios diagnósticos. El límite de edad se ha establecido en 18 años.

Han debido de cumplir alguno de los criterios de inclusión y no se han considerado en el análisis si han presentado alguno de los criterios de exclusión.

##### **8.1.4.1. Criterios de inclusión**

1) Pacientes con indicación de biopsia hepática por signos clínicos y/o analíticos de hepatopatía aguda o crónica (fundamentalmente hipertransaminasemia), tanto si coexiste un patrón de colestasis como si no, incluyendo las que se efectúen por técnica percutánea, laparoscópica o por acceso transyugular.

2) Insuficiencia hepática aguda definida según el consenso del PALFSG (*Pediatric Acute Liver Failure Study Group*): deterioro agudo de las funciones hepáticas en un paciente sin antecedentes de patología hepática, con signos de encefalopatía y con un INR >1,8 que no responde a vitamina K o un factor V por debajo del 50% [218].

#### **8.1.4.2. Criterios de exclusión**

- 1) Trasplante hepático ortotópico o de hepatocitos.
- 2) Diagnóstico previo de cualquier enfermedad hepática (congénita o adquirida).
- 3) Diagnóstico previo de cualquier enfermedad sistémica o metabólica con afectación hepática.
- 4) Lesión ocupante de espacio o masa intrahepática.
- 5) Biopsia hepática indicada para toma de cultivos en situación de fiebre de origen desconocido.
- 6) Debut antes de los primeros 6 meses de vida.

#### **8.1.5. Cálculo del tamaño muestral**

Existen dos factores que definen el número necesario de pacientes a incluir en la muestra. Previo a la recogida de los datos y a la formulación final del protocolo del estudio, se ha llevado a cabo la estimación en base a los procedimientos recomendados según estos dos condicionantes. A continuación se describe el razonamiento que se siguió para el cálculo y la decisión final.

##### **8.1.5.1. Estimación del número de pacientes para la elaboración del modelo predictivo de regresión logística**

En primer lugar, tenemos el objetivo específico de hallar un modelo predictivo de regresión logística con los criterios diagnósticos ESPGHAN/NASPGHAN de 2009 para la HAI pediátrica. El estudio completo del nuevo modelo consta de dos fases.

#### **8.1.5.1.1. Fase de elaboración o desarrollo**

Se asume que para la elaboración de un modelo de regresión hay que utilizar una muestra que contenga al menos diez casos por cada variable que se pretenda incluir en el modelo máximo [219,220].

Dado que los criterios de 2009 contemplan siete ítems (que se pueden consultar rápidamente en los materiales anexos), se deberá de conseguir reclutar 70 HAI. Esta parte de la muestra se utilizaría para el desarrollo del nuevo modelo predictivo, y contendría dos terceras partes del total de casos.

#### **8.1.5.1.2. Fase de validación**

Otra fracción de la muestra, que debería estar constituida por 35 casos de HAI, se utilizará para validar el nuevo modelo construido. Constituirá una tercera parte del número de casos de HAI totales.

Por tanto, el tamaño muestral global debería incluir un total de, por lo menos, 105 pacientes pediátricos con diagnóstico positivo de HAI.

Hasta donde sabemos, no hay estudios que informen de la prevalencia de HAI en población pediátrica con sospecha diagnóstica de presunción o clínica compatible. Sin embargo, un registro de biopsias hepáticas en niños del Hospital La Fe de València, apunta a la presencia de un 50% de HAI entre todas las que se realizaron con intención diagnóstica en niños sin problemas hepáticos de base [221].

En consecuencia, se estimaría necesario un tamaño muestral global final de 210 pacientes.

#### **8.1.5.2. *Estimación del número de pacientes para el estudio de la validez de los criterios diagnósticos***

No existe un consenso sobre la mejor metodología para el cálculo del tamaño muestral en estudios de evaluación de pruebas diagnósticas [222,223]. Las aproximaciones más estrictas requieren disponer de una estimación previa de la

sensibilidad o la especificidad esperable, así como de la prevalencia de la enfermedad [224].

Para hallarlas, se ha llevado a cabo un estudio piloto basado en las aproximaciones más laxas al respecto del tamaño muestral necesario. Éstas consideran como requisitos mínimos el que en la muestra no debe haber un número inferior a 30 pacientes que presenten positivamente la condición de interés en base al patrón oro, y otros 30 pacientes en el grupo de diagnósticos diferenciales (grupo de “no casos”) [225]. Adicionalmente se ha seguido la otra norma clásica de que el tamaño muestral suficiente es aquel que permite un mínimo de 10 sujetos en cada una de las casillas de la tabla tetracórica de contingencia [225,226].

En efecto, el estudio piloto en el que se revisaron 102 casos arrojó una prevalencia del 48,0% (IC95% 38,6% a 57,6%). Para el punto de corte de 6 en los criterios simplificados de 2008 (diagnóstico probable según la bibliografía), se obtuvo una Se y una Sp de 74,0 y 94,2%, respectivamente. Para valores de 7 o superiores (diagnóstico definitivo), la Se fue de 46,0%, la Sp de 98,1%. Estos resultados, con una prevalencia equiparable a la del registro de biopsias hepáticas del Hospital La Fe, fueron presentados en el XXIII congreso de la Sociedad Española de Gastroenterología, Hepatología y Nutrición Pediátrica, en Gijón, el 12 de mayo de 2016 [227].

De forma arbitraria, pero siguiendo el proceder más habitual, se consideró adecuado trabajar con un error  $\alpha$  de 0,05 para una precisión absoluta ( $L$ ) del 10%, es decir, para obtener una estimación de los indicadores de validez que, con una confianza del 95% no se alejen más de un 10% por arriba ni por abajo de los indicadores reales de la población diana.

Utilizando esta información en las fórmulas de Buderer, llegamos a las siguientes propuestas de tamaño muestral, donde  $N_{Se}$  y  $N_{Sp}$  representan el número de pacientes necesario basado en la sensibilidad y la especificada respectivamente [224]:

Para el punto de corte en 6, los criterios simplificados obtuvieron una estimación de sensibilidad de 0,74 y de especificidad de 0,942, de modo que

$$N_{Se} = \frac{Z_{1-\alpha/2}^2 \times Se \times (1 - Se)}{L^2 \times Prevalencia} = \frac{1,96^2 \times 0,74 \times (1 - 0,74)}{0,1^2 \times 0,48} = 153,98$$

$$N_{Sp} = \frac{Z_{1-\alpha/2}^2 \times Sp \times (1 - Sp)}{L^2 \times (1 - Prevalencia)} = \frac{1,96^2 \times 0,942 \times (1 - 0,942)}{0,1^2 \times (1 - 0,48)} = 40,36$$

Para el punto de corte en 7, los criterios simplificados obtuvieron una estimación de sensibilidad de 0,46 y de especificidad de 0,981, de modo que

$$N_{Se} = \frac{Z_{1-\alpha/2}^2 \times Se \times (1 - Se)}{L^2 \times Prevalencia} = \frac{1,96^2 \times 0,46 \times (1 - 0,46)}{0,1^2 \times 0,48} = 198,80$$

$$N_{Sp} = \frac{Z_{1-\alpha/2}^2 \times Sp \times (1 - Sp)}{L^2 \times (1 - Prevalencia)} = \frac{1,96^2 \times 0,981 \times (1 - 0,981)}{0,1^2 \times (1 - 0,48)} = 13,77$$

De los cuatro valores, redondeados al entero superior, el valor más elevado es 199 y por lo tanto representa la estimación más conservadora del tamaño muestral necesario.

Estas cifras coinciden con las obtenidas con el software Epidat y se demuestra que son bien aproximadas a través del nomograma de Malhotra e Indrayan [228,229]. La resolución gráfica del nomograma de Carley es insuficiente para una buena estimación del tamaño muestral en nuestro caso [230].

### **8.1.5.3. Decisión final sobre el tamaño muestral**

El factor limitante para la decisión del tamaño muestral final es aquel que exija un número mayor de pacientes. Para esta tesis, ha resultado ser la elaboración del modelo de regresión logística del objetivo específico 5, que requiere 210 pacientes, discretamente superior a los 199 necesarios para el estudio de validación del sistema diagnóstico. De esta suma, 105 deberán de tener el diagnóstico de hepatitis autoinmune.

No se disponen de elementos que permitan anticipar los indicadores de validez del sistema de puntos basado en los criterios ESPGHAN/NAPSGHAN de 2009, y por lo tanto no se puede estimar según la regla de Buderer el tamaño necesario

para ello. En cualquier caso, su validación externa tendrá lugar sobre una tercera parte de la muestra total, que, aunque será suficiente para validar la ecuación del modelo de regresión logística, no lo será para validar el sistema de puntos si asumimos una sensibilidad y especificidad similares a las de los criterios simplificados de 2008. Así, el intervalo de confianza de los parámetros de validez de la nueva propuesta de criterios se prevé mayor que el que se obtendrá para la evaluación de los criterios del 2008, que se llevará a cabo con la muestra completa. En cualquier caso, se hará cumplimentación de la norma general de incluir un mínimo de 30 pacientes en los grupos de casos y no casos [225].

### **8.1.6. Aplicación de las pruebas a validar**

Existen dos pruebas diagnósticas a validar en este proyecto:

En primer lugar, tenemos los criterios simplificados de 2008. Fueron propuestos por Hennes *et al.* para hacer más sencillo el diagnóstico clínico de la HAI [19]. Se basó en una selección de variables de los criterios clásicos con un mejor potencial discriminador por técnicas de regresión logística. Para ello se trabajó con una población de adultos. El resultado final fue un listado de solo 4 parámetros: autoanticuerpos, elevación de IgG, anatomía patológica hepática y descarte de etiología vírica, todos ellos con comportamiento como variables categóricas. Hasta el momento, los dos estudios de validación de estos criterios en población pediátrica han obtenido conclusiones dispares y en ninguno de ellos se calculan valores predictivos fiables por razones de diseño [210,211].

En segundo lugar, está el nuevo modelo que vamos a proponer, basado en una optimización de la propuesta de criterios diagnósticos para la HAI pediátrica establecidos por la ESPGHAN y la NASPGHAN en 2009. Además de los parámetros del sistema simplificado de 2008 (con modificaciones en los títulos de anticuerpos y las categorías de la histopatología), incluye la presencia de hipertransaminasemia, la exclusión de enfermedad de Wilson y la existencia de una prueba de imagen de vías

biliares normal. El sistema original está construido de manera teórica, no funciona como un sistema de puntuación y no está todavía validado su uso considerándolo como un listado de criterios de presencia obligatoria. Está fundamentado en las principales diferencias de la HAI pediátrica con respecto a los adultos [93].

La medición óptima de todos los criterios obliga a practicar un análisis sanguíneo, una biopsia hepática y una prueba de imagen específica para el estudio de la vía biliar (colangio-RM o colangiografía retrógrada). El protocolo habitual de estudio para las sospechas de HAI incluye la realización de dichas exploraciones desde el año 2009.

La estrategia planeada consistió en emplear una tabla 3x2 para manejar los resultados indeterminados de los criterios de clasificación a validar, y estimar los indicadores de validez *por intención de diagnosticar* [231]. Solo se previó que pudiera haber un número significativo de valores perdidos en la categoría de la normalidad de las pruebas de imagen de la vía biliar en los criterios ESPGHAN/NASPGHAN 2009. Para esta situación se plantearía un tratamiento de los datos perdidos mediante una imputación múltiple.

Las imágenes de la vía biliar, que se hará preferentemente por colangio-RM por ser menos invasiva e igualmente útil, fueron interpretadas por el radiólogo intensificado en resonancia magnética.

El estudio microscópico de las piezas de biopsia lo llevó a cabo el anatómopatólogo encargado de cada centro, también según su práctica estándar.

Los autoanticuerpos se detectaron por inmunofluorescencia indirecta.

### **8.1.7. Aplicación del patrón de referencia**

No existe un patrón oro estricto para el diagnóstico de la hepatitis autoinmune. Para definirla a efectos de investigación se considera HAI todos los casos que cumplan con los criterios clásicos revisados de 1999. La aplicación se hizo post-tratamiento y se consideró como caso positivo el alcanzar una puntuación de

$\geq 12$ . Entre 12 y 17 se considera HAI probable y  $\geq 18$  es indicativo de HAI definitiva. Se utilizó el punto de corte  $\geq 12$  basado en las conclusiones del estudio de Czaja *et al.* del 2011, sobre una muestra de adultos, que demuestra que la designación de HAI probable se basa en diferencias en las manifestaciones clínicas y no refleja necesariamente una distinta validez diagnóstica o cambios en la proporción de respuesta al tratamiento [201]. Para aquellos casos que no recibieron tratamiento inmunosupresor, se consideró HAI probable una puntuación entre 10 y 15 (incluido), y HAI definitiva, una puntuación superior a 15.

Tal como se recomienda para población pediátrica, como excepción a la norma, se asignó +1 punto en el ítem de los autoanticuerpos si se detectaron títulos positivos por debajo de 1:40 [28,61]. Mientras no se especificara en un sentido contrario en la historia clínica, la ingesta de alcohol se consideró no significativa.

Los criterios clásicos revisados de 1999 del IAHG se han utilizado como patrón de referencia en multitud de estudios de validación diagnóstica en población adulta [35,90,232–235]. El trabajo sobre el rendimiento de los criterios simplificados en niños llevado a cabo por Hiejima en 2011 y Mileti en 2012 también los emplean como criterio de referencia diagnóstico [210,211]. Ambos informan de una sensibilidad del 100% para los criterios clásicos (esperable dado que se utilizan como criterio de selección de casos). Hiejima, además, refiere una especificidad del 81%, mientras que Mileti no estima este dato de forma directa, pero se deduce que es del 100% en base al área bajo la curva COR.

Para dar robustez al diagnóstico y evitar falsos positivos, se comprobó que los casos clasificados como HAI probable y definitiva cumplían necesariamente dos requisitos:

- 1) Tener, como mínimo, una anatomía patológica compatible con HAI según la definición del *score* simplificado de 2008, es decir, una hepatitis crónica con infiltración linfocítica [19].



2) Constatarse respuesta al tratamiento farmacológico definida como mejoría de los síntomas, si los hay, y disminución de las transaminasas. Es decir, considerar como criterio necesario el ítem de la respuesta terapéutica.

Los criterios validados del IAIGH y revisados en 1999 no pueden considerarse un patrón oro de forma estricta. Para serlo, deberían de reunir la condición doble de disponer de una sensibilidad y especificidad del 100%. Sin embargo los indicadores de validez que ha demostrado el sistema clásico han resultado muy buenos pero subóptimos [236,237].

Así pues, los criterios clásicos de 1999 se comportan como un patrón oro difuso (*fuzzy gold standard*). Como medida de compensación de este fenómeno, y para minimizar los falsos negativos, se llevó a cabo un análisis de casos discrepantes, que tiene como objetivo evitar la sobredimensión de los indicadores de validez cuando se emplean *gold standard* con errores potencialmente dependientes o relacionados con los errores del sistema diagnóstico a validar [238,239]. El tratamiento consistió en reclasificar, durante la fase prospectiva del estudio, los casos inicialmente no dados como HAI y con una puntuación de 6 o más en los criterios simplificados, pero en los que se diera simultáneamente que el médico refiriera explícitamente en la historia clínica que su juicio diagnóstico iba a favor de esta entidad y, al mismo tiempo, cumpliera los dos requisitos de robustez mencionados anteriormente.

### **8.1.8. Recogida de otras variables**

Con la intención de describir las características de la población muestreada y de que se pueda analizar su representatividad y validez externa, se recogieron otra serie de variables demográficas, como la edad, el sexo y los diagnósticos previos de cada paciente.

También se incluyó información relevante para demostrar el diagnóstico alternativo a HAI cuando lo hubo. Se previó que el resto de las entidades que

configurarían la muestra serían, principalmente, colangitis esclerosante primaria, síndromes *overlap*, hepatitis vírica aguda, enfermedad de Wilson, esteatohepatitis no alcohólica, fibrosis quística, déficit de alfa1-antitripsina, enfermedades de depósito, colestasis intrahepática familiar progresiva, fibrosis hepática congénita y errores innatos del metabolismo no diagnosticados previamente. En el caso de la enfermedad de Wilson, cuya ausencia cuenta en los criterios ESPGHAN/NASPGHAN 2009, se recopilaron los niveles de ceruloplasmina y, si estaban disponibles, el resultado de la cuantificación de la excreción urinaria de cobre (con o sin administración previa de penicilamina), la medición de cobre en tejido hepático y los resultados de los estudios genéticos.

#### **8.1.9. Técnicas de medición y extracción de los datos**

En una primera fase se han recuperado los casos que cumplían los criterios de inclusión de las historias clínicas almacenadas tanto en soporte informático como en el Servicio de Documentación. Las palabras clave empleadas para ello han sido “insuficiencia hepática aguda” y “hepatitis” (y términos relacionados) en el campo de diagnóstico y “biopsia hepática” en el campo de procedimientos. La información relevante se anotó en una hoja de recogida de datos prediseñada.

Desde allí se volcó la información a una base de datos realizada *ex profeso* con el programa Microsoft Access 2010<sup>®</sup>. El acceso estuvo restringido por contraseña al autor de la tesis. Una serie retrospectiva de diez casos de HAI se utilizó de prueba para comprobar el correcto funcionamiento de la base de datos y corregir posibles errores. En los anexos se puede encontrar una captura de pantalla de la interfaz de usuario de la base de datos.

Finalmente se importó la matriz de datos para su tratamiento estadístico a SPSS Statistics<sup>®</sup> versión 21.0. En este entorno se definieron y corrigieron las variables iniciales y se computaron las nuevas.

## 8.2. Aspectos éticos

---

El diseño del estudio no supone ningún tipo de intervención y contempla trabajar con información clínica generada durante la práctica asistencial normal. Para la explotación de esta información, se pediría el consentimiento por escrito a los pacientes maduros y/o a sus tutores legales salvo a aquellos con los que existan dificultades para establecer contacto por corresponder a la fase retrospectiva del estudio o se haya perdido la comunicación. En los anexos se muestra la hoja de información al paciente y el documento de consentimiento informado.

Se les ha garantizado el anonimato y la confidencialidad de su información personal de acuerdo con la Ley Orgánica de Protección de Datos de Carácter Personal (Ley orgánica 15/1999, de 13 de diciembre, BOE 1999; 298, -14-XII:43088-99), tanto en el almacenamiento de resultados como en su exposición y divulgación.

Durante la actividad asistencial de la que se ha generado la información explotada en esta tesis se ha procurado hacer cumplimiento de las normas de buena práctica clínica (Orden SCO 256/2007, BOE 13-II-2007) y se han respetado los principios éticos de la investigación en salud y la legislación al respecto: Ley de Investigación Biomédica 14/2007 (BOE 4-VII-2007).

El protocolo del estudio, que incluye todos los detalles de la metodología referidos en la sección anterior, la organización del trabajo y su cronograma, han sido aprobados por el Comité Ético de Investigación Clínica de la Fundación Sant Joan de Déu de Barcelona en fecha de 12 de diciembre de 2016 (PIC-99-16).

# 9

## Metodología y resultados



## **9.1. Capítulo 1: Estimación de la prevalencia de la HAI y evaluación de la validez de los criterios simplificados de 2008**

---

### **9.1.1. Metodología específica**

Las características de la población de estudio vienen determinadas por el diseño general descrito en el bloque anterior. El objetivo ha sido confeccionar un grupo de casos y no casos representativos del global de niños con disfunción hepática de origen incierto, que incluya una proporción de enfermos de HAI equivalente a la de la población diana. Esta información ha permitido describir todos los indicadores de validez (tanto los independientes como aquellos cuya interpretación depende de la prevalencia de la enfermedad de interés) y completar el análisis de decisiones clínicas del capítulo 6.

El hecho de que se haya incluido una fase prospectiva desde enero de 2016 hasta enero de 2017 ha permitido corregir algunos valores perdidos en los datos recogidos en la fase previa retrospectiva. Por protocolo, aquellos pacientes en los que no fue posible recoger la información suficiente para asegurar la aplicación del estándar de referencia diagnóstico, se excluyeron del análisis. La implementación del patrón de referencia se llevó a cabo con anterioridad a la aplicación de la prueba a validar de modo que los resultados de los criterios simplificados de 2008 no influenciaron en la clasificación diagnóstica de los pacientes.

Por lo que respecta al manejo de los valores indeterminados o desconocidos de los criterios simplificados, la estrategia adoptada consistió en emplear una tabla de 3x2 para efectuar la evaluación de la validez con una aproximación de intención de diagnosticar [231]. Los investigadores que llevaron a cabo el cálculo de la puntuación según el sistema simplificado no estuvieron cegados respecto al diagnóstico según el sistema de referencia.

Dado que el número mínimo de pacientes a recoger se ha elegido en base al objetivo específico 5, se dispone de un número de HAI y diagnósticos alternativos

superior al necesario para obtener la sensibilidad y especificidad con una precisión absoluta máxima de  $\pm 10\%$ .

Una vez conocidos los valores de sensibilidad y especificidad, así como la prevalencia de HAI, se calculó el área de indicación para comprobar la adecuación de la prueba (la aplicación de los criterios de 2008) al contexto epidemiológico que se quiere representar con nuestra muestra. Cuándo hacer y cuándo no hacer una prueba diagnóstica es un dilema recurrente en la práctica clínica. Se reconoce que la capacidad de un *test* de reducir la incertidumbre es una función dependiente de la prevalencia de la enfermedad de interés, que tiende a comportarse inadecuadamente en escenarios extremos de probabilidades preprueba. De hecho, en contextos de prevalencias muy bajas o muy altas, suele existir pérdida de información diagnóstica. En palabras de Knottnerus: “Existe un *área de indicación* para una prueba diagnóstica entre los márgenes de ciertas probabilidades preprueba, más allá de los cuales la ganancia de certeza sobre la presencia de la enfermedad es negativa. La evaluación de pruebas diagnóstica debe incluir un análisis sobre si se pueden ser útiles en las categorías habituales de prevalencia del proceso de interés” [240]. El cálculo de los valores predictivos como solución a este problema es insuficiente dado que se obtienen al final de la evaluación de la prueba y para la prevalencia concreta de la muestra empleada, que no tiene porqué reflejar la probabilidad preprueba real de la población diana, especialmente en estudios con diseños de cohortes o de casos y controles. En efecto, se entiende que en la práctica real es habitualmente más relevante conocer de antemano el valor añadido de realizar el *test* para guiar la actuación ulterior [241].

Para enfermedades categorizables de forma binaria (presencia o ausencia) por un *test* diagnóstico, solo tres parámetros son suficientes para determinar su desempeño: sensibilidad, especificidad y prevalencia. Basado en el análisis de decisiones clínicas, Stalpers *et al.* propusieron un modelo matemático que permitiera calcular el ancho de prevalencias en las que el resultado de una prueba diagnóstica aporta certeza diagnóstica [242].

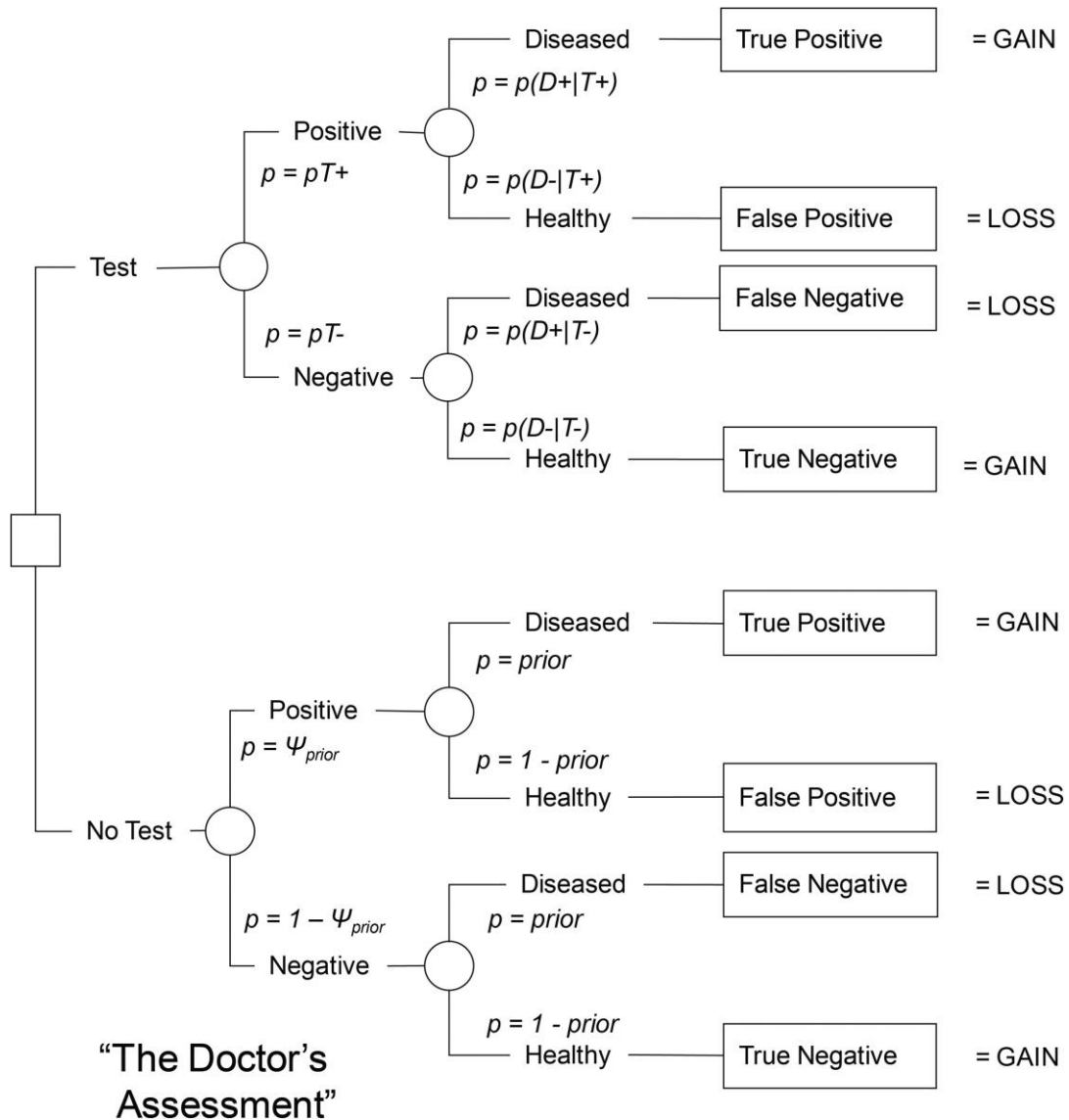


Figura 15: Árbol de decisiones para la elección de aplicar una prueba diagnóstica, en función de la sensibilidad y especificidad propias de la prueba y en base a las probabilidades preprueba reales ( $p$ ) y heurísticas ( $\psi$ ). En el ejemplo, “positivo” implica clasificado como enfermo y “negativo”, clasificado como sano. Reproducido de Stalpers et al. *Journal of Clinical Epidemiology*. 2015;68:1122. Con permiso de Elsevier.

La figura anterior presenta el árbol de decisiones para la evaluación de una prueba diagnóstica estructurado en un conflicto de elección entre sus dos alternativas de aplicación (sí o no). La rama superior muestra las consecuencias si se lleva a cabo el *test* y representa la “certeza diagnóstica” que es la probabilidad de clasificar correctamente (valor predictivo global,  $VP_G$ ) menos la probabilidad de clasificar incorrectamente (inverso aditivo del  $VP_G$ ), es decir:  $2VP_G - 1$ , que es una



función lineal de la prevalencia. La rama inferior del árbol representa las consecuencias de no hacer la prueba, que dependen del juicio médico (*the doctor's assessment*) de si ese paciente potencial podría ser, o no, enfermo. Es pues una decisión cognitiva de naturaleza heurística, imposible de representar matemáticamente dado que sería como asignar un valor a la intuición de un médico. Aun así, es razonable aproximarla asumiendo que la probabilidad de un juicio positivo aumenta con la prevalencia ( $P_{Enfermo} = \Psi_{prior}$ ) y la probabilidad de un juicio negativo aumenta con el opuesto de la prevalencia ( $P_{No\ enfermo} = 1 - \Psi_{prior}$ ). Se asume también que la sensibilidad y la especificidad de la intuición de un médico, sin recibir ningún tipo de información sobre la población de donde proviene el paciente, son ambas del 50%. Como consecuencia, los resultados de no hacer la prueba son iguales para cualquier análisis y solo la rama superior del árbol cambia para cada *test*. Representado en un gráfico la “certeza diagnóstica” de hacer y de no hacer la prueba, resulta una función lineal para la primera (que cambia en función del  $VP_G$  propio de la prueba para cada prevalencia), y una curva cóncava para la segunda.

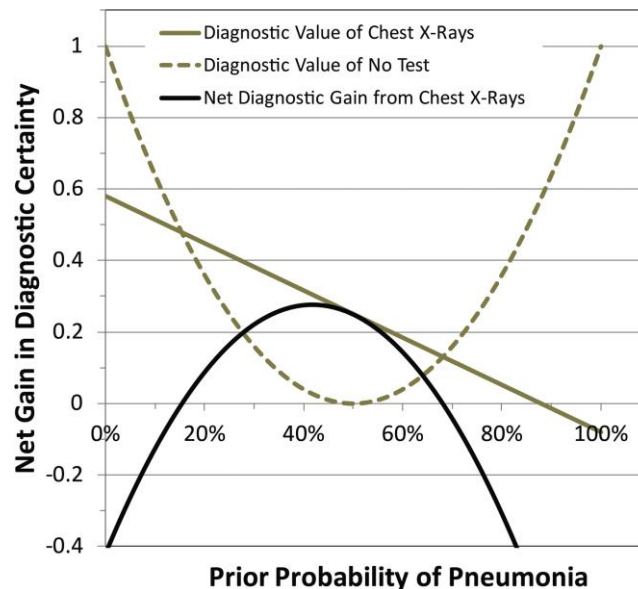


Figura 16: Ejemplo de representación de la ganancia neta en certeza diagnóstica con y sin hacer radiografía de tórax para el diagnóstico de neumonía. Reproducido de Stalpers et al. *Journal of Clinical Epidemiology*. 2015;68;1124. Con permiso de Elsevier.

La ganancia neta en certeza diagnóstica sería pues la diferencia entre las certezas diagnósticas de hacer y no hacer la prueba, que quedaría representada como una curva convexa con una leve asimetría. El ancho entre los puntos de intersección de esta última curva con la línea del 0 es el área de indicación.

### 9.1.2. Análisis estadístico

Los indicadores de validez se estimaron según los cálculos explicados en el apartado 4.2.2. Una vez hallada la prevalencia, la sensibilidad y la especificidad, con el tamaño muestral final alcanzado y en base a las fórmulas de Buderer, se calculó la precisión absoluta resultante para la estimación de dichos parámetros de los criterios simplificados de 2008.

Dado que la aplicación de los criterios de la IAHG arroja una asignación diagnóstica en dos categorías nominales (HAI sí o no) o tres categorías ordinales (HAI no, probable o definitiva), el grado de acuerdo entre los criterios clásicos y simplificados se resumió en base al estadístico kappa. Para el caso de la clasificación no binaria, se calculó aplicando tanto una ponderación lineal como una ponderación cuadrática.

Los resultados binarios se expresaron como porcentajes con su intervalo de confianza al 95% por el método de Wilson. En el caso de muestras pequeñas, el extenso estudio clásico de simulación de Newcombe ha demostrado que la aproximación para su cálculo basada en la distribución normal o en el procedimiento exacto propuesto por Clopper y Pearson, dan un intervalo excesivamente amplio [243]. Entre los diferentes estimadores que se han propuesto como alternativa, Altman *et al.* proponen el método de Wilson como el de elección porque es aplicable a muestras de cualquier tamaño, ofrece un intervalo de confianza con buena cobertura y los límites obtenidos siempre están comprendidos entre 0 y 1 [244]. Por su parte, las variables continuas se resumieron como medianas y su RIC. Para comprobar la hipótesis nula de que, al respecto de estas variables continuas, la

muestra ha sido extraída de una población con distribución de probabilidad normal, se aplicó la prueba de Kolmogorov-Smirnov. El estadístico de Kolmogorov-Smirnov es la máxima diferencia ( $D$ ):

$$D = \text{máx}|F_n(x) - F_0(x)|$$

Siendo  $F_n(x)$  la función de distribución muestral y  $F_0(x)$  la función teórica o correspondiente a la población normal especificada en la hipótesis nula.

La distribución del estadístico de Kolmogorov-Smirnov es independiente de la distribución poblacional especificada en la hipótesis nula y los valores críticos de este estadístico están tabulados. Si la distribución postulada fue normal y se estimaron sus parámetros, los valores críticos se obtuvieron aplicando la corrección de significación propuesta por Lilliefors [245].

El punto de corte óptimo de la puntuación de los criterios simplificados se estimó siguiendo la estrategia de Zweig y Campbell, que maximiza la siguiente función [246]:

$$Se - \left( RC^{-1} \times \frac{1 - P}{P} \right) \times (1 - Sp)$$

Donde  $Se$  es la sensibilidad,  $Sp$  es la especificidad,  $P$  es la prevalencia de HAI en la muestra entera y  $RC$  es la razón de costes, es decir, el cociente entre el coste de los falsos negativos y el de los falsos positivos. Por convención se utilizó una  $RC$  de 1 con la idea de penalizar por igual los errores de tipo I que los de II. Los primeros serían pacientes innecesariamente expuestos a los riesgos del tratamiento y los segundos no se beneficiarían de forma precoz del inicio de la medicación.

Las diferencias entre variables continuas se evaluaron mediante el estadístico  $U$  de Mann-Whitney y el *test* de  $\chi^2$  se empleó para el caso de variables dicotómicas. En el caso de que no se cumplieran las condiciones de tamaño muestral para poder aplicar el *test* de  $\chi^2$ , se empleó el *test* exacto de Fisher.

Se consideró estadísticamente significativo un valor  $p$  inferior a 0,05.

La mayoría de los análisis se realizaron con el paquete estadístico SPSS® de IBM, en su versión 21.0. Algunos procedimientos con SPSS se ejecutaron siguiendo la

sintaxis de las macros !DT, !CIP y !ROC, en concreto los relacionados con el cálculo del intervalo de confianza de proporciones, la obtención de los principales indicadores de validez, su comportamiento en función de la prevalencia de los casos, el análisis de la curva COR y los puntos de corte para la prueba diagnóstica a estudio [247–249]. Esta documentación ha sido obtenida del *Laboratori d'Estadística Aplicada* de la *Universitat Autònoma de Barcelona*. Para la confección de la curva de Lorenz y el cálculo de los índices de Pietra y Gini se empleó Stata® 14.

### **9.1.3. Resultados**

#### **9.1.3.1. Flujo de pacientes y descripción de las categorías diagnósticas**

Un total de 425 pacientes se identificaron como potencialmente elegibles entre los dos centros. De ellos, 35 (un 8,2%) lo fueron en el periodo comprendido entre enero de 2016 y enero de 2017, es decir, durante la fase prospectiva del trabajo. Después de aplicar los criterios de exclusión, fueron seleccionados 218 pacientes, de los cuales solo 6 se tuvieron que retirar del estudio porque no fue posible recuperar información suficiente para asegurar la correcta aplicación de los criterios clásicos de 1999. En la figura de la página siguiente se esquematiza el flujo de pacientes según la recomendación de la iniciativa STARD.

De los 212 que finalmente constituyen la muestra total para esta tesis, 5 se reasignaron a diagnóstico definitivo de HAI después de efectuar el análisis de casos discrepantes. Ninguno de los pacientes con puntuación por debajo del dintel recomendado tanto en los criterios simplificados como en los clásicos, se diagnosticó en la realidad de HAI. En todos los casos finalmente diagnosticados de HAI, los criterios clásicos revisados de 1999 se pudieron aplicar en un escenario post-tratamiento.

A continuación, se describen las características clínicas y demográficas de los pacientes en función del diagnóstico final.

Se han recuperado un total de 100 pacientes con HAI, de los cuales 17 fueron de tipo 2 (por seropositividad para anti-LKM1), lo que representa que la relación de casos entre tipo 1 y 2 en nuestra muestra fue de 4,9:1.

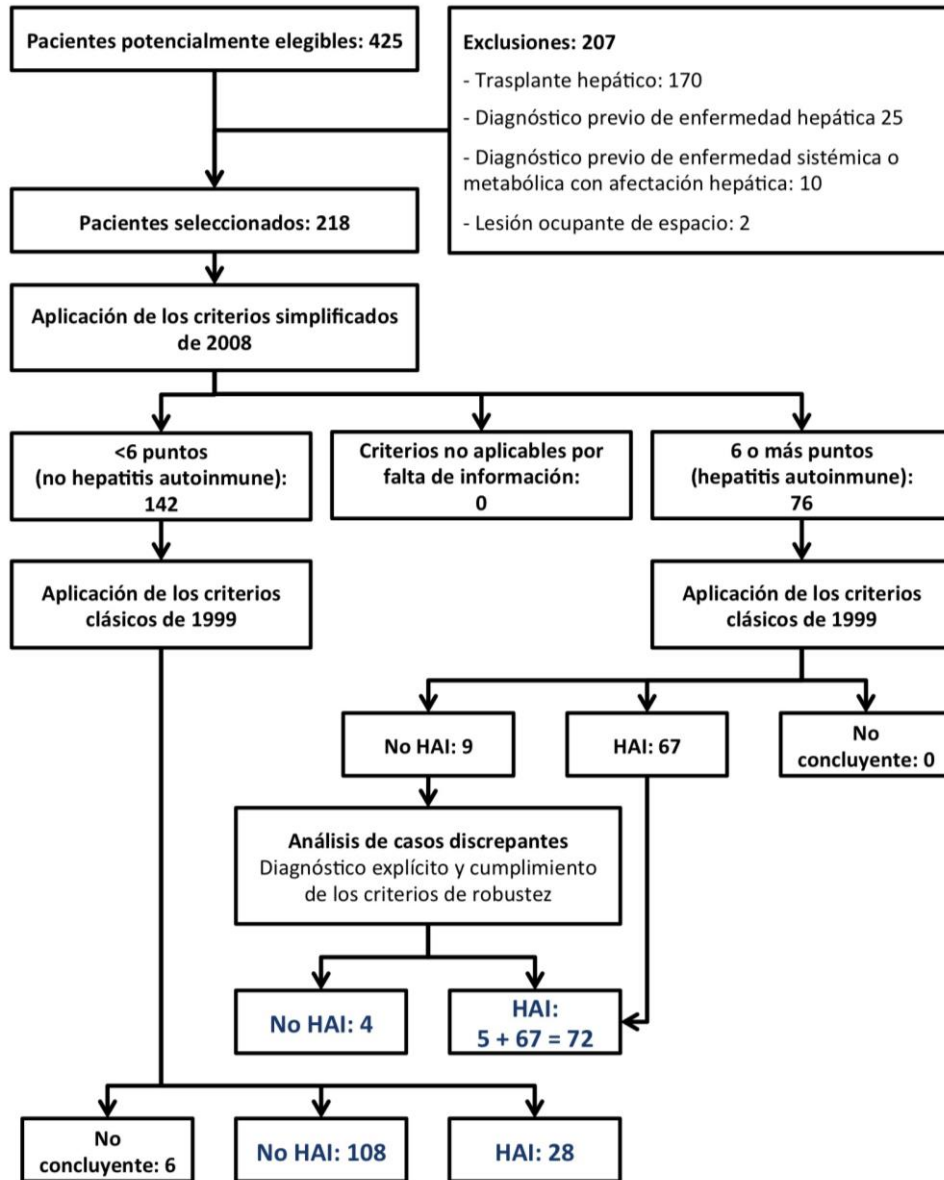


Figura 17: Diagrama de flujo STARD con los resultados de la prueba índice y la asignación diagnóstica final. HAI: hepatitis autoinmune.

Así pues, la prevalencia de HAI *por intención de diagnosticar*, es decir, sobre

la base de los pacientes que quedaron después de aplicar los criterios de exclusión, fue de 45,9% (IC95% 39,4% a 52,5%). Sin contar con los 6 pacientes en los que no se pudo asegurar la correcta exclusión de HAI por falta de información para aplicar el estándar basado en los criterios clásicos, la prevalencia ascendió a 47,2% (IC95% 40,6% a 53,9%). Este será el dato que se contemplará para análisis posteriores.

Cuatro de los casos de HAI tuvieron como forma de presentación una insuficiencia hepática aguda según la definición dada en los criterios de inclusión. Los cuatro tuvieron signos leves de encefalopatía que mejoró con el tratamiento y requirieron de administración de plasma fresco congelado para mejorar la hipocoagulabilidad secundaria. La obtención de la muestra hepática para estudio histológico se llevó a cabo mediante biopsia transyugular.

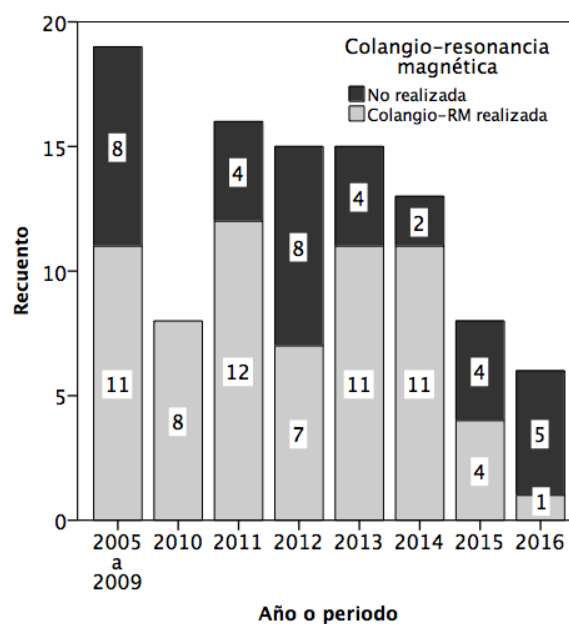


Figura 18: Número de pacientes con hepatitis autoinmune en los que se realizó colangio-resonancia magnética en función del año de diagnóstico.

Ninguno de los niños se diagnosticó de síndrome de solapamiento con CEP. Sin embargo, en no todos los pacientes el estudio incluyó una prueba de imagen biliar. Del total de casos y no casos, a un 45,3% se les realizó una exploración específica de vía biliar. En concreto, consistió en dos CPRE en sendos pacientes con

diagnóstico final de hepatopatía crónica criptogénica con fibrosis, y colangio-RM en el resto. Considerando solo los pacientes con diagnóstico final de HAI, la proporción de enfermos con exploraciones de vía biliar aumenta a 65,0%.

Dentro del grupo de los pacientes con HAI, 23 casos (23%) presentaron además otras condiciones de base autoinmune o autoinflamatoria. En todos ellos el diagnóstico de la HAI fue posterior al de las otras enfermedades. La asociación más frecuente fue entre enfermedad celiaca y HAI (34,8% de los niños con múltiples diagnósticos de naturaleza autoinmune y 8% del total de casos de HAI).

**Tabla 13: Frecuencia de las asociaciones entre hepatitis autoinmune y otras enfermedades autoinmunes.**

Enfermedad asociada	Frecuencia (porcentaje respecto al total de enfermos con asociación)
Celiaquía	8 (34,8%)
Colitis ulcerosa	7 (30,4%)
Diabetes mellitus tipo I	4 (17,4%)
Artritis idiopática juvenil	4 (17,4%)
Inmunodeficiencia primaria XLP-like*	1 (4,3%)

\*La suma total es de uno más del número real de pacientes con asociación porque el paciente con esta entidad también tenía diagnóstico de colitis ulcerosa.

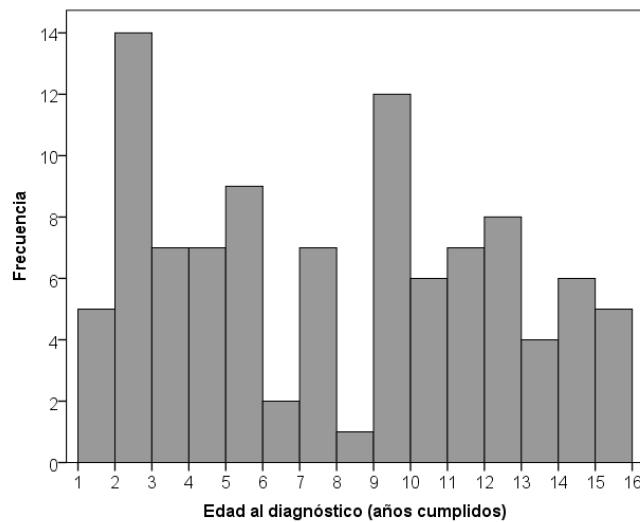
Por lo que respecta al grupo de pacientes con diagnóstico alternativo a hepatitis autoinmune, está constituido por 112 niños con una variedad de enfermedades hepáticas. Tres diagnósticos engloban al 60,7% de los pacientes sin hepatopatía autoinmune: hepatitis criptogénica aguda (sin o con colestasis), enfermedad de Wilson y hepatitis vírica aguda o crónica. De este último grupo, 4 casos fueron infecciones congénitas por CMV, 4 casos fueron hepatitis crónicas (2 por virus B y 2 por virus C) y 7 casos fueron hepatitis víricas agudas (4 por VEB, 2 por CMV y 1 por virus herpes humano 6).

La proporción de mujeres en el grupo de HAI fue significativamente mayor que en el grupo de no HAI: 72,0% (IC95% 62,5% a 79,9%) frente a 42,0% (IC95% 33,2% a 51,2%), con un valor  $p < 0,001$ .

**Tabla 14: Enfermedades y frecuencias que integran el grupo de no casos.**

Hepatitis criptogénica aguda sin colestasis	27
Enfermedad de Wilson	22
Hepatitis vírica	15
Síndrome de Alagille	7
Hepatitis tóxica	6
Colangitis esclerosante primaria	5
Hepatopatía congestiva o hígado de estasis	5
Hepatitis criptogénica aguda con colestasis	4
Enfermedad hepática por depósito	4
Fibrosis hepática congénita	4
Esteatohepatitis no alcohólica	3
Colestasis intrahepática familiar progresiva	3
Enfermedad mitocondrial	3
Enfermedad hepática crónica criptogénica	2
Hepatitis de células gigantes	1
Intolerancia hereditaria a la fructosa	1

La edad mediana al diagnóstico de HAI fue de 7,9 años, con un RIC entre 3,9 y 11,6 años. El paciente más joven fue diagnosticado de esta entidad a los 16 meses y el más mayor lo fue a los 16 años. La distribución de edades en el grupo de HAI no siguió un patrón normal, demostrado por un valor p de 0,006 en la prueba de significación de Kolmogorov-Smirnov.



**Figura 19: Histograma con la distribución de frecuencias por edad al diagnóstico en los pacientes con hepatitis autoinmune.**



La edad en el momento del diagnóstico en el grupo de no HAI (o en el de la primera biopsia hepática si el diagnóstico no está claro al final de seguimiento) tuvo un valor mediano de 8,1 años (RIC 3,8 a 12,5 años). La distribución de edades entre los pacientes con y sin HAI no mostró diferencias estadísticamente significativas (valor  $p = 0,665$ ).

Sí que se constató diferencia en la proporción de pacientes con antecedentes personales o familiares de otras enfermedades con sustrato autoinmune. Mientras que en el 29,0% (IC95% 21,0% a 38,5%) de los niños con HAI se documentaron antecedentes, solo el 9,8% (IC95% 5,6% a 16,7%) de los no casos los presentaron (valor  $p < 0,001$ ).

Los resultados comparados de la bioquímica básica al diagnóstico entre el grupo de pacientes con y sin HAI se resumen en la tabla y figura siguientes:

**Tabla 15: Marcadores bioquímicos relevantes para el diagnóstico acorde a la clasificación final. Para los datos cuantitativos se emplea la mediana y el rango intercuartílico como marcador de tendencia central.**

	HAI (n=100)	No HAI (n=112)	Valor $p$ de la diferencia
Proporción de pacientes con hiper- $\gamma$ -globulinemia	67,0% (IC95% 57,3 a 75,4%)	18,8% (IC95% 12,6 a 27,0%)	<0,001
Niveles de IgG (mg/dL)	1598 (RIC 1130,5 a 2373,5)	960 (RIC 776 a 1143)	<0,001
AST (U/L)	788,5 (RIC 141,5 a 1730,5)	96 (RIC 53,5 a 209,5)	<0,001
ALT (U/L)	678 (RIC 174 a 1833)	105 (RIC 52 a 387,5)	<0,001
Fosfatasa alcalina (U/L)	288 (RIC 215 a 407,5)	290 (RIC 201 a 380,5)	0,935
GGT (U/L)	76 (RIC 36 a 145,5)	42 (RIC 22,5 a 80)	0,001

HAI: Hepatitis autoinmune. IC95%: Intervalo de confianza al 95%. IgG: Inmunoglobulina G. RIC: Rango intercuartílico. AST: Aspartato aminotransferasa. ALT: Alanina aminotransferasa. GGT: Gamma-glutamyl transpeptidasa.

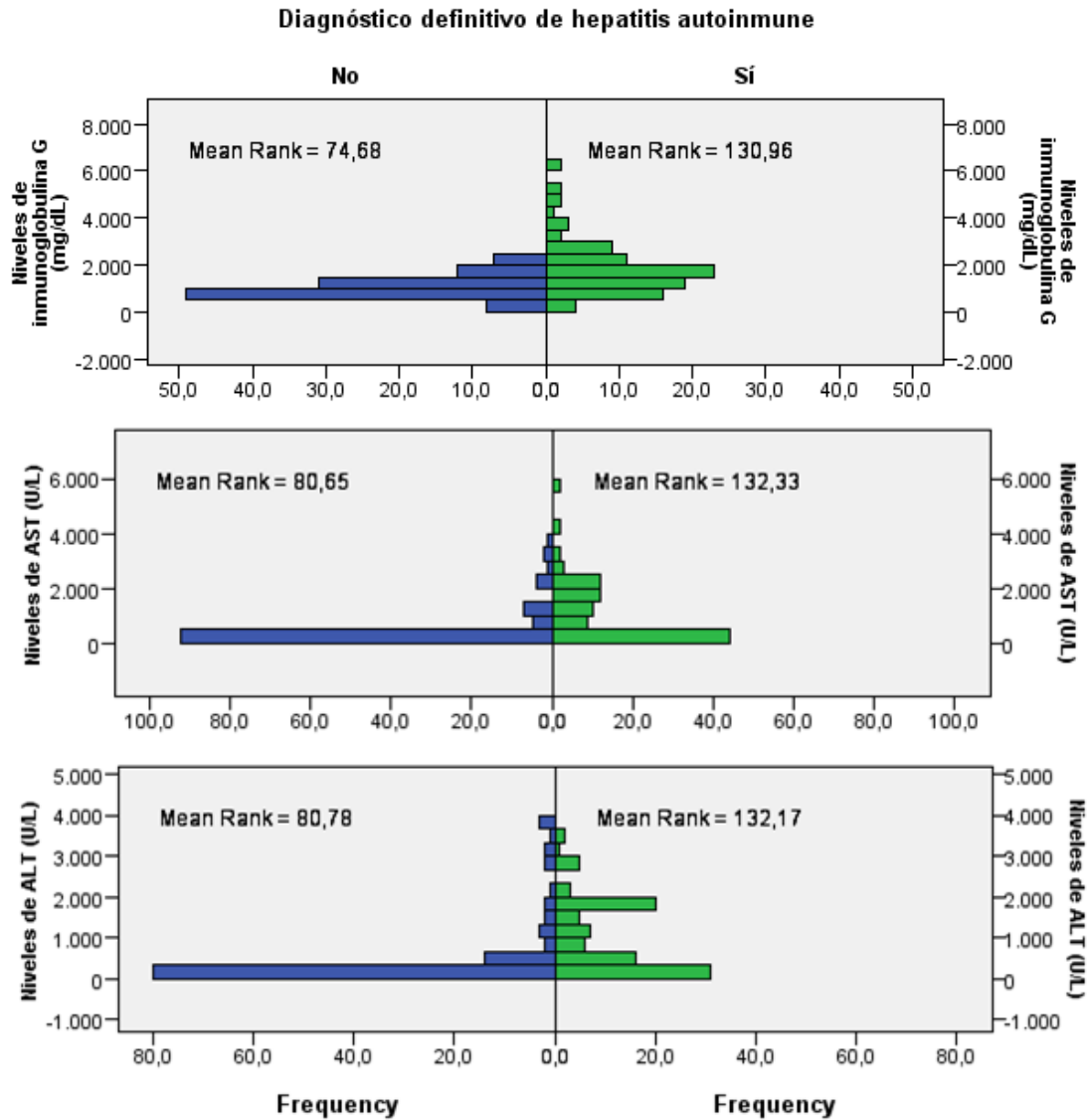


Figura 20: Distribuciones de los parámetros bioquímicos con patrón significativamente diferente entre los dos grupos diagnósticos.

La presencia de autoanticuerpos es el otro rasgo analítico definitorio de la HAI. Ninguno de los niños de la población estudiada presentó unos títulos por debajo de 1:40 de ANA, anti-Sm o anti-LKM1. Por lo tanto, ningún caso se consideró seronegativo por los criterios diagnósticos de 1999 y 2008 a pesar de presentar niveles detectables. Los autoanticuerpos de tipo anti-SLA, anti-célula parietal y PANCA, se hallaron en una menor proporción de pacientes, todos ellos finalmente diagnosticados de HAI.

Tabla 16: Proporción de pacientes con títulos de autoanticuerpos superiores a los dinteles establecidos por los criterios simplificados de 2008 en los grupos de casos y no casos de hepatitis autoinmune (HAI).

	HAI (n=100)	No HAI (n=112)	Valor p de la diferencia
Proporción de pacientes con ANA $\geq$ 1:40	66,0% (IC95% 56,3 a 74,5)	31,3% (IC95% 23,4 a 40,3)	<0,001
Proporción de pacientes con ANA $\geq$ 1:80	60,0% (IC95% 50,2 a 69,1%)	25,9% (IC95% 18,7% a 34,7%)	<0,001
Proporción de pacientes con anti-Sm $\geq$ 1:40	63,0% (IC95% 53,2% a 71,8%)	52,7% (IC95% 43,5% a 61,7%)	0,129
Proporción de pacientes con anti-Sm $\geq$ 1:80	58,0% (IC95% 48,2% a 67,2%)	24,1% (IC95% 17,1% a 32,8%)	<0,001
Proporción de pacientes con anti-LKM1 $\geq$ 1:40	15,0% (IC95% 9,3% a 23,3%)	1,8% (IC95% 0,5% a 6,3%)	<0,001

Tabla 17: Distribución descrita a través de la mediana y el rango intercuartílico (RIC) de los títulos de autoanticuerpos entre los casos de hepatitis autoinmune (HAI) y el grupo de no casos.

	HAI (n=100)	No HAI (n=112)	Valor p de la diferencia
ANA (título 1:X)	240 (80 a 640)	80 (RIC 80 a 320)	0,001
Anti-Sm (título 1:X)	160 (RIC 140 a 640)	40 (RIC 40 a 80)	<0,001
Anti-LKM1 (título 1:X)	160 (RIC 160 a 240)	80 *	0,059
Anti-SLA (título 1:X)	320 (RIC 320 a 640)	-	-

\*RIC no informativo al haber solo dos casos con el mismo valor.

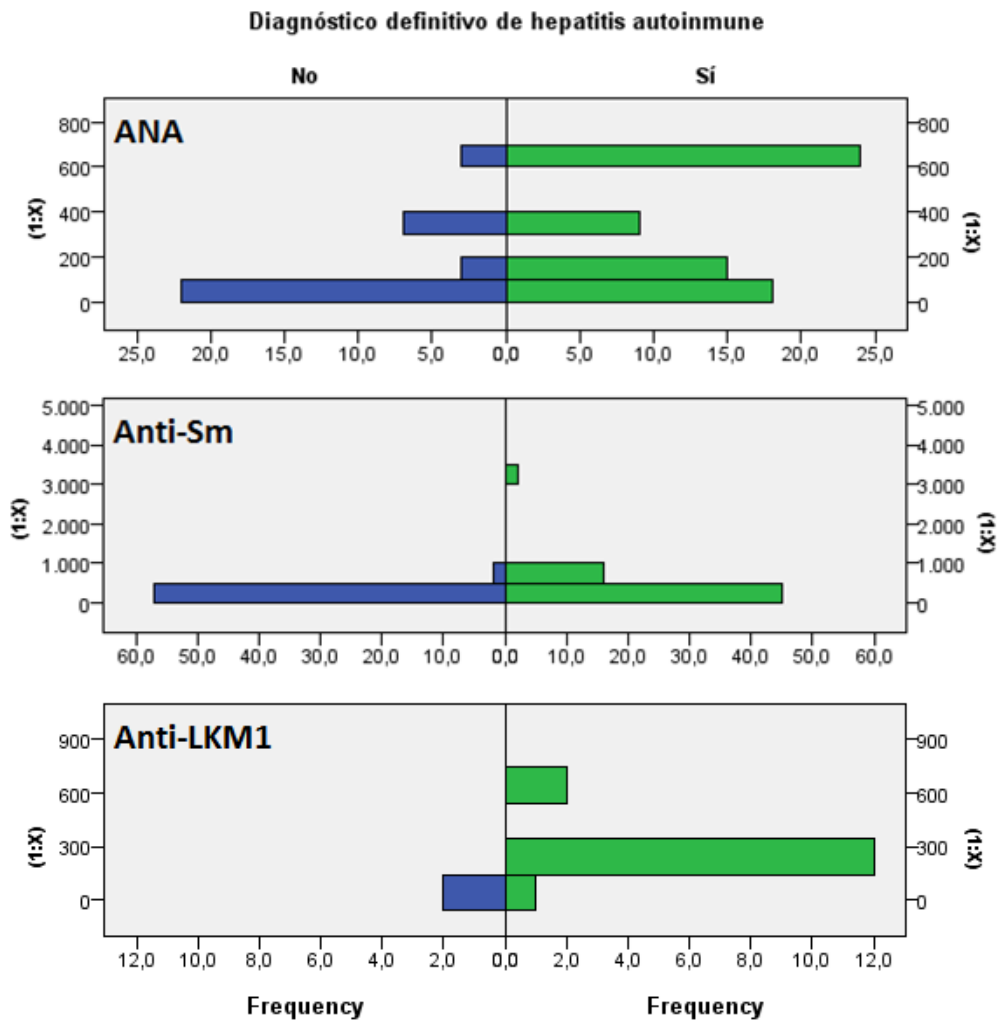


Figura 21: Distribuciones de los títulos de autoanticuerpos entre los dos grupos diagnósticos.

No se encontraron diferencias estadísticamente significativas en la proporción de pacientes con títulos de anticuerpos anti-Sm superiores a 1:40 entre los dos grupos diagnósticos, pero sí en la proporción de pacientes con niveles de más de 1:80. De hecho, dentro del grupo de no HAI se detectaron autoanticuerpos en el 68,7% de los pacientes. De entre ellos, en un 96,1% de los niños, se incluía anti-Sm, de forma aislada en 58 de 77 casos, y en combinación con ANA en el resto. De entre los 100 pacientes con HAI, solo en dos no se detectaron autoanticuerpos. Siguiendo con los niños con HAI, los ANA se presentaron de forma aislada en 19 de los 66 casos que los presentaron; en 36 pacientes en combinación con anti-Sm y en

11 casos en combinación con anti-LKM1. La mayoría de los pacientes con HAI de tipo 2 presentaron los anticuerpos específicos anti-LKM1 en combinación con ANA (68,8%). En todos los casos HAI anti-SLA positivos se detectaron anti-Sm a títulos >1:80.

Al respecto de la distribución de puntos en los sistemas de clasificación clásico y simplificado, se muestra la información agregada por grupos diagnósticos:

Tabla 18: Distribución descrita a través de la mediana y el rango intercuartílico (RIC) de los puntos obtenidos en los criterios diagnósticos para la hepatitis autoinmune (HAI).

	HAI (n=100)	No HAI (n=112)	Valor <i>p</i> de la diferencia
Criterios clásicos revisados de 1999	14 (RIC 12 a 17)	2 (RIC 0 a 4)	<0,001
Criterios simplificados de 2008	6 (RIC 5 a 7)	3 (RIC 2 a 4)	<0,001

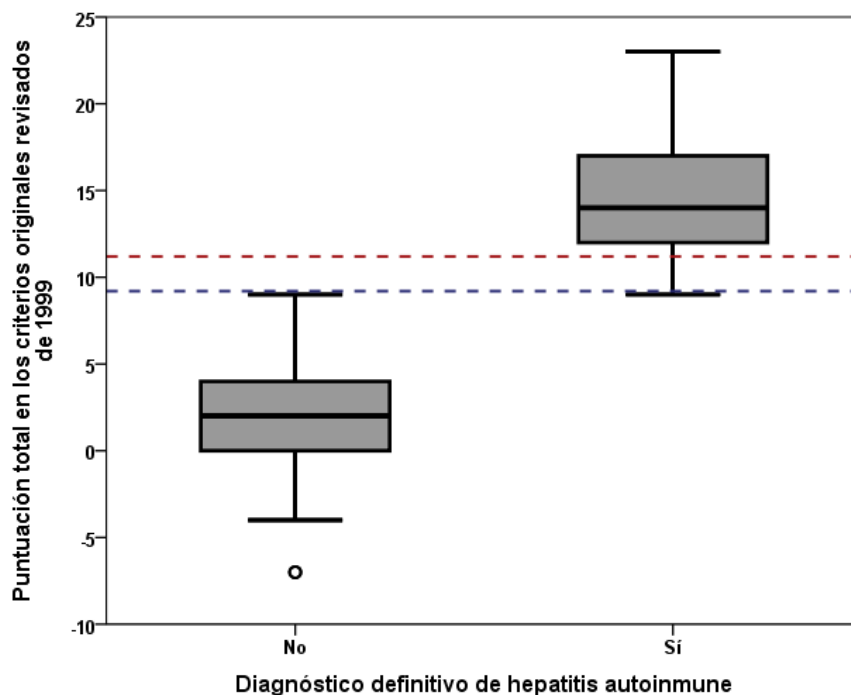


Figura 22: Diagrama de caja de la puntuación obtenida en los criterios clásicos revisados de 1999 por los pacientes con o sin hepatitis autoinmune. Las patillas incluyen el 90% central de la muestra. La línea discontinua azul (en 10 puntos) representa el punto de corte para HAI probable pre-tratamiento. La línea roja (en 12 puntos), para HAI probable post-tratamiento.

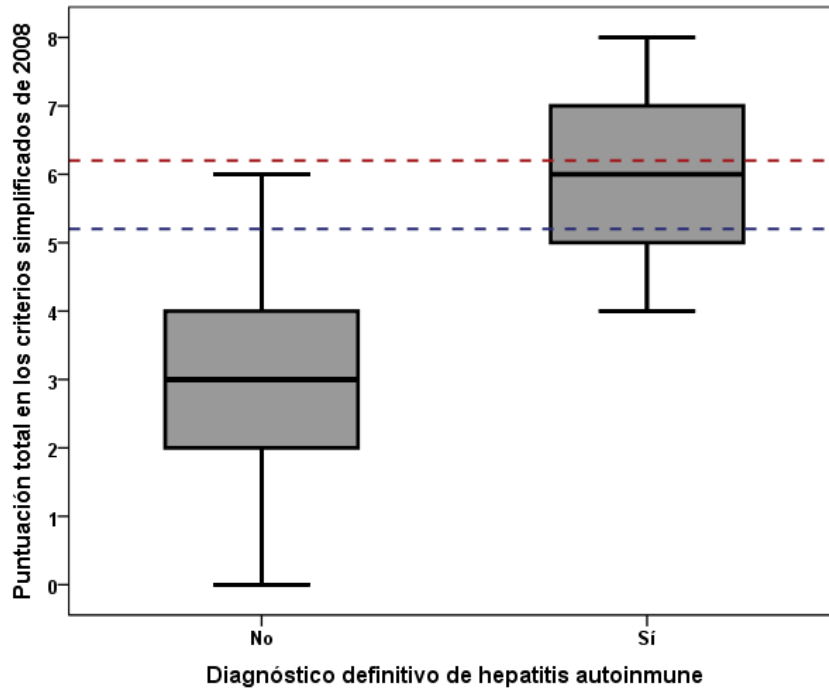


Figura 23: Diagrama de caja de la puntuación obtenida en los criterios simplificados de 2008 por los pacientes con o sin hepatitis autoinmune. Las patillas incluyen el 90% central de la muestra. La línea discontinua azul (en 6 puntos) representa el punto de corte para HAI probable y la roja (en 7 puntos), para HAI definitiva.

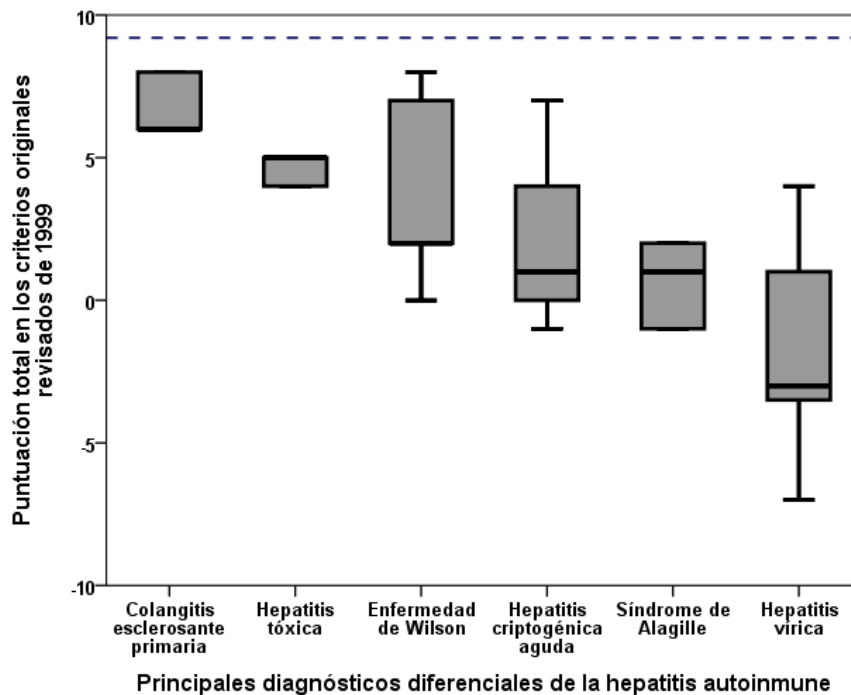


Figura 24: Diagrama de caja de las puntuaciones obtenidas en los criterios clásicos revisados de 1999 por los pacientes con los principales diagnósticos diferenciales. La línea discontinua azul (en 10 puntos) representa el punto de corte para HAI probable pre-tratamiento.

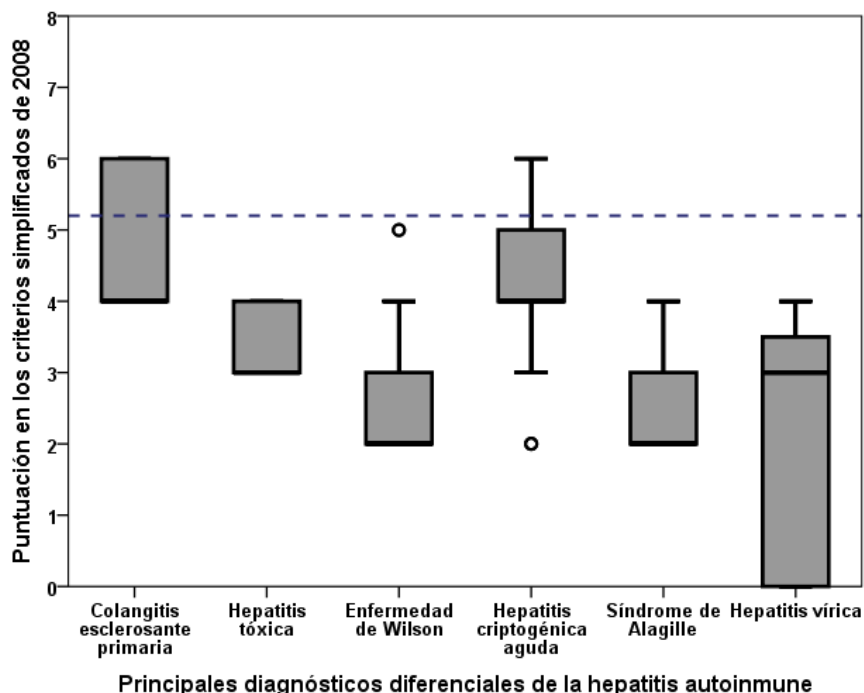


Figura 25: Diagrama de caja de las puntuaciones obtenidas en los criterios simplificados de 2008 por los pacientes con los principales diagnósticos diferenciales. La línea discontinua azul (en 6 puntos) representa el punto de corte para HAI probable.

Dentro del grupo de HAI, los criterios de 1999 categorizaron al 73% de los casos como HAI probable y al 22%, como definitiva. Los criterios simplificados, por su parte, definieron al 27% como HAI probable, al 45% como definitiva y el 28% restante no fue clasificado como HAI. Además, en el grupo de no HAI, los criterios de 2008 asignaron correctamente el diagnóstico de exclusión a 108 de los 112 pacientes (96,4%) pero 4 niños (3,6%) fueron mal clasificados como HAI probable.

Dos de los cuatro casos de HAI que presentaron fallo hepático agudo se diagnosticaron de HAI definitiva por ambos criterios de clasificación. Los otros dos se dieron también como HAI definitiva por los criterios clásicos revisados de 1999, pero se clasificaron como HAI probable por los criterios clásicos.

### 9.1.3.2. Indicadores de validez interna

#### 9.1.3.2.1. Indicadores independientes de la prevalencia de la enfermedad

El cálculo de la sensibilidad y la especificidad por intención de diagnosticar

arrojó los mismos resultados que el análisis por protocolo dado que fue posible obtener información suficiente de todos los pacientes incluidos como para aplicar los criterios simplificados.

Utilizando un punto de corte de 6 (HAI probable) los criterios simplificados de 2008 mostraron una sensibilidad del 72,0% (IC95% 62,5% a 79,9%) y una especificidad del 96,4% (IC95% 91,2% a 98,6%). La precisión absoluta para la obtención de estos indicadores de validez, obtenida por las fórmulas de Buderer fue de  $\pm 8,8\%$  para la sensibilidad y de  $\pm 3,4$  para la especificidad.

La tasa de falsos positivos fue de 3,6% (IC95% 1,4% a 8,8%) y la de falsos negativos de 28,0% (IC95% 20,1% a 37,5%).

Las RV positiva y negativa fueron de 20,2 y 0,3 respectivamente. Traducido a WoE positivo y negativo, los equivalentes logarítmicos de estos valores fueron de +13,0 deciban y -5,4 deciban.

El índice de Youden de 0,68 para el punto de corte en 6 indicó que los criterios simplificados como prueba de clasificación son, de modo global, suficientemente informativos.

La efectividad de la prueba ( $\delta$ ) arrojó un resultado de 4,1, que refleja una magnitud elevada para el diagnóstico si el caso se clasifica como HAI probable por los criterios de 2008 (recuérdese que es la diferencia entre las medias de los resultados entre una población de enfermos y otra de sanos en una escala normalizada).

Como la incorporación de los pacientes al estudio no ha sido a partir de los resultados de la prueba a validar (diseño de cohortes), tenemos una serie no sesgada de casos de HAI y una de no casos que pueden comportarse como controles. Así, hemos podido calcular la *odds ratio* diagnóstica que, para el caso del punto de corte en 6, ha resultado de 67,3. Indica una capacidad discriminante significativa.

Para un punto de corte de 7 puntos (HAI definitiva), se obtuvo una sensibilidad de 45,0% (IC95% 35,6% a 54,8%) y una especificidad del 100% (IC95%



96,7% a 100%), con una precisión absoluta de  $\pm 9,8\%$  y  $\pm 2,4\%$ , respectivamente.

No hubo falsos positivos (IC95% de 0 a 3,3%) y la tasa de falsos negativos aumentó a 55,0% (IC95% 45,2% a 64,4%).

El cálculo de la RV positiva no se puede hacer con una especificidad del 100% y la fórmula descrita en el anexo 13.1. Como medida de aproximación empleamos el punto medio del intervalo de confianza de la especificidad (98,3%), con lo que se obtuvo una RV positiva de 27,1. Por su parte la RV negativa fue de 0,6. Expresadas como WoE positiva y negativa, los resultados fueron de +14,3 deciban y -2,6 deciban respectivamente.

Con el punto de corte en 7, el índice de Youden se degradó a 0,45, lo que indica una capacidad informativa más pobre.

La efectividad de la prueba ( $\delta$ ) también se aproximó con una especificidad no 1, igual que para el cálculo de la RV positiva. El resultado fue de 3,8.

Finalmente, la *odds ratio* diagnóstica para HAI definitiva según los criterios simplificados fue de 45,2.

#### **9.1.3.2.2. Indicadores sensibles a cambios en la prevalencia de la enfermedad y naturaleza de su relación**

La prevalencia de HAI entre los niños sin hepatopatía conocida o potencial de base, con hipertransaminasemia y/o signos de disfunción hepática, a los que se somete a una biopsia hepática diagnóstica, ha resultado de 47,2%. Esta es la probabilidad preprueba de padecer la enfermedad problema.

En este escenario, la probabilidad postprueba de HAI tras un resultado de 6 puntos o más en el sistema simplificado de 2008 (VPP) se ha calculado en 94,7% (IC95% 87,2% a 97,9%). La probabilidad de acertar de un resultado inferior (descartar HAI: VPN), por su parte, ha sido de 79,4% (IC95% 71,9% a 85,4%). El valor predictivo global, la probabilidad general que tienen los criterios de 2008 de acertar con este punto de corte en esta población (proporción de resultados válidos entre la totalidad de las mediciones efectuadas) ha sido de 84,9% (IC95% 79,5% a 85,4%).

Sin embargo, como se discute en el anexo 13.1, la utilidad global de un sistema diagnóstico se mide de forma más adecuada con la ganancia diagnóstica (o contenido diagnóstico) de la prueba. Es un parámetro análogo al índice de Youden que emplea los valores predictivos en vez de la sensibilidad y especificidad. Para el caso de los criterios de 2008 con el punto de corte en 6, ha resultado de 0,74 (o 74%). Ha sido posible calcular la prevalencia para la cual la ganancia diagnóstica es máxima: diagnosticar HAI con un resultado de 6 puntos en los criterios simplificados obtiene su máximo rendimiento en poblaciones con una probabilidad preprueba de 29,3%, que iguala los valores predictivos positivo y negativo en 89,3%. De forma específica, un resultado positivo de la prueba obtiene una ganancia diagnóstica máxima en poblaciones con un 18,2% de HAI; y un resultado negativo, con un 65,0%.

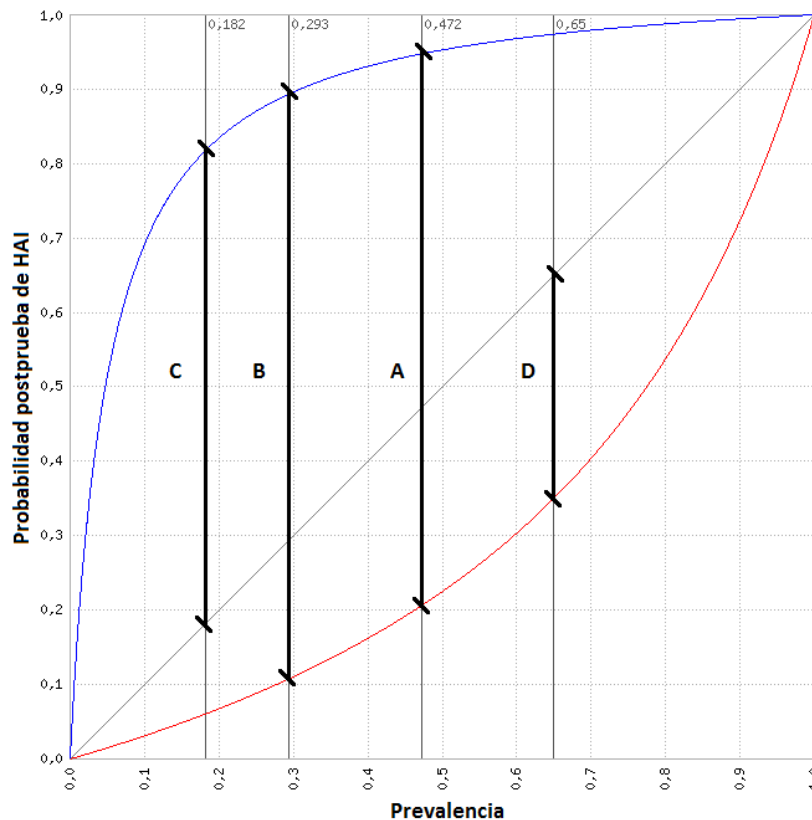


Figura 26: Relación entre la prevalencia de hepatitis autoinmune (HAI) y la probabilidad postprueba con puntuaciones iguales o superiores a 6 (resultado positivo: línea azul) e inferiores a 6 (resultado negativo: línea roja) en los criterios simplificados de 2008. El segmento A representa la ganancia diagnóstica (GD) con la prevalencia de la serie estudiada. El segmento B, la GD global máxima. Los segmentos C y D, la GD máxima para un resultado positivo y negativo, respectivamente.

La probabilidad postprueba de HAI con una puntuación de 7 puntos o más en una población con la prevalencia obtenida con nuestra muestra, ha sido del 100% (IC95% 92,1% a 100%), con lo que el diagnóstico de HAI definitiva por los criterios simplificados es igual de válido que las experiencias comunicadas previamente en la bibliografía. Por otro lado, el VPN ha sido del 67,1% (IC95% 59,6% a 73,7%) y el valor predictivo global ha sido de 74,1% (IC95% 67,8% a 79,5%).

Al evaluar la utilidad global de los criterios de 2008 a través de la ganancia diagnóstica, considerando el punto de corte en 7 puntos, el resultado ha sido inferior al del sistema con el corte en 6 puntos: 67%. Para maximizar esta ganancia diagnóstica con independencia del resultado, la prevalencia de HAI en la población debería de haber sido de 20,6%.

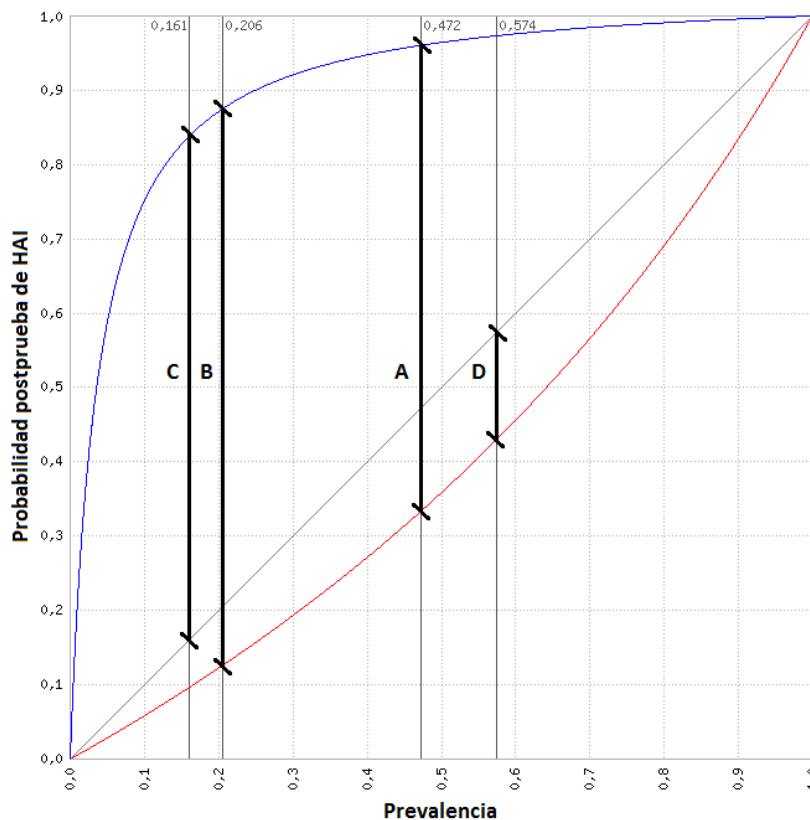


Figura 27: Relación entre la prevalencia de hepatitis autoinmune (HAI) y la probabilidad postprueba con puntuaciones de 7 u 8 (resultado positivo: línea azul) e inferiores a 7 (resultado negativo: línea roja) en los criterios simplificados de 2008. El segmento A representa la ganancia diagnóstica (GD) con la prevalencia de la serie estudiada. El segmento B, la GD global máxima. Los segmentos C y D, la GD máxima para un resultado positivo y negativo, respectivamente.

La ganancia máxima de un resultado positivo se hubiera dado en una población con un 16,1% de HAI y la de un resultado negativo, en una población con una probabilidad preprueba de 57,4%.

Al respecto del área de indicación, cuyo cálculo se ha descrito en el bloque sobre la metodología específica para este capítulo, los resultados obtenidos han sido los siguientes: Para el par sensibilidad/especificidad de los criterios simplificados con un punto de corte en 6, el área de indicación obtenida abarcó el ancho de prevalencias de 2% al 86% (amplitud de 84%). La ganancia neta en certeza diagnóstica máxima sería de 0,7. Empleando los criterios de 2008 con el punto de corte en 7 (puntuaciones superiores indican HAI definitiva), el área de indicación abarcó del 0% al 73% (amplitud del 73%). La ganancia neta en certeza diagnóstica máxima sería de 0,53. Cabe recordar que este concepto denota la diferencia entre la certeza diagnóstica que aportaría una prueba de realizarse respecto a no realizarse, que difiere sensiblemente del de ganancia diagnóstica, que traduce la capacidad de aportar certeza de una prueba una vez se ha realizado.

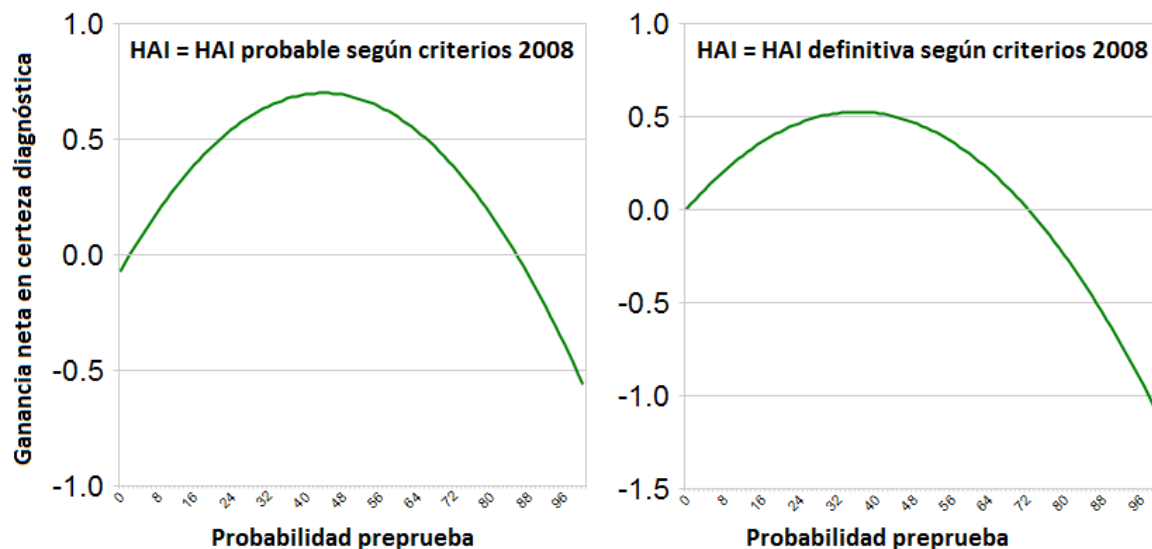


Figura 28: Comportamiento de la ganancia neta en certeza diagnóstica en función de la prevalencia de la enfermedad para los puntos de corte de 6 (izquierda) y 7 (derecha) de los criterios simplificados de 2008. Permite visualizar el área de indicación para cada escenario, que es el ancho de prevalencias en las que la ganancia neta en certeza diagnóstica supera el valor de 0,0. Gráfico obtenido con la calculadora *on-line* de Stalpers [242].

Como referencia para interpretar los posibles resultados de los criterios simplificados de 2008, se simularon las probabilidades postprueba de acierto en función de la prevalencia de la enfermedad, aplicando el teorema de Bayes.

**Tabla 19: Valores predictivos positivos de los criterios simplificados de 2008 según la prevalencia de hepatitis autoinmune en la población. En negrita, prevalencias dentro del área de indicación; en rojo, prevalencias para ganancia diagnóstica máxima; en verde ganancia máxima de un resultado positivo y en azul, prevalencia de nuestra muestra (izquierda para el punto de corte en 6 puntos y derecha para el punto de corte en 7 puntos). Entre paréntesis, intervalo de confianza al 95% (calculado por el método asintótico en la columna de la derecha). Cálculos asumiendo sensibilidad y especificidad constante (esta última aproximada).**

Prevalencia	HAI probable según los criterios simplificados	HAI definitiva según los criterios simplificados	Prevalencia
1%	16,9% (7,2% a 34,9%)	21,5 % (2,1% a 77,8%)	<b>1%</b>
<b>2%</b>	29,2% (13,5% a 52,0%)	35,6% (6,1% a 82,5%)	<b>2%</b>
<b>5%</b>	51,5% (28,7% a 73,7%)	58,8% (19,3% a 89,5%)	<b>5%</b>
<b>10%</b>	69,1% (45,9% a 85,5%)	75,1% (36,5% a 94,0%)	<b>10%</b>
<b>16,1%</b>	79,5% (59,5% a 91,1%)	83,9% (50,5% a 96,4%)	<b>16,1%</b>
<b>18,2%</b>	81,8% (63,0% a 92,2%)	85,8% (54,2% a 96,9%)	<b>18,2%</b>
<b>20,6%</b>	84,0% (66,5 a 93,2%)	87,6% (57,8% a 97,3%)	<b>20,6%</b>
<b>29,3%</b>	89,3% (76,0% a 95,7%)	91,8% (67,6% a 98,4%)	<b>29,3%</b>
<b>40%</b>	93,1% (83,6% a 97,3%)	94,8% (75,4% a 99,1)	<b>40%</b>
<b>47,2%</b>	94,7% (87,2% a 97,9%)	96,0% (79,1% a 99,4%)	<b>47,2%</b>
<b>50%</b>	95,3% (88,4% a 98,2%)	96,4 (80,3% a 99,4%)	<b>50%</b>
<b>60%</b>	96,8% (92,0% a 98,8%)	97,6% (83,9% a 99,7%)	<b>60%</b>
<b>70%</b>	97,9% (94,7% a 99,2%)	98,4% (86,7% a 99,8%)	<b>70%</b>
<b>73%</b>	98,2% (95,4% a 99,3%)	98,7% (87,4% a 99,9%)	<b>73%</b>
<b>80%</b>	98,8% (96,8% a 99,5%)	99,1% (88,9% a 99,9%)	80%
<b>86%</b>	99,2% (97,9% a 99,7%)	99,4% (90,0% a 100%)	86%
90%	99,5% (98,6% a 99,8%)	99,6% (90,6 a 100%)	90%

Tabla 20: Valores predictivos negativos de los criterios simplificados de 2008 según la prevalencia de hepatitis autoinmune en la población. En negrita, prevalencias dentro del área de indicación; en rojo, prevalencias para ganancia diagnóstica máxima; en verde ganancia máxima de un resultado negativo y en azul, prevalencia de nuestra muestra (izquierda para el punto de corte en 6 puntos y derecha para el punto de corte en 7 puntos). Entre paréntesis, intervalo de confianza al 95%. Cálculos asumiendo sensibilidad y especificidad constante.

Prevalencia	No HAI probable según los criterios simplificados	No HAI definitiva según los criterios simplificados	Prevalencia
1%	99,7% (99,6% a 99,8%)	99,4% (99,3% a 99,5%)	<b>1%</b>
<b>2%</b>	99,4% (99,2% a 99,6%)	98,9% (98,7% a 99,1%)	<b>2%</b>
<b>5%</b>	98,5% (97,9% a 98,9%)	97,2% (96,7% a 97,6%)	<b>5%</b>
<b>10%</b>	96,9% (95,8% a 97,7%)	94,2% (93,2% a 95,1%)	<b>10%</b>
<b>20,6%</b>	93,0% (90,6% a 94,8%)	87,5% (85,4% a 89,3%)	<b>20,6%</b>
<b>29,3%</b>	89,3% (85,8% a 91,9%)	81,4% (78,2% a 84,0)	<b>29,3%</b>
<b>40%</b>	83,8% (79,0% a 87,6%)	73,1% (69,6% a 76,5%)	<b>40%</b>
<b>47,2%</b>	79,4% (73,7% a 84,1%)	67,0% (63,0% a 70,8%)	<b>47,2%</b>
<b>50%</b>	77,5% (71,5% a 82,5%)	64,5% (60,4% a 68,5%)	<b>50%</b>
<b>57,4%</b>	71,9% (65,1% a 77,8%)	57,4% (53,1% a 61,7%)	<b>57,4%</b>
<b>60%</b>	69,7% (62,6% a 75,9%)	54,8% (50,4% a 59,1%)	<b>60%</b>
<b>65,0%</b>	65,0% (57,5% a 71,8%)	49,5% (45,1% a 53,9%)	<b>65,0%</b>
<b>70%</b>	59,6% (51,8% a 66,9%)	43,8% (39,5% a 48,2%)	<b>70%</b>
<b>73%</b>	56,0% (48,1% a 63,6%)	40,2% (36,0% a 44,5%)	<b>73%</b>
<b>80%</b>	46,3% (38,6% a 54,2%)	31,3% (27,6% a 35,2%)	<b>80%</b>
<b>86%</b>	35,9% (29,0% a 43,5%)	22,8% (19,9% a 26,1%)	<b>86%</b>
<b>90%</b>	27,7% (21,8% a 34,4%)	16,8 (14,5% a 19,4%)	<b>90%</b>

#### 9.1.3.2.3. Número necesario de pacientes para diagnosticar correcta e incorrectamente a uno

De forma análoga al número de pacientes necesarios a tratar (NNT) de los estudios sobre intervenciones terapéuticas, se ha desarrollado el concepto de número de pacientes necesario para diagnosticar (NND). Si el NNT es el inverso de la

diferencia entre la proporción de pacientes que mejoran con el tratamiento y el número de los que mejoran sin él (o con el tratamiento control), el NND utiliza como denominador la diferencia entre la proporción de sanos bien clasificados por la prueba diagnóstica y la proporción de enfermos mal clasificados por la misma prueba. Se comprueba que esta expresión es equivalente a la del inverso del índice de Youden:  $NND = 1/(Se - (1 - Sp)) = 1/(Se + Sp - 1)$ .

El NND obtenido para el punto de corte en 6 ha sido de 1,5 pacientes y para el punto de corte en 7, de 2,2 pacientes.

Existe también un segundo parámetro relacionado con el anterior, que es el número de pacientes necesario para diagnosticar mal (NNDM). A diferencia del NND, el NNDM sí depende de la prevalencia de la enfermedad de interés dado que se formula como el inverso del valor predictivo global:  $NNDM = 1/(1 - VP_G) = 1/(1 - Sp - (P \times (Se - Sp)))$ . Dónde  $P$  representa la prevalencia.

Con los datos de nuestro estudio, el NNDM para el corte en 6 y 7 puntos ha sido de 6,6 y 3,9 respectivamente.

Tabla 21: Resumen de los principales indicadores de validez de los criterios simplificados de 2008 para el diagnóstico de hepatitis autoinmune en niños. Entre paréntesis, intervalo de confianza al 95%.

Indicador de validez	Corte en 6 puntos (HAI probable)	Corte en 7 puntos (HAI definitiva)
Sensibilidad	72,0% (62,5% a 79,9%)	45,0% (35,6% a 54,8%)
Especificidad	96,4% (91,2% a 98,6%)	100% (96,7% a 100%)
Índice de Youden	0,68	0,45
Valor predictivo positivo*	94,7% (87,2% a 97,9%)	100% (92,1% a 100%)
Valor predictivo negativo*	79,4% (71,9% a 85,4%)	67,1% (59,6% a 73,7%)
Valor predictivo global*	84,9% (79,5% a 89,1%)	74,1% (67,8% a 79,5%)
Razón de verosimilitud positiva	20,2	27,1**
Razón de verosimilitud negativa	0,3	0,6
WoE de un resultado positivo	+13,0 deciban	+14,3 deciban**
WoE de un resultado negativo	-5,4 deciban	-2,6 deciban
Odds ratio diagnóstica	67,3	45,2
Ganancia diagnóstica*	0,74	0,67
Efectividad diagnóstica	4,1	3,8

\*Datos para una prevalencia de 47,2%. \*\*Aproximación con una especificidad de 0,983.

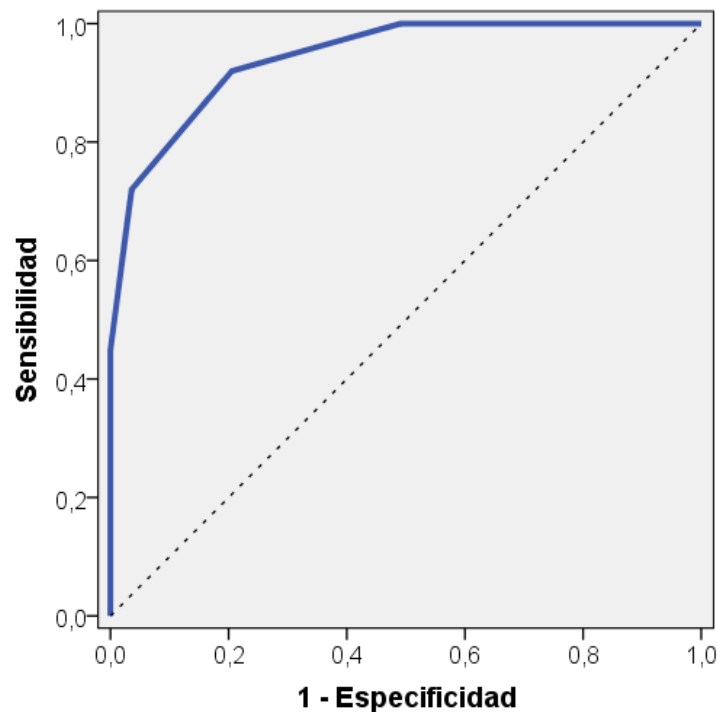
**9.1.3.3. Cálculo del punto de corte óptimo y del poder discriminante global**

Por el método exacto, se estimó un área bajo la curva COR de 94,3% (IC95% 90,3% a 97,0%) considerando los criterios simplificados como una prueba cuantitativa discreta con posibles valores entre 1 y 8.

**Tabla 22: Indicadores de validez para cada uno de los posibles puntos de corte de los criterios simplificados de 2008 para el diagnóstico de hepatitis autoinmune.**

<b>Cut-off</b>	<b>Sensibilidad</b>	<b>Especificidad</b>	<b>Clasificaciones correctas</b>	<b>RV +</b>	<b>RV -</b>	<b>WoE+</b>	<b>WoE-</b>
≥0	100%	0,0%	47,2%	1,0	-	0,0	-
≥2	100%	4,5%	49,5%	1,1	0,0	+0,4	-
≥3	100%	28,6%	62,3%	1,4	0,0	+1,5	-
≥4	100%	50,9%	74,1%	2,0	0,0	+3,0	-
≥5	92,0%	79,5%	85,4%	4,5	0,1	+6,5	-10,0
≥6	72,0%	96,4%	84,9%	20,2	0,3	+13,0	-5,4
≥7	45,0%	100%	74,1%	-	0,6	-	-2,6
8	20,0%	100%	62,3%	-	0,8	-	-1,0

RV: Razón de verosimilitud. WoE: Peso de la evidencia (en deciban)



**Figura 29: Curva de características operativas del receptor de los criterios simplificados de 2008 para el diagnóstico de hepatitis autoinmune.**



Para la prevalencia de nuestra muestra, el punto de corte óptimo de los criterios de 2008 para establecer el diagnóstico de HAI fue de 6, igual que el propuesto en la bibliografía. Incluso para razones de costes de 0,25 (penalizar cuatro veces más un falso positivo que un falso negativo), el punto de corte óptimo se mantuvo en este dintel. Si se hubiera empleado una lógica de evitar de forma predominante los falsos negativos (por ejemplo, con una razón de costes de 2) el punto de corte óptimo hubiera sido 5 en una población con una prevalencia de HAI del 47,2%.

**Tabla 23: Simulación de los puntos de corte óptimos para distintas prevalencias y distintas asunciones respecto a la razón de costes idónea. Encuadrado el contexto más parecido al de la población del estudio.**

Prevalencia		Razón de costes (coste de un falso negativo / coste de un falso positivo)						
		1/8	1/4	1/2	1	2	4	8
5%	Cut-off	7	7	7	7	7	6	6
	Se	45,0%	45,0%	45,0%	45,0%	45,0%	72,0%	72,0%
	Sp	100%	100%	100%	100%	100%	96,4%	96,4%
10%	Cut-off	7	7	7	7	6	6	5
	Se	45,0%	45,0%	45,0%	45,0%	72,0%	72,0%	92,0%
	Sp	100%	100%	100%	100%	96,4%	96,4%	79,5%
20%	Cut-off	7	7	7	6	6	5	5
	Se	45,0%	45,0%	45,0%	72,0%	72,0%	92,0%	92,0%
	Sp	100%	100%	100%	96,4%	96,4%	79,5%	79,5%
30%	Cut-off	7	7	6	6	5	5	5
	Se	45,0%	45,0%	72,0%	72,0%	92,0%	92,0%	92,0%
	Sp	100%	100%	96,4%	96,4%	79,5%	79,5%	79,5%
40%	Cut-off	7	6	6	6	5	5	4
	Se	45,0%	72,0%	72,0%	72,0%	92,0%	92,0%	100%
	Sp	100%	96,4%	96,4%	96,4%	79,5%	79,5%	50,9%
50%	Cut-off	7	6	6	6	5	4	4
	Se	45,0%	72,0%	72,0%	72,0%	92,0%	100%	100%
	Sp	100%	96,4%	96,4%	96,4%	79,5%	50,9%	50,9%

Se: Sensibilidad. Sp: Especificidad.

Los criterios simplificados de 2008, contemplados como un sistema diagnóstico binario (HAI sí o no), también pueden estudiarse a través de la curva COR. En este caso, el trazo solo presenta una angulación y su área bajo la curva da una idea del poder discriminante global del modelo diagnóstico basado en los criterios con el punto de corte elegido.

Para el caso del sistema con el corte óptimo en 6 o más puntos, el área bajo de la curva COR del modelo diagnóstico fue de 84,2% (IC95% 78,6% a 88,8%).

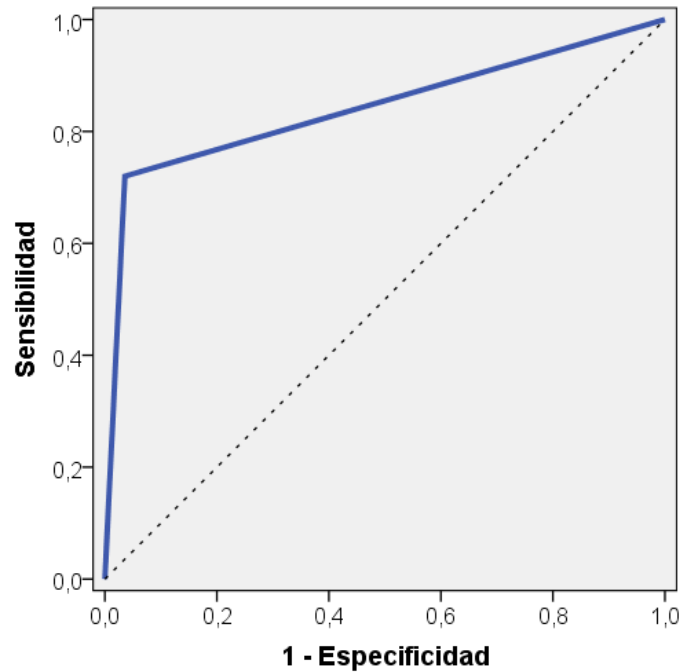


Figura 30: Curva de características operativas del receptor del modelo diagnóstico basado los criterios de 2008 para el diagnóstico de hepatitis autoinmune con el punto de corte en  $\geq 6$ .

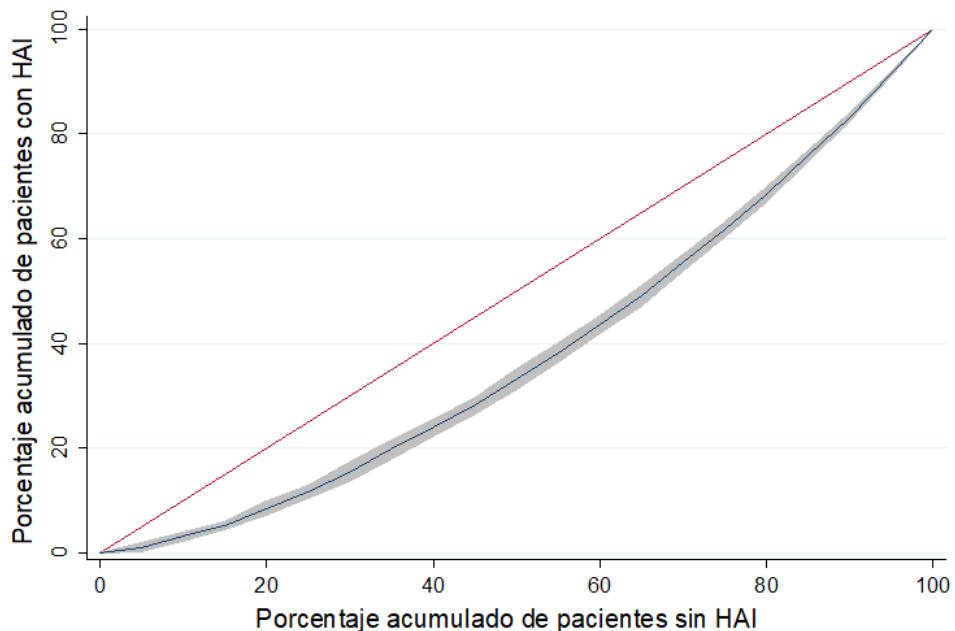


Figura 31: Curva de Lorenz de los criterios simplificados para el diagnóstico de hepatitis autoinmune (HAI). La región sombreada gris representa el intervalo de confianza del 95% de la estimación.

Los indicadores relacionados con la curva de Lorenz se obtuvieron por métodos geométricos mediante el paquete estadístico: El índice de Pietra fue de 0,24 y el índice de Gini, de 0,23.

## **9.2. Capítulo 2: Revisión sistemática y meta-análisis de estudios sobre la validez de los criterios simplificados de 2008**

---

### **9.2.1. Metodología específica**

Al final de la tesis se ha incluido una exposición detallada de los fundamentos de la revisión sistemática de estudios de sistemas diagnósticos, así como del meta-análisis de sus resultados y los métodos de combinación de indicadores de exactitud. La formulación de los cálculos empleados para la resolución de este bloque se puede consultar en el anexo correspondiente.

Lógicamente, a diferencia del resto de capítulos, no se trabajó sobre la población obtenida a través del diseño general. Sin embargo, en el caso de que fuera posible meta-analizar información proveniente de diversos estudios, se añadirían los datos de validez obtenidos en el capítulo 1.

Esta revisión sistemática y meta-análisis se ha registrado en PROSPERO con el código CRD42017081947. Se trata de un directorio prospectivo internacional de revisiones sistemáticas del ámbito de la salud. Está gestionado por la Universidad de York y promovido por el NHS, el sistema de salud del Reino Unido. Los protocolos de la revisión se publican previamente al inicio de esta con el fin de evitar trabajos duplicados y permitir comparar los proyectos con las revisiones completas. Las revisiones sistemáticas admitidas en el registro se pueden recuperar a través de directorios electrónicos como Trip Database.

#### **9.2.1.1. Definición de la pregunta diagnóstica de interés**

Se planteó este bloque con el fin de responder a la cuestión sobre la validez de los criterios simplificados propuestos en 2008 por la IAIHG para el diagnóstico de la HAI en población pediátrica, agregando toda la evidencia disponible en la

literatura. Así pues, los elementos de la pregunta diagnóstica sobre la que se articuló la revisión sistemática son:

- 1) Tipo de estudios. Trabajos observacionales de evaluación de sistemas diagnósticos de corte transversal o de caso-control.
- 2) Participantes. Debían de ser niños y adolescentes con edades comprendidas entre el mes de vida y los 18 años.
- 3) Sistema diagnóstico índice (objeto del estudio). Es la serie de criterios de clasificación descritos en apartados previos y publicados por Hennes en 2008 [19].
- 4) Estándar diagnóstico. No existe un estándar diagnóstico para la HAI que cumpla el requisito de tener una sensibilidad y especificidad perfectas. La aproximación más adecuada son los criterios clásicos revisados en 1999 por la IAIHG, que se han empleado previamente como referencia en multitud de estudios con un diseño análogo [35,90,232–235]. Se consideró óptimo el empleo de los criterios revisados o cualquier sistema de referencia basado en ellos.

#### **9.2.1.2. Búsqueda en fuentes bibliográficas y selección de estudios**

Se llevó a cabo una búsqueda sistemática de publicaciones que informaran de la exactitud o la validez de los criterios de 2008 para el diagnóstico de la HAI en niños, hasta febrero de 2017 (momento en el que se inició la revisión). En un primer paso se consultó el directorio Medline y la librería Cochrane para descartar que ya se hubiera realizado alguna revisión sistemática para la misma pregunta de investigación, tanto en población pediátrica como en adultos. No se obtuvieron resultados.

Los directorios elegidos para la búsqueda electrónica fueron Medline, Embase, Trip Database, Web of Science y Biblioteca Virtual en Salud. Este último, además de en Medline, permite sub-búsquedas en LILACS (Literatura

Latinoamericana y del Caribe en Ciencias de la Salud) y en IBECS (Índice Bibliográfico Español en Ciencias de la Salud).

Las instrucciones de recuperación de datos se basaron en el empleo de los términos “hepatitis autoinmune” Y (“diagnóstico” O “criterios simplificados” O “selección de pacientes” O “clasificación”) tanto como tesauros como texto libre. No se restringió la búsqueda por motivo del idioma de publicación. Se acotó el periodo de búsqueda desde el año 2008 (momento de la publicación de los criterios simplificados) y también se limitó la pesquisa a pacientes menores de 18 años o se introdujeron los términos “niño”, “adolescente” y otros relacionados, con el operador booleano Y, a la estrategia de búsqueda (también como tesauros y como texto libre). Finalmente, en las opciones avanzadas, si el directorio lo permitía, se excluyeron los tipos de publicación de carta al editor y de comunicación de caso clínico. Se permitió la entrada de revisiones en una primera fase para consultar posteriormente su bibliografía y permitir recuperar artículos por esta vía en una instancia siguiente.

Como criterios de inclusión se consideraron estos puntos:

- 1) Diseño transversal o de caso-control de evaluación de la validez de los criterios simplificados de 2008 para el diagnóstico de hepatitis autoinmune en población pediátrica.
- 2) Aportar suficiente información como para poder construir la tabla de contingencia tetracórica de exactitud diagnóstica para cada estudio.

Se excluyeron los estudios que presentaran las siguientes características:

- 1) Diseño de cohortes. El motivo principal es que no aseguran la aplicación del sistema diagnóstico a todos los pacientes. Además, dada la naturaleza de los criterios simplificados, este tipo de estudios no sería necesario, dado que su beneficio principal es la optimización del consumo de recursos económicos o la minimización de posibles efectos perjudiciales de la aplicación del estándar diagnóstico.

- 2) En el caso de los diseños de tipo caso-control, el hecho de que no ofrecieran una descripción detallada de los diagnósticos finales que constituyen el grupo control.
- 3) Tamaño muestral de menos de 20 casos.
- 4) Imposibilidad de obtener datos válidos en los casos que ofrecieran dudas incluso después de contactar con los autores del trabajo original.

Artículos y comunicaciones a congreso adicionales se buscaron manualmente en el listado de la revista *Journal of Pediatric Gastroenterology and Nutrition* (JPGN). Esta publicación se reconoció como una fuente relevante de información durante el desarrollo del estudio piloto para el cálculo del tamaño muestral [227]. Además, en ella aparecen los resúmenes de las ponencias y comunicaciones de las reuniones anuales de la ESPGHAN y la NASPGHAN. También se revisaron las comunicaciones presentadas en la reunión anual de la Sociedad Latinoamericana de Gastroenterología, Hepatología y Nutrición Pediátrica (LASPGHAN) desde el 2009.

Dos personas, incluyendo el autor de la tesis, repasaron de forma independiente el título y el resumen de los artículos recuperados en un primer paso. Se rechazaron los estudios duplicados y los que tuvieran un diseño o un objetivo inapropiados. Los trabajos seleccionados se leyeron en su totalidad para confirmar su elegibilidad. Los posibles desacuerdos se solucionaron tras discusión y revisión conjunta de las fuentes originales. En la última fase se repitió el proceso con las citas de los estudios incluidos y se rechazaron los trabajos de revisión.

A continuación, se detalla la semántica y las operaciones lógicas de la estrategia de búsqueda en cada directorio.

#### **9.2.1.2.1. Medline**

("Hepatitis, Autoimmune"[Mesh] OR "Hepatitis, Autoimmune"[Majr] OR "Hepatitis, Autoimmune/diagnosis"[Majr]) AND ("Patient Selection"[Majr] OR "Patient Selection"[Mesh] OR "Classification"[Majr] OR "Classification"[Mesh] OR (simplified[All Fields] AND ("standards"[Subheading] OR "standards"[All Fields] OR

"criteria"[All Fields])) AND ("2008/01/01"[PDAT] : "3000/12/31"[PDAT]) AND ("infant"[MeSH Terms] OR "infant"[All] OR "child"[MeSH Terms] OR "child"[All] OR "children"[All] OR "adolescent"[MeSH Terms] OR "adolescent"[All]) AND (Classical Article[ptyp] OR Clinical Study[ptyp] OR Journal Article[ptyp] OR Meta-Analysis[ptyp] OR Review[ptyp] OR systematic[sb] OR Multicenter Study[ptyp] OR Practice Guideline[ptyp] OR Guideline[ptyp] OR Congresses[ptyp] OR Comparative Study[ptyp]).

#### **9.2.1.2.2. Embase**

#1: Thesauri (autoimmune hepatitis) OR Title (autoimmune hepatitis) OR Free text (autoimmune hepatitis).

#2: Thesauri (diagnostic) OR Free text (simplified criteria) OR Thesauri (patient selection) OR Thesauri (classification).

#3: #1 AND #2. Filtered by: Age (younger than 18 years old) AND Pub type (Article OR Article in press OR Review OR Conference paper OR Conference abstract OR Conference review) AND Date (2008 – 2017).

#### **9.2.1.2.3. Trip Database**

All of these words: “autoimmune hepatitis” AND “simplified criteria” AND “children” [anywhere in the document].

Timeframe: 2008 – 2017.

Evidence type: Primary Research.

#### **9.2.1.2.4. Web of Science**

#1: Tema: (autoimmune hepatitis) OR Título: (autoimmune hepatitis).

#2: Tema: (diagnostic) OR Tema: (simplified criteria) OR Tema: (patient selection) OR Tema: (classification).

#3: Tema: (pediatrics) OR Tema: (children) OR Tema: (infants) OR Tema: (adolescents) OR Tema: (young adults).



#4: #1 AND #2 AND #3. Refinado por: Áreas de investigación: (GASTROENTEROLOGY HEPATOLOGY OR IMMUNOLOGY OR PEDIATRICS) AND Tipos de documento: (ARTICLE OR ABSTRACT OR REVIEW OR MEETING) AND Período de tiempo=2008-2017 AND Idioma de búsqueda=Auto.

#### **9.2.1.2.5. Biblioteca Virtual en Salud**

tw:(tw:(tw:(("autoimmune hepatitis")) AND ((tw:(("diagnostic")) OR (af:(simplified criteria)) OR (tw:(("classification"))))) AND (instance:"regional") AND (type\_of\_study:(("cohort" OR "case\_control" OR "guideline")) AND limit:(("adolescent" OR "child" OR "child, preschool" OR "infant")) AND year\_cluster:(("2008" OR "2009" OR "2010" OR "2011" OR "2012" OR "2013" OR "2014" OR "2015" OR "2016" OR "2017")) AND type:(("article" OR "thesis")))).

#### **9.2.1.3. Extracción y presentación de los datos de los estudios primarios**

Para cada uno de los estudios primarios finalmente seleccionados, se recogieron los siguientes datos: total de casos de HAI, total de no casos, verdaderos positivos, verdaderos negativos, punto de corte considerado, año de publicación, tipo de diseño del estudio y estándar diagnóstico de referencia. La información se volcó a una hoja de cálculo diseñada expresamente con el programa *Excel* versión 2010 (Microsoft, Redmond, WA).

Eventualmente, la proporción de verdaderos positivos y negativos se estimó a partir de los valores de sensibilidad y especificidad comunicados en cada estudio original. Se expresaron gráficamente en un diagrama con su IC95%.

Para evitar errores en la extracción de la información de los estudios primarios, los mismos dos investigadores que efectuaron la búsqueda comprobaron independientemente los datos. La elección de los mismos se hizo por consenso.

#### **9.2.1.4. Riesgo de errores sistemáticos**

La evaluación de la calidad se llevó a cabo mediante la herramienta QUADAS-2 (*Quality Assessment of Diagnostic Accuracy Studies 2*) [250]. En los anexos se

puede consultar una descripción detallada de este método de evaluación de estudios primarios de sistemas diagnósticos.

En primer lugar, se realizó una exposición narrativa de las cuestiones referentes a las cuatro fases que contempla la herramienta: selección de pacientes, sistema diagnóstico en estudio, prueba de referencia y flujo de pacientes.

Por último, se resumió la evaluación de los ítems sobre el riesgo de sesgo y la aplicabilidad con una figura resumen empleando el programa informático *Review Manager 5.3*.

### 9.2.2. Análisis estadístico

En un primer paso se exploró gráficamente, con un *forest plot*, la sensibilidad y la especificidad de cada estudio primario. Posteriormente se situó cada par sensibilidad / 1 – especificidad en el plano COR y se calculó el coeficiente de correlación de Spearman ( $\rho$ ) entre la sensibilidad y la especificidad. Estas aproximaciones ayudarían en la decisión de establecer si existe efecto umbral. Solo si los estudios recuperados hacen explícito el punto de corte de los criterios de 2008, y es el mismo en todos los trabajos originales, el estudio del efecto umbral sería innecesario.

Siguiendo el ejemplo de la figura 78, utilizando como referencia el punto del plano COR correspondiente a los indicadores obtenidos en el capítulo 1, se comparó la posición relativa de los otros puntos obtenidos del resto de estudios primarios.

Se halló la medida de inconsistencia  $I^2$ , basada en el estadístico  $Q$  de Cochran, para el estudio de la homogeneidad de las razones de verosimilitud y las *odds ratio* diagnósticas. Un valor  $p$  para el estadístico  $Q$  inferior a 0,05 o una  $I^2 \geq 50\%$  se interpretó como indicativo de heterogeneidad significativa, en cuyo caso se emplearon métodos de efectos aleatorios para el cálculo de los promedios ponderados de los indicadores de validez. En caso contrario, se emplearon modelos de efectos fijos.

Para llevar a cabo el análisis, se consideró necesario un número mínimo de estudios originales de 3 [251]. Las decisiones sobre la idoneidad de la combinación de indicadores de validez y la representación de la curva COR resumen se tomaron en base al algoritmo propuesto por Chappell *et al.* [252].

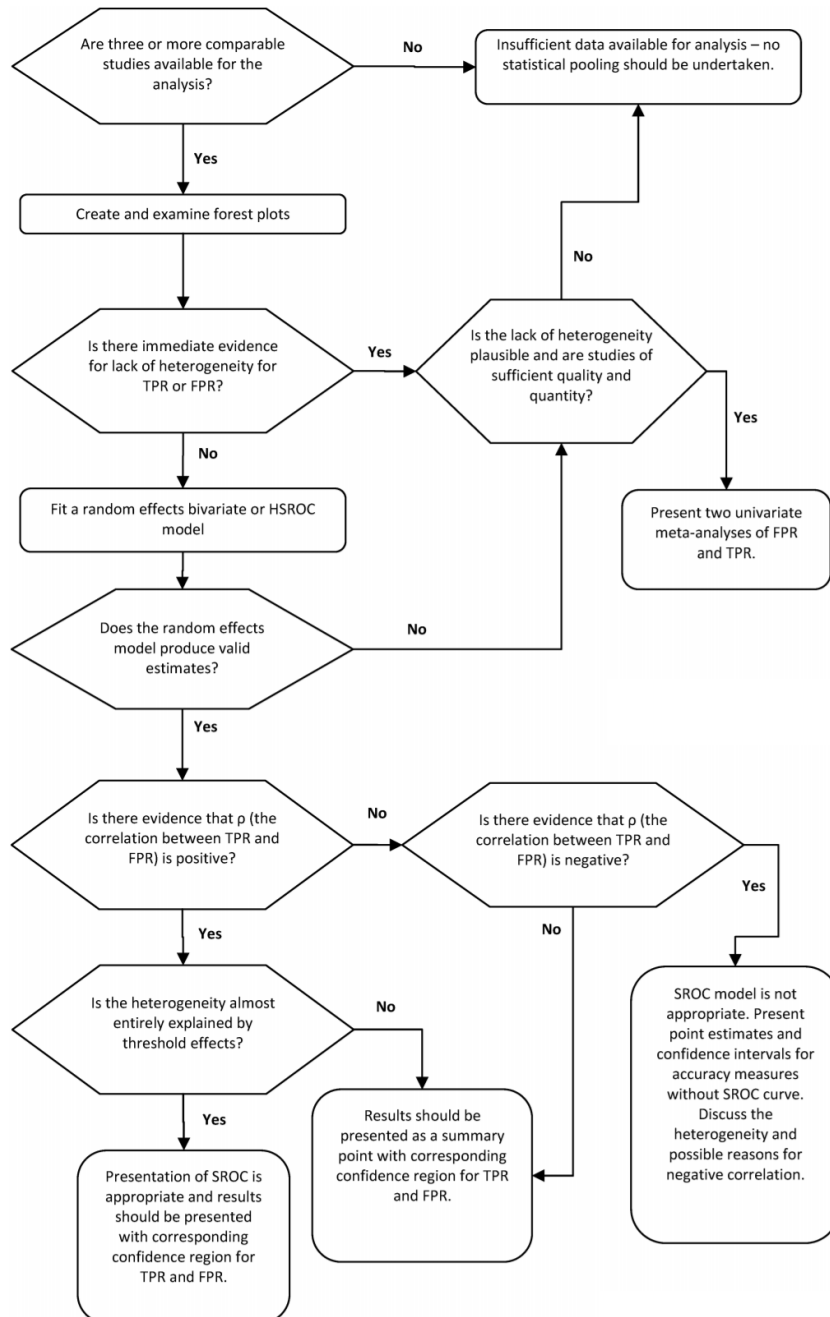


Figura 32: Algoritmo para la toma de decisiones en meta-análisis de sistemas diagnósticos. Adaptado de Chappell *et al.* Reproducido con el permiso de la *European Network for Health Technology Assessment*.

La presencia potencial de sesgo de publicación se estudió con el gráfico en embudo de Deeks, que consiste en representar gráficamente la *odds ratio* diagnóstica frente a una medida de la precisión de los estudios basada en el tamaño muestral efectivo, que depende del número de enfermos y no enfermos. La prueba evalúa la asimetría de este gráfico mediante un valor *p*, que se consideró significativo para valores inferiores a 0,1 [253].

Las revisiones sistemáticas de estudios diagnósticos habitualmente mezclan diseños de tipo caso-control y de cohortes. Los métodos estándar para la evaluación de sistemas diagnósticos habitualmente se centran en la sensibilidad y la especificidad, pero, como se explica en el anexo 13,1 y ha quedado reflejado en el capítulo 1, ello conlleva ignorar las medidas de validez relacionadas con la prevalencia de la condición de interés. Recientemente se ha propuesto un modelo matemático híbrido para el análisis conjunto de la prevalencia estimada en los estudios de cohortes o transversales, con los indicadores de validez obtenidos en cualquier tipo de diseño (Chen *et al.*). Este modelo se fundamenta en un proceso de inferencia basado en la verosimilitud combinada (*composite likelihood*), que ha demostrado ser relativamente efectiva para armonizar las estimaciones de valores predictivos combinados en meta-análisis de estudios de diagnóstico por imagen para la detección de metástasis en pacientes con melanoma [254].

Existe documentación de ejemplo para efectuar las operaciones pertinentes en el entorno informático *Statistical Analysis Software (SAS/STAT)* [254]. En caso de que se encontraran más trabajos que aportaran información válida sobre la probabilidad preprueba de HAI, se llevaría a cabo el meta-análisis mediante el modelo híbrido de Chen *et al.* [254]. En caso contrario, los análisis se realizarían con el programa de distribución libre Meta-Disc, de la Unidad de Bioestadística del Hospital Ramón y Cajal de Madrid [255] y con el entorno Stata (versión 14.0, *Stata Corporation*, Texas, EEUU) empleando las librerías *Metandi* y *Metafunnel*, tampoco sujetas a derechos de autor y desarrolladas de forma colaborativa y abierta [256,257].

## 9.2.3. Resultados

### 9.2.3.1. Características de los estudios recuperados e incluidos

Se identificaron un total de 123 artículos a través de los criterios de búsqueda iniciales en los directorios informáticos y en la literatura gris. Después de retirar los trabajos duplicados o aquellos de tipo inapropiado (cartas al director o comunicación de casos clínicos que no se pudieron filtrar en la primera búsqueda) quedaron 60 estudios. El resumen de estos artículos originales se estudió para comprobar que su contenido era adecuado a la pregunta de interés y con este gesto se pudieron excluir 25 trabajos (4 de diseño adecuado, pero sobre población íntegramente adulta). Los 35 artículos restantes se consiguieron a texto completo para una lectura detallada, tras la cual solo 3 estudios fueron seleccionados finalmente.

Tabla 24: Características de los estudios incluidos en el meta-análisis.

Estudio →	Hiejima [211]	Mileti [210]	Gonçalves [258]	Arcos*
Año de publicación	2011	2012	2016	2017
Proporción de verdaderos positivos	11/20	34/37	40/46	72/100
Proporción de verdaderos negativos	31/36	38/40	45/46	108/112
Sistema diagnóstico de referencia	Criterios revisados de 1999 y propuesta ESPGHAN/NASPGHAN de 2009	Criterios revisados de 1999	Criterios revisados de 1999 y propuesta ESPGHAN/NASPGHAN de 2009	Análisis de casos discrepantes basado en criterios de 1999 y el diagnóstico de la HC
Tipo de diseño	Fase III: Caso-control	Fase III: Caso-control	Fase III: Caso-control	Fase III: Transversal
Ancho de edades de los pacientes	1 a 15 años	1 mes a 19 años	1 a 16 años	16 meses a 16 años
Controles	Predominio de hepatitis crónica C. Incluye CEP	Predominio de metabopatías. Incluye CEP	Predominio de infección crónica por virus B. No incluye CEP	Predominio de hepatitis criptogénica y Wilson. Incluye CEP
Casos	Incluye FHA	Incluye FHA	No incluye FHA	Incluye FHA

\*Estudio propio. HC: Historia clínica. CEP: Colangitis esclerosante primaria. FHA: Fallo hepático agudo.

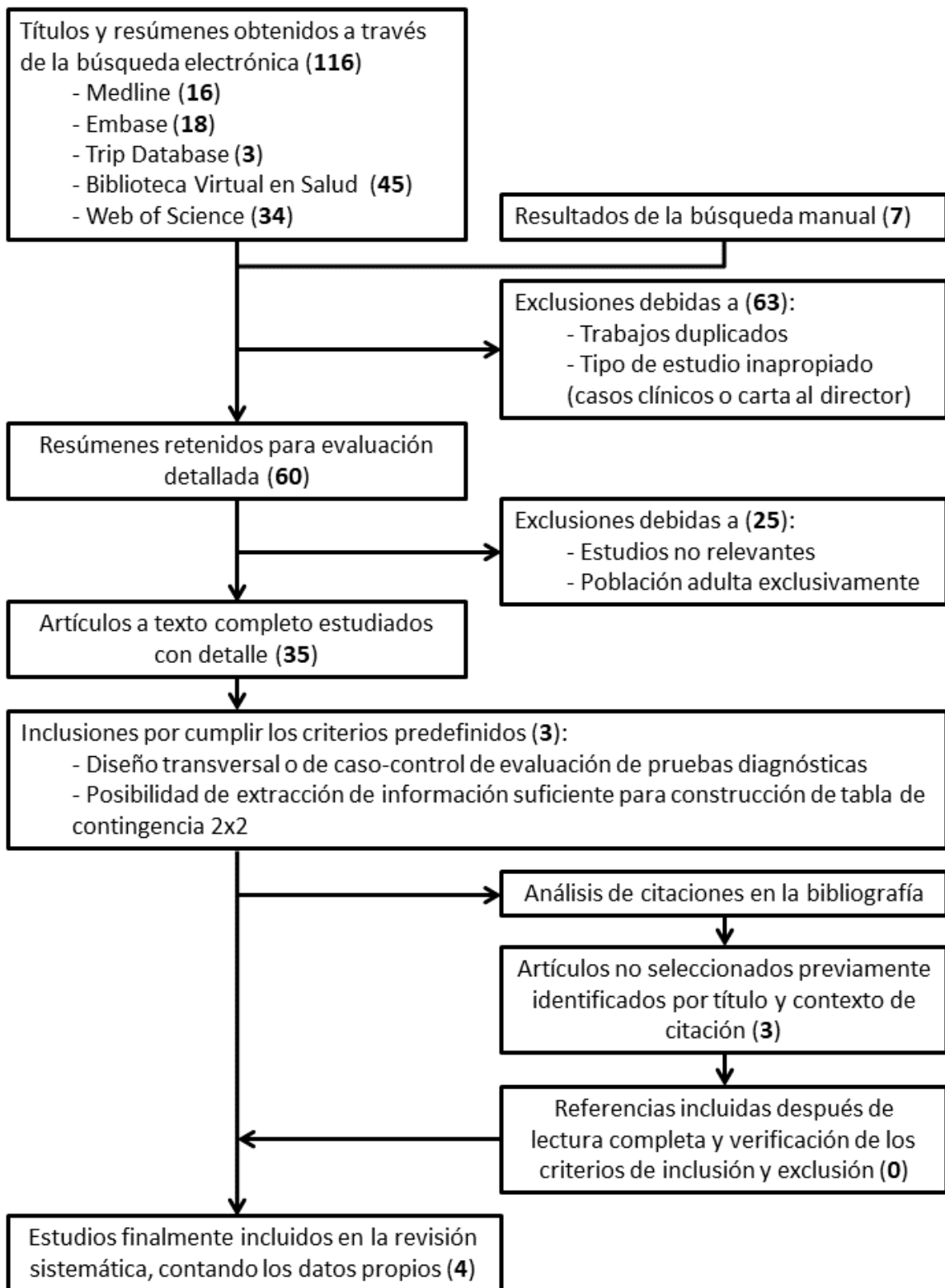


Figura 33: Diagrama de flujo de la selección de estudios primarios para la revisión sistemática y el meta-análisis de la validez de los criterios simplificados de 2008 para el diagnóstico de la hepatitis autoinmune.

### **9.2.3.2. Evaluación de la calidad de los estudios primarios**

El primer estudio de evaluación de los criterios simplificados de 2008 en pediatría fue llevado a cabo por Hiejima *et al.* y publicado en 2011 en JPN [211]. Se trata de un estudio unicéntrico sobre población japonesa, cuyos casos fueron diagnosticados de forma consecutiva de HAI entre 2007 y 2010. Incluye un total de 20 casos y 36 controles. El artículo describe que la selección de los controles se hizo a través del listado de pacientes a los que se les practicó una biopsia hepática en el mismo periodo. Incluyen en el grupo control todos los casos con información suficiente para poder aplicar el estándar diagnóstico basado en los criterios de 1999, sobre un total de 111 pacientes elegibles. No se discute si en la elección de los controles puede haberse incurrido en un sesgo de selección. Tampoco se explica si los investigadores que aplican los criterios simplificados y los clásicos están cegados respecto al resultado del otro sistema diagnóstico. Al respecto del uso de las pruebas, se remite a las fuentes originales que describen su desarrollo y validación, por lo que se entiende que se aplican siguiendo las instrucciones correctas. El análisis se lleva a cabo por protocolo, sin contemplar la influencia de posibles casos de clasificación indeterminada. Finalmente, se aportan datos básicos demográficos sobre los casos y controles, haciendo explícito su diagnóstico definitivo y sin ofrecer análisis agregado por diagnósticos alternativos concretos debido al escaso tamaño muestral. Estos datos permiten deducir que las características de los pacientes abarcan el perfil pediátrico habitual.

El otro trabajo recuperado se publicó en 2012 en *Clinical Gastroenterology and Hepatology*, por Mileti *et al.* [210]. El estudio se realizó en un solo centro de referencia de California (Estados Unidos). El tamaño muestral es similar al del primer estudio original, con 37 casos y 40 controles. El periodo de estudio abarcó de 1991 a 2010. Se recuperaron los casos a través de un registro de pacientes propio, en base a la codificación diagnóstica y a la presencia de información suficiente recuperable en la historia clínica para poder aplicar los criterios clásicos. No se hace explícito que el muestreo sea consecutivo, aunque es posible que sí lo sea. Tampoco se discute si

la selección de casos por esta vía puede imprimir un sesgo de selección ni se contempla como compensarlo a través del diseño ni por el análisis estadístico. La aplicación de ambos criterios se hace de forma abierta y, en el caso de los simplificados, utilizando los niveles de globulina como alternativa a los de IgG cuando esta última no estaba disponible. Se obtuvieron unos resultados similares respecto al análisis considerando solo los casos en los que solo se disponía de IgG. Se describen los datos de sexo y edades de los grupos de HAI y no HAI, que abarca también algunos pacientes menores de un año. En esencia se trata de un estudio con una metodología igual, por lo que respecta a sus elementos básicos, a la del trabajo de Hiejima.

La revisión de la literatura gris permitió encontrar también una comunicación oral al congreso de la LASPGHAN del año 2017 [258]. Se estudió el resumen del trabajo y se amplió la información necesaria tras contactar con los autores. Es un estudio llevado a cabo durante ese año en la Unidad de Trasplante Hepático del Hospital Universitario de Coimbra (Portugal). Se trata de un estudio de evaluación de sistemas diagnósticos con un diseño de caso-control. Incluye 46 pacientes pediátricos en cada grupo. Entre los controles, no se incluyó ningún caso de CEP, pero sí 21 infecciones crónicas por VHB, 9 EHNA, 9 infecciones crónicas por VHC y 7 enfermedad de Wilson. Entre los casos hubo un 28% de HAI tipo 2 y ninguno con presentación clínica al debut de fallo hepático agudo. Dentro de los casos etiquetados de HAI por el estándar de referencia, describen un 30% de casos de CEAI por criterios histológicos y de imagen. Los datos crudos de validez se informan contando estos pacientes como casos reales de HAI, lo que puede imprimir un sesgo de clasificación. Sin embargo, en el análisis efectuado tras excluirlos, no se encontraron diferencias significativas en términos de sensibilidad y especificidad.

A modo de resumen, los tres trabajos seleccionados tienen un diseño de caso-control en los que la recuperación de los pacientes se hizo de forma retrospectiva en base a la codificación diagnóstica y el registro de biopsias hepáticas. Se aplicó la prueba índice (criterios simplificados) por igual todos los pacientes y el



estándar diagnóstico elegido (criterios clásicos revisados), aunque no es un auténtico patrón de referencia, puede comportarse adecuadamente para tal fin. Aun así, es posible que no se hayan incluido pacientes con HAI y resultado dudoso de los criterios clásicos, es decir, que el grupo de casos sea una muestra sobreesleccionada y no represente el abanico de posibles fenotipos. No se describen claramente criterios de exclusión y la población de interés no parece predefinida más allá de catalogarla como pediátrica. Por ejemplo, no se puede saber si el hecho de que el trabajo de Hiejima *et al.* no incluya lactantes menores de un año es por motivos de diseño o por casualidad.

Los tres estudios consideran válido el dintel de positividad en 6 puntos y efectúan el análisis bajo la premisa de que los casos con diagnóstico de HAI probable son igual de válidos que los de diagnóstico definitivo (7 u 8 puntos). Los resultados propios que se intentarán combinar serán los obtenidos con el mismo punto de corte.

Las diferencias principales con el estudio de validación del capítulo 1 no interfieren de forma absoluta en la posibilidad de combinación de indicadores no modificables por la prevalencia de la enfermedad. En primer lugar, nuestro estudio presenta un diseño transversal con recuperación ambispectiva de la información, pero no se ajusta a la definición de un estudio de cohortes, cuyo planteamiento sí que dificultaría la comparabilidad con los resultados de otro tipo de estudios. Además, debido a su mayor tamaño muestral, nuestra estimación de la validez ofrece una mayor potencia estadística. Otra diferencia es que los pacientes se han reclutado en dos centros de tercer nivel, favoreciendo la representatividad de la población diana en la muestra final. Por otra parte, se ha intentado minimizar la posibilidad de sesgo de selección y clasificación para no sobrevalorar la sensibilidad y la especificidad ni aportar una estimación errónea de la prevalencia.

En las tablas siguientes se detallan las respuestas a los ítems del cuestionario QUADAS-2, cuya descripción ampliada se ha incluido como anexo al final de la tesis.

Tabla 25: Respuestas a las preguntas del listado QUADAS-2 sobre el riesgo de sesgo en cada uno de los estudios primarios de la revisión sistemática.

Pregunta	Hiejima [211]	Mileti [210]	Gonçalves [258]	Arcos*
<b>1. Selección de pacientes</b>				
¿Se reclutaron los pacientes de forma consecutiva o aleatoria?	Sí	Sí	No	Sí
¿Se evitó un diseño de caso-control?	No	No	No	Sí
¿Se evitaron exclusiones inapropiadas?	Sí	Dudoso	Dudoso	Sí
<b>2. Realización e interpretación de los criterios simplificados</b>				
¿Se aplicó la prueba índice de forma ciega?	No	No	No	No
¿Se especificó el umbral de positividad o punto de corte?	Sí	Sí	Sí	Sí
<b>3. Estándar diagnóstico para la definición de caso de HAI</b>				
¿El estándar es adecuado para clasificar correctamente la HAI?	Sí	Sí	Sí	Sí
¿Se aplicó el estándar diagnóstico de forma ciega?	No	No	No	Sí
<b>4. Flujo de pacientes y tiempos</b>				
¿Hubo un intervalo de tiempo adecuado entre las pruebas?	Sí	Sí	Sí	Sí
¿A todos los pacientes se les aplicó el mismo estándar?	Sí	Sí	Sí	Sí
¿Se incluyeron todos los pacientes en el análisis?	Dudoso	Dudoso	Sí	Sí

\*Estudio propio. HAI: Hepatitis autoinmune.

Tabla 26: Resumen de la calidad metodológica de los estudios incluidos en la revisión sistemática según la metodología QUADAS-2. Las respuestas hacen referencia al riesgo global de error sistemático en cada una de las cuestiones.

Pregunta	Hiejima [211]	Mileti [210]	Gonçalves [258]	Arcos*
¿Hay sesgo en la selección de los pacientes?	Bajo	Dudoso	Alto	Bajo
Riesgo de sesgo ¿Podría haber sesgos en la realización e interpretación de los criterios simplificados de 2008?	Bajo	Bajo	Bajo	Bajo
¿Podría haber sesgos en la realización e interpretación del estándar diagnóstico basado en los criterios clásicos?	Bajo	Bajo	Bajo	Bajo
¿El flujo de seguimiento del paciente podría haber producido algún sesgo?	Bajo	Bajo	Bajo	Bajo

	Pregunta	Hiejima [211]	Mileti [210]	Gonçalves [258]	Arcos*
Aplicabilidad	¿Hay dudas de que los pacientes incluidos y su ámbito de estudio no se ajusten a la pregunta de la revisión?	Bajo	Bajo	Bajo	Bajo
	¿Hay dudas de que los criterios simplificados (realización e interpretación) difieran de la pregunta de la revisión?	Bajo	Bajo	Bajo	Bajo
	¿Hay dudas de que la HAI definida por los criterios clásicos difiera de la pregunta de la revisión?	Bajo	Dudoso	Dudoso	Bajo

\*Estudio propio. HAI: Hepatitis autoinmune.

### 9.2.3.3. Exactitud diagnóstica y exploración de la heterogeneidad

Los tres estudios recuperados, junto con los datos del capítulo 1, configuran cuatro estimaciones válidas sobre la exactitud de los criterios de 2008 para el diagnóstico de la HAI. Los valores comunicados de sensibilidad y especificidad se representaron en un *forest plot* y en el plano COR.

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Hiejima 2011	11	5	9	31	0.55 [0.32, 0.77]	0.86 [0.71, 0.95]
Mileti 2012	34	2	3	38	0.92 [0.78, 0.98]	0.95 [0.83, 0.99]
Gonçalves 2017	40	1	6	45	0.87 [0.74, 0.95]	0.98 [0.88, 1.00]
Arcos 2017	72	4	28	108	0.72 [0.62, 0.81]	0.96 [0.91, 0.99]

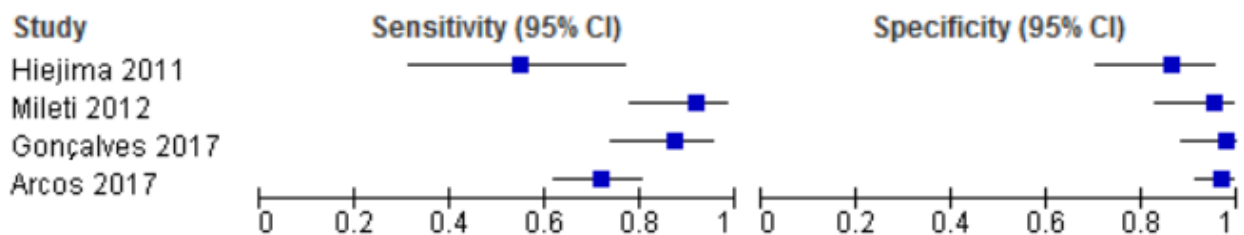
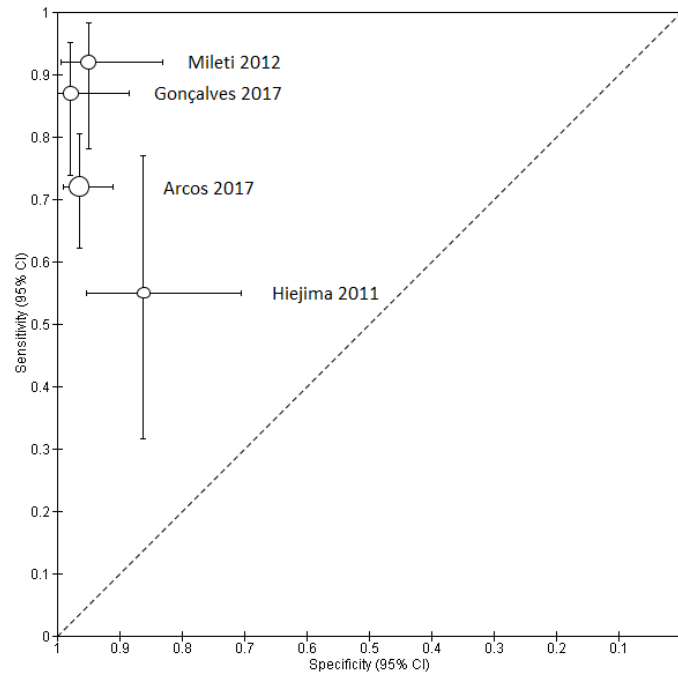


Figura 34: Principales indicadores de validez de los criterios de 2008 comunicados por los estudios primarios de la revisión sistemática. TP y FP: Verdaderos y falsos positivos. FN y TN: Falsos y verdaderos negativos.



**Figura 35: Sensibilidad y especificidad de los trabajos originales de la revisión sistemática representados en el plano de características operativas del receptor con sus intervalos de confianza al 95%.**

Las figuras anteriores no aportan evidencia inmediata de falta de heterogeneidad en las estimaciones de la proporción de verdaderos positivos ni en la de verdaderos negativos. En base a los resultados del análisis de la calidad, se consideró que esta falta de heterogeneidad pudo corresponder de forma plausible a las desviaciones esperables estadísticamente y se decidió continuar con el meta-análisis a pesar del escaso número de trabajos originales encontrados. En un primer paso se calcularon los principales indicadores de validez combinados.

La sensibilidad global redondeada al entero fue de 77%, calculada según un modelo de efectos aleatorios. El IC95% de la estimación fue de 71% a 83%, corrigiendo la sobredispersión a través de una aproximación normal a la binomial.

Por su parte, la especificidad global fue de 95%, estimada por un modelo de efectos fijos dada la no significación estadística de la prueba de la razón de verosimilitud. El IC95% siguiendo el mismo método que para el de la sensibilidad fue de 91% a 97%.

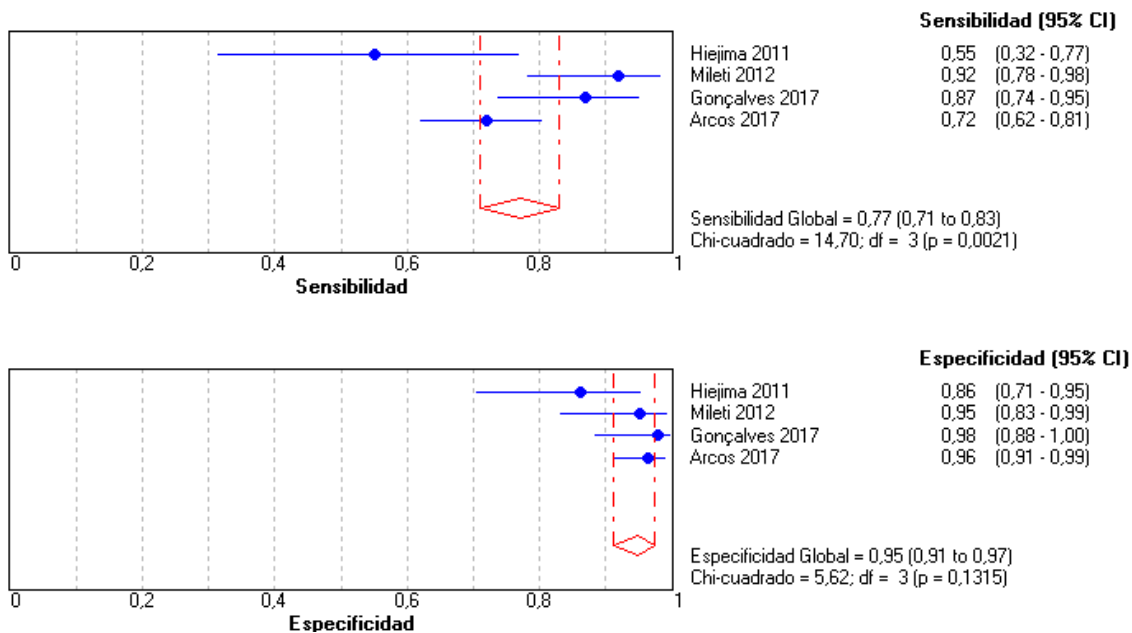


Figura 36: Meta-análisis univariante de la sensibilidad y la especificidad de los criterios simplificados para el diagnóstico de la hepatitis autoinmune.

La razón de verosimilitud combinada positiva y negativa se calculó siguiendo el método de DerSimonian-Laird (de efectos aleatorios) después de comprobar la presencia de heterogeneidad con el contraste basado en la prueba Q de Cochran. La RV positiva global fue de 13,7, con un IC95% de 4,5 a 42,2. El peso del estudio propio para esta estimación fue de 29%. Por su parte, la RV negativa global fue de 0,23 (IC95% 0,12 a 0,45), estimación para la que el estudio propio tuvo un peso de 31%.

Debido a la presencia de un solo trabajo con un diseño transversal adecuado para la estimación de prevalencias (el del capítulo 1 de la tesis), no se calcularon valores predictivos globales mediante el proceso de inferencia basado en las razones de verosimilitud combinadas descrito por Chen *et al.* [254].

Como medida estadística de desempeño global de los criterios simplificados, se calculó la *odds ratio* diagnóstica combinada por el modelo de efectos aleatorios. Su valor fue de 66,8, con un IC95% de 13,2 a 339,0. Además de su interés como indicador de validez diagnóstica, en caso de que fuera posible, la ORD combinada se emplearía en la construcción del intervalo de confianza de la curva COR resumen.

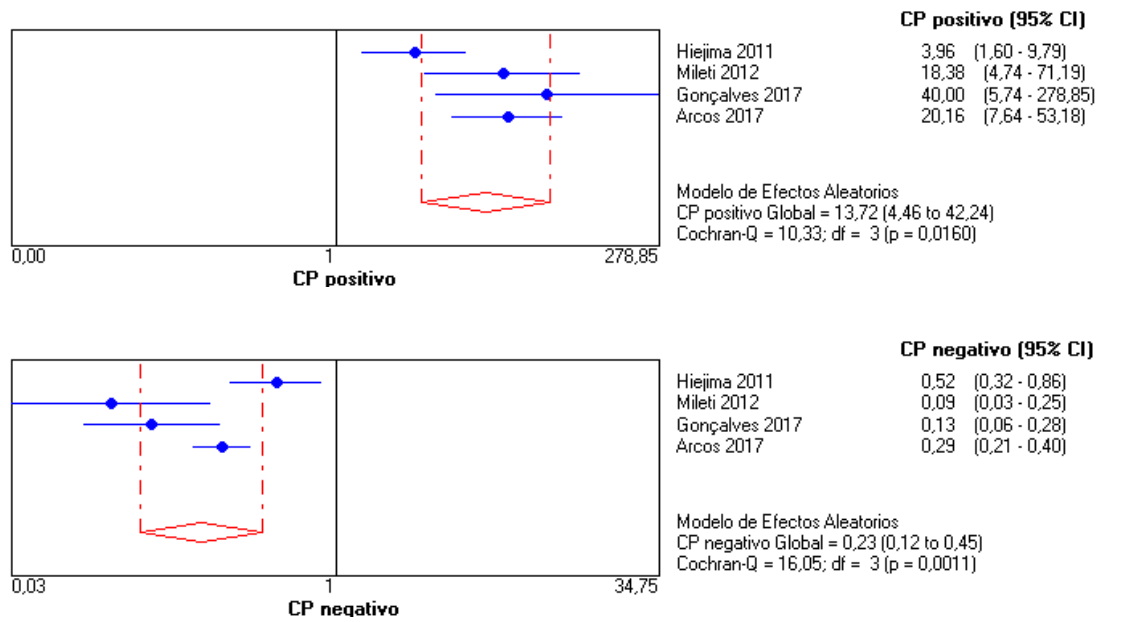


Figura 37: Meta-análisis univariante de las razones de verosimilitud de los criterios simplificados para el diagnóstico de la hepatitis autoinmune.

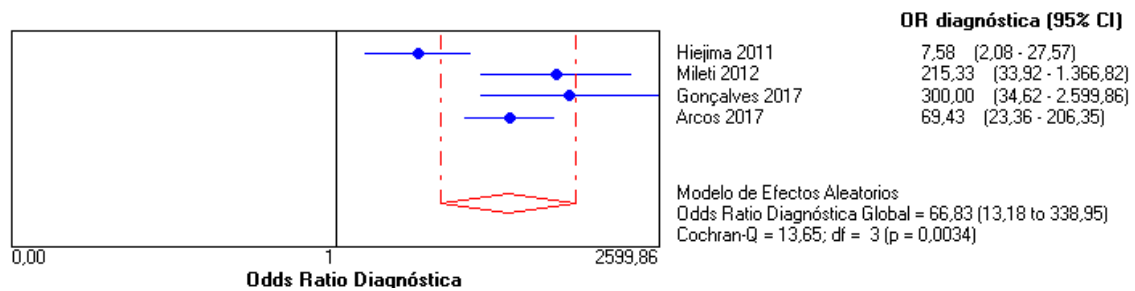


Figura 38: Meta-análisis univariante de la *odds ratio* diagnóstica de los criterios simplificados para el diagnóstico de la hepatitis autoinmune.

Todos los estudios incluidos en el meta-análisis aportan unos datos de validez diagnóstica basados en una concepción binaria de las posibilidades de clasificación de los criterios simplificados (HAI sí o no). Aunque el sistema diagnóstico bajo estudio se puede comportar como una variable cuantitativa, los cuatro trabajos dan sus indicadores para un punto de corte explícito acorde a la recomendación recogida en la bibliografía. Este umbral diagnóstico es el mismo para todos los estudios originales (6 puntos), que además es el que ha demostrado un mejor rendimiento

según los resultados descritos en el capítulo previo. En este escenario, el estudio de la presencia de efecto umbral es innecesario. Por este motivo, el ajuste de una curva COR resumen pierde interés para completar el meta-análisis. Sin embargo, a modo exploratorio, se llevó a cabo la estimación de los parámetros (constante y pendiente –parámetro  $b$ –) de la relación cuasi-lineal entre la ORD resumen y los indicadores de validez diagnóstica para los distintos umbrales. El parámetro  $b$  se calculó en 0,450 por el método de los mínimos cuadrados, con valor  $p$  de 0,788, indicando que no es significativamente diferente de 0, con lo que la curva COR resumen ajustada por Moses-Shapiro-Littenberg resultaría en un trazo simétrico sin la influencia de las ORD combinadas para los varios umbrales potenciales.

Otro argumento a favor de no presentar un modelo de curva COR resumen (y calcular su área bajo la curva) es el hecho de que el coeficiente  $\rho$  de Spearman entre la proporción de verdaderos positivos y la proporción de falsos positivos da un valor negativo de  $-0,400$  con un valor  $p$  de 0,600, con lo que su signo tampoco es significativo. Así, siguiendo las recomendaciones habituales, los resultados se presentaron exclusivamente como un punto resumen del par sensibilidad y especificidad con sus intervalos de confianza al 95% [252].

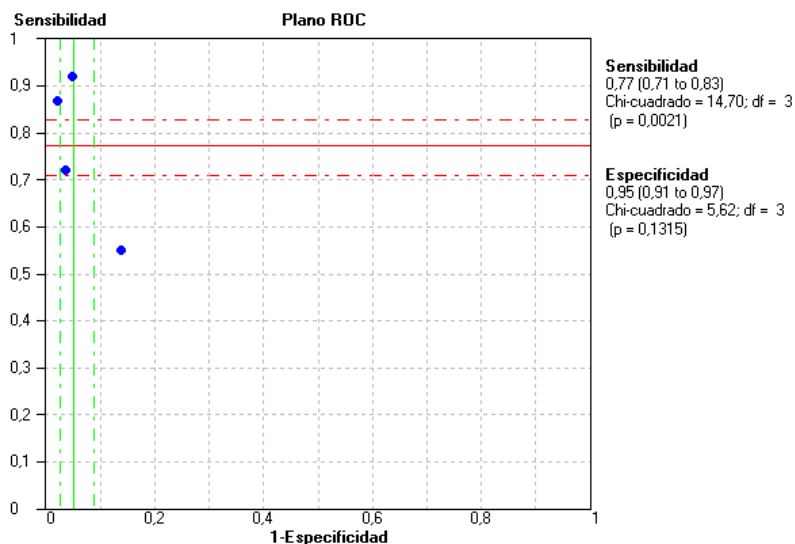


Figura 39: Espacio de características operativas del receptor con la intersección entre la sensibilidad y la especificidad combinadas y sus intervalos de confianza al 95%.

Se observa con claridad la cercanía del punto resumen a la estimación de validez ofrecida por el estudio del capítulo 1.

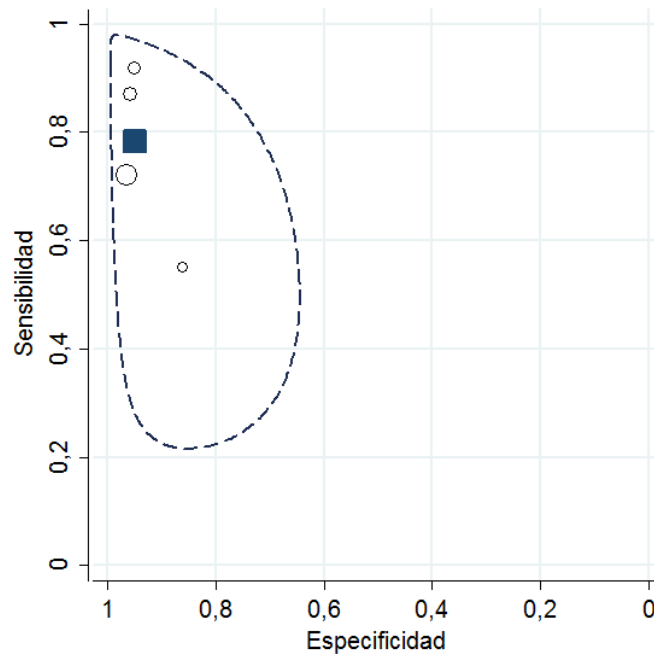


Figura 40: Plano de características operativas del receptor con el punto resumen (cuadrado azul) y su intervalo de confianza al 95% (línea discontinua). Los círculos representan los datos de los estudios primarios (de arriba abajo: Mileti 2012, Gonçalves 2017, Arcos 2017 y Hiejima 2011).

Como se ha discutido previamente, se ha considerado que el trazado de una curva COR resumen es inapropiado en las condiciones de los resultados de esta revisión sistemática. Los estudios incluidos utilizan el mismo punto de corte para un sistema diagnóstico que, por otro lado, admite solo 8 posibles puntuaciones (podría comportarse como una variable cualitativa ordinal más que como una cuantitativa al uso). Las técnicas de modelización trazan un perfil curvilíneo que no es asimilable a la curva COR que darían los estudios primarios. Además, no se cumplen las condiciones de correlación claramente positiva entre la sensibilidad y el complementario de la especificidad.

A pesar de ello y como ejemplo, se ha adjuntado una curva COR resumen simétrica basada en la constante del modelo de Moses y también el ajuste de una



curva basada en un modelo jerárquico (HSROC), que tiene en cuenta la posible imperfección del sistema diagnóstico de referencia [252]. Debido al escaso número de estudios, la zona de predicción con una confianza del 95% se extiende por una parte importante del plano COR y no se ha representado en las figuras siguientes. El área bajo la curva fue de 97,4% (con una desviación estándar de 0,051) y el índice  $Q^*$  se calculó en 0,93 (con una desviación estándar 0,87) pero por los motivos previamente mencionados, no se puede hacer una lectura fiable del significado de estas magnitudes. O expresado de otra manera, su papel como indicadores de una excelente capacidad discriminante global es cuestionable.

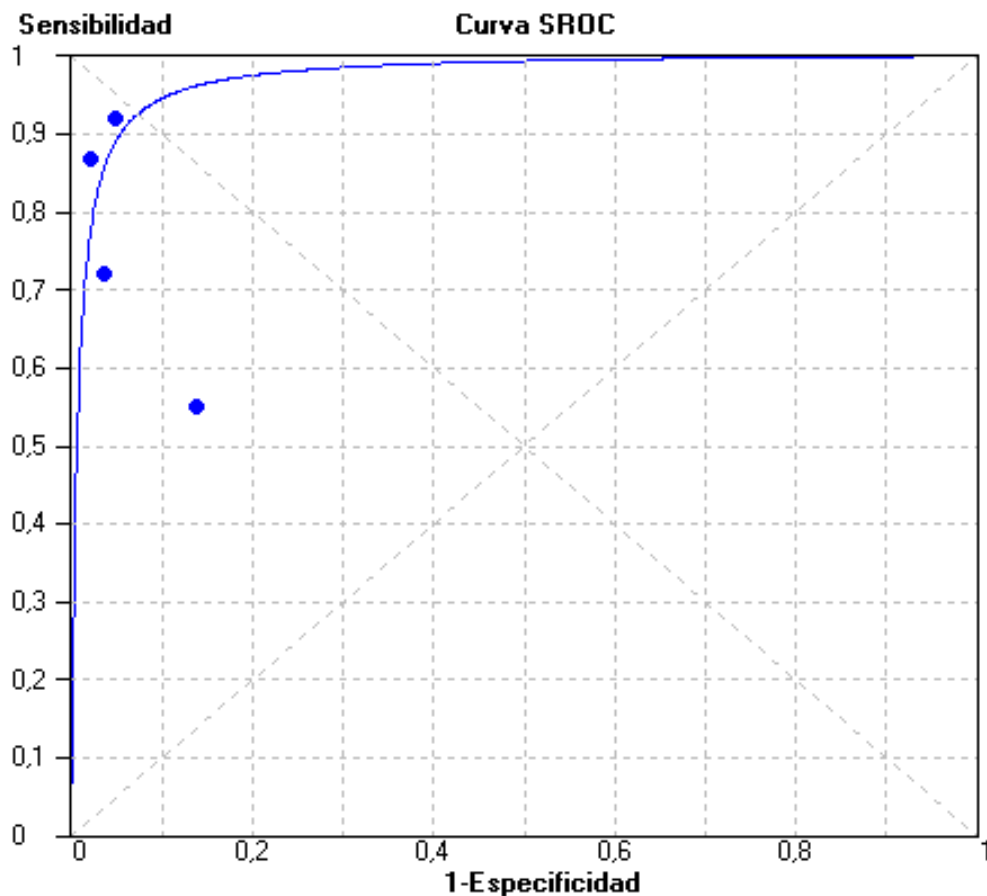


Figura 41: Curva resumen simétrica de características operativas del receptor siguiendo el modelo de Moses-Shapiro-Littenberg. Proyección más allá del tramo calculable con los datos proporcionados por los estudios primarios (puntos azules, de arriba abajo: Mileti 2012, Gonçalves 2017, Arcos 2017 y Hiejima 2011).

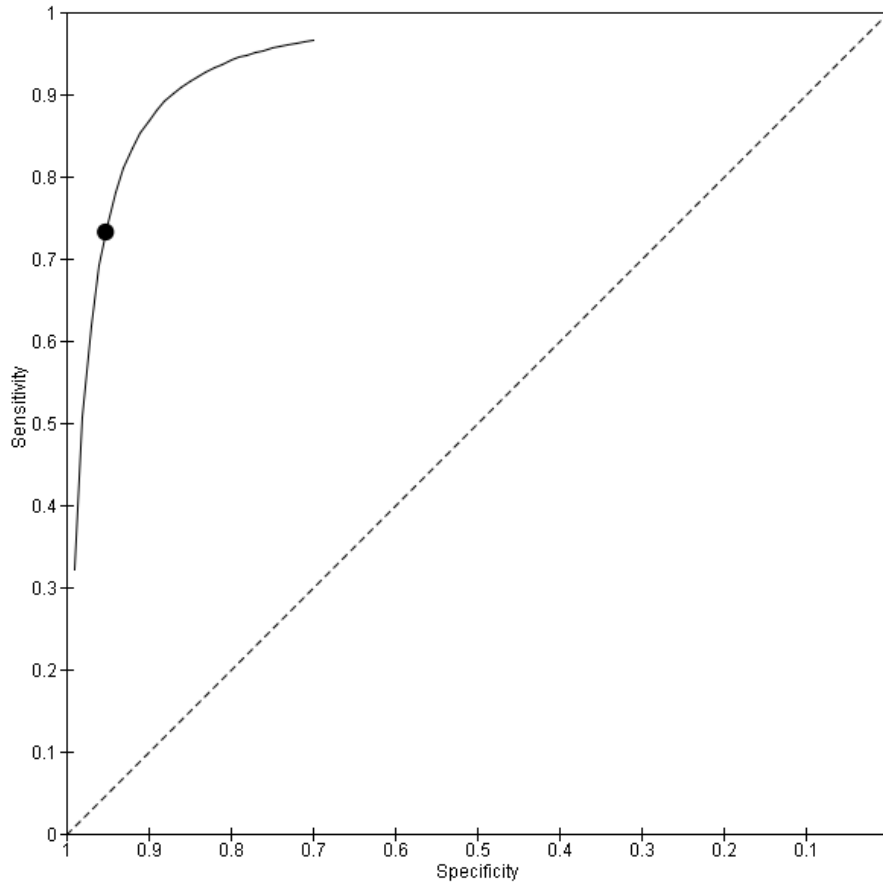


Figura 42: Representación del punto resumen y la curva COR resumen basada en un modelo jerárquico.

#### 9.2.3.4. Estudio de la presencia de sesgo de publicación

De forma análoga al gráfico en embudo (*funnel plot*) de Egger para la evaluación del sesgo de publicación en meta-análisis de medidas de asociación, se representaron los datos de los estudios primarios en un eje bidimensional de la transformación logarítmica de la *odds ratio* diagnóstica con su error estándar. Una inspección de la distribución de los puntos permite constatar que su distribución es aproximadamente simétrica en torno a la *odds ratio* diagnóstica combinada (línea continua central) y englobada dentro de las proximidades de los pseudo-límites de confianza al 95% (líneas intermitentes oblicuas). El hecho de que la nube de puntos esté constituida por escasos elementos resta robustez a leer esta observación como una demostración clara de la ausencia de sesgo de publicación. Sin embargo,

permite sugerir que los estudios recuperados aportan estimaciones de una magnitud similar a la que sería esperable por motivos puramente aleatorios.

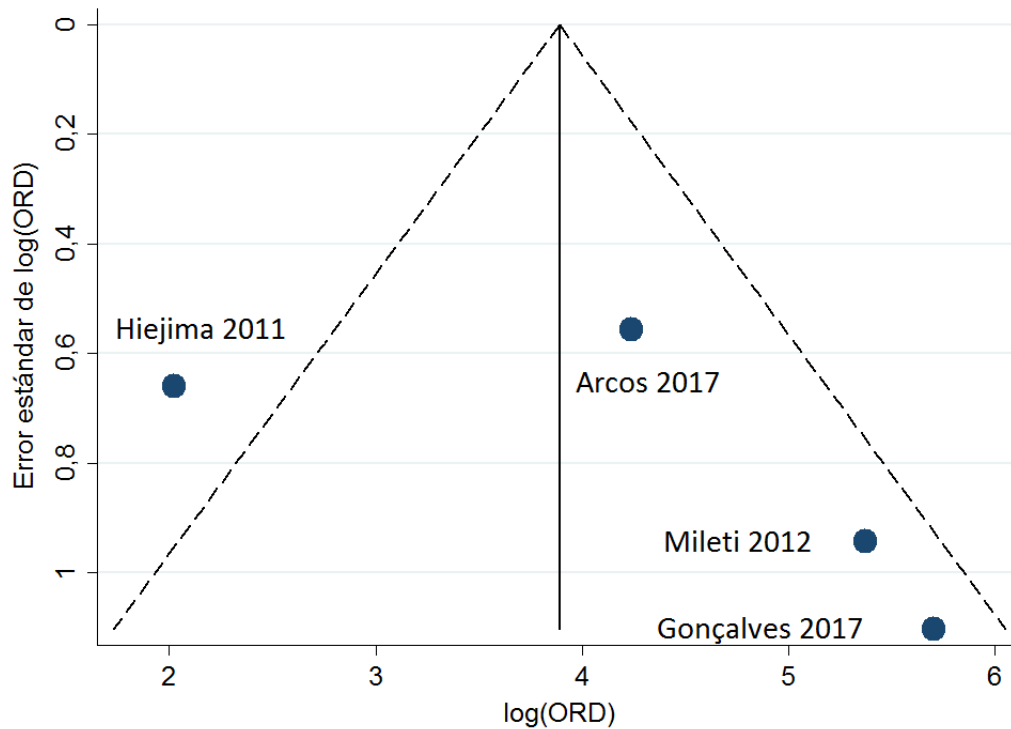


Figura 43: Diagrama en embudo de Deeks que relaciona el logaritmo de la *odds ratio* diagnóstica con su error estándar para cada estudio primario del meta-análisis. La distribución simétrica de los puntos en relación a los pseudo-límites de confianza al 95% no sugiere sesgo de publicación.

El *funnel plot* se considera una buena herramienta para la evaluación de la presencia de sesgo de publicación. Se comprueba que, en ausencia de este tipo de error sistemático, la nube de puntos que representa cada observación en el espacio configurado por una medida de la magnitud del efecto y una medida de su precisión tiene una forma simétrica. Para superar la subjetividad de la apreciación de la simetría, se han desarrollado métodos de regresión lineal paramétricos y no paramétricos que no se comportan de forma adecuada en el caso de los estudios de validez diagnóstica. En este escenario, resulta más conveniente emplear la aproximación de Deeks, en la que la regresión se encarga del ajuste entre el logaritmo natural de la *odds ratio* diagnóstica y el inverso del cuadrado del tamaño

muestral efectivo  $(4 \times n_1 \times n_2)/(n_1 + n_2)$ , siendo  $n_1$  el número de enfermos y  $n_2$ , el número de no enfermos. La significación medida por el valor p de la pendiente de la recta se usa como indicador de la presencia de asimetría y, por tanto, de sesgo de publicación [253].

Tabla 27: Parámetros del ajuste de la recta de regresión en el test de asimetría de Deeks para el estudio de la presencia de sesgo de publicación.

	<b>Coefficiente</b>	<b>Intervalo de confianza al 95%</b>		<b>Error estándar</b>	<b>t</b>	<b>Valor P</b>
Sesgo	-9,59	-121,14	101,96	25,93	-0,37	<b>0,747</b>
Origen	5,23	-5,49	15,96	2,49	2,10	0,171

En nuestro caso, la pendiente de la recta de regresión no es significativa, con lo que el test de Deeks tampoco ha demostrado presencia de sesgo de publicación.

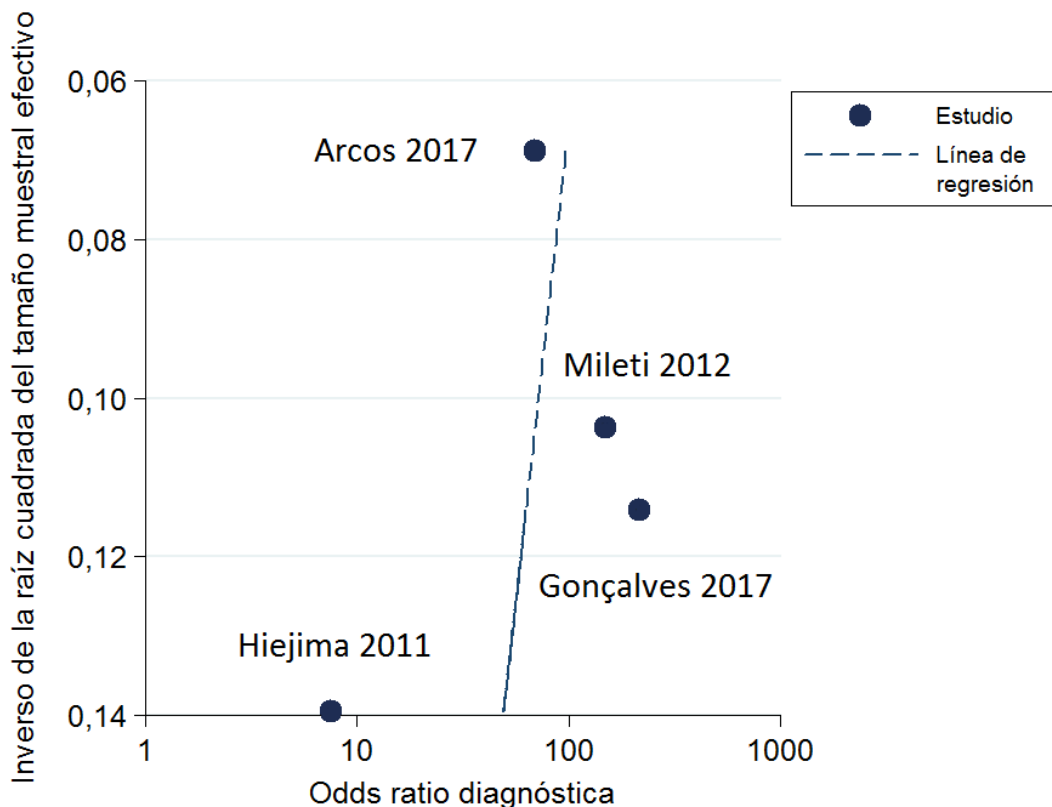


Figura 44: Representación gráfica del test de asimetría sobre el diagrama en embudo de Deeks.



## 9.3. Capítulo 3: Fiabilidad de los criterios simplificados de 2008

---

### 9.3.1. Metodología específica

La fiabilidad de unos criterios diagnósticos depende de la reproductibilidad de sus clasificaciones entre distintos médicos. Por la misma naturaleza de un sistema basado en puntuaciones, el origen de la variabilidad de las mediciones depende fundamentalmente de la interpretación que el operador hace de sus enunciados. Existe además una fuente subordinada de variabilidad para cada uno de los criterios individuales que constituyen el sistema. Por ejemplo, si un sistema diagnóstico por puntos incluye los resultados de una determinación analítica categorizada ordinalmente en anchos de posibles valores, la fiabilidad de las mediciones del laboratorio influye en la fiabilidad del sistema en general. A igualdad de condiciones respecto a la información que se dispone de un caso, sin embargo, la reproductibilidad de los criterios diagnósticos está en función de las características propias de la persona que los aplica.

En un entorno asistencial convencional, los encargados del empleo de unos criterios de clasificación clínica son los médicos, habitualmente con una formación académica similar. Es previsible que en un sistema diagnóstico que incluya una mezcla de elementos clínicos y resultados de pruebas complementarias dependientes del operador, exista el problema potencial de una fiabilidad no perfecta basada en la subjetividad del intérprete.

*A priori*, el parámetro más susceptible de interpretarse incorrectamente es el de la anatomía patológica. En los criterios clásicos de la IAIHG, los hallazgos típicos de la HAI (rosetas, infiltración linfoplasmocitaria, hepatitis de interfase...) reciben, de forma individual, una puntuación concreta. Sin embargo, la variable de la histología de los criterios simplificados solo admite dos categorías: anatomía patológica típica o compatible. Esta última se define como una hepatitis crónica con infiltración linfocítica en ausencia de los rasgos típicos. La presencia simultánea de hepatitis de

interfase, infiltración linfocítica o linfoplasmocitaria en tractos portales con extensión hacia el lóbulo y formaciones de hepatocitos en roseta, por su parte, es la descripción de una histología típica de HAI según los criterios simplificados.

Para la ejecución de esta parte de la tesis, sobre el total de los pacientes reclutados, se aplicaron los criterios simplificados de forma independiente por dos observadores. Por un lado, se consideraron las puntuaciones establecidas por el doctorando. Las clasificaciones basadas en ellas se revisaron de forma consensuada partiendo de la información recogida en la historia clínica y, cuando fue posible, en los hallazgos prospectivos. Son las puntuaciones que se han tenido en cuenta para el estudio de la validez del capítulo 1. Los detalles sobre las características de los criterios simplificados están expuestos en el apartado '*aplicación de las pruebas a validar*' del bloque sobre el diseño general. Por otra parte, también se recogieron las puntuaciones asignadas por un segundo observador. Se trata de un pediatra con entrenamiento en Hepatología infantil, que estableció los puntos en base a lo recogido en los archivos electrónicos de los pacientes de los dos centros de procedencia de la muestra. Las condiciones de aplicación de los criterios fueron las propias de un entorno asistencial normal. Así, no se le restringió la posibilidad de consultar dudas sobre la interpretación de resultados de pruebas complementarias. La elección de qué análisis de sangre concreto elegir para asignar los puntos de las categorías de pruebas de laboratorio fue libre. Es decir, al segundo observador no se le ofreció directamente los resultados de ninguna determinación analítica o biopsia hepática, sino que tuvo que seleccionarlas de entre todas las incluidas en el historial de cada niño.

### **9.3.2. Análisis estadístico**

En un primer paso se obtuvo el índice *kappa* para el acuerdo entre las categorías nominales "no HAI" y "sí HAI" de los criterios simplificados. Estos resultados quedan definidos por el *cut-off* en 6 puntos.

Los resultados relevantes de los criterios simplificados son las categorías diagnósticas “no HAI”, “HAI probable” y “HAI definitiva”. Son resultados discretos ordinales dado que se puede interpretar que siguen cierto orden jerárquico en la probabilidad de HAI que traducen para cada caso. Por este motivo, tal como queda reflejado en el anexo 13.1, es conveniente emplear un índice *kappa* ponderado como estimador del grado de concordancia. No disponemos de ponderaciones objetivas para las discordancias intermedias (“HAI probable” con las otras dos categorías) y máximas (“no HAI” con “HAI definitiva”), así que se utilizó el método de los pesos cuadráticos. Con este cálculo, a cada discrepancia se le asignó un peso correspondiente al cuadrado de la distancia de cada casilla respecto a la diagonal principal en la tabla de contingencia (donde se sitúan las casillas de acuerdo perfecto). De cara a la interpretación y discusión de los resultados se empleó este valor *kappa*. Sin embargo, también se obtuvo el mismo indicador, pero con ponderación lineal a efectos exploratorios.

Los criterios simplificados, además, son una escala que va de 0 a 8 puntos. Esta propiedad permite estudiar la reproductibilidad como una variable cuantitativa continua. Se calculó la desviación estándar intrasujetos basada en los cuadrados medios de los residuos, lo que permitió cuantificar el margen de error de las mediciones con una confianza del 95%. También se estimó el coeficiente de correlación intraclase (CCI), con su error estándar, la prueba de significación y su intervalo de confianza al 95%. La noción que subyace a la formulación del CCI fue introducida por Fisher en los años 20, que propuso una definición especial del coeficiente de correlación de Pearson para distribuciones de igual media y varianza. Se basa en el modelo de análisis de la varianza con medidas repetidas o intrasujeto, tal como se expone con más profundidad en el anexo 13.1. Este estimador del CCI se emplea para medir la *consistencia* y no considera como discrepancias las diferencias de tipo aditivo o constantes. Más allá de esta propiedad, también existe una expresión del CCI para estudiar el *acuerdo* (que sí las tiene en cuenta y, por lo tanto, valora el acuerdo absoluto). Además de los cuadrados medios entre pacientes



( $CMp$ ) y los cuadrados medios de los residuos ( $CMr$ ), también incluye en su formulación los cuadrados medios entre evaluadores ( $CMe$ ). Si  $k$  representa el número de observaciones por paciente y  $n$ , el número de pacientes [259]:

$$CCI_{acuerdo} = \frac{CMp - CMr}{CMp + (k - 1)CMr + (k/n)(CMe - CMr)}$$

El interés máximo de hallar este segundo estimador del CCI se encuentra en la valoración de la fiabilidad de mediciones cuantitativas continua. Aun así, se exploró en el caso de los criterios simplificados (variable cuantitativa discreta) con el fin de disponer de un indicador análogo al índice *kappa* ponderado que contemple cualquier diferencia entre medidas como una discordancia, con independencia de su naturaleza (constante, proporcional u otra).

Por último, se representó un diagrama de Bland-Altman utilizando como puntuación de referencia la empleada en el capítulo 1. Se contrastó la hipótesis de normalidad de la variable diferencia entre las mediciones de los dos observadores con la prueba de Shapiro-Wilk. Si el *test* fuera no significativo, se añadiría en el diagrama el trazado correspondiente a la media de las diferencias y los límites de confianza al 95% basado en la desviación estándar de las diferencias.

Los cálculos y los gráficos asociados se realizaron con el paquete estadístico SPSS® de IBM, en su versión 21.0. El cálculo de los índices *kappa* con SPSS se ejecutó siguiendo la sintaxis de la macro !KAPPA; los datos necesarios para implementar el método de Bland-Altman, con la sintaxis de la macro !AGREE y el cálculo del CCI, con el procedimiento RELIABILITY [260,261].

### 9.3.3. Resultados

El análisis de la fiabilidad de los criterios simplificados se realizó con los mismos 212 pacientes que constituyen la muestra de trabajo para el capítulo sobre la validez. En la tabla siguiente se representan los puntos y categorías diagnósticas asignadas para cada caso por los observadores 1 (doctorando) y 2.

Tabla 28: Relación de resultados de la aplicación de los criterios simplificados por dos observadores independientes.

Paciente	Observador 1			Observador 2			Paciente	Observador 1			Observador 2	
	HAI	Puntos	Clasificación	Puntos	Clasificación	HAI		Puntos	Clasificación	Puntos	Clasificación	
1	No	0	No HAI	0	No HAI	52	No	3	No HAI	3	No HAI	
2	No	0	No HAI	0	No HAI	53	No	3	No HAI	3	No HAI	
3	No	0	No HAI	0	No HAI	54	No	3	No HAI	3	No HAI	
4	No	0	No HAI	0	No HAI	55	No	3	No HAI	3	No HAI	
5	No	0	No HAI	0	No HAI	56	No	3	No HAI	3	No HAI	
6	No	2	No HAI	2	No HAI	57	No	3	No HAI	3	No HAI	
7	No	2	No HAI	2	No HAI	58	No	4	No HAI	4	No HAI	
8	No	2	No HAI	2	No HAI	59	No	4	No HAI	4	No HAI	
9	No	2	No HAI	2	No HAI	60	No	4	No HAI	4	No HAI	
10	No	2	No HAI	2	No HAI	61	No	4	No HAI	4	No HAI	
11	No	2	No HAI	2	No HAI	62	No	4	No HAI	4	No HAI	
12	No	2	No HAI	2	No HAI	63	No	4	No HAI	4	No HAI	
13	No	2	No HAI	2	No HAI	64	No	4	No HAI	4	No HAI	
14	No	2	No HAI	2	No HAI	65	No	4	No HAI	4	No HAI	
15	No	2	No HAI	2	No HAI	66	No	4	No HAI	4	No HAI	
16	No	2	No HAI	2	No HAI	67	No	4	No HAI	4	No HAI	
17	No	2	No HAI	2	No HAI	68	No	4	No HAI	4	No HAI	
18	No	2	No HAI	2	No HAI	69	No	4	No HAI	4	No HAI	
19	No	2	No HAI	2	No HAI	70	No	4	No HAI	4	No HAI	
20	No	2	No HAI	2	No HAI	71	No	4	No HAI	4	No HAI	
21	No	2	No HAI	2	No HAI	72	No	4	No HAI	4	No HAI	
22	No	2	No HAI	2	No HAI	73	No	4	No HAI	4	No HAI	
23	No	2	No HAI	2	No HAI	74	No	4	No HAI	4	No HAI	
24	No	2	No HAI	2	No HAI	75	No	4	No HAI	4	No HAI	
25	No	2	No HAI	2	No HAI	76	No	4	No HAI	4	No HAI	
26	No	2	No HAI	2	No HAI	77	No	4	No HAI	4	No HAI	
27	No	2	No HAI	2	No HAI	78	No	4	No HAI	4	No HAI	
28	No	2	No HAI	2	No HAI	79	No	4	No HAI	4	No HAI	
29	No	2	No HAI	2	No HAI	80	No	4	No HAI	4	No HAI	
30	No	2	No HAI	2	No HAI	81	No	4	No HAI	4	No HAI	
31	No	2	No HAI	2	No HAI	82	No	4	No HAI	4	No HAI	
32	No	2	No HAI	2	No HAI	83	No	4	No HAI	4	No HAI	
33	No	3	No HAI	3	No HAI	84	No	4	No HAI	4	No HAI	
34	No	3	No HAI	3	No HAI	85	No	4	No HAI	4	No HAI	
35	No	3	No HAI	3	No HAI	86	No	4	No HAI	4	No HAI	
36	No	3	No HAI	3	No HAI	87	No	4	No HAI	4	No HAI	
37	No	3	No HAI	3	No HAI	88	No	4	No HAI	4	No HAI	
38	No	3	No HAI	3	No HAI	89	No	4	No HAI	4	No HAI	
39	No	3	No HAI	3	No HAI	90	No	5	No HAI	5	No HAI	
40	No	3	No HAI	3	No HAI	91	No	5	No HAI	5	No HAI	
41	No	3	No HAI	3	No HAI	92	No	5	No HAI	5	No HAI	
42	No	3	No HAI	3	No HAI	93	No	5	No HAI	5	No HAI	
43	No	3	No HAI	3	No HAI	94	No	5	No HAI	5	No HAI	
44	No	3	No HAI	3	No HAI	95	No	5	No HAI	5	No HAI	
45	No	3	No HAI	3	No HAI	96	No	5	No HAI	5	No HAI	
46	No	3	No HAI	3	No HAI	97	No	5	No HAI	5	No HAI	
47	No	3	No HAI	3	No HAI	98	No	5	No HAI	5	No HAI	
48	No	3	No HAI	3	No HAI	99	No	5	No HAI	5	No HAI	
49	No	3	No HAI	3	No HAI	100	No	5	No HAI	5	No HAI	
50	No	3	No HAI	3	No HAI	101	No	5	No HAI	5	No HAI	
51	No	3	No HAI	3	No HAI	102	No	5	No HAI	5	No HAI	

.../...

Paciente	Observador 1			Observador 2			Paciente	Observador 1			Observador 2	
	HAI	Puntos	Clasificación	Puntos	Clasificación	HAI		Puntos	Clasificación	Puntos	Clasificación	
103	No	5	No HAI	5	No HAI	158	Sí	6	HAI probable	6	HAI probable	
104	No	5	No HAI	5	No HAI	159	Sí	6	HAI probable	6	HAI probable	
105	No	5	No HAI	5	No HAI	160	Sí	6	HAI probable	6	HAI probable	
106	No	5	No HAI	5	No HAI	161	Sí	6	HAI probable	6	HAI probable	
107	No	5	No HAI	5	No HAI	162	Sí	6	HAI probable	6	HAI probable	
108	No	5	No HAI	5	No HAI	163	Sí	6	HAI probable	6	HAI probable	
109	No	6	HAI probable	6	HAI probable	164	Sí	6	HAI probable	6	HAI probable	
110	No	6	HAI probable	7	HAI definitiva	165	Sí	6	HAI probable	6	HAI probable	
111	No	6	HAI probable	6	HAI probable	166	Sí	6	HAI probable	6	HAI probable	
112	No	6	HAI probable	6	HAI probable	167	Sí	6	HAI probable	6	HAI probable	
113	Sí	4	No HAI	4	No HAI	168	Sí	7	HAI definitiva	6	HAI probable	
114	Sí	4	No HAI	4	No HAI	169	Sí	7	HAI definitiva	6	HAI probable	
115	Sí	4	No HAI	4	No HAI	170	Sí	7	HAI definitiva	7	HAI definitiva	
116	Sí	4	No HAI	4	No HAI	171	Sí	7	HAI definitiva	7	HAI definitiva	
117	Sí	4	No HAI	4	No HAI	172	Sí	7	HAI definitiva	7	HAI definitiva	
118	Sí	4	No HAI	4	No HAI	173	Sí	7	HAI definitiva	7	HAI definitiva	
119	Sí	5	No HAI	5	No HAI	174	Sí	7	HAI definitiva	7	HAI definitiva	
120	Sí	5	No HAI	5	No HAI	175	Sí	7	HAI definitiva	7	HAI definitiva	
121	Sí	5	No HAI	5	No HAI	176	Sí	7	HAI definitiva	7	HAI definitiva	
122	Sí	5	No HAI	5	No HAI	177	Sí	7	HAI definitiva	7	HAI definitiva	
123	Sí	5	No HAI	5	No HAI	178	Sí	7	HAI definitiva	7	HAI definitiva	
124	Sí	5	No HAI	5	No HAI	179	Sí	7	HAI definitiva	7	HAI definitiva	
125	Sí	5	No HAI	5	No HAI	180	Sí	7	HAI definitiva	7	HAI definitiva	
126	Sí	5	No HAI	5	No HAI	181	Sí	8	HAI definitiva	8	HAI definitiva	
127	Sí	5	No HAI	5	No HAI	182	Sí	8	HAI definitiva	8	HAI definitiva	
128	Sí	5	No HAI	5	No HAI	183	Sí	8	HAI definitiva	8	HAI definitiva	
129	Sí	5	No HAI	5	No HAI	184	Sí	8	HAI definitiva	7	HAI definitiva	
130	Sí	5	No HAI	5	No HAI	185	Sí	8	HAI definitiva	8	HAI definitiva	
131	Sí	5	No HAI	5	No HAI	186	Sí	8	HAI definitiva	8	HAI definitiva	
132	Sí	5	No HAI	5	No HAI	187	Sí	8	HAI definitiva	8	HAI definitiva	
133	Sí	5	No HAI	5	No HAI	188	Sí	8	HAI definitiva	8	HAI definitiva	
134	Sí	5	No HAI	5	No HAI	189	Sí	8	HAI definitiva	7	HAI definitiva	
135	Sí	5	No HAI	5	No HAI	190	Sí	8	HAI definitiva	8	HAI definitiva	
136	Sí	5	No HAI	5	No HAI	191	Sí	8	HAI definitiva	8	HAI definitiva	
137	Sí	5	No HAI	5	No HAI	192	Sí	8	HAI definitiva	7	HAI definitiva	
138	Sí	5	No HAI	5	No HAI	193	Sí	7	HAI definitiva	8	HAI definitiva	
139	Sí	5	No HAI	5	No HAI	194	Sí	7	HAI definitiva	7	HAI definitiva	
140	Sí	5	No HAI	5	No HAI	195	Sí	7	HAI definitiva	8	HAI definitiva	
141	Sí	6	HAI probable	7	HAI probable	196	Sí	7	HAI definitiva	7	HAI definitiva	
142	Sí	6	HAI probable	6	HAI probable	197	Sí	7	HAI definitiva	7	HAI definitiva	
143	Sí	6	HAI probable	6	HAI probable	198	Sí	7	HAI definitiva	7	HAI definitiva	
144	Sí	6	HAI probable	6	HAI probable	199	Sí	7	HAI definitiva	7	HAI definitiva	
145	Sí	6	HAI probable	6	HAI probable	200	Sí	7	HAI definitiva	7	HAI definitiva	
146	Sí	6	HAI probable	6	HAI probable	201	Sí	7	HAI definitiva	7	HAI definitiva	
147	Sí	6	HAI probable	6	HAI probable	202	Sí	7	HAI definitiva	7	HAI definitiva	
148	Sí	6	HAI probable	6	HAI probable	203	Sí	7	HAI definitiva	7	HAI definitiva	
149	Sí	6	HAI probable	6	HAI probable	204	Sí	7	HAI definitiva	7	HAI definitiva	
150	Sí	6	HAI probable	6	HAI probable	205	Sí	7	HAI definitiva	7	HAI definitiva	
151	Sí	6	HAI probable	6	HAI probable	206	Sí	7	HAI definitiva	7	HAI definitiva	
152	Sí	6	HAI probable	6	HAI probable	207	Sí	8	HAI definitiva	7	HAI definitiva	
153	Sí	6	HAI probable	6	HAI probable	208	Sí	8	HAI definitiva	8	HAI definitiva	
154	Sí	6	HAI probable	6	HAI probable	209	Sí	8	HAI definitiva	8	HAI definitiva	
155	Sí	6	HAI probable	6	HAI probable	210	Sí	8	HAI definitiva	8	HAI definitiva	
156	Sí	6	HAI probable	6	HAI probable	211	Sí	8	HAI definitiva	8	HAI definitiva	
157	Sí	6	HAI probable	6	HAI probable	212	Sí	8	HAI definitiva	8	HAI definitiva	

El índice *kappa* que resume la concordancia entre las clasificaciones binarias (HAI sí o no) de los dos observadores ha sido de 1, resultado que demuestra una reproductibilidad perfecta para el uso general de los criterios simplificados.

Al 98,1% de los pacientes se le asignó el mismo diagnóstico categorizado por parte de los dos observadores: no HAI, HAI probable o HAI definitiva. El empleo de los criterios de 2008 como un instrumento de clasificación con tres categorías arrojó un valor *kappa* sin ponderar de 0,964, con un IC95% de 0,929 a 0,998, que refleja una concordancia excelente según la referencia de Landis y la de Fleiss [262,263].

Tabla 29: Tabla de contingencia de las clasificaciones de los dos observadores empleando los criterios simplificados de 2008.

		Observador 2			
		No HAI	HAI probable	HAI definitiva	Total
Observador 1	No HAI	136	0	0	136 (64,2%)
	HAI probable	0	29	2	31 (14,6%)
	HAI definitiva	0	2	43	45 (21,2%)
	Total	136 (64,2%)	31 (14,6%)	45 (21,2%)	212 (100%)

HAI: Hepatitis autoinmune.

Respecto al cálculo ponderado, la concordancia en base al estadístico *kappa* también fue muy buena. La aproximación cuadrática obtuvo el valor máximo, con un límite superior del intervalo de confianza que tiende a la unidad.

Tabla 30: Estudio de la concordancia mediante el estadístico *kappa* ponderado para la clasificación en no hepatitis autoinmune, enfermedad probable o enfermedad definitiva según los criterios simplificados de la IAIHG.

Ponderación	Porcentaje de acuerdo	<i>Kappa</i>	Intervalo de confianza al 95%	Grado de concordancia
Lineal	99,1%	0,976	0,953 a 0,999	Excelente
Cuadrática	99,5%	0,986	0,972 a 1,000	Excelente

El porcentaje de desacuerdos estuvo equilibrado, es decir, no sesgados por la tendencia predominante de ningún observador. La discordancia se dio en 4 casos, entre las categorías de HAI probable y definitiva. Dos pacientes dados como

definitivos por el observador 1 se catalogaron como probables por el observador 2 y los otros dos, en sentido contrario. Tres de estos casos fueron correctamente diagnosticados de HAI por los criterios simplificados y el restante fue un falso positivo.

El cálculo del CCI de consistencia y de acuerdo absoluto obtuvo el mismo valor numérico, de 0,994 (IC95% 0,992 a 0,995). Además de traducir una fiabilidad muy buena de los criterios, descarta que haya podido existir desviaciones de tipo constante entre las interpretaciones de los dos observadores.

Por último, se calculó la variable que representa la diferencia entre los puntos absolutos otorgados por los dos investigadores independientes. La distribución de los valores no siguió un patrón normal atendiendo a la prueba de significación del estadístico de Shapiro-Wilk ( $<0,0001$ ). Como se puede deducir de la tabla de resultados, la mayoría fueron iguales en cada paciente, de modo que, por razones de tamaño muestral, las cuatro discordancias se comportan como valores extremos.

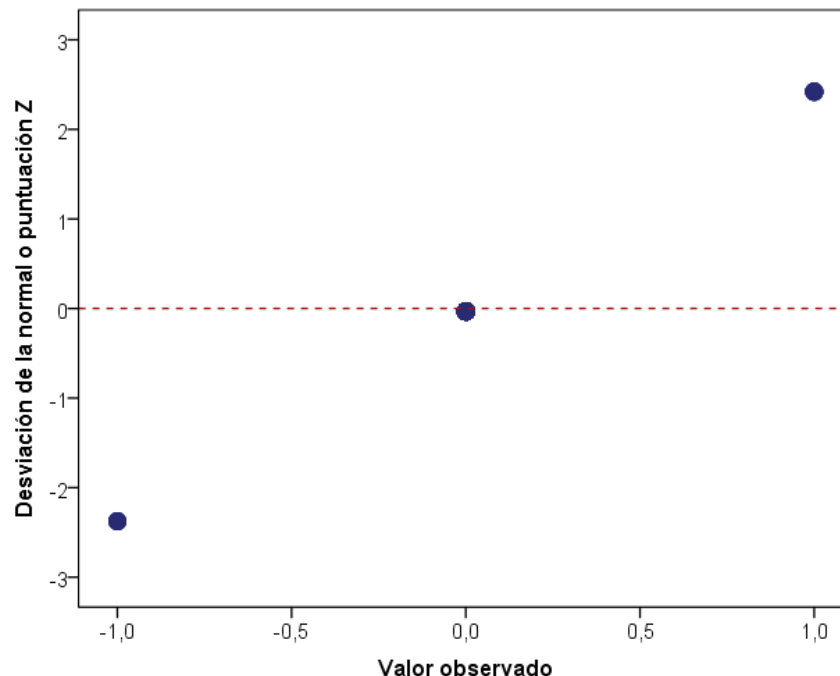


Figura 45: Gráfico Q-Q normal sin tendencia para la diferencia entre las puntuaciones obtenidas por los criterios de 2008 aplicados por dos observadores distintos.

Por este motivo, no se consideró informativo representar la media de las diferencias entre observaciones (ni sus intervalos de confianza) en el diagrama de Bland-Altman, cuyo trazado con datos crudos se ofrece a continuación.

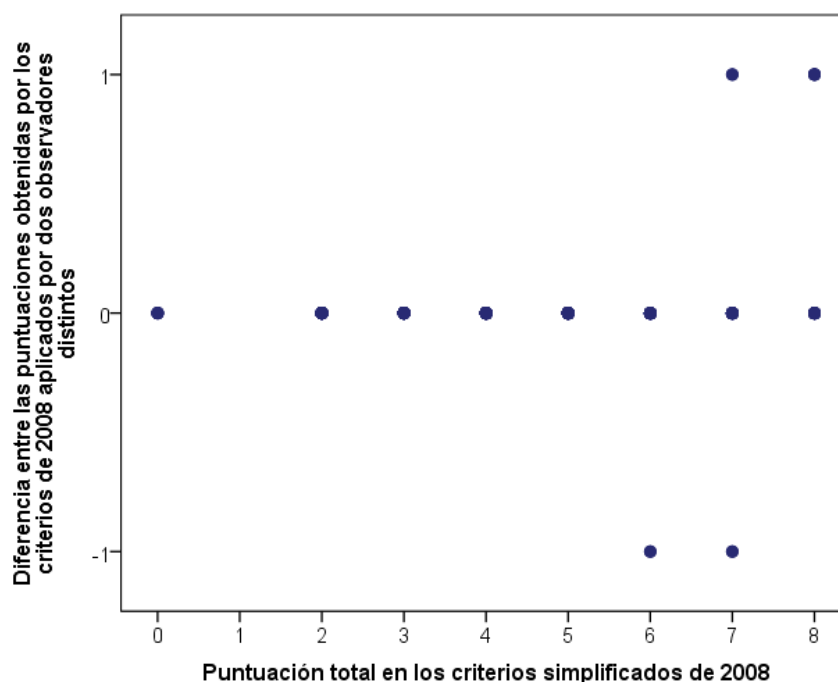


Figura 46: Diagrama de Bland-Altman para las diferencias entre las puntuaciones de los criterios simplificados aplicados por dos observadores independientes.

La dispersión de los puntos no va más allá de 1 unidad respecto a la diferencia nula, que es la única observada en las puntuaciones correspondientes a la categoría diagnóstica “no HAI”. En el extremo de las puntuaciones a favor del diagnóstico de HAI, la dispersión ha resultado ser aproximadamente simétrica.

Atendiendo de forma particular a los criterios del sistema que han sido fuente de discrepancias, el único punto que ha originado conflictos es el de la anatomía patológica. En algunos pacientes, uno de los observadores ha puntuado +1 punto (histología compatible) y el otro +2 (histología típica). En ningún caso se dio la circunstancia de que uno de ellos puntuara 0.



## **9.4. Capítulo 4: Modelización y validación de un nuevo sistema diagnóstico por puntos a partir de los criterios ESPGHAN/NASPGHAN 2009**

---

### **9.4.1. Metodología específica**

Para la ejecución de este bloque se necesita trabajar con la misma muestra que la empleada en el primer capítulo. Igualmente, debe estar compuesta por un grupo de pacientes pediátricos con patología hepática de inicio reciente, entre cuyos diagnósticos finales se encuentre una proporción de casos de HAI equiparable a la que se pueda encontrar en la población diana, así como sus diagnósticos diferenciales. Por este motivo, a efectos de diseño del estudio, esta parte no ha obligado a incorporar ninguna característica metodológica relevante concreta. La recopilación de los casos, como ya es conocido, ha sido consecutiva y ambispectiva. Se pueden consultar más detalles en el apartado sobre el diseño general.

La clasificación diagnóstica de referencia se ha realizado en función del análisis de casos discrepantes basado en los criterios clásicos revisados de la IAIHG y el diagnóstico recogido en la historia clínica más los criterios de robustez. A modo de recordatorio, volvemos a señalar que se consideró como criterio necesario para establecer el diagnóstico de HAI la presencia simultánea de:

1) Tener, como mínimo, una anatomía patológica compatible con HAI según la definición del score simplificado de 2008, es decir, una hepatitis crónica con infiltración linfocítica [19].

2) Constatarse respuesta al tratamiento farmacológico definida como mejoría de los síntomas, si los hay, y disminución de las transaminasas. Es decir, considerar como criterio necesario el ítem de la respuesta terapéutica.

Esto, sumado al análisis de casos discrepantes (con selección de casos en conflicto a través del resultado de los criterios simplificados y resolución del mismo



en base a una revisión consensuada de las historias clínicas y los dos criterios de robustez anteriores) se empleó como medida para reducir posibles falsos negativos y positivos al no poder considerar estrictamente a los criterios de 1999 como un *gold standard* [236].

Respecto a la extracción de los datos, además del diagnóstico final (HAI sí o no), se recogieron las variables correspondientes a los criterios pediátricos propuestos por la ESPGHAN y la NASPGHAN en 2009 [93]. Se trata de siete variables que se comportan todas como categóricas binarias, lo que presumiblemente facilitaría su modelización y transformación en un sistema diagnóstico por puntos.

**Tabla 31: Criterios diagnósticos pediátricos para la hepatitis autoinmune propuestos por Mieli-Vegani et al en 2009 (*J Pediatr Gastroenterol Nutr.* 2009;49:159).**

Características diagnósticas específicas de la hepatitis autoinmune en niños	Posibilidades que admite
Hipertransaminasemia (límites sin definir) Autoanticuerpos (ANA o anti-Sm $\geq 1:20$ , anti-LKM1 $\geq 1:10$ , anti-LC1 o anti-SLA) Hipergammaglobulinemia (límites sin definir) Biopsia hepática con hepatitis de interfase o colapso multilobular Signos analíticos de hepatitis vírica Evidencia de enfermedad de Wilson Colangiografía normal (endoscópica o por resonancia magnética)	Sí o no

ANA: Anticuerpo antinuclear. Anti-Sm: Anticuerpo anti-músculo liso. Anti-LKM1: Anticuerpo anti-microsomal de hígado/riñón de tipo 1. Anti-LC1: Anticuerpo anti-citosol hepático de tipo 1. Anti-SLA: Anticuerpos anti-antígeno soluble hepático.

Dado que en la fuente original de la propuesta no está explícitamente descrito qué se debe considerar como colangiografía normal, se definió previamente a la recogida de datos. El motivo por que se incluye este criterio en la propuesta es por la posibilidad de la existencia de un componente de colestasis que oriente hacia una CEAI o para descartar una CEP. Por este motivo, y de forma consensuada, se señaló como colangiografía anormal, cualquier signo de estenosis o dilataciones saculares en el árbol biliar, tanto en su parte intrahepática como en la extrahepática [264].

Los valores de normalidad de la IgG dependen de la edad, así que para responder al criterio de la presencia de hipergammaglobulinemia se consultó la referencia del laboratorio de cada hospital, basadas en datos propios y en el trabajo clásico de Stoop *et al.* [265].

Se consideró descartada la enfermedad de Wilson en base a unos niveles de ceruloplasmina normales y, si estuvieran disponibles, en función de la normalidad de los resultados de la cuantificación de la excreción urinaria de cobre (con o sin administración previa de penicilamina), la medición de cobre en tejido hepático y el estudio genético.

Respecto a los autoanticuerpos, se tuvieron en cuenta aquellos contemplados en los criterios de 2009 con los puntos de corte dados, inferiores a los de los criterios clásicos y los simplificados de la IAIHG. Por este motivo no se consideró positiva esta variable en presencia de pANCA aisladamente pero sí para títulos de ANA y anti-SM iguales o superiores a 1:20, o de anti-LKM1 iguales o superiores a 1:10. Cualquier valor de anti-LC1 o anti-SLA también fue considerado positivo.

Desde un punto de vista teórico, es posible que el ítem de los anticuerpos funcione mejor comportándose como una variable categórica ordinal, que posiblemente otorgaría más seguridad diagnóstica cuanto más elevado sea el título de anticuerpos. Así, se exploró también cambiar este criterio por un equivalente similar al de los criterios simplificados de 2008. En realidad, se utilizó una versión modificada que tuviera en cuenta la posibilidad de puntuar positivo con títulos bajos de anticuerpos y la presencia de anticuerpos infrecuentes (anti-SLA/LP, LC1, ASGPR, pANCA, anti-actina). De este modo se tuvieron en cuenta los fundamentos clínicos de los criterios pediátricos de 2009. Aunque finalmente el mejor modelo elegido contemplara esta variable con varias categorías, igualmente se trabajaría como si todas las del sistema fueran binarias dado que cada categoría dentro de los autoanticuerpos (título bajo, intermedio o alto) también admite un sí o no como respuesta.

#### **9.4.1.1. Selección de las submuestras de elaboración y validación**

Validar un modelo diagnóstico con la misma población que se ha utilizado para generar el sistema imprime un error sistemático que tiende a sobrevalorar su capacidad discriminante. Por este motivo, evaluar si el modelo predice satisfactoriamente el diagnóstico real se debe de hacer sobre datos procedentes de otras muestras obtenidas de la población de interés [219].

El mejor procedimiento para valorar la exactitud de un modelo es replicar el estudio y comprobar los indicadores de validez en otros datos, pero es costoso y tiene el inconveniente de retardar el establecimiento final del modelo. Por consiguiente, se dispuso emplear muestras partidas (*split-sample*) para los fines de elaborar el modelo y validar su capacidad predictiva diagnóstica. Como se puede consultar en el apartado correspondiente, esta decisión tuvo implicaciones sobre la estimación del tamaño muestral necesario. Se procuró sobredimensionar la muestra del estudio para poder dividirla en un grupo mínimo adecuado para construir el modelo (*derivation set* o *training set*) y otro grupo para validarlo (*validation group*). Se trata de un procedimiento de validación a través de los sujetos de la propia muestra (*cross-validation*) [266].

Siguiendo las indicaciones habituales para cuando no es posible reclutar un número de pacientes abundante, se determinó emplear 2/3 para estimar la ecuación predictiva y 1/3 para el grupo de validación [266].

Con el paquete estadístico SPSS® de IBM (versión 21.0) se generó la variable binaria “set”, que tomó valores 0 o 1, de forma pseudoaleatoria, con probabilidad 1/3 y 2/3, respectivamente. Este *software* genera números binarios al azar en base a una distribución de Bernouilli con un parámetro  $p$ , que arroja 0 con una probabilidad  $p$  y 1 con una probabilidad  $1 - p$ . En lógica, la distribución de Bernouilli resultante tiene una media  $1 - p$  y una varianza  $p(1 - p)$ . Los casos con un valor 1 en la variable “set” constituyeron la muestra de derivación y el resto, la muestra de validación.

## 9.4.2. Análisis estadístico

### 9.4.2.1. *Desarrollo de un modelo predictivo de regresión logística para el diagnóstico de la HAI a partir de los criterios de 2009*

El primer paso para el desarrollo de un modelo de regresión logística es la elección de las variables potenciales a incluir. En nuestro caso, vienen dadas por la propuesta de criterios diagnósticos pediátricos para la HAI de Mieli-Vergani en 2009, en representación de la ESPGHAN y la NASPGHAN [93]. En la introducción se justifica que los ítems que incluye son variables relevantes en la práctica clínica y que existe una buena justificación teórica para poder funcionar como elementos independientes en una ecuación predictiva estable. Todas ellas se comportan como variables predictoras binaras y, por lo tanto, se definieron con los valores 0 y 1 para representar su ausencia o presencia en cada caso. En este contexto, consideramos que el riesgo de seleccionar un modelo poco generalizable es mínimo dado que la naturaleza de las variables potenciales es básicamente clínica.

Además, en nuestra muestra no se da ninguna de las condiciones problemáticas para la precisión de las estimaciones de modelos de regresión logística señaladas por Hosmer y Lemeshow [267]:

- 1) El número de casos no es relativamente pequeño respecto al número de variables. De hecho, la selección del tamaño muestral se ha realizado teniendo en cuenta que el modelo máximo incluye 7 variables.
- 2) La proporción de respuesta (diagnóstico afirmativo de HAI) no es próxima a los extremos del intervalo (0 o 1). En efecto, la proporción de pacientes HAI=1 es 0,472.

#### 9.4.2.1.1. Manejo de los valores perdidos

Por protocolo, tal como se especifica en el capítulo 1, cualquier valor perdido que afecte a la seguridad del diagnóstico definitivo supondrá que el caso se elimine de la matriz de datos. Por lo que respecta a los valores perdidos en las variables

explicativas relacionadas con los criterios de 2009, es previsible que su distribución no sea al azar. El criterio de la prueba de imagen de la vía biliar es, *a priori*, el más susceptible de presentar valores perdidos. Posiblemente los casos más claros de hepatopatía de etiología no autoinmune no dispongan de información a este respecto por considerarse innecesaria. Del mismo modo, cabe esperar que el estrato de los diagnósticos de HAI realizados en la primera parte del periodo de estudio tenga más probabilidad de datos desconocidos en esta variable.

Se empleó el módulo *Missing Value Analysis* de SPSS Statistics® versión 21.0 para describir los valores desconocidos y los alejados de la matriz de datos.

Referente al manejo propiamente dicho, se descartó la estrategia de excluir los sujetos con algún valor desconocido por el riesgo de reducir la muestra útil hasta el punto de disminuir la precisión de las estimaciones del modelo (y de la potencia estadística) por debajo de lo recomendable.

Para evitar este problema, se decidió sustituir los valores desconocidos por el correspondiente valor estimado a través de un método de imputación múltiple. Este procedimiento fue propuesto por D. B. Rubin en los años 80 y consiste en reemplazar cada valor perdido por un vector de  $M \geq 2$  valores estimados de forma adecuada [268]. Para nuestro caso, se utilizó un modelo de regresión logística que hace uso del conjunto del resto de variables independientes del modelo máximo. Así, cada vez que se sustituye el valor desconocido por uno de estos valores predichos se obtiene un conjunto completo de datos. Para esta situación concreta se empleó  $M = 5$  valores para cada valor perdido. Posteriormente, en cada uno de estos conjuntos de datos, se realiza el análisis estadístico. El paso final refleja el promedio de los parámetros estudiados en cada uno de los conjuntos completos de información. Se considera que este tratamiento de los valores desconocidos, alternativo a la exclusión de los casos afectados, es el más idóneo para la situación de datos perdidos de forma no aleatoria [219].

La imputación múltiple se llevó a cabo con el procedimiento *Multiple Imputation* (MI) de SPSS Statistics® versión 21.0. Para completar la matriz de datos

con los valores imputados, cada ejecución de la sintaxis del paquete estadístico emplea una semilla de aleatorización distinta. Por ello, los datos finales suelen ser diferentes en cada implementación del proceso. Para resolver este capítulo de la tesis se consideraron los valores arrojados en la primera ejecución. Se interpretó el valor crudo de la eficiencia relativa de la imputación para darla por buena, considerando como adecuados valores cercanos a 1. Este dato lo ofrece el procedimiento MI. En caso de obtener una eficiencia relativa pobre, se repetiría la imputación múltiple con un número mayor de simulaciones ( $M$ ).

#### 9.4.2.1.2. Selección del mejor modelo a partir de todas las ecuaciones posibles

Se reconoce que la aproximación idónea para la selección de un modelo de regresión es aquella que pasa por estudiar todas las posibles combinaciones de variables explicativas. Ciertamente, ello implica construir todos los submodelos potenciales combinando los términos del modelo máximo, y valorar para cada uno un indicador de ajuste. El modelo máximo, en nuestro caso particular, es el que incluye las 7 variables propuestas en los criterios de la ESPGHAN/NASPGHAN 2009. Por su parte, el indicador de ajuste elegido para la selección del mejor modelo fue el criterio de Akaike corregido, que es tanto mejor cuanto mayor sea su valor [219].

Posibles modelos más parsimoniosos (con menos variables) se estudiaron a través del  $C_p$  de Mallows, que Hosmer y Lemeshow han adoptado para la modelización a través de regresión logística [219,267].

El criterio  $C_p$  de Mallows, habitualmente usado en regresión lineal también se puede calcular de forma aproximada en un modelo de regresión logística con la siguiente fórmula:

$$C_p = S_p - S_q + 2q - p + 1$$

Donde:

$p$  es el número de predictores del modelo máximo (incluyendo las variables ficticias que se puedan generar y los términos de interacción, aunque en nuestro ejercicio no se incluyeron ninguno de los dos).

$q$  es el número de predictores del modelo evaluado (también incluyendo variables ficticias y términos de interacción).

$S_p$  es la tasa de discriminación (*Score tests*) del modelo máximo.

$S_q$  es la tasa de discriminación (*Score tests*) del modelo evaluado.

Bajo el supuesto de que un modelo restringido en variables independientes ajusta igual que el modelo máximo, su valor de  $C_p$  es  $q + 1$ . Por consiguiente, en caso de que un modelo con un conjunto inferior de variables obtuviera un  $C_p$  muy cercano a  $q + 1 = 8$ , se seleccionaría también bajo el principio de conseguir una ecuación predictiva lo más sencilla posible. Si varios modelos fueran de interés, el desempate se llevaría a cabo en función de los resultados de los diagnósticos del modelo (detallado en el apartado siguiente).

Además del criterio de información de Akaike y el  $C_p$ , también se calcularon el área bajo la curva COR y el valor de  $-2LL$ , que es el logaritmo neperiano, multiplicada por  $-2$ , del valor de la función de verosimilitud del modelo. Mediante esta transformación, se da un valor positivo que permite comparar modelos de forma más intuitiva. Aunque la selección del mejor modelo estuvo basada en el criterio de Akaike y el  $C_p$ , la obtención de los otros indicadores se llevó a cabo a título informativo.

Estas operaciones se llevaron a cabo con el *script* AllSetsReg de SPSS Statistics® versión 21.0 [269]. Se trata de unas instrucciones prediseñadas por el *Laboratori d'Estadística Aplicada* de la *Universitat Autònoma de Barcelona*, que construye en primer lugar todos los modelos con un término, luego todos los modelos con dos términos, y así sucesivamente hasta el modelo máximo que contiene todos los términos. Para cada ecuación estima los índices  $C_p$  de Mallows, área bajo la curva COR, sensibilidad, especificidad,  $-2LL$  y el valor p del estadístico de ajuste de Hosmer-Lemeshow [219]. El algoritmo implementado soluciona los problemas que presentan los métodos automáticos por pasos ya que permite incorporar el principio jerárquico (cuando el modelo máximo incluye términos de interacción) y los conocimientos previos (al permitir fijar variables que deben estar

presentes en todos los subconjuntos analizados por razones teóricas). En cualquier caso, para resolver este capítulo no se necesitó ninguna de estas características. Para el cálculo del criterio de información de Akaike se empleó el comando *estat ic* en el programa Stata (versión 14.0, *Stata Corporation*, Texas, EEUU).

#### **9.4.2.1.3. Diagnósticos del mejor modelo seleccionado**

Sobre el/los modelo/s seleccionado/s en el paso previo, se comprobaron los supuestos que rigen la correcta aplicación de las técnicas de regresión logística con finalidad tanto explicativa como predictiva.

##### *9.4.2.1.3.1. Detección de valores alejados que afecten a las estimaciones*

Se revisaron aquellas observaciones (pacientes) que afectaban de forma individual a la magnitud de los parámetros estimados o a sus errores estándar. Para ello se calculó la distancia de Cook, que es un dato que indica cuánto cambiaría en promedio la estimación de los parámetros si un determinado caso fuera excluido del análisis. No existe una forma objetivamente óptima para establecer matemáticamente un punto de corte para la distancia de Cook. Por ello, se decidió emplear el método visual, que consiste en graficar la distancia de Cook contra la probabilidad predicha en una nube de puntos y seleccionar el valor que mejor separe a simple vista los valores elevados [219]. Los cálculos y el trazado del gráfico se hicieron con SPSS Statistics<sup>®</sup> versión 21.0. En caso de encontrar valores muy alejados, se descartarían de la matriz de datos y se repetiría el ejercicio de modelización.

##### *9.4.2.1.3.2. Comprobación del supuesto de equidispersión*

Cuando en un modelo de regresión la varianza de la distribución de probabilidad de las respuestas queda determinada por la media, como en el caso de la regresión logística, se debe de cumplir el supuesto de equidispersión. Para la regresión logística este supuesto exige que la varianza de respuesta observada en los datos sea igual a la varianza esperada bajo el supuesto de distribución binomial.



Cuando la equidispersión no se cumple es debido a que se presenta el fenómeno de la infradispersión si la varianza empírica es menor que la teórica, o el de sobredispersión si es mayor. La principal consecuencia de la sobredispersión, que es más frecuente que la infradispersión, es que subestima los errores estándar, lo que incrementa el error de tipo I [219].

En modelos de regresión logística con todos los predictores categóricos (como es el caso de la ecuación que se prevé obtener en este capítulo), y por tanto con varias observaciones en cada patrón de valores predictivos, la variación total  $-2LL$  sigue una distribución  $\chi^2$  con los grados de libertad residuales ( $df_{Res}$ ), que se obtienen restando al tamaño muestral el número de parámetros estudiados (7 en el modelo máximo de nuestro caso). Se comprobó la equidispersión dividiendo el valor  $-2LL$  entre los  $df_{Res}$ , para obtener una variación media residual. Si el resultado es estadísticamente superior a 1 indica sobredispersión y si es inferior a la unidad, infradispersión. Para obtener el grado de significación de esta comparación debe situarse el valor  $-2LL$  en su distribución de probabilidad  $\chi^2$  de referencia. El valor  $-2LL$  se obtuvo con SPSS Statistics® versión 21.0, el cálculo de la varianza media residual y el valor p de significación se llevó a cabo manualmente con ayuda de una calculadora científica. Se contempló el umbral de significación en  $p = 0,05$ .

La existencia de sobredispersión puede deberse a diferentes motivos, como un error de especificación del modelo por exclusión de variables relevantes, la existencia de valores influyentes, un muestreo por conglomerados que da lugar a elevadas correlaciones entre sujetos de un mismo grupo, o una distribución de probabilidad incorrecta de la variable respuesta. Si la sobredispersión fuera moderada, con valores de la varianza media residual significativa y discretamente por encima de 1, se corregirían los errores estándar del parámetro B de cada variable predictiva multiplicándolo por la varianza media residual. El error estándar así corregido construye los intervalos de confianza más amplios, dificulta la obtención de resultados significativos y, de este modo, reduce los errores de tipo I.

Si la sobredispersión fuera elevada (y si fuera posible) se probaría a añadir variables que hayan podido ser omitidas en la especificación del primer modelo elegido.

#### 9.4.2.1.3.3. *Bondad de ajuste*

En un modelo de regresión logística múltiple, estimado por mínimos cuadrados, la bondad de ajuste viene dada por el coeficiente  $R^2$ , que representa la proporción de la variabilidad de la variable respuesta (la probabilidad de un diagnóstico de HAI en nuestro modelo con intención predictiva) explicada verdaderamente por la ecuación de regresión. Se obtuvo  $R^2$  a través de la corrección de Nagelkerke de la fórmula original de Cox y Snell, con SPSS Statistics® versión 21.0 [270].

#### 9.4.2.1.3.4. *Significación global del modelo*

Puesto que el modelo se estimó por métodos de máxima verosimilitud, su significación global (es decir, la significación del conjunto de variables predictoras incluidas) se efectuó con la prueba de la razón de verosimilitud. Consiste en comparar el logaritmo de la verosimilitud del modelo estimado con las  $p$  variables predictoras  $L(M)$  con el modelo que solo contiene la constante  $L(0)$ , a través del cociente dado por la siguiente fórmula, que en muestras grandes sigue una ley de  $\chi^2$  con  $p$  grados de libertad [219]:

$$\chi_M^2 = -2 \ln \frac{L(0)}{L(M)}$$

Por convención, se interpretó como significativo un valor  $p$  inferior a 0,05. El dato se obtuvo con SPSS Statistics® versión 21.0, a través del procedimiento LOGISTIC REGRESSION, dentro del recuadro *Omnibus Tests of Model Coefficients* de la ventana de resultados.

#### 9.4.2.1.3.5. *Calibración del modelo*

La calibración del modelo es un aspecto del ajuste que consiste en valorar la concordancia entre las probabilidades observadas en la muestra ( $p_i$ ) y las predichas por el modelo ( $\pi_i$ ). Esta valoración se puede realizar con el estadístico  $\chi^2$  de bondad

de ajuste. Si el modelo es adecuado, las proporciones esperadas no presentarán grandes diferencias con las observadas en la muestra.

Para modelos con varias variables predictoras, si la muestra está formada por  $n$  sujetos diferentes, tendremos  $J \leq n$  estimaciones con  $\pi_i$  diferentes. En efecto, si el modelo contiene  $p$  variables predictoras ( $x = (x_1; x_2; \dots; x_p)$ ), cada combinación  $x_j$  diferente de valores de los predictores dará una estimación diferente de  $\pi_j$ , de manera que  $J$  tomará el valor máximo  $n$  si ninguna combinación se repite.

En este trabajo, es esperable que el número  $J$  de combinaciones sea muy grande debido al tamaño muestral. Por ello, siguiendo la propuesta de Hosmer y Lemeshow, se ordenaron los  $n$  sujetos según las predicciones  $\pi_i$  y se dividieron en  $g = 10$  grupos de aproximadamente igual tamaño (deciles de riesgo). Esta estrategia es especialmente adecuada cuando muchas de las probabilidades estimadas  $\pi_i$  son pequeñas [267]. Se compararon las distribuciones predichas ( $e$ ) y observadas ( $o$ ) en estos  $g = 10$  grupos con el estadístico  $\chi^2$  de bondad de ajuste:

$$\chi^2 = \sum_{j=1}^g \sum_{i=1}^2 \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

Se calculó el estadístico con la ley de  $\chi^2$  con  $g-2 = 8$  grados de libertad. Valores de significación  $p$  cercanos a 1 son indicativos de un buen ajuste. Todos estos cálculos se realizaron también con SPSS Statistics® versión 21.0, empleando el procedimiento LOGISTIC REGRESSION.

#### 9.4.2.1.4. Construcción de una tabla de índices pronósticos

Para valorar la probabilidad diagnóstica de HAI en un paciente individual a través de las variables incluidas en el mejor modelo seleccionado, es útil el empleo de una tabla de índices pronósticos. Esta tabla debe recoger los riesgos relativos para cada uno de los patrones correspondientes a todas las posibles combinaciones de las variables predictoras. El modelo máximo con las 7 variables binarias de los criterios ESPGHAN/NASPGHAN 2009, permite  $2^7 = 128$  posibles patrones distintos ( $o$   $2^9 = 512$  patrones con un ítem ordinal con tres categorías para los anticuerpos). El

paciente tipo que se utilizó como referencial para el cálculo de los riesgos relativos fue aquel con hipertransaminasemia, sin marcadores virales ni enfermedad de Wilson, colangiografía normal y resto de variables negativas.

Mediante el entorno de sintaxis de SPSS Statistics<sup>®</sup> versión 21.0, se realizaron predicciones añadiendo sujetos ficticios con un determinado patrón de variables predictoras y sin valor en la variable respuesta que se generó durante la modelización. Con un número suficiente de bucles anidados en la instrucción INPUT PROGRAM del *software* se pudo automatizar este proceso. La tabla se dibujó con la instrucción TABLES y se dio formato con capas para facilitar su lectura.

#### **9.4.2.1.5. Punto de corte óptimo del mejor modelo seleccionado**

En primer lugar, se generó el valor respuesta para cada caso de la muestra de derivación a partir de la ecuación seleccionada. Posteriormente, se estimó el punto de corte óptimo de estos valores a partir de la fórmula de Zweig y Campbell empleada en el capítulo 1 [246]. Dado que uno de los términos que se consideran por este método es la prevalencia de la enfermedad, se empleó la hallada con la muestra entera (elaboración más validación) para minimizar desviaciones respecto a la prevalencia global de la muestra por reducción del tamaño muestral en la selección de las submuestras. La razón de costes empleada para el cálculo fue de 1, para penalizar por igual los errores de tipo I y II. Para este punto de corte se calculó su sensibilidad y especificidad y también se generó una curva COR y se calculó su área bajo la curva. Estos análisis se efectuaron con SPSS Statistics<sup>®</sup> con ayuda de la sintaxis de la macro !ROC [247].

#### **9.4.2.2. Transformación del modelo en un sistema de puntos**

Una explicación detallada del desarrollo de criterios de clasificación por puntos a través de modelos de regresión con intención predictiva se puede encontrar en el apartado 4.1.3 de la introducción.

Dado que todos los criterios ESPGHAN/NASPGHAN 2009 se comportan como variables binarias, la transformación en un sistema de puntos es sencilla. A la ausencia del criterio se le asignaron 0 puntos; y a una respuesta positiva,  $\beta_1/B$  puntos, siendo  $\beta_1$  el coeficiente de regresión de esa variable en el modelo, y  $B$  el  $\beta_1$  de la variable con menos peso en la ecuación. De este modo, se dio 1 punto a ésta y valores superiores a las demás. La posible existencia de variables con puntuación negativa se corregiría redefiniendo el criterio para hacerlo opuesto al de los criterios de 2009 de Mieli-Vergani [93].

Se obtuvo la puntuación total de cada caso y se calculó el punto de corte siguiendo la misma metodología descrita en el punto anterior. Todo ello se llevó a cabo solamente en la muestra de derivación.

#### **9.4.2.3. Validación del nuevo sistema de puntos basado en los criterios de 2009**

Por último, sobre la submuestra de validación (los casos restantes a los empleados en los puntos anteriores), se efectuaron los análisis destinados a obtener los indicadores de validez del nuevo sistema de puntos basado en los criterios ESPGHAN/NASPGHAN de 2009 [93].

Igual que en el capítulo 1, los indicadores de validez se estimaron según las fórmulas descritas en el apartado 4.2.2. Con la prevalencia de la muestra completa, la sensibilidad y la especificidad, se aplicaron las fórmulas de Buderer con el fin de obtener la precisión absoluta resultante para la estimación de los parámetros de validez con el tamaño de la submuestra.

Los resultados binarios se expresaron como porcentajes con su intervalo de confianza al 95% por el método de Wilson [244]. Las variables continuas se resumieron como medianas y RIC.

Finalmente, se estimó la pérdida de predicción del nuevo sistema diagnóstico entre la muestra de derivación y la de validación. A tal fin se calculó la diferencia del área bajo la curva COR del rendimiento de los criterios entre las dos submuestras.

### 9.4.3. Resultados

Solo existieron valores perdidos en la variable sobre los resultados de la colangiografía. El motivo de la ausencia de información fue la no realización de la prueba en pacientes sin colestasis clínica ni analítica (HAI y no HAI). Recordemos, como se expuso en los resultados del capítulo 1, que solamente a un 45,3% de todos los pacientes de la muestra se les estudió específicamente la vía biliar. Considerando solo los pacientes con diagnóstico final de HAI, la proporción de enfermos con colangiografía fue de 65,0%.

Todos los casos con valor desconocido en esta variable se les asignó un valor de "0" (colangiografía normal). La eficiencia relativa del procedimiento con  $M = 5$  simulaciones fue de 0,843, por lo que se dio como válida la imputación.

De forma aleatoria, se seleccionaron 142 pacientes para la muestra de derivación. Los restantes 70 se reservaron para la validación externa del modelo y el sistema de puntos resultante. La última submuestra representa exactamente una tercera parte de la muestra global.

El criterio de los autoanticuerpos se consideró desde dos aproximaciones: La binaria propuesta en los criterios ESPGHAN/NASPGHAN 2009 con el límite reducido para ANA, anti-Sm y anti-LKM1, y la modificada. Esta segunda asume el umbral de los criterios 2009 pero permite dar más peso a los títulos más elevados, teniendo en cuenta la filosofía de los criterios clásicos y simplificados (ver tabla).

**Tabla 32: Definición del criterio de los autoanticuerpos en la propuesta ESPGHAN/NASPGHAN 2009 y su versión modificada para la modelización**

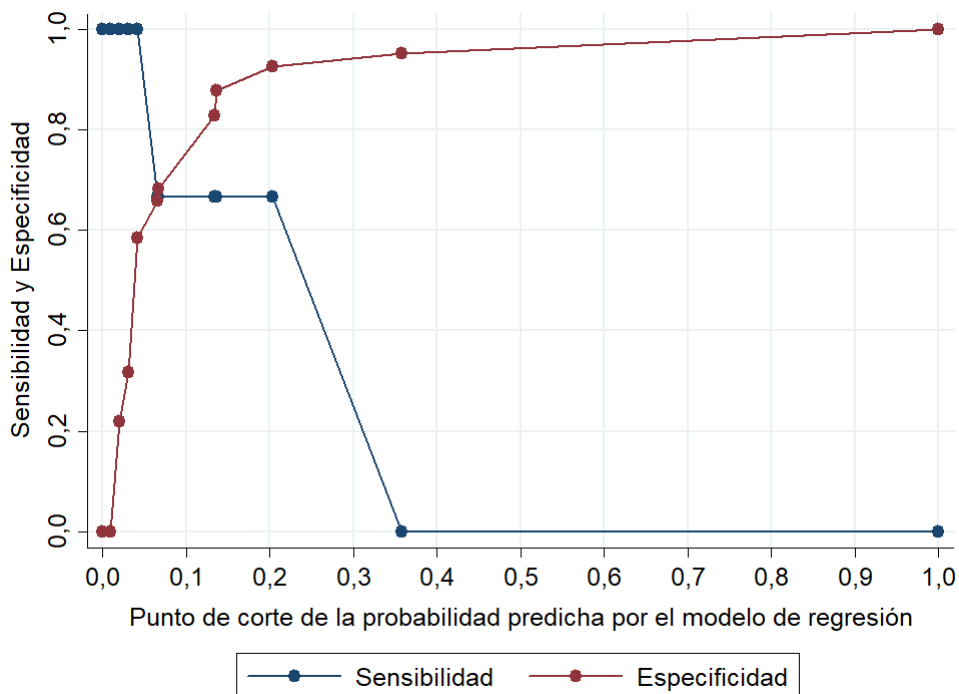
Autoanticuerpos	Criterio en ESPGHAN/ NASPGHAN 2009		Criterio en versión modificada (categórico ordinal)					
			Categoría 1		Categoría 2		Categoría 3	
ANA o anti-SM	$\geq 1:20$		$< 1:20$		$\geq 1:20$ y $< 1:80$		$\geq 1:80$	
Anti-LKM1	$\geq 1:10$	Alguno sí = Sí	$< 1:10$	Todo sí = Sí	$\geq 1:10$ y $< 1:80$	Ni 1 ni 3 = Sí	$\geq 1:80$	Alguno sí = Sí
Anti-LC1, anti-SLA u otros	Positivos		Negativo		Negativo		Positivo	

A continuación, se reproducen los indicadores de ajuste de los modelos máximos con el criterio de los autoanticuerpos original y el modificado.

**Tabla 33: Calidad del ajuste de los modelos máximos basados en los criterios ESPGHAN/NASPGHAN 2009 para el diagnóstico de hepatitis autoinmune.**

	Modelo máximo con el criterio de los autoanticuerpos original	Modelo máximo con el criterio de los autoanticuerpos modificado
$C_p$ de Mallows	8	8
Criterio de Akaike corregido	361,8	362,5
-2LL	16,5	13,2
Área bajo la curva COR	0,996	0,998
Valor p de Hosmer-Lemeshow	1,000	1,000

Los dos modelos máximos muestran un rendimiento diagnóstico similar, discretamente a favor de aquel con el criterio de los autoanticuerpos categorizado, tal como señala un criterio de Akaike corregido superior y una mayor área bajo la curva COR. En cualquier caso, la magnitud de la diferencia es mínima.



**Figura 47: Sensibilidad y especificidad obtenidas para cada punto de corte en la probabilidad de HAI predicha por el modelo máximo basado en los criterios de 2009 con el ítem de los autoanticuerpos categorizado.**

Una vez demostrada la leve superioridad de la categorización del ítem de los autoanticuerpos, se exploraron todos los posibles submodelos. Entre los mismos, los indicadores de ajuste de los que arrojaron un mejor rendimiento fueron los siguientes:

Tabla 34: Características de los mejores submodelos restringidos en variables.

Modelo	R <sup>2</sup>	C <sub>p</sub>	Akaike	Akaike corregido	-2LL	ABC COR	Valor p del estadístico de ajuste de Hosmer-Lemeshow
1	0,967	7,8	-145,7	261,2	13,2	0,998	1,000
2	0,963	6,4	-146,8	259,9	22,0	0,996	1,000
3	0,961	6,2	-146,8	259,8	23,1	0,997	1,000
4	0,920	4,9	-148,1	258,3	24,1	0,995	1,000
5	0,947	5,5	-148,3	258,2	31,0	0,988	0,938

- 1) Hipergammaglobulinemia, Wilson, Colangiografía, Virus, Autoanticuerpos, Anatomía patológica
- 2) Hipergammaglobulinemia, Wilson, Virus, Autoanticuerpos, Anatomía patológica
- 3) Hipergammaglobulinemia, Wilson, Colangiografía, Autoanticuerpos, Anatomía patológica
- 4) Hipergammaglobulinemia, Wilson, Autoanticuerpos, Anatomía patológica
- 5) Hipergammaglobulinemia, Wilson, Colangiografía, Anatomía patológica

Las diferencias de los submodelos de la tabla anterior con el modelo máximo son despreciables en términos de R<sup>2</sup> y área bajo la curva COR. Aun así, desde un punto de vista teórico, resulta difícil excluir variables como la ausencia de marcadores de hepatitis vírica de un modelo predictivo, por ejemplo. Más teniendo en cuenta que algunos de los falsos negativos encontrados en la validación de los criterios simplificados fueron debidos a que la HAI fue posiblemente desencadenada a raíz de una hepatitis vírica. La colangiografía también puede excluirse del modelo sin que tenga una repercusión matemática relevante sobre las predicciones efectuadas con el mismo. Pese a ello, tampoco resulta lógico construir un sistema diagnóstico sin ella dado que se clasificarían erróneamente muchos casos de síndrome de solapamiento o CEAI.

Con todo, la decisión provisional fue quedarnos con el modelo máximo con el ítem de los autoanticuerpos en tres categorías. Expresado de otra forma, se sacrificó parsimonia a favor de una mayor capacidad predictiva.



Tabla 35: Modelo provisional de regresión logística con intención diagnóstica (predictiva) para hepatitis autoinmune a partir de los criterios ESPGHAN/NASPGHAN de 2009 con los autoanticuerpos categorizados.

Criterio	Coefficiente $\beta$	Error estándar	Significación
Constante	-57,595	16345,9	0,997
Hipertransaminasemia	2,453	15647,0	1,000
Autoanticuerpos negativos	<i>referencia</i>		
Autoanticuerpos positivos*	(1) 18,150	3995,4	0,996
Autoanticuerpos elevados*	(2) 36,300	3995,4	0,996
Hipergammaglobulinemia	18,176	3995,4	0,996
Anatomía patológica	52,090	5070,0	0,992
Descarte hepatitis vírica	16,644	10657,4	0,999
Descarte enfermedad Wilson	17,997	8096,9	0,998
Colangiografía normal	1,267	1,322	0,338

\*Autoanticuerpos positivos: ANA y anti-SM  $\geq 1:20$  y  $< 1:80$  con anti-LKM1  $\geq 1:10$  y  $< 1:80$ , y anti-LC1, anti-SLA u otros negativos (categoría 2 de la tabla 32). Autoanticuerpos elevados: ANA y anti-SM  $\geq 1:80$  o anti-LKM1  $\geq 1:80$  o anti-LC1, anti-SLA u otros positivos (categoría 3 de la tabla 32).

#### 9.4.3.1. Diagnósticos del modelo elegido y tabla de índices pronósticos

El análogo de la distancia de Cook del modelo lineal se representó con relación a la probabilidad predicha en el modelo. En la mayoría de los casos se obtuvo una probabilidad de 0 o 1, lo que explica la baja densidad de la nube de dispersión.

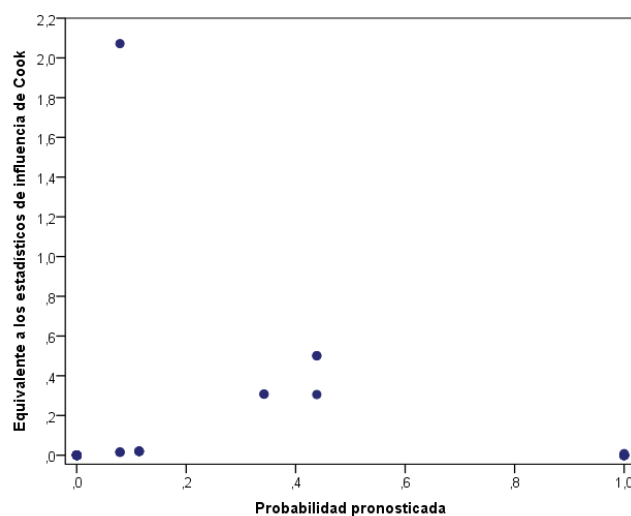


Figura 48: Detección de casos influyentes con la distancia de Cook.

Se aprecia con claridad la presencia de solamente un caso extremo, con una probabilidad predicha según el modelo máximo de 0,08 pero que en realidad es un caso confirmado de HAI. Los tres puntos comprendidos entre valores de 0,2 y 0,6 de distancia de Cook representan 5 casos, también bastante separados del resto. Sus probabilidades diagnósticas según el modelo son 0,34 (tres casos de hepatitis aguda no filiada) y 0,44 (dos casos de HAI).

Retirando el caso extremo de la matriz de datos y repitiendo la ejecución del modelo, los coeficientes B de cada variable no presentaron modificaciones significativas.

En relación con el problema de la sobredispersión, se calculó la varianza media residual siguiendo la expresión:

$$-2LL/df_{Res} = 13,2/(142 - 7) = 0,1$$

Este resultado, inferior a la unidad, es indicativo de infradispersión y su prueba de significación arroja un valor  $p < 0,0001$ . Por consiguiente, no existe necesidad de corregir los parámetros B de las variables predictivas del modelo.

El coeficiente  $R^2$  (con la corrección de Nagelkerke), como indicador de la bondad de ajuste, fue de 0,967. La interpretación de este valor es que el 96,7% de la variabilidad de la variable respuesta del modelo (diagnóstico de HAI sí o no) queda explicada por el conjunto de los criterios ESPGHAN/NASPGHAN 2009.

Por su parte, el valor  $p$  de la significación global del modelo predictivo de HAI fue inferior a 0,0001.

La prueba de bondad de ajuste entre las predicciones elaboradas por el modelo y las probabilidades observadas en la muestra (Hosmer y Lemeshow) arrojó un valor  $p$  de 1,000, indicativo de una calibración excelente.

Todos estos cálculos han permitido explorar problemas potenciales que puedan afectar al análisis de la regresión, además de determinar si los supuestos del modelo son razonables con la matriz de datos empleada. En general, los diagnósticos del modelo han superado la evaluación de la aplicación de la técnica de

regresión logística. Por ello confirmamos la idoneidad que mantener el modelo máximo con el ítem de los autoanticuerpos en tres categorías.

Finalmente, se muestra una tabla de índices pronósticos para cada combinación de variables explicativas, expresados como el riesgo relativo frente al de un paciente de referencia (presencia de los criterios de descarte de hepatitis vírica, descarte de enfermedad de Wilson, colangiografía normal, hipertransaminasemia y ausencia del resto de criterios).

					Riesgo relativo						
					Autoanticuerpos						
					Negativos		Títulos bajos		Títulos altos		
					Anatomía patológica		Anatomía patológica		Anatomía patológica		
Hipertransaminasemia	Hipergammaglobulinemia	Ausencia de marcadores virales	Descarte de Wilson	Colangiografía normal	No compatible	Compatible	No compatible	Compatible	No compatible	Compatible	
No	No	No	No	No	,00	,00	,00	,01	,00	6,00	
				Sí	,00	,09	,00	6,00	,00	6,00	
			Sí	No	,00	,40	,00	6,00	2,41	6,00	
				Sí	,00	6,00	,60	6,00	,00	6,00	
			Sí	No	,00	6,00	,00	6,00	6,00	,00	
				Sí	,00	6,00	6,00	,00	,00	,00	
	Sí	No	No	No	,00	,05	,00	6,00	,47	6,00	
				Sí	,00	6,00	,08	6,00	,00	6,00	
				Sí	,00	6,00	,37	6,00	,00	6,00	
		Sí	No	No	,00	6,00	,00	6,00	6,00	6,00	
				Sí	,00	6,00	6,00	6,00	,00	6,00	
				Sí	,00	6,00	3,01	6,00	6,00	,00	
Sí	No	No	No	No	,00	,00	,00	,01	,00	6,00	
				Sí	,00	,00	,00	6,00	,00	,01	
			Sí	No	,00	,07	,00	6,00	,00	6,00	
				Sí	,00	6,00	,00	6,00	,00	,05	
			Sí	No	No	,00	,31	,00	6,00	2,05	6,00
					Sí	,00	6,00	,47	6,00	,00	6,00
	Sí	No	No	No	,00	6,00	,00	6,00	6,00	,00	
				Sí	,00	6,00	6,00	,00	,00	,00	
				Sí	,00	,04	,00	6,00	,37	6,00	
		Sí	No	No	,00	6,00	,00	6,00	1,42	6,00	
				Sí	,00	6,00	,29	6,00	,00	6,00	
				Sí	,00	6,00	6,00	6,00	,00	6,00	
Sí	No	No	No	,00	6,00	,00	6,00	6,00	6,00		
			Sí	,00	6,00	6,00	6,00	,00	6,00		
			Sí	,00	6,00	2,63	6,00	6,00	6,00		
	Sí	No	No	,00	6,00	,00	6,00	6,00	,01	6,00	
			Sí	,00	6,00	6,00	6,00	,00	6,00		
			Sí	,68	6,00	6,00	6,00	,00	6,00		

Figura 49: Riesgos relativos de cada posible combinación de criterios ESPGHAN/NASPGHAN 2009 respecto al riesgo de un paciente con hipertransaminasemia, sin marcadores virales ni enfermedad de Wilson, colangiografía normal y resto de variables negativas.

La probabilidad de diagnóstico de HAI arrojada por el modelo de regresión reproducido en la tabla 41 se presentó en el espacio COR, así como el punto óptimo calculado para una prevalencia de 47,2%. El área bajo la curva del modelo fue de 99,8% (IC95% 99,3% a 100%) y el del punto óptimo, de 98,6% (IC95% 96,3% a 100%).

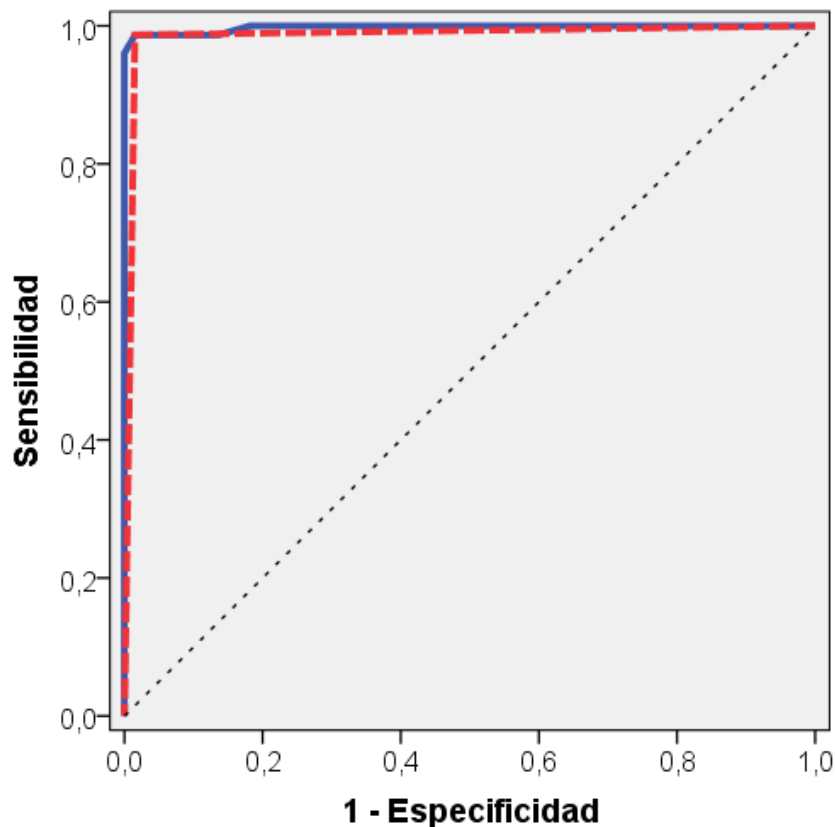


Figura 50: Curva de características operativas del receptor. En azul, para el modelo predictivo de regresión logística, basado en todas las variables de los criterios ESPGHAN/NASPGHAN 2009. En rojo y discontinua, para el punto de corte óptimo con una razón de costes de 1.

#### 9.4.3.2. Sistema de puntos basado en los criterios de 2009

Los coeficientes para cada criterio de 2009 se transformaron en un número entero con el fin de confeccionar el nuevo sistema de puntos. La constante del modelo y su coeficiente  $\beta$  no se tuvieron en cuenta. Dos variables presentaron unos coeficientes  $\beta$  próximos a 0: el criterio de la hipertransaminasemia (2,453) y el de la colangiografía normal (1,267). Por este motivo se decidió excluirlos del conjunto de

ítems final. Sin embargo, atendiendo al conocimiento teórico que tenemos de la HAI y la posibilidad de clasificaciones erróneas en casos de CEP o CEAI, consideramos obligatoria la presencia simultánea de estos dos criterios para efectuar un diagnóstico de HAI basado en el nuevo sistema de puntos.

Sin contar con los criterios de la hipertransaminasemia y de la colangiografía, el coeficiente con el valor inferior fue el del descarte de la hepatitis vírica, con  $\beta = 16,64$ . Este valor sirvió como multiplicador fijo o constante B, que es el número de unidades de regresión que reflejan 1 punto en el sistema de puntos final. En relación con esta constante B, se ajustaron los coeficientes  $\beta$  de cada criterio y, a continuación, se aproximó el cociente al entero más cercano. Los resultados se resumen en la tabla siguiente.

**Tabla 36: Proceso de transformación de los coeficientes  $\beta$  de cada variable del modelo de regresión seleccionado en los puntos de los nuevos criterios diagnósticos de hepatitis autoinmune, basados en la propuesta pediátrica de 2009.**

<b>Criterio</b>	<b>Coficiente <math>\beta</math></b>	<b>Coficiente <math>\beta</math> / constante B</b>	<b>Aproximación al entero más cercano del cociente <math>\beta/B</math></b>
Autoanticuerpos negativos	–	–	<b>0</b>
Autoanticuerpos positivos*	18,150	1,09	<b>1</b>
Autoanticuerpos elevados*	36,300	2,18	<b>2</b>
Hipergammaglobulinemia	18,176	1,09	<b>1</b>
Anatomía patológica	52,090	3,13	<b>3</b>
Descarte hepatitis vírica	16,644	1,00	<b>1</b>
Descarte enfermedad Wilson	17,997	1,08	<b>1</b>

\*Autoanticuerpos positivos: ANA y anti-SM  $\geq 1:20$  y  $< 1:80$  con anti-LKM1  $\geq 1:10$  y  $< 1:80$ , y anti-LC1, anti-SLA u otros negativos (categoría 2 de la tabla 38). Autoanticuerpos elevados: ANA y anti-SM  $\geq 1:80$  o anti-LKM1  $\geq 1:80$  o anti-LC1, anti-SLA u otros positivos (categoría 3 de la tabla 38).

El sistema resultante es una combinación de 5 criterios: autoanticuerpos (categorizado en tres escalones: negativos, positivos moderados o positivos elevados), hipergammaglobulinemia, anatomía patológica (hepatitis de interfase o colapso multilobular), descarte de hepatitis vírica y descarte de enfermedad de Wilson.

Según el nuevo sistema de puntos, para considerar positivos moderados los autoanticuerpos se tienen en cuenta dos dinteles:  $\geq 1:20$  para los ANA y los anti-SM, y  $\geq 1:10$  para los anti-LKM1. Títulos de  $\geq 1:80$  de estos anticuerpos, o la presencia de cualquier valor de anti-LC1, anti-SLA u otros, se contempla como autoanticuerpos positivos elevados. Las categorías son excluyentes entre sí, de manera que la puntuación máxima para este criterio es de 2.

A diferencia de los criterios clásicos de 1999, la hipergammaglobulinemia no está categorizada y, por lo tanto, valores más altos, no se traducen en una mayor asignación de puntos. Cualquier valor por encima del ancho de referencia del laboratorio da 1 punto.

El criterio más importante es el de la anatomía patológica. Su aplicación es incluso más sencilla que la del mismo ítem de los criterios simplificados. Se considera positiva, y por lo tanto asigna 3 puntos, a la descripción de hepatitis de interfase o de colapso multilobular.

Finalmente, el descarte de hepatitis vírica y de enfermedad de Wilson, dan 1 punto cada una.

Con este nuevo sistema diagnóstico por puntos, se puede obtener hasta un máximo de 8.

Para cada caso de la muestra de derivación (*training set*) se calculó la puntuación que obtendrían con los nuevos criterios si se hubiesen aplicado en base a la información más cercana a la fecha de su biopsia hepática. Esta submuestra incluye 76 casos de HAI y 66 pacientes con diagnósticos alternativos, entre los que están representados todos los diagnósticos diferenciales de la muestra completa. De los pacientes con HAI, 17 (22,4%) no hubieran puntuado como HAI según los criterios clásicos de 1999 aplicados sin tener en cuenta la respuesta al tratamiento. De estos 17, tras tener en cuenta el ítem de la respuesta al tratamiento, solo 2 casos fueron finalmente diagnosticados de HAI sin cumplir con los criterios clásicos (reclasificados durante el análisis de casos discrepantes).

Asignando una razón de costes de 1, el punto de corte óptimo para considerar como positivo el resultado de los nuevos criterios basados en la propuesta 2009, fue de 6 puntos.

**Tabla 37: Nuevos criterios diagnósticos de hepatitis autoinmune pediátrica basada en la propuesta ESPGHAN / NASPGHAN de 2009.**

Parámetro y discriminador	Puntuación
Autoanticuerpos	
ANA o anti-SM $\geq 1:20$ y $< 1:80$ , o anti-LKM1 $\geq 1:10$ y $< 1:80$	+1
ANA o anti-SM $\geq 1:80$ o anti-LKM1 $\geq 1:80$ o anti-LC1, anti-SLA u otros positivos	+2
Hipergammaglobulinemia	+1
Anatomía patológica	
Hepatitis de interfase	+3
Colapso multilobular	
Ausencia de marcadores de hepatitis vírica	+1
Descarte de enfermedad de Wilson	+1
HAI = 6 a 8 puntos, si aplicados sobre un paciente con hipertransaminasemia y colangiografía normal.	

#### **9.4.3.3. Validación interna del nuevo sistema diagnóstico basado en los criterios de 2009**

Igual que en el análisis de los indicadores de validez llevado a cabo en el capítulo 1 para los criterios de 2008, el cálculo de la sensibilidad y la especificidad por intención de diagnosticar y por protocolo llevó a los mismos resultados. Se obtuvo información suficiente de todos los pacientes excepto en la variable de los resultados de la prueba de imagen de vía biliar, cuyos valores perdidos se solucionaron por imputación múltiple a través de modelización por regresión logística. El procedimiento exacto está explicado en la metodología específica de este bloque.

En la submuestra de derivación, con el punto de corte en 6 para considerar que se cumplen los nuevos criterios ESPGHAN/NASPGHAN 2009, su sensibilidad fue de 96,1% (IC95% 89,0% a 98,7% por el método de Wilson) y su especificidad fue de 100% (IC95% 94,5% a 100%). Solo 3 casos de HAI no fueron correctamente

clasificados por el nuevo sistema diagnóstico (tasa de falsos negativos de 3,9%, con un IC95% 1,4% a 11,0%).

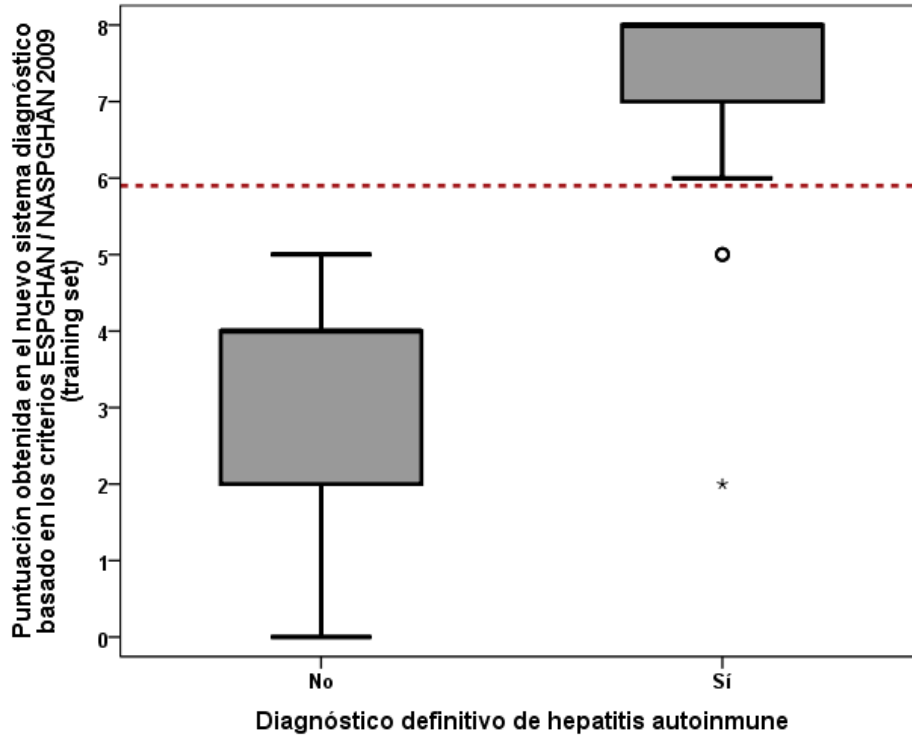


Figura 51: Diagrama de caja de la puntuación obtenida por los nuevos criterios diagnósticos ESPGHAN/NASPGHAN 2009 para la hepatitis autoinmune pediátrica en la muestra de derivación (*training set*). Las patillas incluyen el 90% central de la muestra y los símbolos representan valores alejados. La línea discontinua (en 6) refleja el mejor punto de corte.

La mediana de puntos del sistema ESPGHAN/NASPGHAN 2009 de los casos de HAI fue de 8 (RIC 7 a 8), mientras que la del grupo de no HAI fue de 4 (RIC 2 a 4).

Las RV positiva y negativa fueron de 64,51 y 0,04 respectivamente. Dado que no hubo falsos positivos, el cálculo de la RV positiva se aproximó sumando media unidad a cada casilla de la tabla de contingencia 2x2. Traducido a WoE positivo y negativo, los equivalentes logarítmicos de estos valores fueron de +18,1 y -14,0 deciban.

El índice de Youden para este mismo punto de corte fue de 0,96; por encima, incluso, del obtenido con los criterios simplificados para el corte que indica HAI probable, lo que significa una capacidad informativa global superior.



La efectividad de los nuevos criterios diagnósticos ( $\delta$ ), o la diferencia estandarizada entre las medias de los resultados entre los pacientes de HAI y los controles con otras hepatopatías, fue de 7,2. Como referencia, se considera que una  $\delta > 3$  es indicadora de alta efectividad diagnóstica.

Por lo que respecta a la *odds ratio* diagnóstica, su valor fue de 1239 (IC95% 135 a 11368), calculada también tras la corrección del valor 0 en la tabla de contingencia.

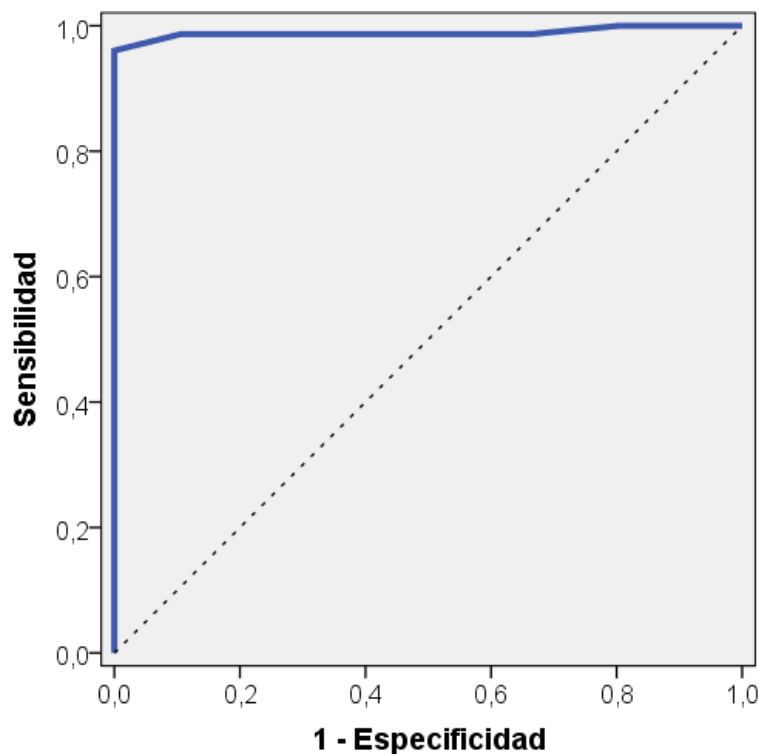


Figura 52: Curva de características operativas del receptor del nuevo sistema diagnóstico por puntos basado en los criterios ESPGHAN/NASPGHAN 2009 para la hepatitis autoinmune pediátrica en la muestra de derivación.

El área bajo la curva COR del sistema de puntos con sus 8 posibles valores (todos ellos representados en casos de la submuestra para el cálculo de la validez interna) fue de 98,9% (IC95% 95,5% a 99,9%). Por su lado, el área bajo la curva COR del modelo diagnóstico con el dintel seleccionado en 6 puntos, fue de 98,0% (IC95% 94,2% a 99,6%). Con todo, los indicadores de validez independientes de la

prevalencia de la enfermedad son indicativos de una muy buena capacidad discriminante en la muestra de derivación.

**Tabla 38:** Indicadores de validez para cada uno de los posibles cortes de los nuevos criterios ESPGHAN/NASPGHAN 2009, concebidos como un sistema de puntos, para el diagnóstico de la hepatitis autoinmune pediátrica (muestra de derivación).

<b>Cut-off</b>	<b>Sensibilidad</b>	<b>Especificidad</b>	<b>Clasificaciones correctas</b>	<b>RV +</b>	<b>RV –</b>	<b>WoE+</b>	<b>WoE–</b>
≥1	100%	6,1%	50,4%	1,1	-	+0,4	-
≥2	100%	19,7%	57,6%	1,3	-	+1,1	-
≥3	98,7%	33,3%	64,2%	1,5	0,04	+1,8	-14,0
≥4	98,7%	48,5%	72,2%	1,9	0,03	+2,8	-15,2
≥5	98,7%	89,4%	93,8%	9,3	0,01	+9,7	-20,0
≥6	96,1%	100%	98,2%	64,6*	0,04	+18,1*	-14,0
≥7	94,7%	100%	97,5%	-	0,05	-	-13,0
8	51,3%	100%	77,0%	-	0,49	-	-3,1

\*Aproximado con corrección de valor nulo en tabla de contingencia.

RV: Razón de verosimilitud. WoE: Peso de la evidencia (en deciban)

**Tabla 39:** Simulación de los puntos de corte óptimos para distintas prevalencias y distintas asunciones respecto a la razón de costes idónea. Encuadrado el contexto más parecido al de la población del estudio y los indicadores de la muestra de derivación.

<b>Prevalencia</b>		<b>Razón de costes (coste de un falso negativo / coste de un falso positivo)</b>						
		<b>1/8</b>	<b>1/4</b>	<b>1/2</b>	<b>1</b>	<b>2</b>	<b>4</b>	<b>8</b>
5%	<b>Cut-off</b>	6	6	6	6	6	6	6
	<b>Se</b>	96,1%	96,1%	96,1%	96,1%	96,1%	96,1%	96,1%
	<b>Sp</b>	100%	100%	100%	100%	100%	100%	100%
10%	<b>Cut-off</b>	6	6	6	6	6	6	6
	<b>Se</b>	96,1%	96,1%	96,1%	96,1%	96,1%	96,1%	96,1%
	<b>Sp</b>	100%	100%	100%	100%	100%	100%	100%
20%	<b>Cut-off</b>	6	6	6	6	6	6	6
	<b>Se</b>	96,1%	96,1%	96,1%	96,1%	96,1%	96,1%	96,1%
	<b>Sp</b>	100%	100%	100%	100%	100%	100%	100%
30%	<b>Cut-off</b>	6	6	6	6	6	6	5
	<b>Se</b>	96,1%	96,1%	96,1%	96,1%	96,1%	96,1%	98,7%
	<b>Sp</b>	100%	100%	100%	100%	100%	100%	89,4%
40%	<b>Cut-off</b>	6	6	6	6	6	5	5
	<b>Se</b>	96,1%	96,1%	96,1%	96,1%	96,1%	98,7%	98,7%
	<b>Sp</b>	100%	100%	100%	100%	100%	89,4%	89,4%
50%	<b>Cut-off</b>	6	6	6	6	6	5	5
	<b>Se</b>	96,1%	96,1%	96,1%	96,1%	96,1%	98,7%	98,7%
	<b>Sp</b>	100%	100%	100%	100%	100%	89,4%	89,4%

Se: Sensibilidad. Sp: Especificidad.

Como se demuestra en la tabla anterior, en la mayoría de los escenarios, el corte en 6 puntos para considerar un caso como HAI ofrece una buena robustez dado que se mantiene como dintel óptimo en una amplia variedad de escenarios de probabilidad *pre-test* y razones de coste.

Con los coeficientes de probabilidad de la muestra de derivación, y para una prevalencia de HAI de 47,2%, el VPP fue de 100% (IC95% de 95,0% a 100%) y el VPN, de 96,6% (IC95% de 89,5% a 98,7%). Además, la gran distancia entre los trazos superior e inferior del gráfico siguiente demuestra que la ganancia diagnóstica de los nuevos criterios, para un ancho de probabilidades preprueba de HAI importante, es excelente.

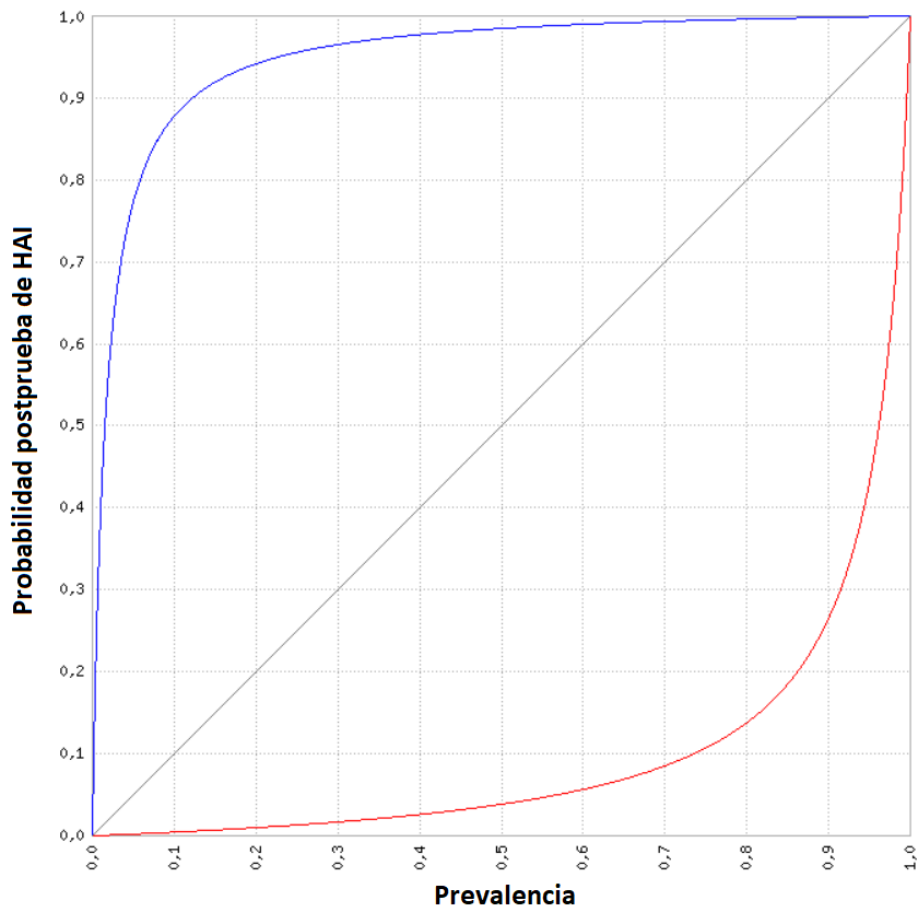


Figura 53: Relación entre la prevalencia de hepatitis autoinmune (HAI) y la probabilidad postprueba con un resultado positivo (línea azul) y con un resultado negativo (línea roja) en el nuevo sistema diagnóstico por puntos basado en los criterios ESPGHAN/NASPGHAN 2009. Obtenido por inferencia bayesiana a partir de las razones de verosimilitud del sistema sobre la submuestra de derivación.

Por último, los indicadores relacionados con la curva de Lorenz, obtenidos por métodos geométricos, se estimaron en un valor de 0,26 tanto para el índice de Pietra como para el de Gini en la muestra de derivación.

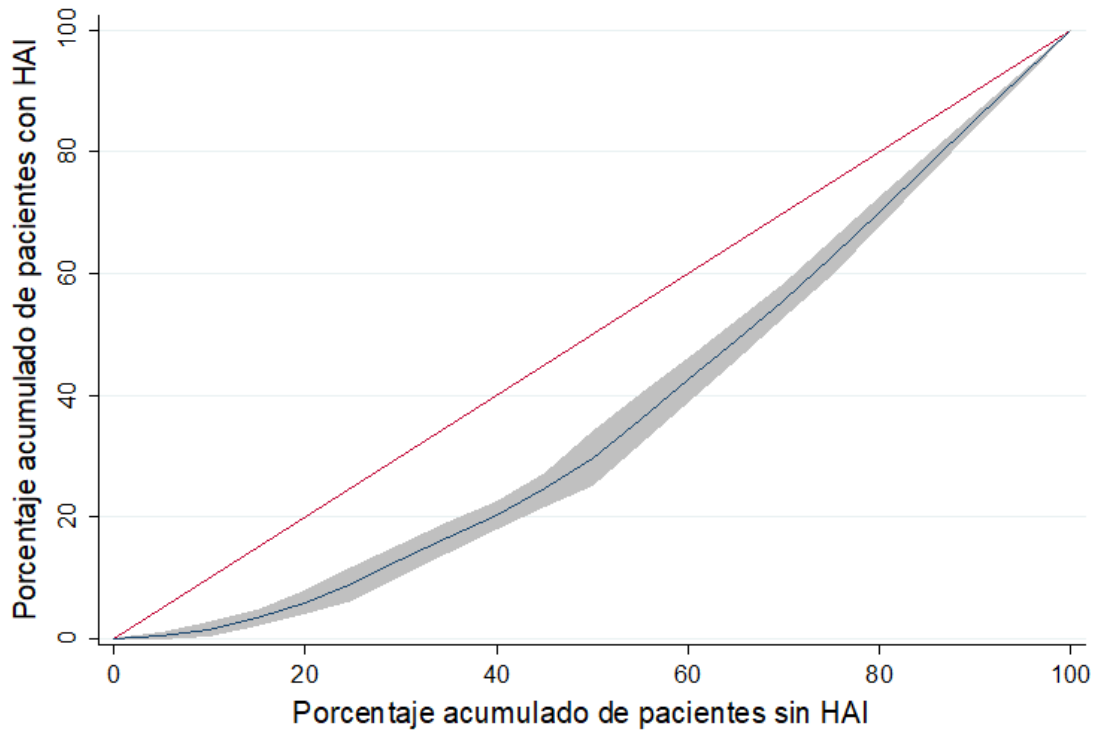


Figura 54: Curva de Lorenz del nuevo sistema diagnóstico basado en los criterios ESPGHAN/NASPGHAN 2009 para la hepatitis autoinmune (HAI) pediátrica, sobre la muestra de derivación. La región sombreada gris representa el intervalo de confianza del 95% de la estimación.

En este apartado, el análisis de los indicadores de validez de los nuevos criterios, incluyendo aquellos sensibles a la prevalencia de la enfermedad en la población sobre la que se aplican, se llevó a cabo sobre la submuestra de validación. La información anterior, al estar obtenida con la misma población sobre la que se ha generado el sistema de puntos, es susceptible de sobredimensionar su bondad diagnóstica. Los resultados ofrecidos en el siguiente apartado son los aportados por el nuevo sistema diagnóstico sobre una submuestra diferente y, por lo tanto, constituyen sus auténticos parámetros de validez.

**9.4.3.4. Validación externa del nuevo sistema diagnóstico basado en los criterios de 2009**

La distribución de los puntos obtenidos por los nuevos criterios en la muestra de validación queda resumida en la tabla y las figuras que se reproducen a continuación. La diferencia entre la mediana de puntos obtenida entre los distintos grupos diagnósticos es significativa.

Tabla 40: Distribución descrita a través de la mediana y el rango intercuartílico (RIC) de los puntos obtenidos por el nuevo sistema de puntos ESPGHAN/NASPGHAN 2009 para la hepatitis autoinmune (HAI).

	HAI (n=24)	No HAI (n=46)	Valor p de la diferencia
Nuevo sistema basado en los criterios de 2009	8 (RIC 7 a 8)	3 (RIC 2 a 4)	<0,001

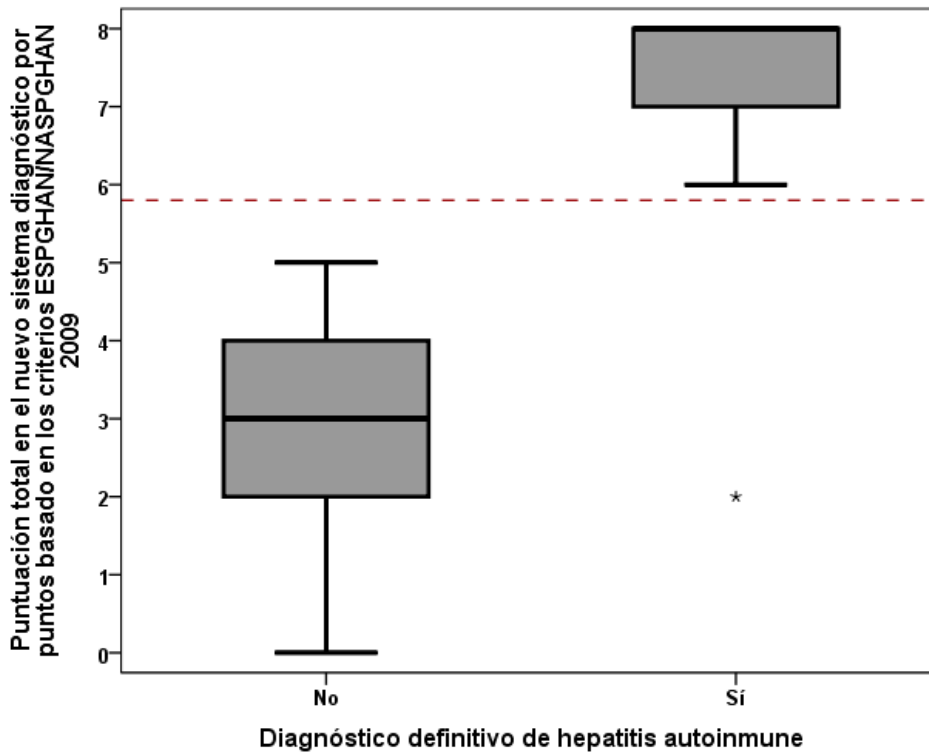


Figura 55: Diagrama de caja de la puntuación obtenida en los nuevos criterios ESPGHAN/NASPGHAN 2009 por los pacientes con y sin hepatitis autoinmune en la muestra de validación (*validation set*). Las patillas incluyen el 90% central de la muestra. La línea roja discontinua (en 6 puntos) representa el mejor punto de corte calculado en la muestra de derivación del sistema.

En la muestra de validación, el grupo de no HAI estuvo compuesto por 46 pacientes finalmente diagnosticados de hepatitis criptogénica aguda (16), enfermedad de Wilson (8), hepatitis vírica (6), hepatitis tóxica (3), hepatitis intrahepática familiar progresiva (3) y 10 otros casos que representan una miscelánea de 1 o 2 casos de CEP, enfermedad mitocondrial, hepatitis de células gigantes, hígado de estasis y síndrome de Alagille.

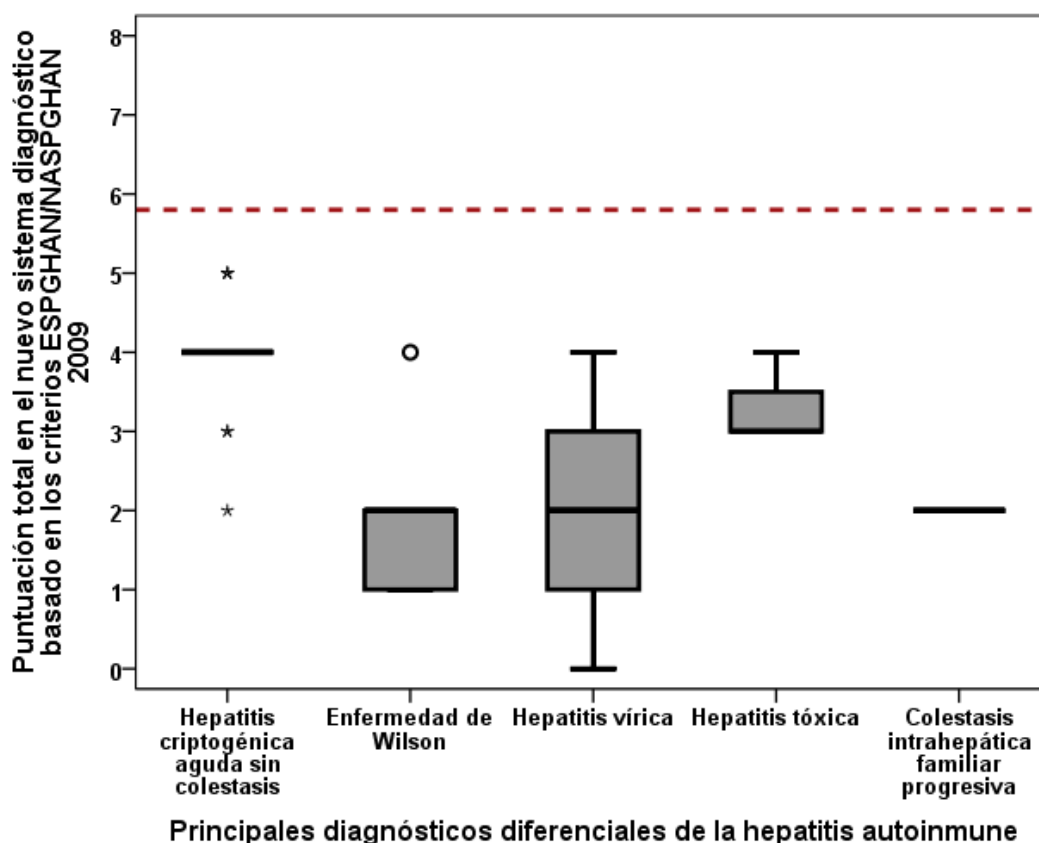


Figura 56: Diagrama de caja de las puntuaciones obtenidas en el nuevo sistema diagnóstico basado en los criterios ESPGHAN/NASPGHAN 2009 por los pacientes con los principales diagnósticos diferenciales. La línea roja discontinua (en 6 puntos) representa el mejor punto de corte calculado en la muestra de derivación del sistema.

Tras la aleatorización para asignar el caso al grupo de elaboración o validación, ningún paciente con fallo hepático agudo como clínica de presentación inicial quedó en esta última submuestra.

Dentro del grupo de HAI, los nuevos criterios categorizaron correctamente al 95,8% de los pacientes (solo un caso puntuó 2, por debajo del corte óptimo), mientras que el 100% de los casos de no HAI fueron bien clasificados por estos mismos criterios. El paciente falso negativo por el nuevo sistema diagnóstico también fue mal clasificado por los criterios simplificados de 2008 y se trata de una niña seronegativa y con valores normales de gammaglobulina, pero con histología hepática compatible y respuesta completa al tratamiento inmunosupresor. La tasa de falsos negativos fue, por consiguiente, de 4,2% (IC95% 0,7% a 20,2%). Por su parte, la tasa de falsos positivos fue de 0,0% (IC95% 0,0% a 7,7%).

Con todo, el par sensibilidad/especificidad del nuevo sistema diagnóstico basado en los criterios de 2009 fue de 95,8% (IC95% 79,8% a 99,3%) y 100% (IC95% 92,3% a 100%), respectivamente. La precisión absoluta con la que se estimaron estos indicadores, obtenida por las fórmulas de Buderer (mencionadas en el bloque sobre los estudios de validación de pruebas diagnósticas del anexo 13.1), fue de  $\pm 6,5$  para la sensibilidad y de  $\pm 6,0$  para la especificidad. El cálculo de la precisión absoluta se llevó a cabo teniendo en cuenta la prevalencia de 47,2% de la muestra entera y la especificidad de 96,1%, que es el punto medio del intervalo de confianza de la especificidad real (no es posible aplicar las fórmulas de Buderer con Se o Sp de 0 o 1).

Las RV positiva y negativa fueron de 24,9 y 0,04, respectivamente (aproximando la especificidad también a 96,1%). Traducido a WoE positivo y negativo, sus equivalentes logarítmicos fueron +14,0 deciban y -13,6 deciban.

El índice de Youden, para el punto de corte en 6 del nuevo sistema diagnóstico se estimó en 0,96 y refleja que los criterios de 2009 tienen una excelente capacidad informativa global.

La efectividad de la prueba ( $\delta$ ) resultó en 6,2, que es la diferencia, en una escala normalizada, entre las puntuaciones medias de los grupos de enfermos y no enfermos.

Finalmente, se calculó la *odds ratio* diagnóstica. Como se ha señalado previamente, se trata de un indicador de validez independiente de la prevalencia de la enfermedad problema y su valor guarda una relación positiva con la capacidad discriminante global de la prueba índice. Para el caso del rendimiento en la submuestra de validación del nuevo sistema diagnóstico basado en los criterios ESPGHAN/NASPGHAN, la ORD fue de 593,1.

En la última parte de la validación externa del nuevo sistema de puntos, se exploraron los indicadores sensibles a los cambios en la prevalencia de la enfermedad. En el contexto para el que se ha diseñado el estudio, la frecuencia de HAI ha sido de 47,2%. En este escenario, la probabilidad postprueba de HAI para un paciente que obtenga 6 o más puntos en los nuevos criterios (VPP) ha sido de 100% (IC95% 89,2% a 100%). Por su parte, la probabilidad de acertar de un resultado inferior (descartar HAI: VPN), fue de 96,4% (IC95% 85,2% a 99,2%).

Sin embargo, como se explica en el anexo 13.1, la utilidad global de una prueba diagnóstica se mide de forma más adecuada con la ganancia diagnóstica (o contenido diagnóstico) de la prueba. Es un parámetro análogo al índice de Youden que emplea los valores predictivos en vez de la sensibilidad y especificidad. Para el caso del nuevo sistema de puntos, ha resultado de 0,96 (o 96%). Ha sido posible calcular la prevalencia para la cual la ganancia diagnóstica es máxima: diagnosticar HAI con un resultado de  $\geq 6$  puntos en los criterios ESPGHAN/NASPGHAN 2009 obtiene su máximo rendimiento en poblaciones con una probabilidad preprueba de 49,1%, que aproxima los valores predictivos positivo y negativo en valores superiores a 96,1%. De forma específica, un resultado positivo del nuevo sistema diagnóstico obtiene una ganancia diagnóstica máxima en poblaciones con un 16,7% de HAI; y un resultado negativo, con un 82,8%.

El área de indicación de los nuevos criterios amplió el segmento de probabilidades preprueba para los que existe ganancia neta en certeza diagnóstica respecto a la de los criterios simplificados de 2008. En este caso, fue de 2% hasta 92% (amplitud de 90%), con una ganancia neta máxima de 0,84.



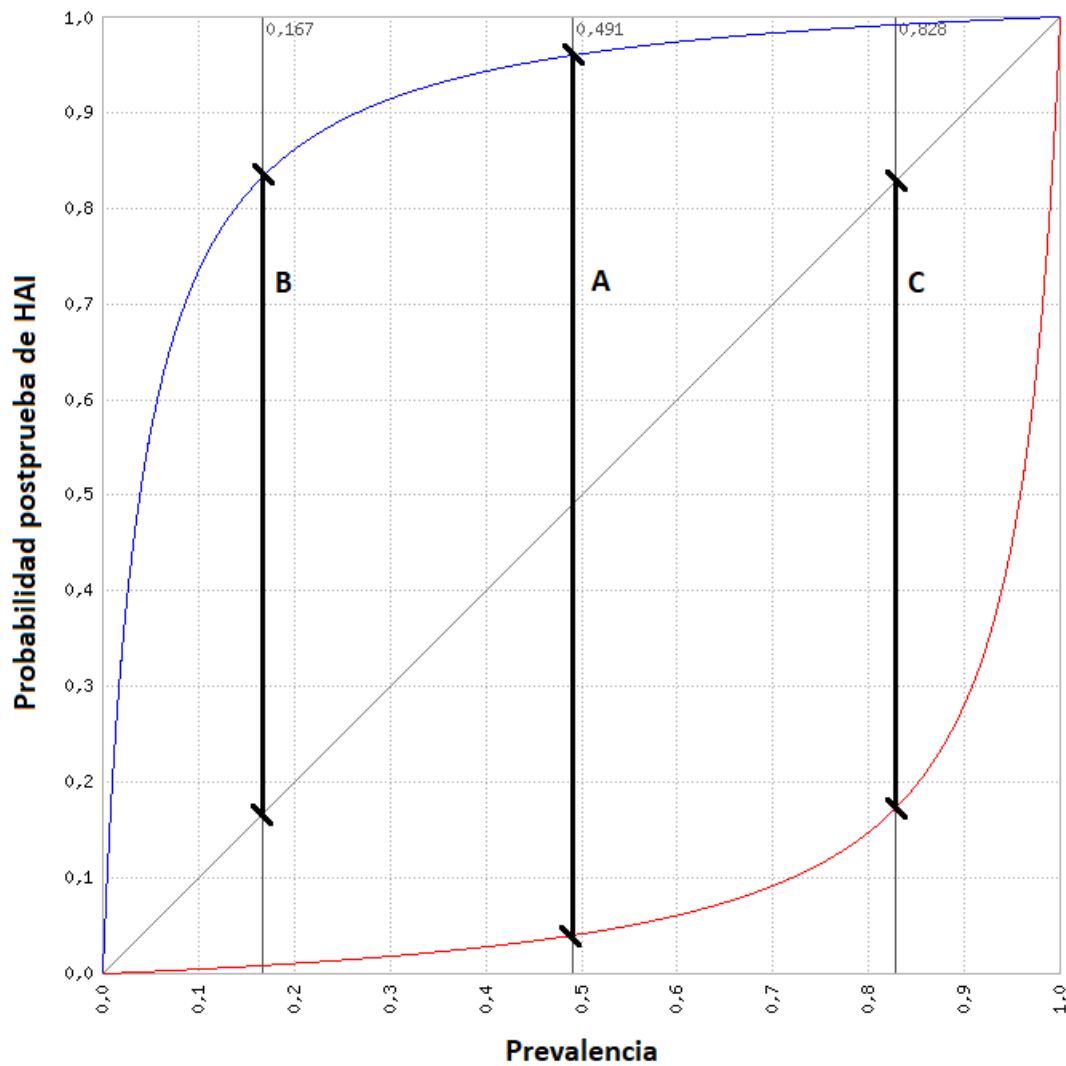


Figura 57: Relación entre la prevalencia de hepatitis autoinmune (HAI) y la probabilidad postprueba con puntuaciones iguales o superiores a 6 (resultado positivo: línea azul) e inferiores a 6 (resultado negativo: línea roja) en el nuevo sistema diagnóstico por puntos basado en los criterios ESPGHAN/NASPGHAN 2009. El segmento A representa la ganancia diagnóstica (GD) global con la prevalencia en la que esta es máxima. El segmento B y C, la GD máxima para un resultado positivo y negativo, respectivamente. Obtenido por inferencia bayesiana a partir de las razones de verosimilitud del sistema sobre la submuestra de validación.

Por inferencia bayesiana, se efectuó una simulación de las probabilidades postprueba de acierto en función de la prevalencia de HAI, tanto para un resultado positivo como negativo del nuevo sistema diagnóstico con los criterios 2009. Con ayuda de los resultados de la tabla siguiente, se puede estimar el riesgo de que el caso sea una HAI tras la aplicación del sistema de puntos desarrollado en este capítulo, en base a cualquier asunción de probabilidad *a priori* de HAI.

Tabla 41: Valores predictivos de los nuevos criterios ESPGHAN/NASPGHAN según la prevalencia de hepatitis autoinmune en la población. En negrita, prevalencias dentro del área de indicación; en rojo, prevalencia para ganancia diagnóstica máxima; en verde, ganancia máxima de un resultado positivo; en naranja, ganancia máxima de un resultado negativo y en azul, prevalencia de nuestra muestra. Entre paréntesis, intervalo de confianza al 95% calculado por el método asintótico. Se ha asumido sensibilidad y especificidad constante (esta última aproximada al punto medio de su intervalo de confianza).

Prevalencia	HAI según los criterios ESPGHAN/NASPGHAN 2009 (6 o más puntos del sistema)	No HAI según los criterios ESPGHAN/NASPGHAN 2009 (5 o menos puntos del sistema)
1%	20,1% (11,7% a 34,3%)	100% (99,7% a 100%)
<b>2%</b>	33,7% (21,1% a 45,6%)	99,9% (99,4% a 100%)
<b>5%</b>	56,7% (45,9% a 67,0%)	99,8% (98,5% a 100%)
<b>10%</b>	73,4% (65,7% a 80,9%)	99,5% (96,9% a 99,9%)
<b>16,7%</b>	83,3% (78,1% a 88,6%)	99,2% (94,6% a 99,9%)
<b>20%</b>	86,2% (81,6% a 90,6%)	99,0% (93,4% a 99,8%)
<b>30%</b>	91,4% (88,3% a 94,5%)	98,2% (89,2% a 99,7%)
<b>40%</b>	94,3% (92,1% a 96,6%)	97,3% (84,1% a 99,6%)
<b>47,2%</b>	95,7% (94,0% a 97,6%)	96,4% (79,8% a 99,5%)
<b>49,1%</b>	96,1% (94,2% a 97,7%)	96,1% (78,5% a 99,4%)
<b>50%</b>	96,1% (94,5% a 97,9%)	96,0% (77,9% a 99,4%)
<b>60%</b>	97,4% (96,2% a 98,7%)	94,1% (70,1% a 99,1%)
<b>70%</b>	98,3% (97,3% a 99,2%)	91,1% (60,2% a 98,6%)
<b>80%</b>	99,0% (98,3% a 99,7%)	85,7% (46,8% a 97,6%)
<b>82,8%</b>	99,2% (98,5% a 99,8%)	83,3% (42,3% a 97,1%)
<b>90%</b>	99,6% (99,1% a 100%)	72,7% (28,1% a 94,8%)
<b>92%</b>	99,7% (99,3% a 100%)	67,6% (23,5% a 93,4%)

El número de pacientes necesarios para diagnosticar (NND), cuya formulación e interpretación se puede consultar en la exposición de resultados específicos del capítulo 1, se calculó en 1,1 para el nuevo sistema de puntos basado en los criterios ESPGHAN/NASPGHAN 2009. Igualmente, el número de pacientes necesarios para diagnosticar mal (NNDM) fue de 25,0. En comparación con sus

respectivos valores para el punto de corte en 6 de los criterios simplificados de 2008 (1,5 y 6,6), estos indicadores reflejan una mayor confiabilidad a favor de los nuevos criterios.

**Tabla 42: Resumen de los principales indicadores de validez de los nuevos criterios ESPGHAN/NASPGHAN 2009 para el diagnóstico de hepatitis autoinmune en niños. Calculados en la submuestra de validación (*validation set*). Entre paréntesis, intervalo de confianza al 95%.**

<b>Indicador de validez</b>	<b>Nuevo sistema diagnóstico por puntos basado en los criterios ESPGHAN/NASPGHAN 2009</b>
Sensibilidad	95,8% (79,8% a 99,3%)
Especificidad	100% (92,3% a 100%)
Índice de Youden	0,96
Valor predictivo positivo*	100% (89,2% a 100%)
Valor predictivo negativo*	96,4% (85,2% a 99,2%)
Valor predictivo global*	98,6% (92,3% a 99,7%)
Razón de verosimilitud positiva**	24,9
Razón de verosimilitud negativa	0,04
WoE de un resultado positivo**	+14,0 deciban
WoE de un resultado negativo	-13,6 deciban
<i>Odds ratio</i> diagnóstica	593,1
Ganancia diagnóstica*	0,96
Efectividad diagnóstica	6,2

\*Datos para una prevalencia de 47,2%. \*\*Aproximación con una especificidad de 0,961.

El poder discriminante global se estimó a través del área bajo la curva COR del nuevo sistema de puntos sobre la muestra de validación, y resultó en 97,1% (IC95% por el método exacto de Wald: 89,9% a 99,6%). Todos los posibles valores que admite el sistema estuvieron representados en esta submuestra y sus indicadores de validez concretos se pueden consultar más adelante.

Para el caso del modelo binario con el punto de corte en 6 (que fue el umbral óptimo estimado empleando la muestra de derivación) el área bajo la curva COR fue de 97,9% (IC95% 91,2% a 99,9%), también al efectuar los cálculos sobre la muestra de validación.

La pérdida de poder discriminante, basado en el área bajo la curva ROC, entre los cálculos efectuados en la muestra de validación con respecto a la de elaboración fue del 2%.

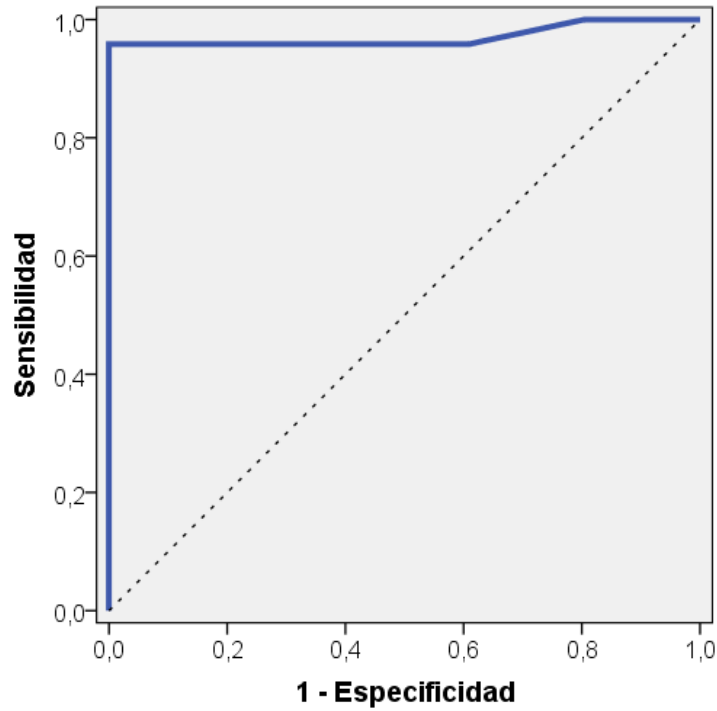


Figura 58: Curva de características operativas del receptor del nuevo sistema diagnóstico por puntos basado en los criterios ESPGHAN/NASPGHAN 2009 para la hepatitis autoinmune pediátrica en la muestra de validación.

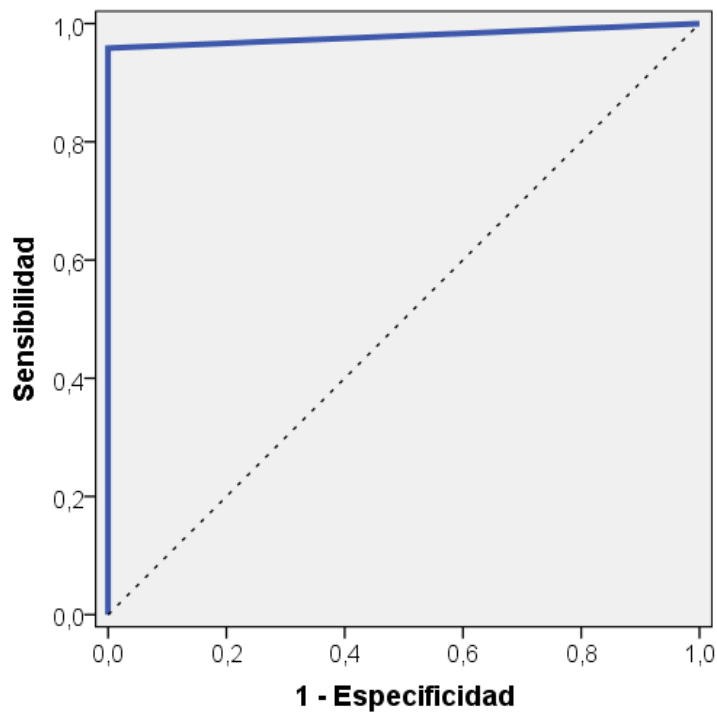


Figura 59: Curva de características operativas del receptor del modelo diagnóstico basado en los criterios ESPGHAN/NASPGHAN 2009 para la hepatitis autoinmune pediátrica en la muestra de validación, con el punto de corte en  $\geq 6$ .

Tabla 43: Indicadores de validez para cada uno de los posibles puntos de corte de los nuevos criterios ESPGHAN/NASPGHAN 2009 para el diagnóstico de hepatitis autoinmune (muestra de validación).

Cut-off	Sensibilidad	Especificidad	Clasificaciones correctas*	RV +	RV -	WoE+	WoE-
≥1	100%	2,2%	47,2%	1,0	-	0,0	-
≥2	100%	19,6%	57,5%	1,2	-	+0,8	-
≥3	95,8%	39,1%	68,9%	1,6	0,11	+2,0	-9,6
≥4	95,8%	63,0%	78,5%	2,6	0,07	+4,1	-11,5
≥5	95,8%	91,3%	93,4%	11,0	0,05	+10,4	-13,0
≥6	95,8%	100%	98,0%	24,9 <sup>†</sup>	0,04	+14,0 <sup>†</sup>	-13,6
≥7	91,7%	100%	96,1%	-	0,08	-	-11,0
8	58,3%	100%	80,3%	-	0,42	-	-3,8

RV: Razón de verosimilitud. WoE: Peso de la evidencia (en deciban). \*Calculado para una prevalencia del 47,2%. †Calculado con una especificidad aproximada a la media de los límites del intervalo de confianza de la especificidad (0,96).

El corte en 6 puntos, fijado en la muestra de derivación, se comportó como el más idóneo también en la muestra de validación, incluso para una variedad amplia de probabilidades iniciales de HAI y de razones de costes.

Tabla 44: Simulación de los puntos de corte óptimos para distintas prevalencias y distintas asunciones respecto a la razón de costes idónea. Encuadrado el contexto más parecido al de la población del estudio y los indicadores de la muestra de validación.

Prevalencia		Razón de costes (coste de un falso negativo / coste de un falso positivo)						
		1/8	1/4	1/2	1	2	4	8
5%	Cut-off	6	6	6	6	6	6	6
	Se	95,8%	95,8%	95,8%	95,8%	95,8%	95,8%	95,8%
	Sp	100%	100%	100%	100%	100%	100%	100%
10%	Cut-off	6	6	6	6	6	6	6
	Se	95,8%	95,8%	95,8%	95,8%	95,8%	95,8%	95,8%
	Sp	100%	100%	100%	100%	100%	100%	100%
20%	Cut-off	6	6	6	6	6	6	6
	Se	95,8%	95,8%	95,8%	95,8%	95,8%	95,8%	95,8%
	Sp	100%	100%	100%	100%	100%	100%	100%
30%	Cut-off	6	6	6	6	6	6	6
	Se	95,8%	95,8%	95,8%	95,8%	95,8%	95,8%	95,8%
	Sp	100%	100%	100%	100%	100%	100%	100%
40%	Cut-off	6	6	6	6	6	6	6
	Se	95,8%	95,8%	95,8%	95,8%	95,8%	95,8%	95,8%
	Sp	100%	100%	100%	100%	100%	100%	100%
50%	Cut-off	6	6	6	6	6	6	6
	Se	95,8%	95,8%	95,8%	95,8%	95,8%	95,8%	95,8%
	Sp	100%	100%	100%	100%	100%	100%	100%

Se: Sensibilidad. Sp: Especificidad.

En la muestra de validación, el índice de Gini, calculado por métodos geométricos sobre el trazo de la curva de Lorenz, fue de 0,32. Por otro lado, el índice de Pietra fue de 0,33.

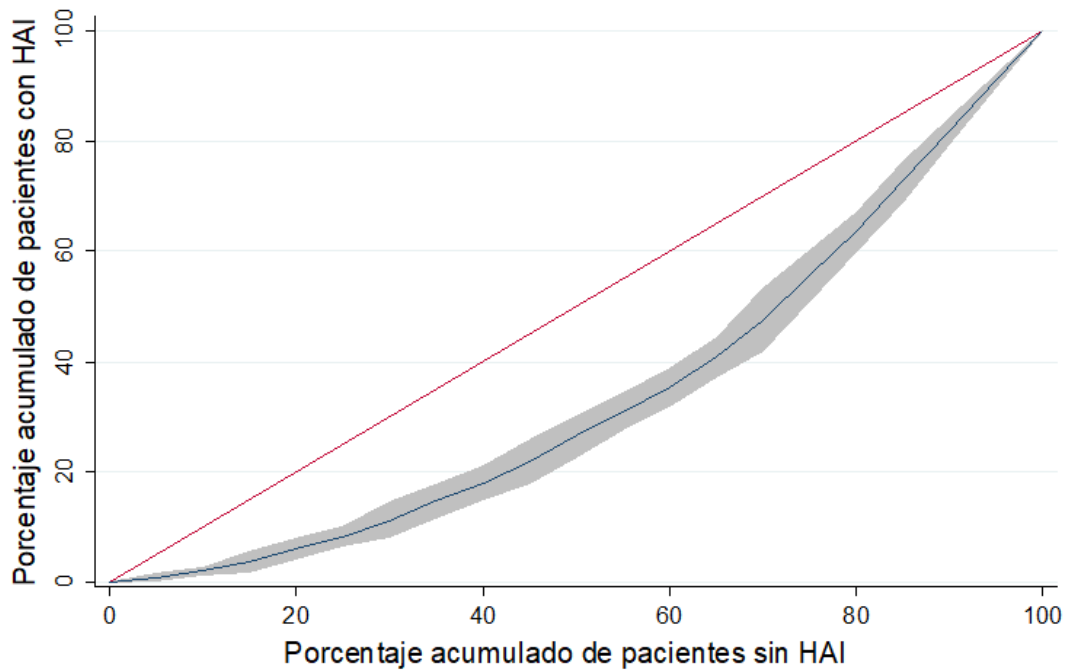


Figura 60: Curva de Lorenz del nuevo sistema diagnóstico basado en los criterios ESPGHAN/NASPGHAN 2009 para la hepatitis autoinmune (HAI) pediátrica, sobre la muestra de validación. La región sombreada gris representa el intervalo de confianza del 95% de la estimación.



## 9.5. Capítulo 5: Evaluación de la concordancia entre las clasificaciones de los criterios diagnósticos

---

### 9.5.1. Metodología específica

En este capítulo se trabajó con la matriz de datos previamente elaborada para el desarrollo de los dos capítulos previos y el primero. La información necesaria estaría contenida en las variables referentes a las clasificaciones de los tres criterios diagnósticos estudiados: los clásicos revisados en 1999, su versión simplificada de 2008 y los nuevos criterios pediátricos basados en la propuesta de Mieli-Vergani de 2009. El objetivo de este bloque es llevar a cabo un estudio de la concordancia de los diagnósticos efectuados por los dos criterios simplificados en comparación con el referente basado en la propuesta original de la IAHG (pre-tratamiento y post-tratamiento). A través de este análisis, complementado con un estudio individual de los casos en los que ha habido discrepancia diagnóstica, se discutirían las bondades y limitaciones de los criterios reducidos.

### 9.5.2. Análisis estadístico

La concordancia entre las clasificaciones establecidas por el estándar de los criterios clásicos de la IAHG, en su revisión de 1999, frente a los criterios simplificados de 2008 y los nuevos criterios ESPGHAN/NASPGHAN 2009 se estudió a través del estadístico *kappa*. Dado que los dos primeros sistemas diagnósticos admiten clasificaciones en 3 categorías (no HAI, HAI probable y HAI definitiva), el estadístico *kappa* se calculó con una ponderación tanto lineal como cuadrática. La clasificación básica en HAI sí o no, se analizó a través de la formulación de *kappa* original, sin ponderar. Al igual que en el capítulo 3, se calcularon los intervalos de



confianza al 95% del estadístico y su interpretación se hizo en base a la propuesta de Landis y Koch [262].

A tal fin se empleó el paquete estadístico SPSS® de IBM, en su versión 21.0, con las ejecuciones necesarias de la sintaxis de la macro !KAPPA.

### 9.5.3. Resultados

El grado de acuerdo global entre las clasificaciones diagnósticas binarias de los criterios clásicos aplicados pre-tratamiento y los criterios simplificados fue del 88,7% (con un 91,0% de acuerdo en el descarte de HAI y un 84,6% de acuerdo en la confirmación de HAI). El estadístico *kappa* se estimó en 0,757 (IC95% 0,665 a 0,848). Dado que los criterios de 2008 no incluyen el ítem de la respuesta al tratamiento, se considera que este contexto es el que ofrece una mejor comparabilidad entre los dos sistemas diagnósticos. De acuerdo con los resultados obtenidos en el bloque anterior, el punto de corte elegido para efectuar este cálculo ha sido el de 6 puntos o más para considerar como positivos los criterios simplificados de 2008.

El mismo análisis, pero empleando la puntuación obtenida por los criterios clásicos añadiendo la información correspondiente a la respuesta al tratamiento, obtuvo un acuerdo global del 87,3% (acuerdo en el descarte de 89,7%, y en la confirmación de 83,2%). Este mínimo descenso en la proporción de acuerdos se debió principalmente a 4 casos que pasaron a diagnosticarse de HAI por los criterios clásicos y que no cambiaron su clasificación en base al resultado de los criterios simplificados. El estadístico *kappa* fue, en este escenario, de 0,730 (IC95% 0,636 a 0,825).

Considerando las posibles categorías ordinales de ambos criterios diagnósticos (no HAI, HAI probable y HAI definitiva), el grado de acuerdo ponderado de forma lineal entre los criterios clásicos pre-tratamiento y los criterios simplificados fue del 88,4%, con un estadístico *kappa* de 0,689 (IC95% 0,610 a 0,768). Empleando una ponderación cuadrática para penalizar de forma

predominante las discordancias no adyacentes en la tabla de contingencia 3x3, el grado de acuerdo de los dos criterios en este mismo escenario pre-tratamiento aumentó a 93,5%, con un *kappa* de 0,775 (IC95% 0,704 a 0,846).

Con los resultados post-tratamiento de los criterios revisados de 1999, la proporción de acuerdo observado con una ponderación lineal ha sido de 87,3%, y el *kappa* de 0,657 (IC95% 0,577 a 0,738). Los valores hallados con una ponderación cuadrática fueron, respectivamente, de 92,9% y 0,750 (IC95% 0,676 a 0,824).

Se ha considerado que, en este caso, la mejor ponderación es la cuadrática. Con ello se pretende otorgar un mayor peso a la discrepancia 'no HAI' *versus* 'HAI definitiva'. Así, la interpretación de estos valores, según la referencia de Landis y Koch, es que se trata de una magnitud de concordancia buena (segundo mejor grado de la escala de *malo* a *excelente*), tanto para la lectura binaria como para la ordinal de los criterios clásicos y los simplificados de la IAIGH.

Respecto a las clasificaciones erróneas de los criterios simplificados, para empezar, se obtuvieron 28 falsos negativos.

La proporción de niños con niveles normales de inmunoglobulina G fue significativamente superior en este grupo respecto a aquellos correctamente clasificados como HAI por los mismos criterios (78.6% frente a 15,3%, con un valor p para la diferencia entre proporciones inferior a 0,001).

Este grupo de pacientes también demostró títulos de autoanticuerpos significativamente inferiores al de los verdaderos positivos. La proporción de niños con valores superiores a 1/80 en los ANA o los anti-SM fue de 50,0% frente a 84,7%, respectivamente (valor p <0,001). Dos casos, de entre los falsos negativos de los criterios de 2008, no cursaron con elevación de autoanticuerpos, mientras que solo un paciente seronegativo se encontró entre los verdaderos positivos.

Se detectaron marcadores virales positivos en el 28,6% de los pacientes con HAI mal clasificados por los criterios de 2008: 4 casos de infección aguda por virus de Epstein-Barr, 3 infecciones agudas por Citomegalovirus y 1 infección aguda por virus A. Estos casos evolucionaron posteriormente hacia una HAI, de modo que la

infección inicial posiblemente actuó de desencadenante de la respuesta autoinflamatoria. La proporción de pacientes con marcadores virales positivos entre los casos de HAI correctamente clasificados fue de 8,3%, con un valor p para la diferencia de 0,009.

No se observaron diferencias en la proporción de mujeres o en las características histológicas entre los falsos negativos y los verdaderos positivos.

Finalmente, los únicos 4 casos falsos positivos de los criterios simplificados de 2008 fueron mujeres, con hipergammaglobulinemia discreta y títulos de ANA superiores a 1/80, que fueron clasificados como HAI probable. Dos de ellas se diagnosticaron finalmente de CEP durante el seguimiento y las otras dos, de hepatitis criptogénica aguda transitoria sin colestasis. A todas se las estudió con colangio-RM y obtuvieron una puntuación negativa en los criterios clásicos.

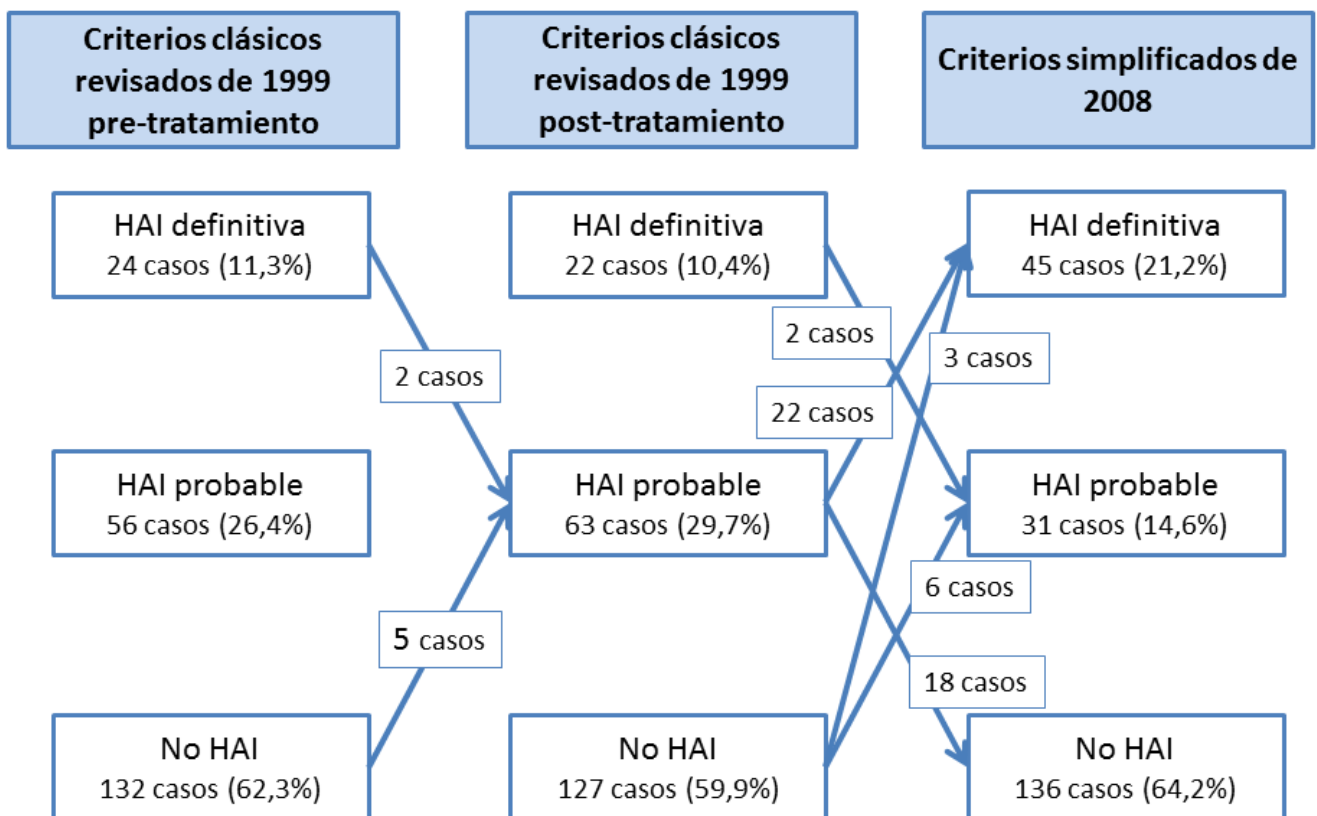


Figura 61: Cambios en las categorías diagnósticas basadas en los dos criterios del Grupo Internacional para el Estudio de la Hepatitis Autoinmune (HAI).

Las mismas operaciones, pero con las clasificaciones del nuevo sistema basado en los criterios pediátricos de 2009, resultaron en unos valores que traducen una concordancia más estrecha. En este caso, el acuerdo se pudo estudiar solo para la clasificación diagnóstica binaria (HAI sí o no) porque es la única que admiten las categorías de los nuevos criterios.

El grado de acuerdo global entre los criterios clásicos pre-tratamiento y los criterios de 2009 fue de 88,7% (con un 90,3% de acuerdos en el descarte de HAI y un 86,4% de acuerdo en la confirmación de HAI). *Kappa* se estimó en 0,768 (IC95% 0,682 a 0,854). Del mismo modo que para los criterios simplificados, este escenario pre-tratamiento es el más asimilable al contexto en el que los nuevos criterios ofrecerían una mayor aplicabilidad en caso de buen rendimiento diagnóstico.

El mismo análisis una vez conocida la respuesta al tratamiento, y con la información diagnóstica que ello aporta a los criterios clásicos revisados de la IAIHG, el acuerdo global aumentó a 91,0% (con un 92,2% de acuerdos en el descarte y un 89,5% de acuerdos en la confirmación). El motivo de este aumento en la proporción de acuerdos se debió a 5 casos que pasaron a clasificarse como HAI por los criterios clásicos y que habían sido correctamente clasificados por los nuevos criterios basados en la propuesta ESPGHAN/NASPHAN pediátrica. El estadístico *kappa* en este escenario fue de 0,817 (IC95% 0,739 a 0,895).

Las únicas 4 clasificaciones erróneas del nuevo sistema diagnóstico generado a partir de los criterios 2009 fueron 4 falsos negativos. Dos de ellos presentaron niveles normales de inmunoglobulina G y los otros dos, hipergammaglobulinemia con valores dos veces por encima del límite superior de normalidad ofrecido por el laboratorio. Curiosamente, estos dos casos con hipergammaglobulinemia, tuvieron títulos de anticuerpos por debajo de 1:40, mientras que los otros dos falsos negativos presentaron ANA >1:80. Los cuatro pacientes fueron mujeres con marcadores de hepatitis vírica negativos. Respecto de la histología, dos casos presentaron solo infiltración leucocitaria con predominio linfoplasmocitario, mientras que en los otros dos se describió un infiltrado inflamatorio inespecífico.

Dos de estos casos seronegativos fueron correctamente clasificados como HAI por los criterios simplificados de 2008 porque se les asignaron 2 puntos en base a los criterios de los valores de IgG y de la ausencia de marcadores de hepatitis vírica, además de presentar una histología compatible con HAI por infiltración inespecífica en la que no se detectó una hepatitis de interfase clara ni colapso multilobular. Se trata de los únicos pacientes bien diagnosticados por los criterios simplificados de 2008 y no clasificados como HAI por los nuevos criterios pediátricos. Los otros dos pacientes falsos negativos por los nuevos criterios también fueron falsos negativos por los criterios simplificados. En el resto de los pacientes mal diagnosticados por los criterios de 2008 (un total de 28), el nuevo sistema de puntos basado en los criterios de 2009 acertó el diagnóstico.

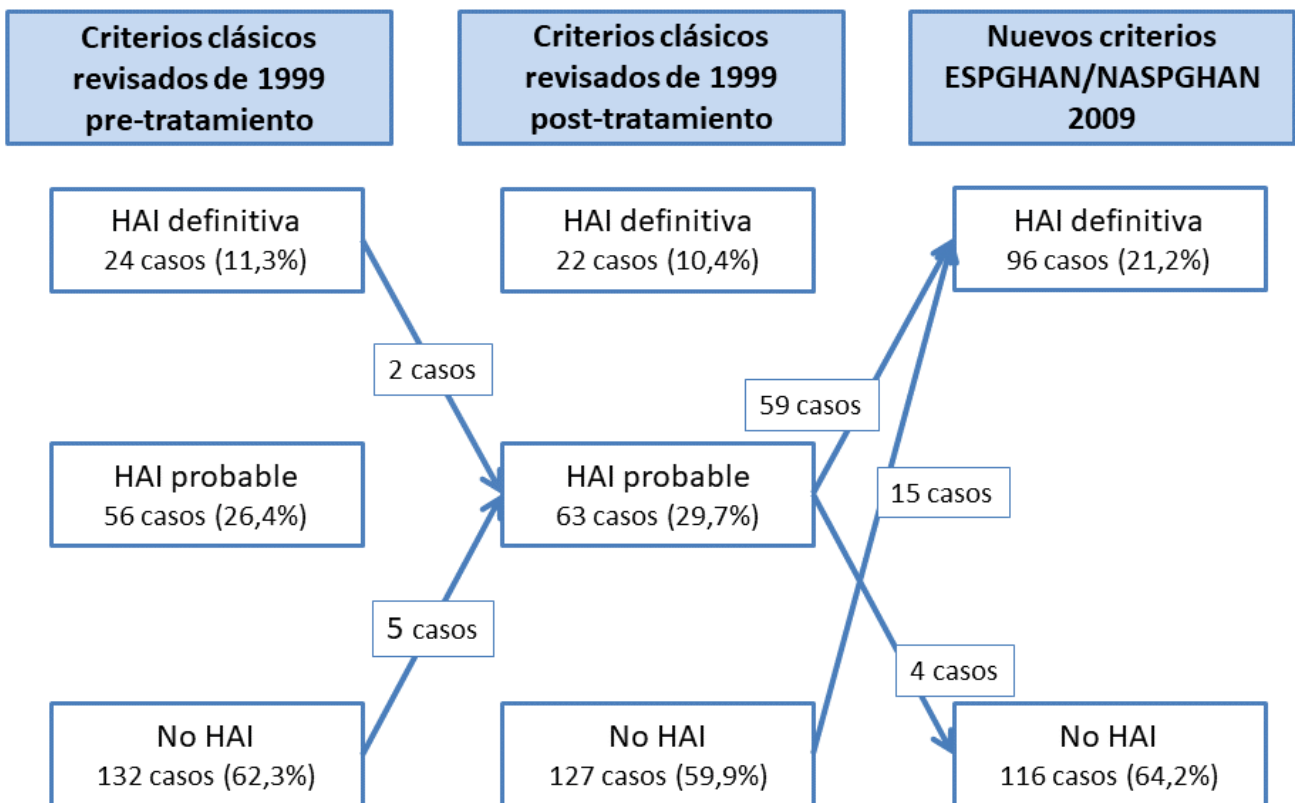


Figura 62: Cambios en las categorías diagnósticas basadas en los criterios clásicos, revisados en 1999, del Grupo Internacional para el Estudio de la Hepatitis Autoinmune (HAI), y en el nuevo sistema de puntos basado en los criterios pediátricos propuestos por la ESPGHAN/NASPGHAN en 2009.

Los 15 casos clasificados como “no HAI” por los criterios clásicos post-tratamiento, que pasan a clasificarse como HAI por los nuevos criterios, incluyen:

1) Los 5 pacientes reclasificados en el análisis de casos discrepantes.

2) Diez casos en los que se confirmó el diagnóstico durante la fase prospectiva, a pesar de no cumplir los criterios clásicos cuando se aplicaron inicialmente. Se catalogaron como HAI, tras la revisión de la historia clínica, por cumplir los dos criterios de robustez consensuados para el diseño del estudio, es decir, además de presentar un diagnóstico explícito de HAI en los informes médicos, tuvieron una anatomía patológica compatible con el diagnóstico y se constató respuesta clínica y analítica al tratamiento farmacológico.



## 9.6. Capítulo 6: Evaluación de la utilidad clínica de los criterios diagnósticos

---

### 9.6.1. Metodología específica

En este último capítulo se comprueba la utilidad de los criterios simplificados de la IAIHG y del nuevo sistema diagnóstico basado en los criterios ESPGHAN/NASPGHAN 2009 para la toma de decisiones terapéuticas en niños con sospecha de HAI.

Parte de la información necesaria para tal fin se ha generado en los capítulos anteriores. Se trata de los indicadores de validez que permiten inferir la probabilidad diagnóstica de HAI una vez conocidos los resultados de la aplicación de los criterios, concebidos como una prueba diagnóstica.

El resto de los parámetros que deben de conocerse o decidirse son la prevalencia de HAI, o probabilidad *a priori* de que el caso sea una HAI, y aquellos que permiten estimar un umbral terapéutico. Esto último se refiere a la probabilidad de estar enfermo por encima de la cual es conveniente que el paciente reciba el tratamiento porque, globalmente, las ventajas de hacerlo (tanto si el diagnóstico está realmente presente o no) superan a las de no hacerlo.

#### 9.6.1.1. Descripción del contexto clínico para el análisis de los criterios diagnósticos simplificados como índices de predicción

##### 9.6.1.1.1. Probabilidad preprueba de hepatitis autoinmune

El diseño del trabajo se ha construido con la intención de reclutar una cohorte representativa de niños con trastornos hepáticos en los que la HAI es una posibilidad diagnóstica dentro del listado de explicaciones o enfermedades potenciales. La elección de los criterios de inclusión y exclusión se ha llevado a cabo



bajo este prisma. En los resultados del capítulo 1 ya se expuso que la prevalencia de HAI en el grupo de niños en riesgo fue de 45,9%, sin contar con los casos rechazados por no disponer de información suficiente para confirmar su diagnóstico final, pero de 47,2% contando todos los pacientes elegibles. Este último dato, por *intención de diagnosticar*, es el que se tuvo en cuenta en primer lugar para la interpretación de la utilidad de los criterios.

En relación con la prevalencia de la HAI, además, se estimó el ancho de probabilidades que van desde el umbral diagnóstico al umbral diagnóstico-terapéutico. La amplitud entre los umbrales de acción incluye las prevalencias de las situaciones clínicas en las que los posibles resultados de los criterios añaden información diagnóstica suficiente como para permitir justificar una acción terapéutica. Se puede definir la utilidad de una prueba también en el sentido de si es apropiada para una variedad relevante de prevalencias y, por lo tanto, empleable en situaciones que generen en el médico diferentes grados de sospecha. Se interpretaría que las dos versiones simplificadas de los criterios diagnósticos de la HAI son tanto más útiles en niños cuanto más amplia sea la distancia entre los dos umbrales de acción.

Análogamente a lo anterior, sobre el nomograma de Fagan y conociendo el umbral terapéutico, se estimaron los límites de prevalencia a partir de los cuales ni un resultado positivo ni uno negativo suponen un cambio de lado de “tratar” a “no tratar” entre las probabilidades preprueba y postprueba. Los cálculos se hicieron a partir de la formulación del teorema de Bayes.

Para una mejor comprensión de los fundamentos teóricos de estas simulaciones, su puede consultar el apartado *Análisis de decisiones clínicas* del anexo 13.1.

Existen datos de que la enfermedad hepática no alcohólica es cada vez más frecuente en niños y adultos. La esteatohepatitis, de hecho, va escalando puestos en el listado de principales etiologías de hipertransaminasemia, tanto en países ricos como en regiones en vías de desarrollo [271,272]. En consecuencia, la frecuencia

con la que se encuentra HAI durante el estudio por una disfunción hepática aguda o crónica está disminuyendo a favor de esta otra entidad. De acuerdo con esta hipótesis, se llevaron a cabo las evaluaciones de la utilidad de los criterios teniendo en cuenta, especialmente, escenarios con probabilidades preprueba inferiores a 47,2%.

#### 9.6.1.1.2. La prueba diagnóstica: Criterios simplificados para la hepatitis autoinmune pediátrica (propuestas de 2008 y 2009)

##### 9.6.1.1.2.1. Riesgo neto de los criterios simplificados

El riesgo neto de una prueba diagnóstica son las consecuencias negativas que, en término medio, se derivan de las complicaciones secundarias al uso del sistema diagnóstico. *Per se*, la aplicación de unos criterios diagnósticos no conlleva un riesgo directo para el paciente. Sin embargo, los procedimientos e intervenciones a partir de las se obtienen algunos resultados de exploraciones complementarias sí que pueden ocasionar problemas. En el caso de los criterios simplificados se ha considerado que la principal fuente potencial de complicaciones es la biopsia hepática.

Complications (minor and major)	Incidence (adult and child)
Pain	84% adults (79)
Bleeding	0%–18% adults (reviewed in (63)), 2.8% children (80)
Arteriovenous fistula	No data
Pneumothorax/haemothorax	0.2% (80)
Organ perforation	0.07%–1.25% (81,82)
Biliary leak/haemobilia	0.6% children (80)
Infection	12.5% in choledochojejunostomy (83)
Death	0%–0.4% adults (reviewed in (63)), 0.6% children (80)

Figura 63: Complicaciones menores y mayores de la biopsia hepática y su correspondiente incidencia. Reproducido de Dezsófi *et al.* J Pediatr Gastroenterol Nutr. 2015;60;413. Con permiso de Lippincott Williams & Wilkins.

La complicación más frecuente después de una biopsia hepática es el dolor. En la mayoría de las ocasiones es de intensidad leve y se maneja con analgesia de primer escalón [273]. Para el cálculo del riesgo neto de la prueba diagnóstica despreciamos esta complicación menor y solo se tuvo en cuenta el sumatorio de la incidencia de complicaciones mayores. Dado que la mayoría de biopsias hepáticas se obtienen por punción percutánea, tampoco se contabilizó el riesgo específico de las biopsias por vía transyugular. Existen datos concretos en población pediátrica de la frecuencia con la que ocurren las complicaciones más relevantes. El sangrado con compromiso hemodinámico o necesidad de transfusión, la fuga biliar o hemobilia y la muerte son los que se tuvieron en cuenta por su especial gravedad. Está descrito que ocurren en el 2,8%, 0,6% y 0,6% de los casos respectivamente [273,274].

Las unidades en las que se estiman las utilidades y, por lo tanto, los riesgos y beneficios de la prueba y el tratamiento, deben de ser los mismos. Se decidió contabilizarlas como proporciones de pacientes. Así, el riesgo de la aplicación de los criterios, homologado al riesgo neto de la biopsia hepática en niños, resultó en:

$$\begin{aligned} R_{\text{criterios}} &= U(\text{sangrado}) + U(\text{hemobilia}) + U(\text{muerte}) = \\ &= 0,028 + 2 \times 0,006 = 0,04 \end{aligned}$$

Cabe especificar que estos riesgos están obtenidos de una fuente de alta calidad metodológica pero que corresponde a una serie antigua, de principios de los años 80 [274]. Actualmente, con una adecuada selección de pacientes y el desarrollo de protocolos que incluyen la vigilancia específica de estas complicaciones, es posible que el riesgo de 4% sobreestime la proporción real de eventos adversos con consecuencias graves. No obstante, se efectuaron los cálculos aceptando este valor porque existe justificación bibliográfica y por considerar más perjudicial el error contrario: infraestimar el riesgo.

#### 9.6.1.1.2.2. Razones de verosimilitud de un resultado positivo y uno negativo

Para el caso de los criterios de 2008, se emplearon las razones de verosimilitud combinadas obtenidas en el meta-análisis del capítulo 2. Representan

la evidencia combinada de todas las fuentes, con calidad metodológica suficiente, que se han podido recuperar en la bibliografía sobre los indicadores de validez de los criterios simplificados de la IAIHG en población pediátrica.

$$RV (+)_{2008} = 13,72$$

$$RV (-)_{2008} = 0,23$$

Por su lado, para el nuevo sistema de puntos basado en los criterios de Mieli-Vergani, se hizo uso de las estimaciones calculadas en la submuestra de validación en el capítulo 4.

$$RV (+)_{2009} = 24,90$$

$$RV (-)_{2009} = 0,04$$

#### 9.6.1.1.2.3. *Umbrales de acción para los criterios simplificados*

De acuerdo con las definiciones expuestas en la introducción, se calculó el umbral diagnóstico y el umbral diagnóstico-terapéutico para cada criterio simplificado: el de 2008 de la IAIHG y la nueva propuesta basada en los criterios pediátricos de 2009.

La formulación matemática propuesta por Djulbegovic y Desoky pone en relación los valores de estos umbrales (que representan una probabilidad preprueba de la enfermedad de interés) con las razones de verosimilitud ( $RV_{2008}$  y  $RV_{2009}$ ), el riesgo neto de la prueba diagnóstica ( $R_{criterios}$ ), el riesgo neto de un tratamiento inapropiado ( $R_t$ ) y el beneficio neto de un tratamiento apropiado ( $B_t$ ) [275,276].

El riesgo y el beneficio asociados al tratamiento se obtuvieron de la revisión de Manns, Lohse y Vergani sobre diagnóstico y tratamiento de la HAI [125]. Es una actualización publicada en *Journal of Hepatology* en 2015 en la que se sintetizan las conclusiones de los estudios mejor diseñados sobre el tema, con una orientación generalista, sin separar los resultados en niños respecto a los adultos. Aun así, se trata de un excelente trabajo de revisión que aporta información suficiente para estimar las utilidades del tratamiento o no, en enfermos o sanos. Siguiendo la misma

lógica que para el cálculo del riesgo neto de la prueba diagnóstica, todas las utilidades se expresaron como proporciones de pacientes.

Se sabe que un 9% de población general, incluyendo adultos, evolucionan negativamente a pesar del tratamiento inmunosupresor estándar o de primera línea. La evolución negativa se define como la ausencia de mejoría clínica, la persistencia de marcadores de citólisis hepática en sangre o la no mejoría histológica a pesar de la corticoterapia. Este 9% representaría el opuesto a la utilidad de tratar a un enfermo (es decir, el 91% de los pacientes con HAI que sí mejoran con el tratamiento). La utilidad de no tratar a un enfermo, en consonancia con la de tratar a un enfermo, sería la proporción de pacientes que mejoran sin recibir farmacoterapia. Manns *et al.* informan de una tasa de resolución espontánea (parcial o completa) de un 12% de pacientes con enfermedad leve. Si bien el espectro clínico completo de HAI también incluye formas graves o incluso FHF, se empleó este dato como la utilidad de no tratar a un enfermo. Si el beneficio neto ( $B_t$ ) de un tratamiento apropiado se define como la diferencia entre las utilidades de tratar y no tratar a un caso,

$$B_t = U(T + |E +) - U(T - |E +) = 0,91 - 0,12 = 0,79$$

El tratamiento de la HAI se basa es una estrategia farmacológica de inmunosupresión para controlar la naturaleza autoinflamatoria de la hepatitis. El primer escalón de tratamiento son los corticoides, que tienen efectos tóxicos y deletéreos cuando se emplean a largo plazo. Especialmente importantes son los efectos secundarios a nivel de crecimiento y mineralización ósea en niños [277]. Además, la decisión de interrumpir el tratamiento por completo se lleva a cabo en una proporción muy pequeña de niños debido a la alta tasa de recaídas, por lo que la gran mayoría de pacientes necesitan dosis bajas de prednisona o análogos a muy largo plazo [278]. Por ello, la utilidad de tratar a un sano se estimó en el opuesto del 100% de pacientes con efectos adversos a corticoides. Paralelamente, la utilidad de

no tratar a un sano se asumió del 100%. En consecuencia, el riesgo neto ( $R_t$ ) de un tratamiento inapropiado se calculó en 1.

$$R_t = U(T - |E -) - U(T + |E -) = 1 - 0 = 1$$

A partir de las anteriores asunciones y cálculos, se aplicaron las fórmulas de Djulbegovic y Desoky. Los umbrales de acción para cada criterio fueron los siguientes:

1) Umbral diagnóstico (UD)

$$UD_{2008} = 0,137$$

$$UD_{2009} = 0,053$$

2) Umbral diagnóstico-terapéutico (UDT)

$$UDT_{2008} = 0,804$$

$$UDT_{2009} = 0,929$$

**9.6.1.1.3. Efecto de la decisión clínica basada en los criterios simplificados: ¿Hay que tratar a este paciente?**

Los umbrales de acción delimitan los conjuntos de probabilidades preprueba para los que es conveniente aplicar el *test* diagnóstico a un paciente con el fin de tratar o no obedeciendo a su resultado. De forma parecida al área de indicación, cuanto más alejados estén los valores del umbral diagnóstico y el diagnóstico-terapéutico, más útil es una prueba diagnóstica, dado que es aplicable en escenarios clínicos más diversos por lo que respecta al riesgo apriorístico de enfermedad para un paciente del que no sabemos el resultado de ningún *test*. La diferencia principal respecto al área de indicación es que ambos umbrales de acción tienen en cuenta los efectos adversos de la aplicación de la prueba diagnóstica.

Existe otro parámetro computable a partir de las utilidades asumidas. Se trata del umbral terapéutico, que, según la teoría del análisis de decisiones, marca la probabilidad a partir de la que se debe de tratar a un paciente porque los beneficios

de hacerlo, aun asumiendo el riesgo de error tipo II, superan los inconvenientes de no hacerlo.

#### *9.6.1.1.3.1. Umbral terapéutico según el modelo de Pauker-Kassirer modificado por Latour*

Siguiendo las definiciones y fórmulas expuestas en el anexo 13.1, se calculó el umbral terapéutico para llevar a cabo el estudio de la utilidad de ambos criterios simplificados, siguiendo el modelo de Pauker-Kassirer modificado por Jaime Latour [279,280].

El umbral terapéutico no depende de los indicadores de validez de la prueba diagnóstica y, por lo tanto, es el mismo tanto para la versión de la IAIHG de 2008 como para la nueva propuesta de sistema diagnóstico por puntos basada en los criterios pediátricos de 2009.

$$\text{Umbral terapéutico} = \frac{R_t}{R_t + B_t} = \frac{1}{1 + 0,79} = 0,56$$

#### *9.6.1.2. Simulaciones de la toma de decisiones en varios contextos clínicos plausibles*

Se representó el nomograma de Fagan, considerando las razones de verosimilitud de un resultado positivo y uno negativo, de las dos versiones simplificadas de criterios diagnósticos de HAI estudiadas. El origen del trazo representa la probabilidad preprueba que se asume de HAI en la población de la que procede el paciente. La representación más adecuada para nuestro contexto clínico es aquella en la que el origen se sitúa en el valor de la prevalencia de HAI obtenida en el capítulo 1, por intención de diagnosticar. Sin embargo, esta asunción puede no ser verdad para poblaciones con prevalencias de HAI distintas en el grupo de niños con disfunción hepática sin antecedentes de trasplante hepático. Manteniendo las razones de verosimilitud calculadas, se adoptaron diversas probabilidades preprueba y se comprobó gráficamente a partir de cuáles se desplaza la probabilidad postprueba más allá del umbral terapéutico. Se estudió

específicamente si, con prevalencias inferiores de HAI a la de nuestro medio, la probabilidad postprueba supera dicho umbral, lo que se interpretaría como un buen indicador de utilidad de los criterios.

Se representó, asimismo, el gráfico de relación entre la probabilidad pre y postprueba (*probability-modifying plot*) en base a las razones de verosimilitud más robustas metodológicamente de cada criterio simplificado (los combinados por meta-análisis para los de 2008, y los obtenidos en la submuestra de validación para los de 2009). Sobre los mismos, se señaló el área de prevalencias o probabilidades preprueba incluidas entre los umbrales de acción y, también, el umbral terapéutico.

### **9.6.1.3. Estudio de la estabilidad diagnóstica de los criterios reducidos**

#### **9.6.1.3.1. Índice de reclasificación neta (*net reclassification improvement*)**

Los índices de predicción están adquiriendo una importancia creciente en la literatura médica. La predicción de un diagnóstico es un caso particular de la predicción de un pronóstico, que también se puede anticipar a través de ecuaciones de riesgo basadas en modelos de regresión, tal como se ha visto en el capítulo 4. Existe un interés específico en cómo se puede mejorar la predicción de un riesgo o un diagnóstico con el empleo de nuevos marcadores que puedan mejorar la información contenida en estas funciones de riesgo [281]. En efecto, gracias a los avances tecnológicos en la investigación básica, incluidas la genómica, la proteómica y las técnicas de imagen no invasivas, aparece la necesidad de evaluar la utilidad de un nuevo marcador para la toma de mejores decisiones, cuestión que ya se ha formulado formalmente en el caso de nuevos marcadores pronósticos para el riesgo cardiovascular [282]. En un sentido inverso, algunas técnicas estadísticas diseñadas para evaluar globalmente las ventajas que aportan nuevos marcadores, o criterios, en un sistema diagnóstico modelizado por regresión, pueden emplearse para estudiar la no inferioridad de la retirada o sustracción de una o varias variables del modelo convencional. Nos propusimos realizar el ejercicio de aplicar el cálculo del índice de reclasificación neta (*net reclassification improvement* o NRI) para dirimir si



la simplificación de los criterios clásicos mantiene realmente su capacidad diagnóstica, y si el hecho de no considerar la exclusión de la enfermedad de Wilson como uno de los criterios de 2008, confiere peor capacidad de acierto al modelo diagnóstico basado en los mismos respecto al basado en los nuevos criterios ESPGHAN/NASPGHAN 2009.

La reclasificación neta ha sido propuesta por Pencina *et al.*, que señalan que se debe de tener en cuenta solo la reclasificación en un sentido correcto, es decir, la recategorización como sano o enfermo, de un sano o un enfermo, respectivamente. Así, proponen que el NRI se calcule como la suma dos porcentajes. El de la diferencia entre el porcentaje de reclasificaciones correctas dentro del grupo de los enfermos con el nuevo modelo o sistema diagnóstico (enfermos dados como tales menos sanos dados como tales dividido entre el total de enfermos), y el de la diferencia entre el mismo porcentaje pero dentro del grupo de los no enfermos (no enfermos dados como tales menos enfermos dados como tales dividido entre el total de no enfermos) [283].

Predicted Risk With LVEF Alone, %	Predicted Risk With LVEF Plus Midwall Fibrosis Status, %		Total
	0-15	>15	
<b>Patients With Arrhythmic Event</b>			
Predicted risk with LVEF alone			
0-15	12	23	35
>15	11	19	30
<b>Total</b>	<b>23</b>	<b>42</b>	<b>65</b>
<b>Patients Without Arrhythmic Event</b>			
Predicted risk with LVEF alone			
0-15	218	46	264
>15	89	54	143
<b>Total</b>	<b>307</b>	<b>100</b>	<b>407</b>

Figura 64: Tabla de resultados para el ejemplo del cálculo del índice de reclasificación neta. Reproducido de Gulati *et al.* JAMA. 2013;309:905. Con permiso de la American Medical Association.

El cálculo se entiende mejor con un ejemplo real, obtenido de un trabajo de 2013, por Gulati *et al.*, sobre el efecto predictor para futuros eventos de arritmia,

más allá del que anticipa la fracción de eyección del ventrículo izquierdo, de la fibrosis de septo ventricular en casos de cardiomiopatía dilatada no isquémica [284]. En este caso, atendiendo a la información proporcionada en la figura anterior, el NRI se expresa como:

$$NRI = \left[ \frac{23 - 11}{65} \right] + \left[ \frac{89 - 46}{407} \right] = 18\% + 11\% = 29\%$$

Conviene señalar que la interpretación del NRI no es equivalente a la del porcentaje neto de pacientes correctamente reclasificados. Se ha señalado que esta es una confusión habitual que conviene evitar [285]. En el ejemplo, se comprueba que este porcentaje se calcularía siguiendo la siguiente formulación:

$$\frac{[(23 - 11) + (89 - 46)]}{65 + 407} = 12\%$$

Como resultado, el NRI puede ser un estadístico de interpretación poco intuitiva, cuyo principal valor sería comparar el efecto pronóstico extra que varias variables independientes extra tienen sobre un mismo modelo clásico [286].

Aun así, llevamos a cabo el cálculo del NRI sobre los modelos de regresión con intención predictiva para el diagnóstico de HAI. Las variables extra analizadas en nuestro caso fueron, en una primera aproximación, las de los criterios clásicos que no se incluyen ni en los criterios simplificados de 2008 ni en los criterios pediátricos: sexo femenino, ingesta de alcohol, presencia de AMA y relación FA/AST. En un segundo análisis, se comprobó el efecto de la adición del descarte de la enfermedad de Wilson sobre los criterios simplificados de la IAIHG.

#### **9.6.1.3.2. Análisis por curvas de decisión (*decision curve analysis*)**

En la metodología específica del presente capítulo se ha hecho un esfuerzo por definir los umbrales de acción y el umbral terapéutico para la aplicación de los criterios diagnósticos simplificados. A la luz de la información más actualizada que se

dispone, y algunas asunciones tomadas por convención de forma razonada, los umbrales han quedado definidos en los valores de probabilidad de enfermedad ya expuestos. Sin embargo, el punto de corte para la aplicación clínica de un modelo de predicción diagnóstica a menudo no se puede definir de manera precisa. La ponderación relativa de daños y beneficios puede no ser conocida a causa de falta de datos científicos o debido a apreciaciones diferentes de los médicos y los pacientes. Por este motivo, Vickers y Elkin han propuesto utilizar una gamma de valores de corte y calcular el beneficio neto para cada valor. El resultado puede representarse gráficamente en una o varias curvas de decisión [287].

El análisis por curvas de decisión evalúa modelos predictivos incorporando las consecuencias del diagnóstico, que es la diferencia esencial con la representación del área bajo la curva COR. Se puede aplicar directamente a los datos de una muestra de validación y no requiere la recogida de información extra para cada caso. Es, por tanto, una herramienta factible de aplicar para el estudio de la utilidad de los criterios diagnósticos de HAI. Dado que se ha desarrollado para las predicciones realizadas por modelos de regresión logística, se estimó la probabilidad de HAI en cada caso a partir de las ecuaciones de regresión. El resultado final es un valor de 0 a 1 para cada paciente de la submuestra de validación, y para cada sistema diagnóstico estudiado. Con el fin de asegurar la comparabilidad se calculó para sendos modelos basados en los criterios clásicos de 1999 pre-tratamiento, los criterios simplificados de 2008 y el nuevo sistema basado en los criterios pediátricos de 2009. Así, se evitó el uso de la puntuación de cada sistema diagnóstico (en números enteros de escalas diferentes) en favor de una escala normalizada, en la que 0 representa la ausencia de HAI y 1, su presencia.

En una curva de decisión, el eje horizontal representa el umbral de probabilidad en tanto por ciento. En el caso de la evaluación de un modelo diagnóstico cuya aplicación tiene derivadas en la decisión de tratar o no a un paciente, es lógico asimilar las ordenadas al umbral terapéutico. En las abscisas se

representa el beneficio neto del tratamiento que, según la fórmula propuesta por Peirce, se calcula del siguiente modo [288]:

$$\text{Beneficio neto} = \frac{\text{Verdaderos positivos}}{n} - \frac{\text{Falsos positivos}}{n} \times \left( \frac{U_t}{1 - U_t} \right) - R_d^*$$

Donde  $n$  es el tamaño muestral,  $R_d^*$  es el riesgo de la prueba diagnóstica y  $U_t$  es el umbral terapéutico, es decir, el punto de corte de la probabilidad diagnóstica predicha por el modelo de regresión a partir del cual diagnosticaríamos al paciente con suficiente seguridad como para tratarlo porque los beneficios de este gesto clínico superarían los inconvenientes. Según Peirce,  $R_d^*$  es la estimación “holística” de las consecuencias negativas de tener que aplicar la prueba diagnóstica (coste, efectos secundarios, inconveniente subjetivo para el paciente...) en las unidades de un resultado verdadero positivo [288]. Por tanto, no es igual al riesgo de la prueba para el cálculo de los umbrales de acción. En este caso, cabría decidir cuántas veces más es peor dejarse un paciente enfermo por diagnosticar que el riesgo implícito en la prueba diagnóstica. En consecuencia, y para mantener la coherencia con las estimaciones previas de los umbrales de acción, calculamos  $R_d^*$  como el cociente entre el riesgo de la prueba diagnóstica ( $R_d$ ) y el beneficio del tratamiento:

$$R_d^* = \frac{R_d}{U(T + |E +) - U(T - |E +)} = \frac{0,04}{0,91 - 0,12} = 0,05$$

Dado que el riesgo de la prueba, equiparado al riesgo de complicaciones de la biopsia hepática, es relativamente bajo en relación con el peso que se le puede asignar a un falso negativo,  $R_d^*$  es esencialmente despreciable en la fórmula de Peirce en nuestro caso.

El ancho de posibles valores del beneficio neto para el análisis por curvas de decisión va desde  $-\infty$  hasta la prevalencia de la enfermedad en la muestra de validación.

El resultado es una gráfica como la de la figura, en la que el modelo diagnóstico que mejor se comporta para cada umbral terapéutico es el representado por el trazo que sigue un recorrido más alto en el espacio beneficio neto–umbral de probabilidad. Aunque se esté seguro del beneficio y el riesgo asumidos para un tratamiento, representar las curvas de decisión permite comparar los beneficios diagnósticos de varios modelos para umbrales de probabilidad diferentes.

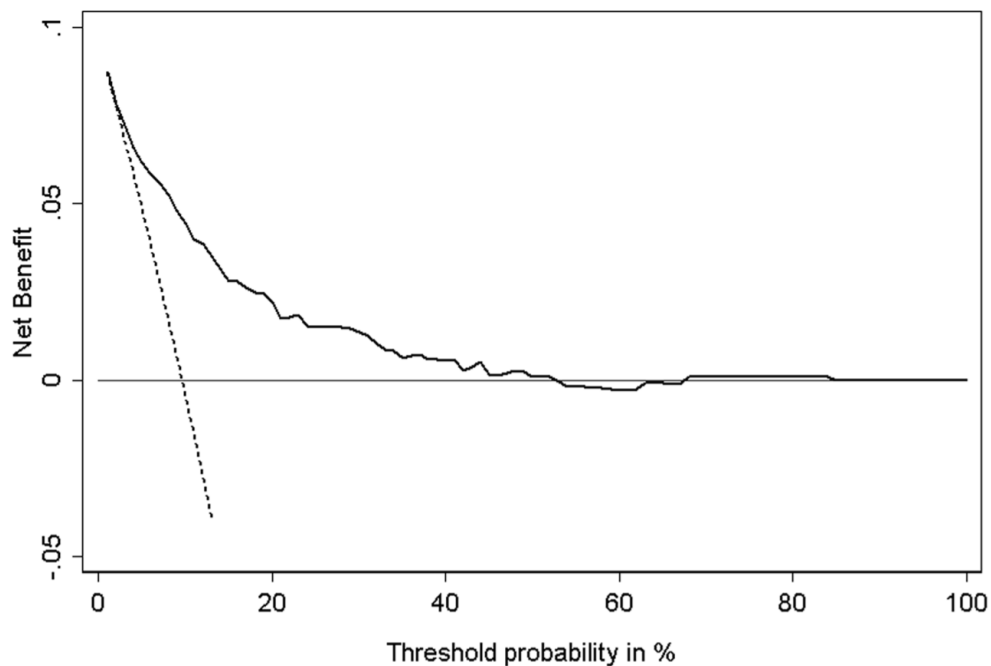


Figura 65: Curva de decisión de ejemplo de un modelo de predicción de la invasión de vesículas seminales (IVS) en el cáncer de próstata. La línea de puntos asume que todos los pacientes tienen IVS. El gráfico representa el beneficio neto esperado por paciente, relativo a no extirpar las vesículas seminales en ningún caso que se opera por cáncer de próstata. La unidad es el beneficio neto asociado a un paciente con IVS y vesículas seminales debidamente extirpadas. Reproducido de Vickers y Elkin. *Med Decis Making*. 2008;26;569. Con permiso de SAGE Journals.

Tal como explican Vickers y Elkin, llevamos a cabo una representación gráfica de las curvas de decisión mediante la repetición de los siguientes pasos para diferentes valores de  $U_t$  [289]:

- 1- Calcular el número de verdaderos y falsos positivos empleando  $U_t$  como el punto de corte del resultado de la ecuación del modelo diagnóstico que separa considerar un caso como negativo o positivo.

- 2- Calcular el beneficio neto.
- 3- Variar  $U_t$  en un ancho apropiado de valores y repetir los pasos 1 y 2.
- 4- Representar  $U_t$  en el eje  $x$  y el beneficio neto en el eje  $y$ .
- 5- Repetir los pasos 1 a 4 para cada modelo diagnóstico que interese analizar o comparar.
- 6- Trazar una recta paralela al eje  $x$  en el valor  $y=0$  que represente el beneficio neto asociado a la estrategia de asumir que todos los pacientes son negativos o no casos.

Por inferencia bayesiana, simulamos los indicadores de validez de cada sistema de criterios diagnósticos para una población con una prevalencia de 47,2%. Se aprecia en las curvas de decisión que las líneas que representan las estrategias de tratar a todos los pacientes y no tratar a ninguno se cruzan en el umbral terapéutico cuyo valor es igual al de la prevalencia de la enfermedad de interés.

En vista de la futura discusión de los resultados, conviene señalar que un modelo diagnóstico realmente útil mantendrá beneficios netos elevados en todo el ancho de valores posibles de umbrales terapéuticos. Por el contrario, para los umbrales (o situaciones clínicas) en los que el beneficio neto es próximo a 0, la prueba diagnóstica no es útil porque su resultado no se traduce en una decisión terapéutica racional.

A modo de ejemplo, en la figura siguiente, la prevalencia de la enfermedad es del 20%. La línea fina horizontal asume que ningún paciente tiene la enfermedad de interés. La línea de puntos asume que todos los pacientes tienen la enfermedad. La línea de trazo grueso horizontal representa un modelo de predicción diagnóstica perfecto, cuya aplicación arroja un beneficio neto máximo en cualquier contexto clínico. Finalmente, se representan otras dos líneas para modelos diagnósticos realistas: la gris continua simula una prueba con un 99% de sensibilidad y un 50% de especificidad y la discontinua, una con un 50% de sensibilidad y un 99% de especificidad.

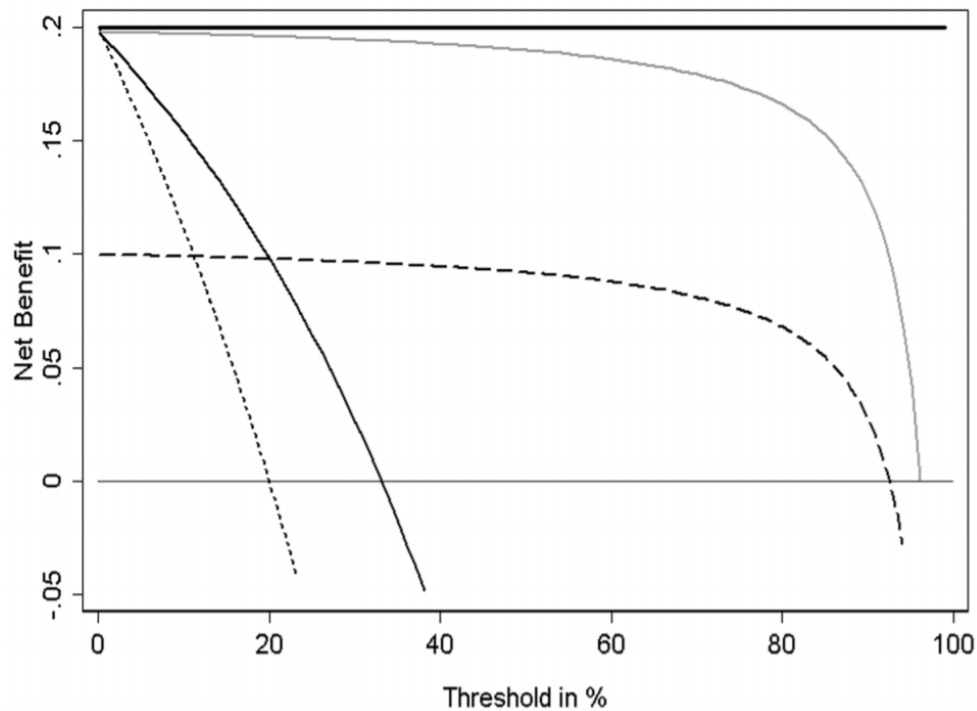


Figura 66: Curvas de decisión para una distribución teórica con una prevalencia del 20%. Reproducido de Vickers y Elkin. *Med Decis Making*. 2008;26;571. Con permiso de SAGE Journals.

El análisis por curvas de decisión se ha recomendado claramente en algunas editoriales de revistas médicas de impacto, como complemento a la representación de la curva COR del modelo, porque añade información relevante sobre el beneficio clínico de la toma de decisiones para cada punto de corte de probabilidad predicha por el sistema diagnóstico [290,291].

### 9.6.2. Recursos para los análisis

El estudio de la utilidad de los dos criterios simplificados se llevó a cabo con la calculadora de la red CASP (*critical appraisal skills programme*) para análisis decisional, creada por Joaquín Primo y actualizada en noviembre de 2015 [292]. Se trata de una hoja de cálculo para *Excel* que se abrió y modificó en una versión de 2010 del programa (Microsoft, Redmond, WA). Está libre para descarga desde la URL <http://www.redcaspe.org/herramientas/calculadora> y el acceso para la misma se

efectuó en julio de 2017. Los cálculos de los umbrales de acción y el umbral terapéutico llevados a cabo con la calculadora coincidieron con los realizados manualmente. Se reprodujeron los gráficos que correlacionan la probabilidad de enfermedad con las utilidades de tratamiento, de no tratamiento y de la prueba, en una escala porcentual.

Adicionalmente, se recurrió a la ejecución de los comandos *Metandi*, *Midas* y *Fagan*, en el entorno estadístico Stata (versión 14.0, *Stata Corporation*, Texas, EEUU) para la representación del nomograma de Fagan y el *probability-modifying plot* [256,293–295]. Por último, el cálculo del índice de reclasificación neta y el análisis por curvas de decisión se ejecutó a partir de la sintaxis de los comandos *Nri*, *Adpred* y *Dca*, también en Stata [296].

### 9.6.3. Resultados

Se presenta, en primer lugar, el análisis de los umbrales de probabilidad según el modelo de Pauker-Kassirer para el empleo de los criterios simplificados de 2008 [279,297,298]. Se han tenido en cuenta las consideraciones expuestas previamente sobre sus utilidades en un escenario clínico real y con la prevalencia estimada de 47,2% de HAI. En adelante, “enfermo” se refiere a un caso con HAI. En consonancia, “sano” es un paciente con un diagnóstico alternativo a HAI.

Tabla 45: Análisis de la utilidad de los criterios simplificados de 2008. Estimaciones numéricas sobre una base 100.

Utilidades	De no tratar	De tratar	
A un sano	100	0	→ $[(1 - E) \times U(T+   E-)] + [E \times U(T-   E-)] + R_d = 95$
A un enfermo	12	91	→ $[(S \times U(T+   E+)] + [(1 - S) \times U(T-   E+)] + R_d = 73$

E: Especificidad. S: Sensibilidad. U: Utilidad. R<sub>d</sub>: Riesgo de la prueba.

El beneficio y el coste o riesgo del tratamiento serían iguales tanto para los criterios simplificados de 2008 como para el nuevo sistema de puntos basado en los



criterios pediátricos de 2009. Se han calculado con anterioridad y se han establecido en 0,79 y en 1, respectivamente.

La prevalencia de HAI en el perfil de paciente pediátrico con hepatopatía definido por los criterios de inclusión y exclusión (no antecedentes de interés e hipertransaminasemia) queda entre los umbrales diagnóstico (*UD*) y diagnóstico-terapéutico (*UDT*) de los criterios simplificados. En efecto, con una probabilidad preprueba de 47,2%, la decisión lógica es plantear el tratamiento de HAI en función del resultado de los criterios de 2008, dado que por encima de 13,7% (*UD*<sub>2008</sub>) el sistema de puntos concebido como herramienta diagnóstica empieza a ser rentable en términos de utilidad clínica. Esta rentabilidad se pierde por encima de una probabilidad preprueba de 80,4% (*UDT*<sub>2008</sub>), a partir de la cual se debería de tratar el paciente sin necesidad de aplicar los criterios simplificados.

La otra lectura de esta situación es que los criterios serían igualmente útiles en poblaciones con una prevalencia bastante diferente de HAI, dentro de los márgenes delimitados por los umbrales de acción (de 13,7 a 80,4%).

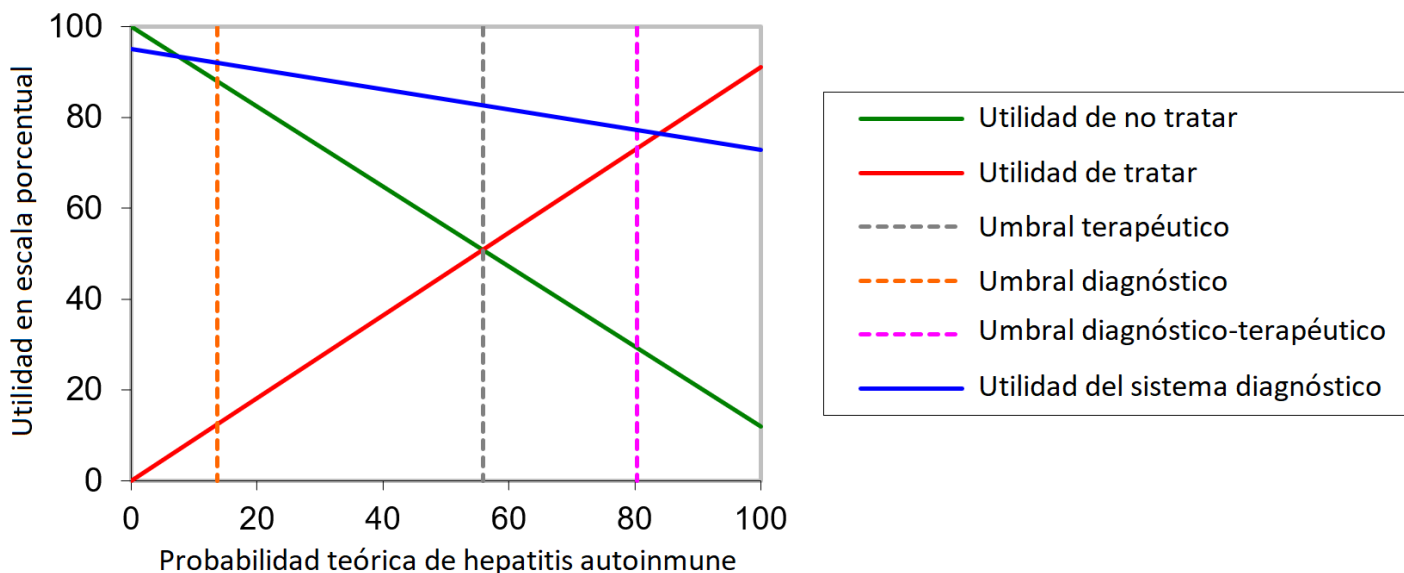


Figura 67: Representación gráfica del cambio de las utilidades de los criterios simplificados de 2008 en función de la probabilidad teórica de HAI.

Gráficamente, se aprecia como el umbral terapéutico representa la probabilidad teórica de enfermedad a partir de la cual la utilidad de tratar supera a la de no tratar (56%). Sin tener en cuenta el riesgo inherente a la aplicación del sistema diagnóstico, el umbral diagnóstico sería la probabilidad de enfermedad en la que se cruzan las utilidades de la prueba y la de no tratar. Cuando la segunda es superior, no tiene sentido decidir aplicar la prueba. Análogamente, el umbral diagnóstico-terapéutico se localiza en el punto en el que la utilidad de tratar supera la propia del sistema diagnóstico. Con probabilidades de enfermedad superiores, habría que tratar directamente al paciente sin necesidad de emplear la prueba. El riesgo del *test* corrige la posición de los umbrales de acción acortando la distancia entre ellos. Los cálculos descritos en la tabla anterior arrojan los puntos de intersección de la utilidad de la prueba con la probabilidad absoluta y nula de hepatitis autoinmune. En el caso de los criterios simplificados de 2008, el segmento que representa la utilidad de estos va desde 95% a 73%. Esta propiedad de la recta de la utilidad de los criterios es de comprensión poco intuitiva, No obstante, observar que el trazo discurre por una posición superior en el espacio de utilidades y probabilidades teóricas de enfermedad permite visualizar mejor el concepto de umbrales de acción. Se comprende fácilmente la necesidad de que todas las utilidades estén medidas en las mismas unidades observando la lógica reflejada en el gráfico.

Se aplicó el nomograma de Fagan para estudiar la relación de los valores predictivos de HAI de los criterios de 2008 respecto al umbral terapéutico, según la propuesta de Pauker-Kassirer y Latour. La situación clínica que representa constituye un ejemplo de prueba muy adecuada para la toma de decisiones terapéuticas. La prevalencia de la enfermedad (47,2%) se sitúa en las inmediaciones del umbral calculado por la fórmula de Djulbegovic y Desoky (56%). Con la información que añade el resultado del sistema simplificado de puntos de la IAHG, la probabilidad de enfermedad se aleja del umbral terapéutico en un sentido ascendente (con

puntuaciones diagnósticas de HAI), o en un sentido descendente (con puntuaciones no diagnósticas de HAI).

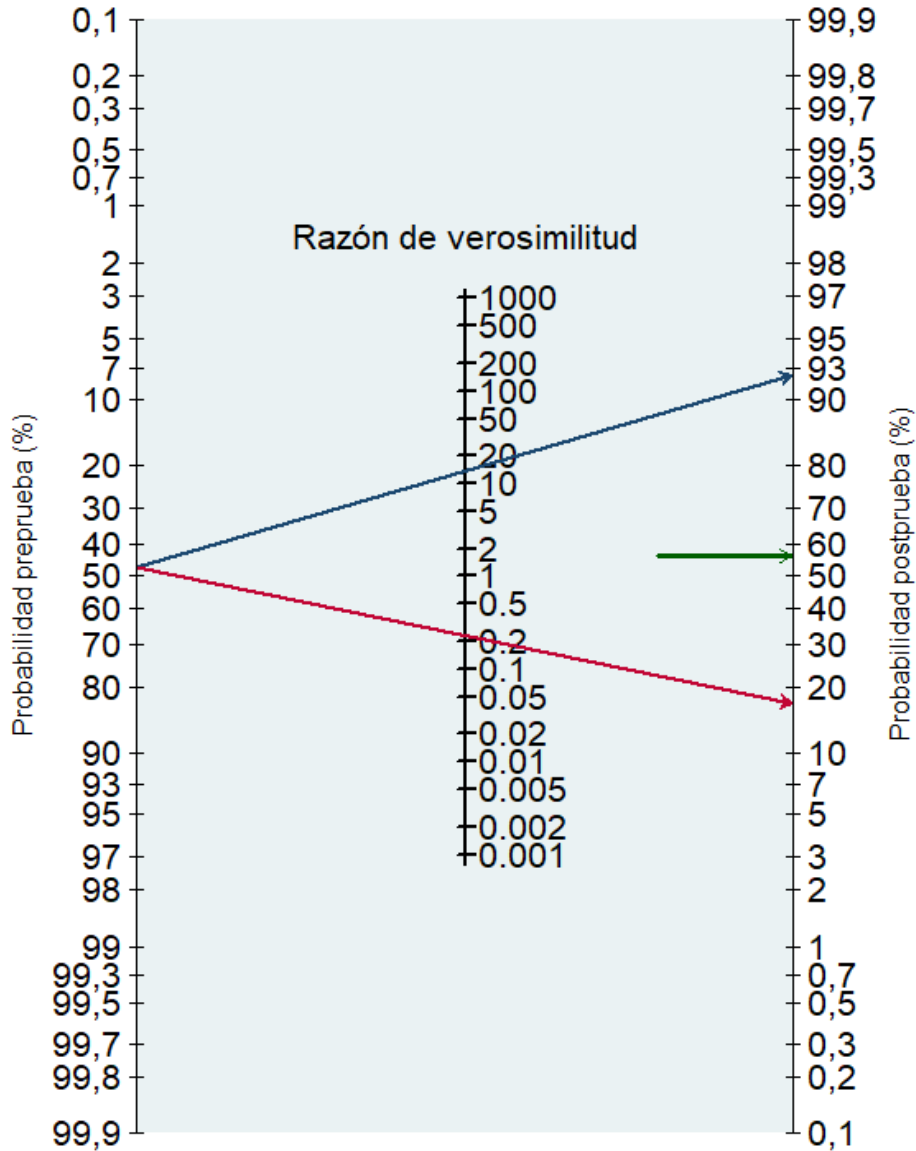


Figura 68: Nomograma de Fagan para los criterios simplificados de 2008. Las razones de verosimilitud positiva y negativa empleadas son 13,72 y 0,23. Prevalencia asumida de hepatitis autoinmune: 47,2%. La flecha azul representa el cambio en la probabilidad de enfermedad con 6 o más puntos en el sistema. La flecha roja, el cambio con un resultado inferior. La línea verde señala el umbral terapéutico según la fórmula de Djulbegovic y Desoky para el modelo de Pauker-Kassirer (56%).

Este comportamiento continuaría siendo válido para prevalencias de HAI entre 8,5% y 84,7%. Más allá de este ancho de valores, si aplicáramos la prueba en

un entorno clínico con una probabilidad *a priori* de HAI de menos de 8,5%, un resultado positivo de los criterios de 2008 no conseguiría que la probabilidad postprueba de HAI superase el umbral terapéutico y, por lo tanto, su lectura no añadiría información suficiente a un caso concreto como para tomar la decisión de tratar al paciente. En el otro extremo del nomograma, con prevalencias de HAI de superiores a 84,7%, un resultado negativo tampoco permitiría decidir racionalmente no tratar al paciente.

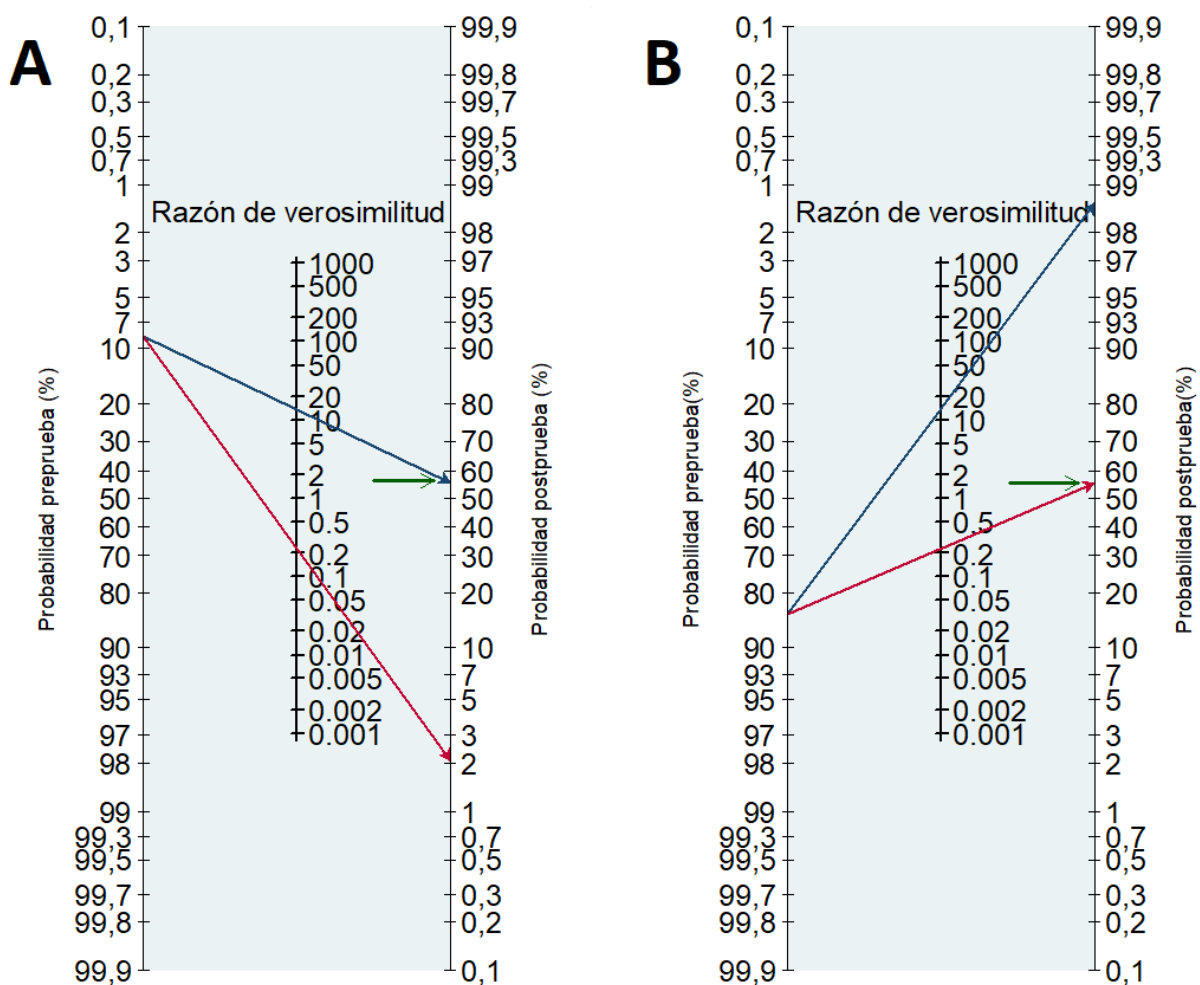


Figura 69: Simulaciones con el nomograma de Fagan para los criterios simplificados de 2008 con las razones de verosimilitud positiva y negativa de 13,72 y 0,23. Panel A: Prevalencia asumida de hepatitis autoinmune: 8,5%. Panel B: Prevalencia asumida de hepatitis autoinmune: 84,7%. La flecha azul representa el cambio en la probabilidad de enfermedad con 6 o más puntos en el sistema. La flecha roja, el cambio con un resultado inferior. La línea verde señala el umbral terapéutico según la fórmula de Djulbegovic y Desoky para el modelo de Pauker-Kassirer (56%).

Como complemento a la exploración de la relación entre la probabilidad preprueba y postprueba de HAI con los criterios simplificados, se trazó el *probability-modifying plot*, indicando los umbrales de acción en el eje de las prevalencias y el umbral terapéutico en el eje de los valores predictivos. Permite visualizar, igual que con el nomograma de Fagan, las prevalencias a partir de las que un resultado positivo o negativo cruzan el umbral terapéutico (8,5% y 84,7% respectivamente). Se observa que estas prevalencias no coinciden exactamente con las de los umbrales de acción, que vienen definidas, no solo por la capacidad de acierto diagnóstico del sistema en cuestión, sino por la efectividad del tratamiento disponible, la historia natural de la enfermedad y el riesgo de la prueba.

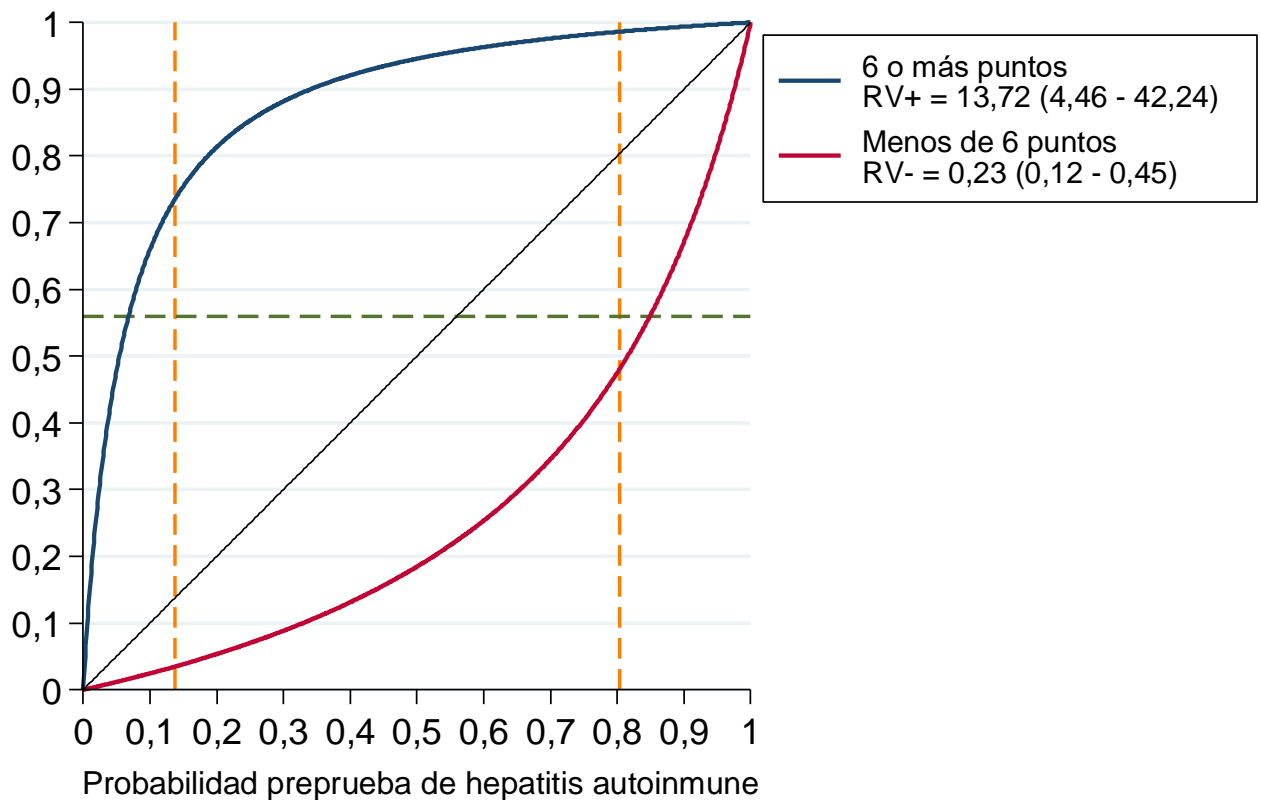


Figura 70: Gráfico de modificación probabilística (*probability-modifying plot*) de la enfermedad construido con las razones de verosimilitud del meta-análisis de la validez de los criterios simplificados de 2008. La marca horizontal verde señala el umbral terapéutico (56%), mientras que las marcas verticales amarillas indican los umbrales de acción (diagnóstico: 13,7% y diagnóstico-terapéutico: 80,4%).

Respecto a la utilidad del nuevo sistema diagnóstico basado en los criterios ESPGHAN/NASPGHAN 2009, se obtuvieron los resultados que se describen seguidamente.

Tabla 46: Análisis de la utilidad de los nuevos criterios pediátricos de 2009. Estimaciones numéricas sobre una base 100.

Utilidades	De no tratar	De tratar	
<b>A un sano</b>	100	0	$\rightarrow [(1 - E) \times U(T+   E-)] + [E \times U(T-   E-)] + R_d = 100$
<b>A un enfermo</b>	12	91	$\rightarrow [(S \times U(T+   E+)] + [(1 - S) \times U(T-   E+)] + R_d = 88$

E: Especificidad. S: Sensibilidad. U: Utilidad. R<sub>d</sub>: Riesgo de la prueba.

Dado que las utilidades de tratar y no tratar a un sano y a un enfermo dependen de la enfermedad y sus posibilidades terapéuticas, al evaluarse dos criterios diagnósticos para una misma situación clínica, se mantienen invariables. Tan solo cambian los parámetros de la utilidad de la prueba, que en este caso son 100% y 88%. Asimismo, también hay que tener en cuenta la posición concreta de los umbrales de acción para los nuevos criterios:  $UD_{2009} = 5,3\%$  y  $UDT_{2009} = 92,9\%$ .

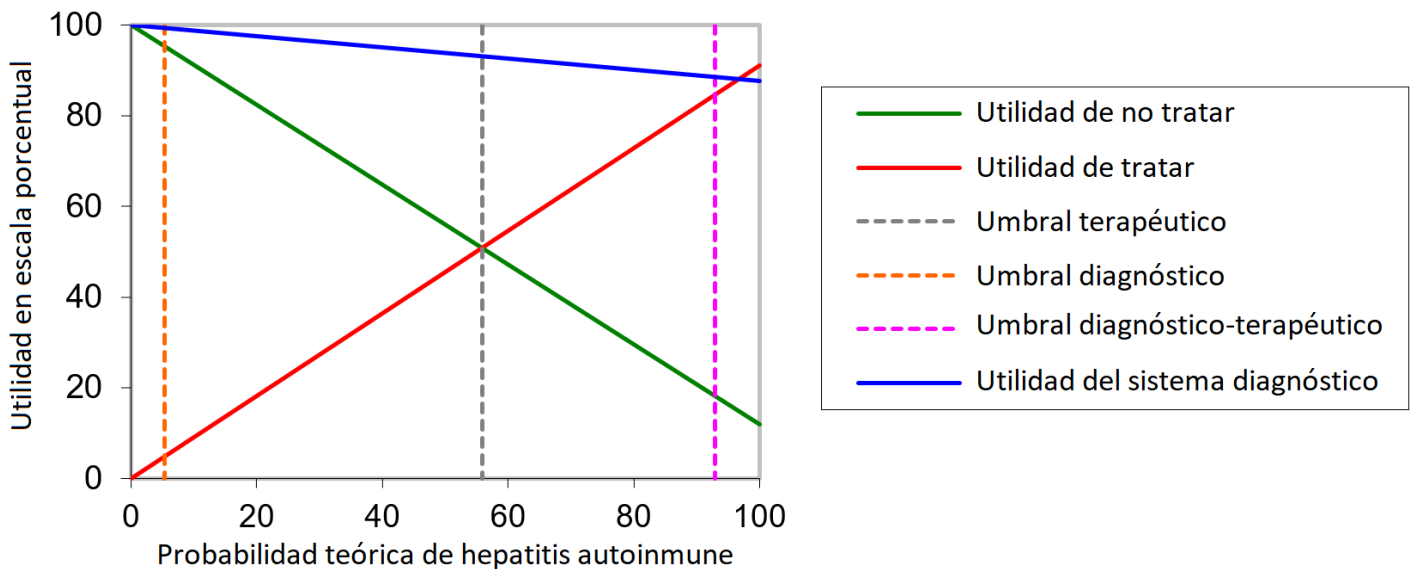


Figura 71: Representación gráfica del cambio de las utilidades del nuevo sistema diagnóstico por puntos basado en los criterios ESPGHAN/NASPGHAN 2009 en función de la probabilidad teórica de HAI.

El resultado es un espacio de utilidades y probabilidades teóricas de HAI en el que el ancho entre los umbrales de acción abarca un sector más grande o, lo que es lo mismo, que los nuevos criterios aportan información clínicamente relevante en más situaciones clínicas (por lo que respecta a la probabilidad previa de HAI) que los criterios simplificados de 2008.

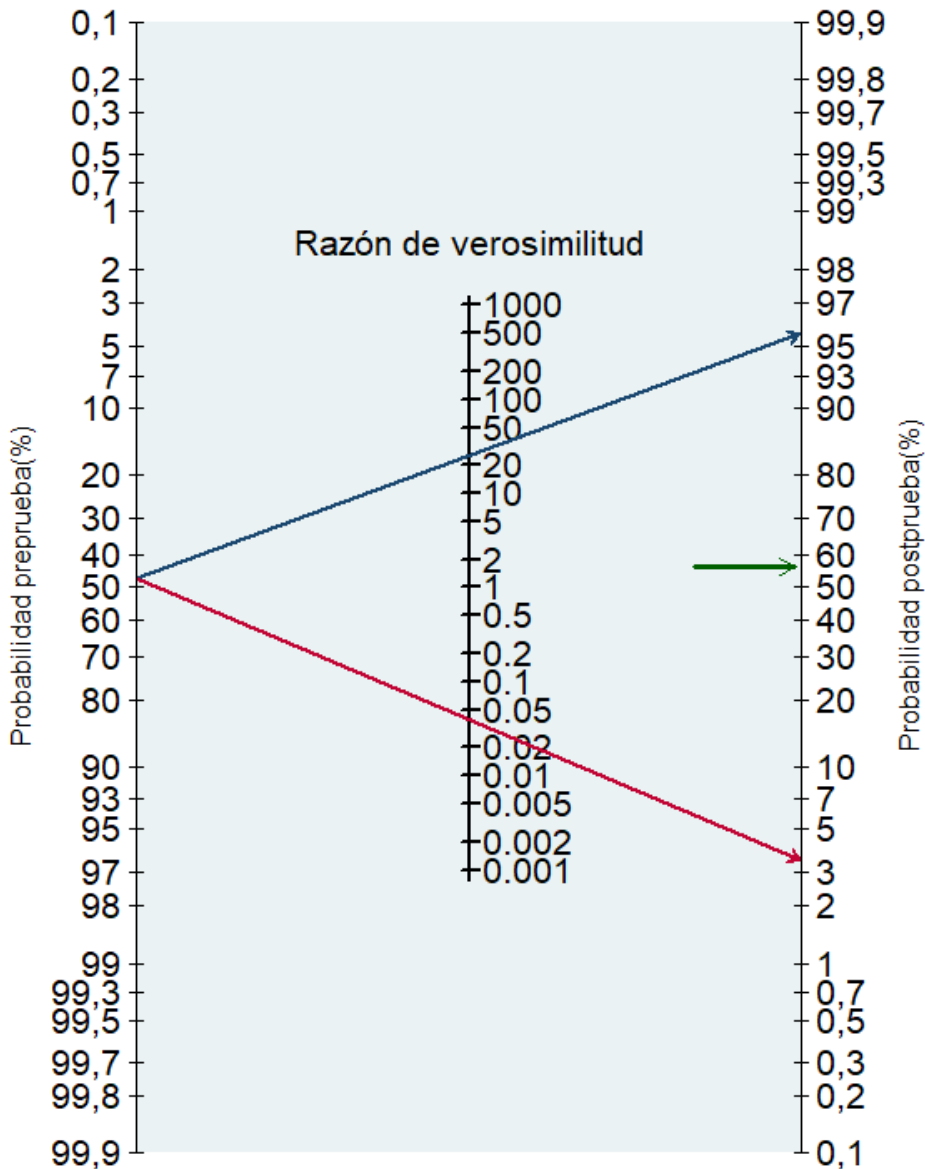


Figura 72: Nomograma de Fagan para los nuevos criterios diagnósticos ESPGHAN/NASPGHAN 2009. Las razones de verosimilitud positiva y negativa empleadas son 24,9 y 0,04. Prevalencia asumida de hepatitis autoinmune: 47,2%. La flecha azul representa el cambio en la probabilidad de enfermedad con 6 o más puntos en el sistema. La flecha roja, el cambio con un resultado inferior. La línea verde señala el umbral terapéutico según la fórmula de Djulbegovic y Desoky para el modelo de Pauker-Kassirer (56%).

La posición relativa de la prevalencia de HAI (47,2%) y el umbral terapéutico (56%) se mantiene, pero los valores predictivos positivo y negativo del nuevo sistema de puntos basado en los criterios de 2009 se alejan sensiblemente en virtud de unas razones de verosimilitud que traducen una mayor capacidad informativa.

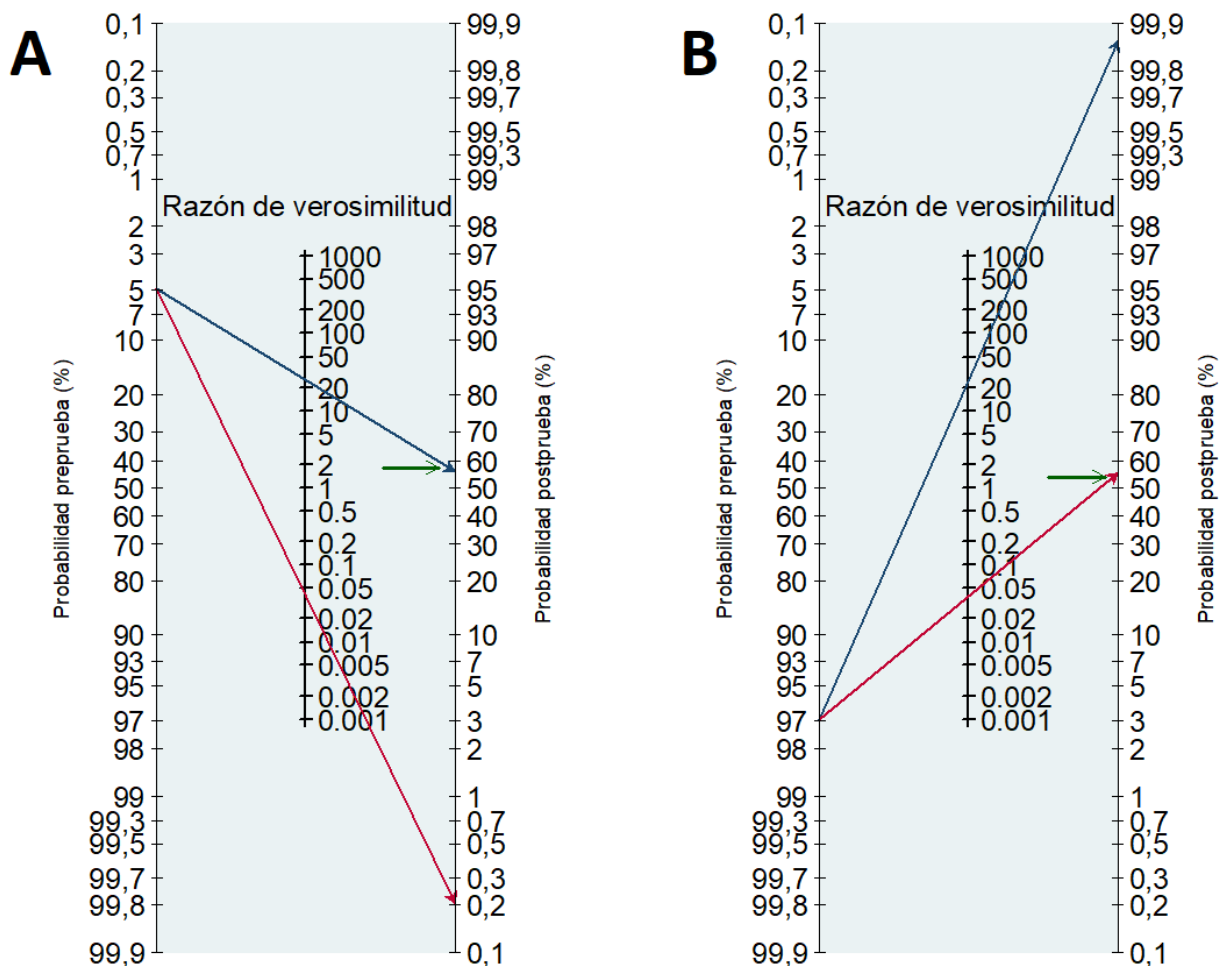


Figura 73: Simulaciones con el nomograma de Fagan para los nuevos criterios ESPGHAN/NASPGHAN 2009, transformados en un sistema diagnóstico por puntos, con las razones de verosimilitud positiva y negativa de 24,9 y 0,04. Panel A: Prevalencia asumida de hepatitis autoinmune: 4,9%. Panel B: Prevalencia asumida de hepatitis autoinmune: 96,9%. La flecha azul representa el cambio en la probabilidad de enfermedad con 6 o más puntos en el sistema. La flecha roja, el cambio con un resultado inferior. La línea verde señala el umbral terapéutico según la fórmula de Djulbegovic y Desoky para el modelo de Pauker-Kassirer (56%).

Con la mayoría de prevalencias de HAI, la aplicación de los nuevos criterios implica una decisión terapéutica porque se cruza el umbral terapéutico en alguno de



los dos posibles sentidos (a favor o en contra del tratamiento). Solo con prevalencias inferiores a 4,9% o superiores a 96,9% este fenómeno deja de observarse. En efecto, en el primer caso, ni tan siquiera un resultado positivo de los nuevos criterios hace que el valor predictivo positivo supere el umbral de 56%. De igual forma, en contextos con una probabilidad teórica de HAI de más de 96,9%, un resultado negativo de los criterios no rebaja la probabilidad postprueba de HAI por debajo del umbral terapéutico. El ancho de prevalencias posible que abarcan estos dos límites es del 92%, lo que otorga al nuevo sistema diagnóstico una capacidad de resistencia importante a errores en el índice de sospecha inicial de HAI.

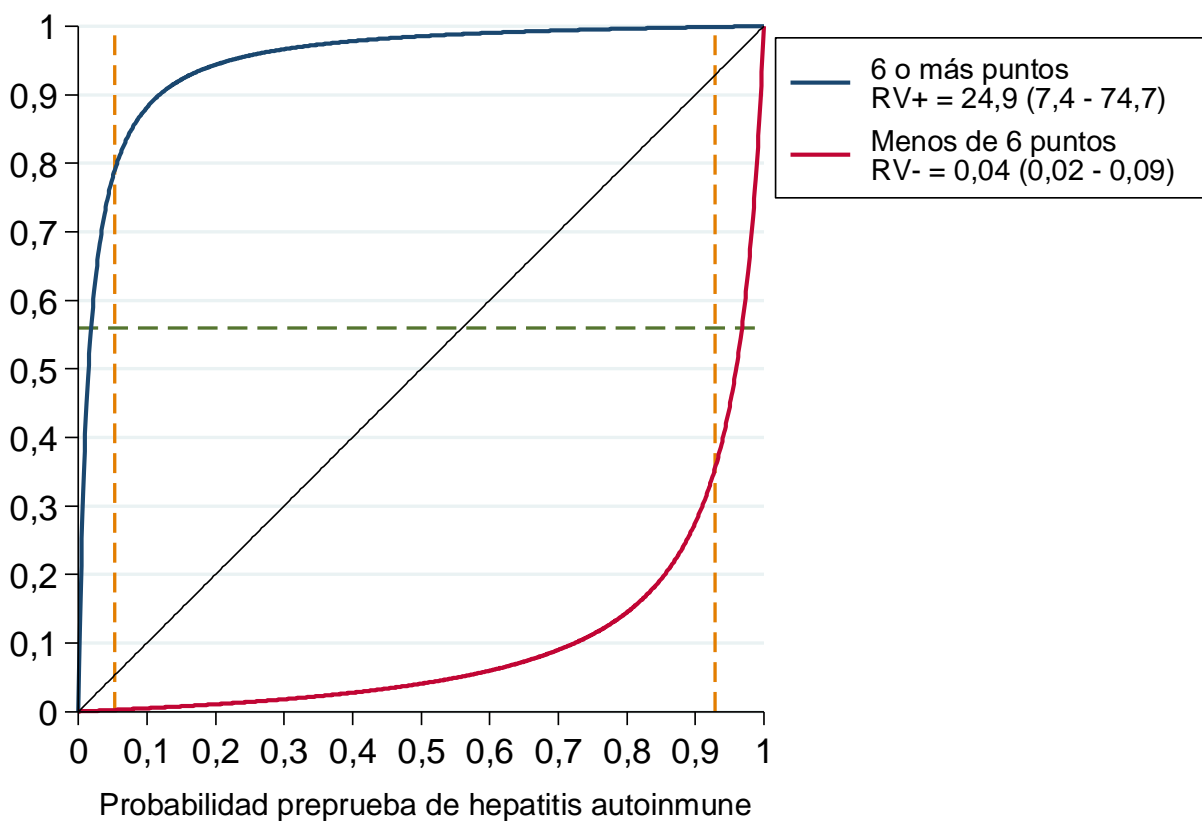


Figura 74: Gráfico de modificación probabilística (*probability-modifying plot*) de la enfermedad construido con las razones de verosimilitud obtenidas en la validación externa del nuevo sistema diagnóstico por puntos basado en los criterios pediátricos 2009 de hepatitis autoinmune. La marca horizontal verde señala el umbral terapéutico (56%), mientras que las marcas verticales amarillas indican los umbrales de acción (diagnóstico: 5,3% y diagnóstico-terapéutico: 92,9%).

Con el fin de estudiar la contribución individual de los criterios extra de la versión clásica de la IAIHG, sobre los incluidos en la versión simplificada, se llevó a cabo el cálculo de los índices de reclasificación neta (NRI).

**Tabla 47: Evaluación a través del índice de reclasificación neta (NRI) de la contribución de la variable sexo (masculino o femenino) al modelo con los criterios simplificados de 2008.**

Reclasificaciones predichas por el modelo con los criterios simplificados IAIHG 2008 + Sexo	En contra de HAI	A favor de HAI
En pacientes con HAI	28,0%	72,0%
En pacientes sin HAI	58,0%	42,0%
<b>NRI</b>	<b>60,1%</b>	<b>Valor P &lt;0,0001</b>

HAI: Hepatitis autoinmune. NRI: *Net reclassification improvement*.

**Tabla 48: Evaluación a través del índice de reclasificación neta (NRI) de la contribución de la variable alcohol (<25 o >60 g/día) al modelo con los criterios simplificados de 2008.**

Reclasificaciones predichas por el modelo con los criterios simplificados IAIHG 2008 + Alcohol	En contra de HAI	A favor de HAI
En pacientes con HAI	0,0%	0,0%
En pacientes sin HAI	0,0%	0,0%
<b>NRI</b>	<b>0,0%</b>	<b>-</b>

**Tabla 49: Evaluación a través del índice de reclasificación neta (NRI) de la contribución de la variable AMA (sí o no) al modelo con los criterios simplificados de 2008.**

Reclasificaciones predichas por el modelo con los criterios simplificados IAIHG 2008 + AMA	En contra de HAI	A favor de HAI
En pacientes con HAI	10,0%	90,0%
En pacientes sin HAI	10,7%	89,3%
<b>NRI</b>	<b>1,4%</b>	<b>Valor P = 0,917</b>

AMA: Anticuerpos antimitocondriales.

**Tabla 50: Evaluación a través del índice de reclasificación neta (NRI) de la contribución de la variable FA/AST (<1,5, de 1,5 a 3 o >3) al modelo con los criterios simplificados de 2008.**

Reclasificaciones predichas por el modelo con los criterios simplificados IAIHG 2008 + FA/AST	En contra de HAI	A favor de HAI
En pacientes con HAI	27,1%	72,9%
En pacientes sin HAI	64,3%	35,7%
<b>NRI</b>	<b>74,4%</b>	<b>Valor P &lt;0,0001</b>

FA/AST: Relación fosfatasa alcalina/alanina aminotransferasa.

Las variables *sexo* y *FA/AST*, al incluirlas en un modelo de regresión con intención predictiva del diagnóstico de HAI a partir de los criterios simplificados de 2008, favorecen el número de clasificaciones correctas. Esta observación es imposible de cuantificar en términos absolutos a través del NRI de una forma comprensible, pero los valores obtenidos sugieren que *FA/AST* se comporta discretamente mejor que *sexo*. Los resultados del NRI con las variables *AMA* y *alcohol* apoyan el hecho de no considerarlas en unos criterios diagnósticos aplicables a población pediátrica.

**Tabla 51: Evaluación a través del índice de reclasificación neta (NRI) de la contribución de la variable descarte de enfermedad de Wilson (sí o no) al modelo con los criterios simplificados de 2008.**

Reclasificaciones predichas por el modelo con los criterios simplificados IAIHG 2008 + EW		En contra de HAI	A favor de HAI
En pacientes con HAI		2,0%	98%
En pacientes sin HAI		26,8%	73,2%
	<b>NRI</b>	<b>49,6%</b>	<b>Valor P = 0,0002</b>

EW: Enfermedad de Wilson.

**Tabla 52: Evaluación a través del índice de reclasificación neta (NRI) de la contribución de la variable antecedentes personales o familiares de autoinmunidad (sí o no) al modelo con los criterios simplificados de 2008.**

Reclasificaciones predichas por el modelo con los criterios simplificados IAIHG 2008 + AI		En contra de HAI	A favor de HAI
En pacientes con HAI		58,7%	66,7%
En pacientes sin HAI		41,3%	33,3%
	<b>NRI</b>	<b>15,9%</b>	<b>Valor P = 0,275</b>

AI: Antecedentes personales o familiares de autoinmunidad.

Incluir como criterio el descarte de la enfermedad de Wilson en la nueva propuesta diagnóstica de HAI para población pediátrica, también parece contribuir a la capacidad predictiva, pero con un papel inferior al de *sexo* y *FA/AST*. A diferencia de estas dos últimas variables, considerar la exclusión de la enfermedad de Wilson, aunque mejora la clasificación de pacientes con HAI (98% frente a 2%), perjudica el descarte de HAI en los casos con diagnósticos alternativos (27% frente a 73%).

La existencia del fenómeno de autoinmunidad en los antecedentes familiares del paciente o en otras enfermedades concomitantes extrahepáticas es también un criterio que potencialmente podría aportar información diagnóstica. Sin embargo, su estudio por NRI no da un resultado significativo a pesar de que incluirlo como variable independiente en un modelo predictivo mejora discretamente su capacidad de acierto tanto en niños con HAI (67% frente a 59%) como en pacientes con otras hepatopatías (41% frente a 33%).

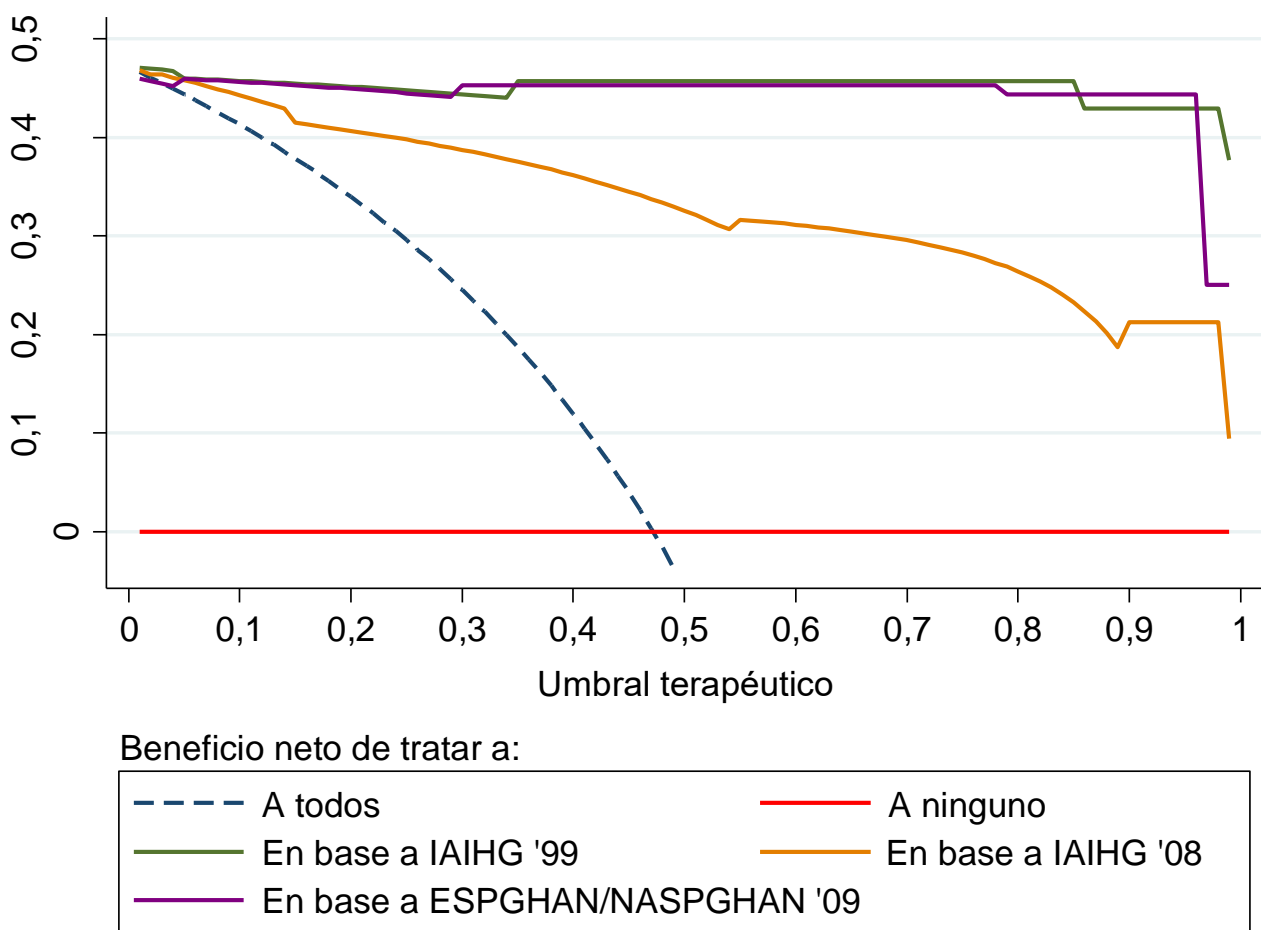


Figura 75: Curvas de decisión de los distintos criterios diagnósticos para la hepatitis autoinmune. Prevalencia de la enfermedad de 47,2%. En verde, criterios clásicos de la IAIHG sin contar el ítem de la respuesta al tratamiento. En amarillo, criterios simplificados de 2008. En morado, nuevo sistema diagnóstico basado en los criterios pediátricos propuestos por la ESPGHAN/NASPGHAN en 2009. La unidad es el beneficio neto asociado a un tratamiento correcto de un enfermo.

Finalmente, los resultados del análisis por curvas de decisión (DCA o *decision curve analysis*) se plasmaron gráficamente. Mediante el nomograma de Fagan se ha estudiado como cambiaría la utilidad clínica de los criterios diagnósticos simplificados en contextos clínicos con prevalencias diferentes de HAI. A través de DCA, se exploró como cambiaría esta utilidad (medida en términos de beneficio neto de tratamiento) considerando umbrales terapéuticos variables con estimaciones flexibles del coste y el beneficio del tratamiento.

El trazo de los nuevos criterios basados en la propuesta de 2009 prácticamente se solapa en el de los criterios clásicos pre-tratamiento. Ambos transcurren por la parte alta del gráfico en todo el universo de umbrales de probabilidad a partir del que se podría decidir tratar a los pacientes. Esto refleja que ambos criterios pueden llegar a ser homologables como sistema diagnóstico en un contexto clínico de sospecha de HAI en un niño y necesidad de decidir el inicio de la farmacoterapia en base a su resultado.

# 10

Discusión



## Discusión conjunta de los resultados

---

### 10.1. De la validez

La medicina es una disciplina científica basada en riesgos, además del arte de manejar la incertidumbre que conlleva, precisamente, su naturaleza probabilística. Dicha incertidumbre se extiende no sólo a las actividades preventivas, terapéuticas y pronósticas sino también a las diagnósticas [299,300]. En las fases del proceso diagnóstico intervienen la historia clínica, la exploración física y la realización de pruebas complementarias. A partir de estas fuentes de información es posible confeccionar sistemas diagnósticos basados en criterios, en los que se asigna un peso específico (o capacidad de aportar evidencia a favor de una enfermedad) a cada elemento constituyente [301,302]. Los criterios diagnósticos son, en realidad, unos criterios de clasificación con intención diagnóstica [12]. Por este motivo, es posible aplicar la teoría del análisis de la exactitud de un *test* a unos criterios diagnósticos en su conjunto. Bajo este prisma, han sido objeto de la tesis dos criterios diagnósticos para la hepatitis autoinmune pediátrica: los criterios simplificados de la IAIHG (2008) [19] y el nuevo sistema diagnóstico por puntos basado en la propuesta de criterios de la ESPGHAN/NASPGHAN (2009) [93].

La metodología del estudio se ha diseñado teniendo en cuenta la necesidad de obtener la frecuencia con la que se diagnostica la HAI en una situación clínica concreta, que es la del paciente en edad infantil o juvenil con signos de disfunción hepática y sin ninguna condición predisponente conocida. Se ha considerado que este es el escenario en que conviene estudiar la validez de los criterios por ser aquel en el que estos son más útiles: como una herramienta diagnóstica para la práctica clínica habitual, antes de conocer la respuesta al tratamiento inmunosupresor e, incluso, para justificarlo. Consiguientemente, los criterios se han aplicado en un sentido transversal en una población de pacientes pediátricos con hepatopatía todavía por diagnosticar. Para ello, por facilidad logística, el reclutamiento de los



casos se ha basado en la necesidad de llevar a cabo una biopsia hepática, entre cuyas indicaciones se recoge, justamente, la del marco clínico descrito [273]. La *American Association for the Study of Liver Diseases* recomienda actualmente someter a una biopsia hepática, siempre y cuando no exista una contraindicación absoluta, para confirmar la presunción de HAI antes de iniciar el tratamiento [40]. Respecto a esto, todos los casos elegibles que se presentaron como fallo hepático agudo se pudieron biopsiar después de administrar plasma fresco congelado o por vía transyugular. Aunque se ha sugerido que la toma de una muestra de tejido hepático puede no ser necesaria en algunos casos, se consideró que todavía no existen pruebas suficientes para elegir qué pacientes podrían ser candidatos para diagnosticarse sin el criterio histológico. Específicamente, Björnsson señala que en un número significativo de pacientes (tanto adultos como niños) con signos analíticos típicos, existe colinealidad con la presencia de datos histológicos compatibles. Añade, además, que aquellos pacientes con una histopatología atípica no son tratados de forma diferente, por lo que la necesidad de conocer con certeza la tipología de la infiltración inflamatoria en el parénquima puede ser redundante [92]. Ante la falta de más experiencias que reproduzcan estos hallazgos, la práctica habitual es seguir la recomendación de biopsiar siempre que sea posible. En esta línea, la elección de los criterios de inclusión y exclusión serían coherentes con el objetivo general de la tesis.

El interés por conocer la exactitud de los criterios diagnósticos simplificados de la IAIHG existe desde el momento de su propuesta. Diversos estudios se han llevado a cabo al respecto, tanto en población no seleccionada por un criterio etario, como en niños exclusivamente. Algunos emplean de referencia diagnóstica los primeros criterios descriptivos codificados [19,123,233]. Otros, en cambio, utilizan la versión revisada de 1999, argumentado que su uso clínico está más establecido, su especificidad es mayor y también está consolidado su empleo a efectos de investigación básica y clínica en niños [210,211]. Aun así, hay consenso sobre el hecho de que los criterios clásicos no son, por definición, un *gold standard* [236].

Bajo una concepción semántica de la definición de la enfermedad, asumir un 100% de sensibilidad y especificidad para los criterios clásicos es claramente erróneo o, como mínimo, indemostrable a la luz del sistema popperiano del racionalismo crítico [303]. De hecho, cinco de los casos de HAI se diagnosticaron en base a los hallazgos histológicos, presentación clínica y respuesta completa al tratamiento con corticoides a pesar de no cumplir los criterios de clasificación. Todos fueron varones sin historia personal ni familiar de enfermedades autoinmunes, con niveles normales de IgG y con autoanticuerpos a títulos bajos. En estos pacientes, de hecho, los criterios publicados por Mieli-Vergani en 2009 podrían haber funcionado mejor que los criterios clásicos [93].

En el estudio de validación de Mileti se demuestra que los niveles de IgG y los de globulinas séricas se pueden intercambiar sin afectar a la sensibilidad ni a la especificidad, manteniendo las mismas categorías en relación al límite de normalidad de cada laboratorio [210]. En nuestros análisis se usó solamente la cuantificación de IgG debido a que es la que se determina de forma habitual. El resto de los ítems de los criterios de 2008 se pudieron completar sin pérdidas en todos los pacientes.

El grupo de casos de HAI y de no HAI se ha constituido mediante reclutamiento consecutivo, con una definición de enfermedad no exclusivamente basada en los criterios clásicos y con unos criterios de inclusión laxos. El fin ha sido que el resultado fuera una muestra no sesgada en ambos grupos, que no sobreseleccione los casos más típicos de HAI y que permita comparar el rendimiento de los criterios frente a los diagnósticos alternativos reales que se pueden presentar en un entorno asistencial real. Más de la mitad de los pacientes en el grupo de no HAI fueron diagnosticados de hepatitis aguda criptogénica, enfermedad de Wilson y hepatitis vírica. Esto representa una diferencia importante con los dos estudios publicados hasta el momento sobre la exactitud de los criterios de 2008. En el caso del trabajo de Hiejima, el grupo control estuvo integrado principalmente por hepatitis crónicas por virus C [211]. Por su parte, Mileti estudió el grado de acierto

de los criterios en controles con hepatopatías metabólicas, EHNA y CEP. En consecuencia, se podría haber producido un sesgo de clasificación en presencia de casos de solapamiento o CEAI mal diagnosticados como CEP, dado que el componente de HAI no siempre se reconoce. Está descrito que cerca de la mitad de los pacientes con características compatibles con HAI, y que cumplen los criterios clásicos y los simplificados de la IAIHG, muestran enfermedad de la vía biliar con colestasis ya desde el momento del diagnóstico de la hepatopatía [30]. Las guías actuales ya contemplan la recomendación de realizar una prueba de imagen de la vía biliar (colangio-RM o CPRE en concreto) en niños con posible HAI para descartar la CEAI o la CEP [40,93,178]. Sin embargo, la limitación más relevante de este trabajo ha sido que solo se ha estudiado radiológicamente la vía biliar en un 48% de los niños incluidos, especialmente los reclutados en la segunda mitad del periodo de estudio. En el estudio de Hiejima, los cinco pacientes con CEP se clasificaron como HAI con los criterios simplificados, mientras que en el de Mileti, este error se produjo en tres casos de ocho [210,211]. Asimismo, en nuestro trabajo, dos de cinco pacientes con CEP se diagnosticaron erróneamente como HAI por los criterios simplificados, sin que durante el seguimiento evolucionaran como una CEAI. Además, aunque no necesariamente implique un tratamiento farmacológico distinto (más allá de añadir ácido ursodesoxicólico para la colestasis), la falta de colangiograma en algunos casos clasificados como HAI puede conllevar que se dejen de diagnosticar CEAI. Algunos estudios que evalúan la capacidad de acierto diagnóstico de los criterios de la IAIHG en adultos con CEP y CEAI muestran que la versión clásica de 1999 y la versión simplificada de 2008 muestran una especificidad similar [89,126,199]. Nuestros resultados, añaden más evidencia a favor de la consistencia de la recomendación de realizar una colangio-RM o una CPRE ante la sospecha de una HAI, con independencia del resultado de cualquier sistema de clasificación por puntos. Sería la forma de evitar los posibles errores diagnósticos tipo I y II.

Los parámetros de validez de los sistemas diagnósticos por puntos en pacientes con HAI de debut como fallo hepático ha sido también objeto de discusión. Tanto la versión clásica revisada de 1999 como la simplificada, de la IAIHG, han demostrado una peor capacidad discriminante en este tipo de pacientes [35,211]. Sin embargo, los cuatro casos de la cohorte que cumplían la definición de fallo hepático agudo (dada por el grupo de estudio PALFSG) fueron correctamente clasificados por ambos criterios. Cabe señalar que todos mostraron signos leves de encefalopatía y se recuperaron de la repercusión neurológica de la insuficiencia hepática al empezar la corticoterapia. Es posible que la gravedad clínica del fallo hepático pudiera influir en la rentabilidad diagnóstica de los criterios. Sería la hipótesis de que más insuficiencia hepática implica más encefalopatía y, al mismo tiempo, más características clínicas o analíticas atípicas que rebajen el ajuste de los criterios. Seguramente no exista suficiente validez externa para inferir a nuestra muestra los resultados de los estudios que describen esta menor exactitud de los criterios [35,211].

Con todo, incluso en esta muestra no sesgada por sobreselección, los criterios simplificados han demostrado una validez buena en términos generales. La sensibilidad de 96,4% es incluso más alta que la obtenida por Hiejima (86%) y por Mileti (95%), en parte gracias a una mejor clasificación de algunos pacientes con CEP [210,211]. Dentro de este último grupo diagnóstico, incluso a pesar de haber encontrado una mejor validez que en el estudio de Hiejima, los criterios simplificados no son adecuados para descartar la HAI. Tal y como también lo interpretan Mileti et al, nuestros resultados también apuntan a que el diagnóstico y manejo de los niños con posibilidad de padecer una HAI no solo debe hacerse basado en los criterios diagnósticos, sino en la histopatología de la muestra hepática obtenida por biopsia si existe, además, la sospecha fundada de CEP. En términos de sensibilidad, obtuvimos un resultado intermedio (72,0%) entre el ofrecido por Hiejima (55,0%) y el de Mileti (91,9%) [210,211]. Además, en el entorno clínico general definido por los criterios de inclusión y exclusión (con una probabilidad

preprueba de cerca del 50%), la seguridad con la que un resultado positivo de los criterios simplificados acierta una HAI es casi máxima, del 94,7%. Por ello parece razonable que se pueda fundamentar el inicio del tratamiento en base al resultado de los criterios simplificados con un resultado positivo, en la medida que la respuesta al tratamiento añade evidencia adicional al diagnóstico y la seguridad de acertar es excelente. Así y todo, en medios con una prevalencia de HAI menor, como por ejemplo en entornos con una mayor incidencia de hepatitis víricas o de EHNA, la confiabilidad del diagnóstico en un paciente que cumpla los criterios puede ser menor. Por ello se recomienda estimar la prevalencia de la entidad en cada población para poder calcular los valores predictivos por inferencia bayesiana por parte de cada observador individual.

Los indicadores obtenidos en el estudio de validación de los criterios simplificados permiten efectuar un análisis básico de decisiones clínicas si incorporamos una estimación fiable de la probabilidad preprueba de la enfermedad. Se discutirán los resultados del análisis de decisiones más adelante, pero para poder realizarlo, arrastrando el menor riesgo de sesgo posible, se llevó a cabo una revisión sistemática y meta-análisis de los estimadores de validez diagnóstica de los criterios de 2008 obtenidos en estudios realizados en población exclusivamente pediátrica. Hasta donde se ha podido comprobar, se trata de la primera revisión sistemática realizada nunca con este objeto. A fecha actual, no existe ningún registro en los principales directorios médicos con este fin. Se incluyó el plan de trabajo en PROSPERO para dar visibilidad al proyecto, dificultar la duplicidad de estudios y reducir la probabilidad de sesgo al permitir a la comunidad científica comparar la revisión completa, una vez se publique, con lo contemplado en el protocolo. PROSPERO es, en esencia, un registro internacional de revisiones sistemáticas sobre temas de salud incluidas de forma prospectiva. A diferencia de la base de datos Cochrane de revisiones sistemáticas, no existe la figura del editor/coordinador vinculado al *Cochrane Review Group*. La progresión de registros en PROSPERO ha sido significativa en los últimos años y están empezando a plantearse estudios bien

diseñados para evaluar, dentro de los protocolos, qué predictores son más adecuados para la eventual futura publicación de revisiones sistemáticas no dependientes del *Cochrane Review Group* [304,305].

El número de estudios recuperados en la revisión, e incluidos en el meta-análisis, ha sido bastante restringido, incluso a pesar de haber consultado también la literatura gris. Todo el proceso se ejecutó con ayuda de la oficina de Documentación de la Biblioteca Sant Joan de Déu. Si bien solo se incorporó información sobre cuatro estudios de pruebas diagnósticas, no se encontró evidencia de sesgo de publicación, lo que otorga consistencia a los indicadores de validez combinados o agrupados. Los estudios primarios fueron evaluados críticamente y la información se extrajo de forma independiente por el doctorando y un colaborador. Sin embargo, el diseño de los originales obliga a plantear algunas cuestiones sobre el significado de los indicadores de validez combinados.

En primer lugar, se sabe que los estudios de tipo caso-control de pruebas diagnósticas son particularmente susceptibles al sesgo de selección [306]. La inclusión de casos demasiado típicos o evidentes, especialmente en enfermedades para las que no existe un *gold standard*, pueden conllevar una sobrevaloración de la sensibilidad y la especificidad al excluirse casos dudosos o que no ajusten perfectamente a los criterios establecidos. A este respecto, los tres estudios de caso-control (Hiejima 2011, Mileti 2012 y Gonçalves 2017) basaron la confirmación de la HAI en los resultados de los criterios clásicos de 1999 [210,211,258], que, si bien son la mejor aproximación a un patrón de referencia real, pueden generar dudas sobre el ajuste de la enfermedad-problema a la pregunta de revisión (o pregunta diagnóstica de interés) [236]. Acertadamente, dos autores citan el uso de los criterios pediátricos de la ESPGHAN/NASPGHAN, que dan por positivos títulos de anticuerpos inferiores a los de los criterios de la IAIHG y apuntan a la necesidad de realizar un estudio de imagen de la vía biliar para descartar CEP y CEAI (Hiejima 2011 y Gonçalves 2017) [211,258]. Por una razón de tamaño muestral, el estudio primario a cuyos indicadores se les ha otorgado un peso mayor para el cálculo de los datos

combinados ha sido el del capítulo 1 de la tesis, que se ha publicado en marzo de 2018 en la revista *Pediatric Gastroenterology, Hepatology and Nutrition* [307]. Ya se ha mencionado que su diseño incluye una fase retrospectiva con algunos pacientes a los que no se estudió mediante colangio-RM ni CPRE, lo que puede influir en la estimación de los criterios simplificados al haberse clasificado como HAI casos de CEAI.

Otro de los aspectos que contempla el listado QUADAS-2 se refiere a la aplicación ciega de la prueba índice. En ninguno de los trabajos se menciona que se hiciera un esfuerzo metodológico para cumplir con esta recomendación. Sin embargo, los criterios de 2008 no dejan espacio para la interpretación por parte del observador, por lo que no se ha considerado que esto suponga una fuente significativa de sesgo [250].

Con todo, el par sensibilidad/especificidad combinado se ha estimado en 77%/95% por un modelo de efectos aleatorios para el punto de corte en 6. En un medio con una 47% de probabilidad de HAI después de una biopsia hepática por hipertransaminasemia en un niño no trasplantado, un paciente que cumpla criterios tiene un 92,5% de posibilidades de ser realmente una HAI. Sin embargo, aun con un resultado negativo, el riesgo de HAI mal clasificada es de un 17,1% (el opuesto del VPN). Como cualquier prueba diagnóstica con una buena especificidad y una sensibilidad moderada, un resultado positivo es más interesante, desde un punto de vista práctico, que uno negativo. En la medida en que el sistema diagnóstico por puntos de 2008 tiene solo 4 variables, es más fácil de aplicar que los criterios clásicos de 1999. Obtener 6 o más puntos da una buena seguridad diagnóstica pero el caso contrario no parece ser suficiente para descartar la HAI.

Las razones para los diagnósticos erróneos con los criterios simplificados son diversas. Algunas de ellas ya se han descrito como características de la HAI pediátrica. Concretamente, la necesidad de excluir la CEP, la CEAI, la enfermedad de Wilson, y el hecho de que los niños pueden exhibir títulos bajos de autoanticuerpos, se han incluido en la propuesta de criterios de la ESPGHAN/NASPGHAN (2009) [93].

Hasta el momento actual no existe ninguna publicación indexada que trate de su transformación en un sistema de puntos y su validación. Desde el momento del planteamiento del diseño de la tesis, el cálculo del mínimo tamaño muestral necesario ya se hizo teniendo en cuenta los requerimientos de la modelización por regresión logística y la validación externa del sistema de puntos generado. La representatividad del grupo de casos de HAI y el resto de los diagnósticos alternativos también ha sido idónea para el objeto del capítulo 4.

Cuando la intención con la que se construye un modelo de regresión es predictiva, a diferencia de cuando es explicativa, la lógica de las variables independientes no es relevante siempre y cuando el modelo efectúe predicciones acertadas [219,296]. Uno de los criterios de causalidad de Bradford-Hill es la intensidad de la asociación entre la causa en estudio y su efecto teórico. Dado que la magnitud de esta asociación es medible con medidas como la *odds ratio*, la regresión logística con intención explicativa sirve de fundamento matemático para el análisis de la posible relación [308,309]. Dado que este no es el caso, la no significación estadística de los criterios de 2009 en el modelo elegido no es interpretable como un defecto de la modelización. Sí que es posible, sin embargo, que se haya producido un fenómeno de sobrestimación de la validez de los nuevos criterios por un efecto de solapamiento con los criterios elegidos como referencia diagnóstica de la HAI, que origina un sesgo de clasificación [306,310]. En efecto, aunque con algunas diferencias, todos los criterios ESPGHAN/NASPGHAN 2009 están reconocidos en la propuesta clásica de 1999 [93]. Se ha intentado minimizar este efecto con el análisis de casos discrepantes y los criterios de robustez, para minimizar los posibles falsos negativos y positivos (respectivamente) que tendría usar los criterios de la IAIHG como patrón único de referencia diagnóstico.

Otra limitación del bloque dedicado a la nueva propuesta de criterios diagnósticos pediátricos concierne a las formas clínicas especiales de HAI. Además de la HAI *pura*, se reconoce la existencia de la HAI *de novo* en el paciente trasplantado hepático y la HAI con solapamiento con colangitis esclerosante (*overlap*



*syndrome* o CEAI) [311]. Se carece de validez externa suficiente para inferir los resultados a niños sometidos a trasplante hepático dado que esta población fue excluida durante el reclutamiento de pacientes. En este contexto, se debería de evitar basar el diagnóstico en el nuevo *score*, si bien es posible que también pueda funcionar correctamente dada la similitud clínica y analítica entre la HAI en un hígado nativo y la enfermedad recurrente en el injerto [311,312]. La validación en niños trasplantados del nuevo sistema diagnóstico por puntos basado en la propuesta ESPGHAN/NASPGHAN podría tratarse en un estudio futuro. Tampoco se ha podido validar la bondad de los nuevos criterios en diferenciar HAI de la hepatopatía colestásica de naturaleza autoinmune. En comparación con la rutina asistencial de otros medios, los pacientes con CEAI están claramente infrarrepresentados en nuestra muestra, lo que puede haber llevado a estimar unos indicadores de validez superiores a los reales. Para compensar, se decidió forzar la inclusión del criterio de la colangiografía, basada en la evidencia clínica actual [101,188,311,313]. Aunque no se asignó ningún número de puntos a este ítem, se contempló como un criterio de presencia obligatoria, en línea con las recomendaciones de las guías vigentes [87,311,314]. Como resultado, una puntuación positiva, tal y como se propone el nuevo sistema de puntos, puede descartar la CEAI, aunque no se pueda decir que la presencia de 6 o más puntos en un paciente con alteraciones en la prueba de imagen de la vía biliar tenga CEAI.

Recientemente, durante la redacción de la tesis, se ha publicado una nueva propuesta de criterios diagnósticos por puntos para las formas juveniles de la HAI y de la CEAI. Se ha confeccionado sobre una base teórica a propuesta del comité de Hepatología de la ESPGHAN y todavía no ha sido validada con un estudio específico. Se diferencia de la propuesta de 2009 en que sugiere una puntuación para cada criterio (en función de si la sospecha es una HAI o una CEAI) y en que incorpora más criterios: La exclusión de EHNA y hepatitis tóxica, la presencia de antecedentes personales de enfermedad autoinmune extrahepática y la existencia de historia familiar de cualquier enfermedad de naturaleza autoinmune [311].

Variable	Cut-off	Points	
		AIH	ASC
ANA and/or SMA*	≥1:20 <sup>†</sup>	1	1
	≥1:80	2	2
Anti-LKM-1* or	≥1:10 <sup>†</sup>	1	1
	≥1:80	2	1
Anti-LC-1	Positive <sup>†</sup>	2	1
Anti-SLA	Positive <sup>†</sup>	2	2
pANNA	Positive	1	2
IgG	>ULN	1	1
	>1:20 ULN	2	2
Liver histology	Compatible with AIH	1	1
	Typical of AIH	2	2
Absence of viral hepatitis (A, B, E, EBV), NASH, Wilson disease, and drug exposure	Yes	2	2
Presence of extrahepatic autoimmunity	Yes	1	1
Family history of autoimmune disease	Yes	1	1
Cholangiography	Normal	2	-2
	Abnormal	-2	2

Score ≥7: probable AIH; ≥8: definite AIH. Score ≥7: probable ASC; ≥8: definite ASC. AIH = autoimmune hepatitis; ANA = anti-nuclear antibody; anti-LC-1 = anti-liver cytosol type 1; anti-LKM-1 = anti-liver kidney microsomal antibody type 1; anti-SLA = anti-soluble liver antigen; ASC = autoimmune sclerosing cholangitis; EBV = Epstein-Barr virus; IgG = immunoglobulin G; NASH = nonalcoholic steatohepatitis; pANNA = peripheral anti-nuclear neutrophil antibodies; SMA = anti-smooth muscle antibody; ULN = upper limit of normal.

\*Antibodies measured by indirect immunofluorescence on a composite rodent substrate (kidney, liver, stomach).

<sup>†</sup>Addition of points achieved for ANA, SMA, anti-LKM-1, anti-LC-1, and anti-SLA autoantibodies cannot exceed a maximum of 2 points.

**Figura 76: Propuesta de criterios diagnósticos por puntos para la hepatitis autoinmune pediátrica (ESPGHAN Hepatology Committee Position Statement de 2018). Reproducido de Mieli-Vergani et al. J Pediatr Gastroenterol Nutr. 2018;66:355. Con permiso de Lippincott Williams & Wilkins.**

Una puntuación de 7 puntos (sobre un total de 13) da el diagnóstico de HAI probable, mientras que 8 puntos o más, de HAI definitiva. Es interesante comprobar que el ítem de los autoanticuerpos se ha categorizado de la misma forma que en nuestra propuesta, así como que su asignación de puntos ha sido equivalente, con la salvedad de los pANNA, que se incluyen como dato a favor de la CEAI. También ha sido muy similar la suma de puntos que otorgan los criterios de la anatomía patológica y el de la ausencia de diagnósticos alternativos. Se puede considerar que la homología entre los criterios que se han planteado en la tesis y los de la ESPGHAN dan robustez a nuestra propuesta. Dado que estos últimos todavía no han sido validados, se podría diseñar un estudio transversal de pruebas diagnósticas para comparar el rendimiento de ambos sistemas diagnósticos. También cabría ampliar el análisis de decisiones clínicas incorporando los indicadores de validez de la propuesta de la ESPGHAN de 2018.

El rendimiento de los nuevos criterios diagnósticos de 2009 en las formas de HAI con presentación como FHA también ha sido excelente. Igual que con los criterios simplificados, los cuatro casos se diagnosticaron correctamente. Sería también deseable comprobar su adecuación en pacientes con formas más graves de enfermedad o hepatitis fulminante.

Finalmente, la nueva propuesta basada en los criterios de 2009 permite, teóricamente, que los casos seronegativos puedan ser diagnosticados si el resto de los criterios se cumplen, con lo que se llegaría a 6 puntos justos. No obstante, más de la mitad de los niños con HAI sin autoanticuerpos no presentan hipergammaglobulinemia, como ha sido el caso de nuestros falsos negativos tanto en la muestra de desarrollo como en la de validación. Incluso aunque asocien rasgos habituales como la presencia de anemia aplásica o trombocitopenia periférica, los criterios simplificados de 2008 y 2009 tienen un papel muy limitado en este fenotipo particular de HAI [315]. Ante una sospecha de HAI seronegativa, los criterios originales de la IAIHG tendrían una validez superior al contemplar la respuesta los inmunosupresores, aunque continúe siendo probable que su exactitud pre-tratamiento sea más pobre que en casos típicos.

## **10.2. De la fiabilidad**

Una prueba diagnóstica es fiable cuando existe concordancia suficiente entre los resultados obtenidos, frente a una misma realidad, tras una sucesión mínima de repeticiones cuando lo aplica un único observador o entre las aplicaciones llevadas a cabo por varios observadores [259,262]. El análisis de la fiabilidad se planteó para comprobar que la claridad con la que están expresados los criterios simplificados de 2008 sea suficiente como para que varios investigadores asignen la misma puntuación (y, por lo tanto, diagnostiquen de la misma forma) a los mismos niños en riesgo de tener HAI. Es decir, para evaluar la reproductibilidad inter-observador.

La consistencia de las puntuaciones emitidas por los dos observadores que participaron en el capítulo 3 ha sido excelente. Los indicadores de concordancia

arrojan unos valores muy buenos y el diagrama de Bland-Altman muestra una mínima dispersión en el extremo derecho de posibles puntuaciones de los criterios. Cabe destacar la limitación que supone el hecho de que el número de investigadores involucrados es mínimo. No obstante, ha sido suficiente para comprobar dos fenómenos esperables. El primero ha sido que la posible diversidad de valores observados no imprime una clasificación diferencial clínicamente relevante, es decir, entre presencia o ausencia de HAI. Solo unos pocos casos se puntuaron de forma distinta, y fue en un sentido de clasificar como HAI probable una HAI definitiva y viceversa. El segundo fenómeno es que la única fuente de discrepancias ha sido la interpretación de la anatomía patológica, como fuente de dudas entre una histología compatible o típica. A este respecto, interesa dejar claro qué se debe de entender por típico y compatible, para lo cual es difícil encontrar referencias en la bibliografía [19,93].

En los criterios originales revisados en 1999, la máxima puntuación por el criterio histológico la da la presencia de hepatitis de interfase (3 puntos), seguida por la infiltración con predominio linfoplasmocitario (2 puntos) y, finalmente, por la formación de hepatocitos *en roseta* (1 punto) [61]. Tanto los criterios de 2008 como el *score* de la ESPGHAN de 2018, hablan de hallazgos típicos y compatibles, siguiendo la doctrina de Dienes y Lohse, según la cual la anatomía patológica de la HAI puede ser atípica, compatible o típica [19,93,203]. Las observaciones que definen esta última son la hepatitis de interfase, el infiltrado en tracto portal con extensión hacia el lóbulo de células linfocíticas o linfoplasmocíticas, la formación de rosetas hepáticas (transformación microacinar de los hepatocitos) y la emperipolesis (penetración activa de una célula en otra célula de mayor tamaño). Según ambos criterios simplificados, para que una biopsia hepática pueda ser considerada típica de HAI deben de estar presentes las tres primeras observaciones, sin que sea necesaria la emperipolesis. En caso de que se observe una hepatitis crónica con infiltración linfocítica, sin estar presentes todos los datos típicos, se cataloga la anatomía patológica como compatible. Una descripción microscópica parcial en los

informes del patólogo puede dificultar la aplicación del criterio histológico, cuya importancia es capital atendiendo a los resultados del capítulo 4. Para compensar esta potencial fuente de errores, el resultado de la biopsia debe mencionar explícitamente si la pieza obtenida muestra unas características típicas completas o, simplemente, compatibles, además de dejar claro si existen formaciones *en roseta* (para que la aplicación de los criterios originales no genere dudas), hepatitis de interfase o colapso multilobular (para lo propio en los nuevos criterios pediátricos basados en la propuesta ESPGHAN 2009).

Al hilo de esto, finalmente, se consideraría atípico el hallazgo de cualquier otro dato sugestivo de diagnósticos alternativos, como las inclusiones grasas en la esteatohepatitis no alcohólica (EHNA) y la presencia de conductillos biliares escasos en la enfermedad de Alagille. No es un aspecto de la fiabilidad de los criterios diagnósticos, pero, ciertamente, una de las objeciones que se puede hacer al nuevo *score* de la ESPGHAN es el hecho de que una histología típica (a la que da 2 puntos para la sospecha tanto de HAI como de CEAI), por definición, prácticamente excluye las otras hepatopatías del criterio de descarte de diagnósticos alternativos (hepatitis vírica, hepatitis tóxica EHNA y enfermedad de Wilson, todas con hallazgos histológicos característicos). Así, los pacientes que puntúen +2 en el criterio histológico, automáticamente también puntúan +1 en el criterio de exclusión de otras causas. Se trata de un fenómeno similar al de colinealidad en la modelización por regresión y, al igual que en ésta, debería de evitarse porque falsea el peso real que tiene cada criterio afectado en la explicación de la variable respuesta (o del diagnóstico, en nuestro caso) [219]. Indirectamente, según la propuesta diagnóstica de 2018, la puntuación máxima por los criterios histológicos es +3 (en vez de +2), que es igual que el número de puntos que da nuestro sistema de puntos basado en los criterios pediátricos de 2009.

### 10.3. De la utilidad clínica

Una vez conocidas las razones de verosimilitud de los criterios simplificados de la IAIHG y nuestra propuesta de criterios ESPGHAN/NASPGHAN 2009, se ha podido llevar a cabo un análisis básico de toma de decisiones siguiendo el modelo de Pauker-Kassirer con el esquema de aplicación de J. Latour [279,280,316]. Con ayuda de los nomogramas de Fagan se ha observado que la probabilidad postprueba de un resultado positivo en ambos criterios desplaza la probabilidad de HAI desde la prevalencia cruzando el umbral terapéutico. Esto construye una base teórica para justificar un ensayo terapéutico en niños con puntuaciones más allá del corte óptimo. Tanto más cuando la prevalencia de HAI obtenida en la muestra del estudio queda justo entre los dos umbrales de acción, es decir, dentro del margen en el que es rentable la aplicación de la prueba diagnóstica. Como limitación, se debe de señalar que la información para fijar las utilidades clínicas se obtuvo de informes y estudios realizados sobre población general (incluyendo adultos) y que el riesgo del tratamiento se estableció arbitrariamente con la idea de evitar la inmunosupresión farmacológica a largo plazo en niños en crecimiento [40]. Sin embargo, existe un resultado que otorga consistencia a la observación de que está justificado un ensayo terapéutico en función del resultado de ambos criterios simplificados: el importante ancho de valores de prevalencia para los que funcionan en el modelo de Pauker-Kassirer. En efecto, esto es válido para probabilidades preprueba de HAI de entre 8,5% y 84,7% para los criterios IAIHG de 2008, y entre 4,9% y 96,9% para los nuevos criterios basados en la propuesta ESPGHAN/NASPGHAN 2009. Incluso teniendo en cuenta el riesgo inherente a la prueba, la amplitud entre estas prevalencias solo se acorta a entre 13,7% y 80,4% en el caso de los criterios de 2008, y a entre 5,3% y 92,9% en el caso de los criterios de 2009. Estos últimos anchos de probabilidades *a priori* de enfermedad son equivalentes a los que delimitan los umbrales de acción. Dado que contemplan los posibles perjuicios del test, son los que proponemos que se tengan en cuenta para el análisis efectivo de la utilidad de la prueba. En este caso concreto, además, es posible que su distancia esté estimada a la baja porque con las

medidas de vigilancia post-biopsia estándar que se usan actualmente, el riesgo de sangrado clínicamente significativo puede ser menor que el asumido en las simulaciones. Expresado de otra forma, en entornos clínicos con prevalencias muy diferentes de HAI es muy posible que la aplicación de ambos criterios simplificados sea clínicamente útil.

El estudio individual de la aportación de los criterios que faltan en la versión simplificada de la IAIHG, respecto a los originales, informa de que solo el sexo y la relación FA/AST aumentan el NRI. Por lo tanto, parece razonable que una nueva propuesta de criterios pediátricos los incluyera. No ocurre lo mismo con el descarte de la enfermedad de Wilson ni con la presencia de antecedentes personales o familiares de autoinmunidad, que, aunque también son informativos desde un punto de vista teórico, no se comportan como tales de forma significativa en el análisis por NRI. Este tipo de aproximación no debe usarse como un contraste de hipótesis, sus valores numéricos no son de interpretación intuitiva y es claramente inferior al estudio de las características operativas del modelo para la evaluación de la capacidad discriminante de una prueba diagnóstica [286,317]. De ahí la prudencia con la que se debe de interpretar este resultado. Estos dos criterios sí que aparecen, no obstante, en la nueva propuesta de la ESPGHAN de 2018. El descarte de enfermedad de Wilson está incluido en el criterio de exclusión de diagnósticos alternativos (aporta 2 puntos) y, según el comité de Hepatología de la ESPGHAN, se debería de asignar un punto si existen antecedentes personales de autoinmunidad extrahepática, y otro más si hay familiares afectos de HAI (sin indicar el grado de parentesco necesario para darlo como positivo) [311]. La publicación de esta última propuesta tuvo lugar en la última fase de redacción de la tesis y quedó fuera de los objetivos del trabajo. Aunque lo ideal sería valorar su validez y utilidad al igual que se ha hecho con los dos sistemas de criterios simplificados, la información disponible hace pensar que posiblemente no aporten mucho más que el nuevo sistema de puntos basado en la propuesta de 2009. De hecho, no existe mucho margen de mejora si se atiende a los resultados del análisis de decisiones clínicas. En efecto, en

términos de utilidad clínica, los criterios desarrollados en el capítulo 4, son comparables incluso a los criterios originales revisados en 1999. Aunque se consideren utilidades diferentes del tratamiento, el beneficio de basarlo en los resultados de los nuevos criterios es prácticamente el mismo que el de los criterios clásicos. Por esto, y en la medida en que contemplan las particularidades de los pacientes pediátricos con HAI y su simplicidad es similar, su empleo es más aconsejable que el de los criterios de 2008.

Finalmente, sería deseable comparar las dos nuevas propuestas de sistema de puntos (la del 2009 y la del 2018), no solo en exactitud diagnóstica para la HAI, sino en capacidad para diferenciarla del síndrome de solapamiento. Como planteamiento para un futuro estudio, cabría verificar si la hipótesis de que son equivalentes en exactitud y utilidad clínica es cierta. En este caso, el empleo de nuestra propuesta de sistema diagnóstico basada en los criterios de la ESPGHAN y la NASPGHAN de 2009 arrojaría la ventaja de ser más sencilla y de basar el posible desempate entre HAI y CEAI en el resultado de la prueba de imagen de la vía biliar, sin necesidad de asignar puntos diferentes en función de la sospecha principal.

En cualquier caso, queda demostrada la necesidad de realizar una colangiografía ante la posibilidad de hepatopatía de base autoinmune, así como la indicación justificada de iniciar el tratamiento inmunosupresor ante un resultado positivo de las versiones simplificadas de los criterios.





11

Conclusiones



## Conclusiones

---

1) Los criterios simplificados de la IAIHG (2008) muestran una sensibilidad moderada y una especificidad alta para el diagnóstico de HAI en niños y adolescentes con signos de disfunción hepática adquirida en los que no se ha tomado todavía ninguna medida terapéutica.

2) Una descripción exhaustiva de los hallazgos histológicos permite que las puntuaciones de los criterios simplificados pueden ser perfectamente reproducibles por investigadores o clínicos diferentes.

3) Para la prevalencia de HAI de un entorno clínico realista, un resultado de 6 o más puntos en los criterios simplificados ofrece una alta probabilidad de que se trate de una HAI y, además, resulta clínicamente útil basar la decisión de iniciar el tratamiento inmunomodulador en este resultado. Sin embargo, ni un resultado negativo permite descartar con suficiente seguridad el diagnóstico, ni un resultado positivo excluye la posibilidad de que se trate de una CEAI.

4) Los criterios pediátricos propuestos por la ESPGHAN y la NASPGHAN (2009) han podido transformarse en un sistema diagnóstico por puntos. Sus indicadores de validez externa superan los de los criterios de la IAIHG de 2008 y son igualmente simplificados.

5) La concordancia entre las clasificaciones diagnósticas del nuevo sistema de criterios pediátrico basado en la propuesta de 2009 y el patrón de referencia supera a la de los criterios simplificados de la IAIHG.

6) El principal inconveniente, tanto de los criterios simplificados de 2008 como de la propuesta basada en los criterios de 2009, es la elevada proporción de falsos negativos que arroja, fundamentalmente, en las HAI seronegativas.

7) El nuevo sistema diagnóstico por puntos es clínicamente útil en una mayor diversidad de escenarios clínicos. Su capacidad de acierto es igualmente excelente tanto con un resultado positivo como con uno negativo. Además, debería de permitir diferenciar entre HAI y CEAI.





12

Bibliografía



## Bibliografía

---

1. Porta Serra M. La observación clínica y el razonamiento epidemiológico. *Med Clin.* 1986;87:816-9.
2. Corral Corral C. El razonamiento médico. Madrid: Díaz de Santos; 1994. 79-121.
3. Ochoa Sangrador C, Orejas G. Epidemiología y metodología científica aplicada a la pediatría (IV): Pruebas diagnósticas. *An Esp Pediatr.* 1999;50:301-14.
4. Reznik L. The nature of disease. New York: Routledge & Keagen Paul; 1987.
5. International Statistical Classification of Diseases and Related Health Problems. 10.<sup>a</sup> ed. Ginebra: Organización Mundial de la Salud; 1992.
6. Bernard C. An introduction to the study of experimental Medicine. New York: Dover; 1927.
7. King L. What is a disease? *Philos Sci.* 1954;21:193-203.
8. Nordenfelt L. On the nature of health. 1.<sup>a</sup> ed. Dordrecht: Kluwer Academic Publishers; 1987.
9. Scadding JG. Essentialism and nominalism in medicine: logic of diagnosis in disease terminology. *Lancet (London, England).* 1996;348:594-6.
10. Sade RM. A theory of health and disease: The objectivist-subjectivist dichotomy. *J Med Philos.* 1995;20:513-25.
11. Borlowsky T, Friedman C, Lussier YA. Generating executable knowledge for evidence-based medicine using natural language and semantic processing. *AMIA Annu Symp Proc.* 2006;56-60.
12. Belmonte-Serrano MA. The myth of the distinction between classification and diagnostic criteria. *Reumatol Clin.* 11:188-9.
13. Aggarwal R, Ringold S, Khanna D, Neogi T, Johnson SR, Miller A, et al. Distinctions between diagnostic and classification criteria? *Arthritis Care Res (Hoboken).* 2015;67:891-7.
14. Yazici H. Diagnostic versus classification criteria - a continuum. *Bull NYU Hosp Jt Dis.* 2009;67:206-8.



15. Fries JF, Hochberg MC, Medsger TA, Hunder GG, Bombardier C. Criteria for rheumatic disease. Different types and different functions. The American College of Rheumatology Diagnostic and Therapeutic Criteria Committee. *Arthritis Rheum.* 1994;37:454-62.
16. Miettinen OS. The modern scientific physician: 3. Scientific diagnosis. *CMAJ.* 2001;165:781-2.
17. Taylor WJ, Fransen J. Distinctions Between Diagnostic and Classification Criteria: Comment on the Article by Aggarwal et al. *Arthritis Care Res (Hoboken).* 2016;68:149-50.
18. Núñez E, Steyerberg EW, Núñez J. [Regression modeling strategies]. *Rev española Cardiol.* 2011;64:501-7.
19. Hennes EM, Zeniya M, Czaja AJ, Parés A, Dalekos GN, Krawitt EL, et al. Simplified criteria for the diagnosis of autoimmune hepatitis. *Hepatology.* 2008;48:169-76.
20. Annesi I, Moreau T, Lellouch J. Efficiency of the logistic regression and Cox proportional hazards models in longitudinal studies. *Stat Med.* 1989;8:1515-21.
21. Harrell FE. Regression modeling strategies: with application to linear models, logistic regression, and survival analysis. New York: Springer-Verlag; 2001.
22. Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. New York: Springer; 2009.
23. Royston P, Moons KGM, Altman DG, Vergouwe Y. Prognosis and prognostic research: Developing a prognostic model. *BMJ.* 2009;338:b604.
24. Marshall A, Altman DG, Holder RL. Comparison of imputation methods for handling missing covariate data when fitting a Cox proportional hazards model: a resampling study. *BMC Med Res Methodol.* 2010;10:112.
25. Sullivan LM, Massaro JM, D'Agostino RB. Presentation of multivariate data for clinical use: The Framingham Study risk score functions. *Stat Med.* 2004;23:1631-60.
26. Vergani D, Mieli-Vergani G. Pharmacological management of autoimmune hepatitis. *Expert Opin Pharmacother.* 2011;12:607-13.
27. Mieli-Vergani G, Vergani D. Paediatric autoimmune liver disease. *Arch Dis Child.* 2013;98:1012-7.

28. Mieli-Vergani G, Vergani D. Autoimmune hepatitis in children: what is different from adult AIH? *Semin Liver Dis.* 2009;29:297-306.
29. Gregorio G V, Portmann B, Reid F, Donaldson PT, Doherty DG, McCartney M, et al. Autoimmune hepatitis in childhood: a 20-year experience. *Hepatology.* 1997;25:541-7.
30. Gregorio G V, Portmann B, Karani J, Harrison P, Donaldson PT, Vergani D, et al. Autoimmune hepatitis/sclerosing cholangitis overlap syndrome in childhood: a 16-year prospective study. *Hepatology.* 2001;33:544-53.
31. Waldenström J. Blutprotein und Nahrungseiweiss. *Dtsch Ges Z Verdau Stoffwechselkr.* 1950;15:113-9.
32. Mackay IR. Toward diagnostic criteria for autoimmune hepatitis. *Hepatology.* 1993;18:1006-8.
33. Johnson PJ, McFarlane IG. Meeting report: International Autoimmune Hepatitis Group. *Hepatology.* 1993;18:998-1005.
34. Liberal R, Grant CR, Longhi MS, Mieli-Vergani G, Vergani D. Diagnostic criteria of autoimmune hepatitis. *Autoimmunity Reviews.* 2014. p. 435-40.
35. Yeoman AD, Westbrook RH, Al-Chalabi T, Carey I, Heaton ND, Portmann BC, et al. Diagnostic value and utility of the simplified International Autoimmune Hepatitis Group (IAIHG) criteria in acute and chronic liver disease. *Hepatology.* 2009;50:538-45.
36. Boberg KM, Aadland E, Jahnsen J, Raknerud N, Stiris M, Bell H. Incidence and prevalence of primary biliary cirrhosis, primary sclerosing cholangitis, and autoimmune hepatitis in a Norwegian population. *Scand J Gastroenterol.* 1998;33:99-103.
37. Hurlburt KJ, McMahon BJ, Deubner H, Hsu-Trawinski B, Williams JL, Kowdley K V. Prevalence of autoimmune liver disease in Alaska Natives. *Am J Gastroenterol.* 2002;97:2402-7.
38. Ngu JH, Bechly K, Chapman BA, Burt MJ, Barclay ML, Gearry RB, et al. Population-based epidemiology study of autoimmune hepatitis: a disease of older women? *J Gastroenterol Hepatol.* 2010;25:1681-6.
39. Liberal R, Grant CR, Mieli-Vergani G, Vergani D. Autoimmune hepatitis: a comprehensive review. *J Autoimmun.* 2013;41:126-39.

40. Manns MP, Czaja AJ, Gorham JD, Krawitt EL, Mieli-Vergani G, Vergani D, et al. Diagnosis and management of autoimmune hepatitis. *Hepatology*. 2010;51:2193-213.
41. Grønbaek L, Vilstrup H, Jepsen P. Autoimmune hepatitis in Denmark: incidence, prevalence, prognosis, and causes of death. A nationwide registry-based cohort study. *J Hepatol*. 2014;60:612-7.
42. Schramm C, Lohse AW. Autoimmune hepatitis on the rise. *J Hepatol*. 2014;60:478-9.
43. Liberal R, Vergani D, Mieli-Vergani G. Paediatric Autoimmune Liver Disease. *Dig Dis*. 2015;33 Suppl 2:36-46.
44. Mieli-Vergani G, Vergani D. Autoimmune hepatitis. *Nat Rev Gastroenterol Hepatol*. 2011;8:320-9.
45. Krawitt E-L. Clinical features and management of autoimmune hepatitis. *World J Gastroenterol*. 2008;14:3301-5.
46. Vergani D, Mieli-Vergani G. Cutting edge issues in autoimmune hepatitis. *Clin Rev Allergy Immunol*. 2012;42:309-21.
47. Ferrari R, Pappas G, Agostinelli D, Muratori P, Muratori L, Lenzi M, et al. Type 1 autoimmune hepatitis: patterns of clinical presentation and differential diagnosis of the «acute» type. *QJM*. 2004;97:407-12.
48. Kessler WR, Cummings OW, Eckert G, Chalasani N, Lumeng L, Kwo PY. Fulminant hepatic failure as the initial presentation of acute autoimmune hepatitis. *Clin Gastroenterol Hepatol*. 2004;2:625-31.
49. Krawitt EL. Autoimmune hepatitis. *N Engl J Med*. 2006;354:54-66.
50. Codoñer P. Hepatitis autoinmune. *An Pediatr Contin*. 2003;1:80-5.
51. Heneghan MA, Norris SM, O'Grady JG, Harrison PM, McFarlane IG. Management and outcome of pregnancy in autoimmune hepatitis. *Gut*. 2001;48:97-102.
52. Samuel D, Riordan S, Strasser S, Kurtovic J, Singh-Grewel I, Koorey D. Severe autoimmune hepatitis first presenting in the early post partum period. *Clin Gastroenterol Hepatol*. 2004;2:622-4.

53. Yeoman AD, Al-Chalabi T, Karani JB, Quaglia A, Devlin J, Mieli-Vergani G, et al. Evaluation of risk factors in the development of hepatocellular carcinoma in autoimmune hepatitis: Implications for follow-up and screening. *Hepatology*. 2008;48:863-70.
54. Gatselis NK, Zachou K, Koukoulis GK, Dalekos GN. Autoimmune hepatitis, one disease with many faces: etiopathogenetic, clinico-laboratory and histological characteristics. *World J Gastroenterol*. United States: 2015;21:60-83.
55. Panayi V, Froud OJ, Vine L, Laurent P, Woolson KL, Hunter JG, et al. The natural history of autoimmune hepatitis presenting with jaundice. *Eur J Gastroenterol Hepatol*. 2014;26:640-5.
56. Longhi MS, Mieli-Vergani G, Vergani D. Autoimmune hepatitis. *Curr Pediatr Rev*. 2014;10:268-74.
57. Bogdanos DP, Mieli-Vergani G, Vergani D. Autoantibodies and their antigens in autoimmune hepatitis. *Semin Liver Dis*. 2009;29:241-53.
58. Liberal R, Mieli-Vergani G, Vergani D. Clinical significance of autoantibodies in autoimmune hepatitis. *J Autoimmun*. 2013;46:17-24.
59. Vergani D, Alvarez F, Bianchi FB, Cançado ELR, Mackay IR, Manns MP, et al. Liver autoimmune serology: a consensus statement from the committee for autoimmune serology of the International Autoimmune Hepatitis Group. *J Hepatol*. 2004;41:677-83.
60. Tan L, Zhang Y, Peng W, Chen J, Li H, Ming F. Detection of anti-lactoferrin antibodies and anti-myeloperoxidase antibodies in autoimmune hepatitis: a retrospective study. *J Immunoassay Immunochem*. 2014;35:388-97.
61. Alvarez F, Berg PA, Bianchi FB, Bianchi L, Burroughs AK, Cancado EL, et al. International Autoimmune Hepatitis Group Report: review of criteria for diagnosis of autoimmune hepatitis. *J Hepatol*. 1999;31:929-38.
62. Yuksel M, Wang Y, Tai N, Peng J, Guo J, Beland K, et al. A novel «humanized mouse» model for autoimmune hepatitis and the association of gut microbiota with liver inflammation. *Hepatology*. 2015;62:1536-50.
63. Floreani A, Liberal R, Vergani D, Mieli-Vergani G. Autoimmune hepatitis: Contrasts and comparisons in children and adults - a comprehensive review. *J Autoimmun*. 2013;46:7-16.

64. Bogdanos D-P, Invernizzi P, Mackay I-R, Vergani D. Autoimmune liver serology: current diagnostic and clinical challenges. *World J Gastroenterol*. 2008;14:3374-87.
65. Couto CA, Bittencourt PL, Porta G, Abrantes-Lemos CP, Carrilho FJ, Guardia BD, et al. Antismooth muscle and antiactin antibodies are indirect markers of histological and biochemical activity of autoimmune hepatitis. *Hepatology*. 2014;59:592-600.
66. Czaja AJ, Cassani F, Cataleta M, Valentini P, Bianchi FB. Frequency and significance of antibodies to actin in type 1 autoimmune hepatitis. *Hepatology*. 1996;24:1068-73.
67. Maggiore G, Veber F, Bernard O, Hadchouel M, Homberg JC, Alvarez F, et al. Autoimmune hepatitis associated with anti-actin antibodies in children and adolescents. *J Pediatr Gastroenterol Nutr*. 1993;17:376-81.
68. Zachou K, Oikonomou K, Renaudineau Y, Chauveau A, Gatselis N, Youinou P, et al. Anti- $\alpha$  actinin antibodies as new predictors of response to treatment in autoimmune hepatitis type 1. *Aliment Pharmacol Ther*. 2012;35:116-25.
69. Ferri Liu PM, de Miranda DM, Fagundes EDT, Ferreira AR, Simões e Silva AC. Autoimmune hepatitis in childhood: the role of genetic and immune factors. *World J Gastroenterol*. 2013;19:4455-63.
70. Bailloud R, Bertin D, Roquelaure B, Roman C, Ballot E, Johanet C, et al. Anti-mitochondrial-2 antibodies (anti-PDC-E2): a marker for autoimmune hepatitis of children? *Clin Res Hepatol Gastroenterol*. 2012;36:e57-9.
71. Miyakawa H, Kitazawa E, Abe K, Kawaguchi N, Fuzikawa H, Kikuchi K, et al. Chronic hepatitis C associated with anti-liver/kidney microsome-1 antibody is not a subgroup of autoimmune hepatitis. *J Gastroenterol*. 1997;32:769-76.
72. Johanet C, Ballot E. Autoantibodies in autoimmune hepatitis: anti-liver kidney microsome type 1 (anti-LKM1) and anti-liver cytosol type 1 (anti-LC1) antibodies. *Clin Res Hepatol Gastroenterol*. 2013;37:216-8.
73. Invernizzi P, Alessio MG, Smyk DS, Lleo A, Sonzogni A, Fabris L, et al. Autoimmune hepatitis type 2 associated with an unexpected and transient presence of primary biliary cirrhosis-specific antimitochondrial antibodies: a case study and review of the literature. *BMC Gastroenterol*. 2012;12:92.

74. Vogel A, Strassburg CP, Obermayer-Straub P, Brabant G, Manns MP. The genetic background of autoimmune polyendocrinopathy-candidiasis-ectodermal dystrophy and its autoimmune disease components. *J Mol Med (Berl)*. 2002;80:201-11.
75. Beaune P, Dansette PM, Mansuy D, Kiffel L, Finck M, Amar C, et al. Human anti-endoplasmic reticulum autoantibodies appearing in a drug-induced hepatitis are directed against a human liver cytochrome P-450 that hydroxylates the drug. *Proc Natl Acad Sci U S A*. 1987;84:551-5.
76. Crivelli O, Lavarini C, Chiaberge E, Amoroso A, Farci P, Negro F, et al. Microsomal autoantibodies in chronic infection with the HBsAg associated delta (delta) agent. *Clin Exp Immunol*. 1983;54:232-8.
77. Strassburg CP, Obermayer-Straub P, Alex B, Durazzo M, Rizzetto M, Tukey RH, et al. Autoantibodies against glucuronosyltransferases differ between viral hepatitis and autoimmune hepatitis. *Gastroenterology*. 1996;111:1576-86.
78. Liberal R, Grant CR, Longhi MS, Mieli-Vergani G, Vergani D. Diagnostic criteria of autoimmune hepatitis. *Autoimmun Rev*. 2014;13:435-40.
79. Ma Y, Okamoto M, Thomas MG, Bogdanos DP, Lopes AR, Portmann B, et al. Antibodies to conformational epitopes of soluble liver antigen define a severe form of autoimmune liver disease. *Hepatology*. 2002;35:658-64.
80. Kirstein MM, Metzler F, Geiger E, Heinrich E, Hallensleben M, Manns MP, et al. Prediction of short- and long-term outcome in patients with autoimmune hepatitis. *Hepatology*. 2015;62:1524-35.
81. Volkmann M, Luithle D, Zentgraf H, Schnölzer M, Fiedler S, Heid H, et al. SLA/LP/tRNP((Ser)Sec) antigen in autoimmune hepatitis: identification of the native protein in human hepatic cell extract. *J Autoimmun*. 2010;34:59-65.
82. Wies I, Brunner S, Henninger J, Herkel J, Kanzler S, Meyer zum Büschenfelde KH, et al. Identification of target antigen for SLA/LP autoantibodies in autoimmune hepatitis. *Lancet (London, England)*. 2000;355:1510-5.
83. Liberal R, Grant CR, Longhi MS, Mieli-Vergani G, Vergani D. Diagnostic criteria of autoimmune hepatitis. *Autoimmun Rev*. 2014;13:435-40.

84. Jarrot P-A, Kaplanski G. Pathogenesis of ANCA-associated vasculitis: An update. *Autoimmun Rev.* 2016;
85. Csernok E, Damoiseaux J, Rasmussen N, Hellmich B, van Paassen P, Vermeersch P, et al. Evaluation of automated multi-parametric indirect immunofluorescence assays to detect anti-neutrophil cytoplasmic antibodies (ANCA) in granulomatosis with polyangiitis (GPA) and microscopic polyangiitis (MPA). *Autoimmun Rev.* 2016;
86. Stravitz RT, Lefkowitz JH, Fontana RJ, Gershwin ME, Leung PSC, Sterling RK, et al. Autoimmune acute liver failure: proposed clinical and histological criteria. *Hepatology.* 2011;53:517-26.
87. Gleeson D, Heneghan MA, British Society of Gastroenterology. British Society of Gastroenterology (BSG) guidelines for management of autoimmune hepatitis. *Gut.* 2011;60:1611-29.
88. Zachou K, Rigopoulou E, Dalekos GN. Autoantibodies and autoantigens in autoimmune hepatitis: important tools in clinical practice and to study pathogenesis of the disease. *J Autoimmune Dis.* 2004;1:2.
89. Papamichalis PA, Zachou K, Koukoulis GK, Veloni A, Karacosta EG, Kypri L, et al. The revised international autoimmune hepatitis score in chronic liver diseases including autoimmune hepatitis/overlap syndromes and autoimmune hepatitis with concurrent other liver disorders. *J Autoimmune Dis.* 2007;4:3.
90. Gatselis NK, Zachou K, Papamichalis P, Koukoulis GK, Gabeta S, Dalekos GN, et al. Comparison of simplified score with the revised original score for the diagnosis of autoimmune hepatitis: a new or a complementary diagnostic score? *Dig Liver Dis.* 2010;42:807-12.
91. Tiniakos DG, Brain JG, Bury YA. Role of Histopathology in Autoimmune Hepatitis. *Dig Dis.* 2015;33 Suppl 2:53-64.
92. Björnsson E, Talwalkar J, Treeprasertsuk S, Neuhauser M, Lindor K. Patients with typical laboratory features of autoimmune hepatitis rarely need a liver biopsy for diagnosis. *Clin Gastroenterol Hepatol.* 2011;9:57-63.
93. Mieli-Vergani G, Heller S, Jara P, Vergani D, Chang M-H, Fujisawa T, et al. Autoimmune hepatitis. *J Pediatr Gastroenterol Nutr.* 2009;49:158-64.

94. Grønbaek L, Vilstrup H, Jepsen P. Autoimmune hepatitis in Denmark: incidence, prevalence, prognosis, and causes of death. A nationwide registry-based cohort study. *J Hepatol.* 2014;60:612-7.
95. Muratori P, Granito A, Quarneti C, Ferri S, Menichella R, Cassani F, et al. Autoimmune hepatitis in Italy: the Bologna experience. *J Hepatol.* 2009;50:1210-8.
96. Werner M, Prytz H, Ohlsson B, Almer S, Björnsson E, Bergquist A, et al. Epidemiology and the initial presentation of autoimmune hepatitis in Sweden: a nationwide study. *Scand J Gastroenterol.* 2008;43:1232-40.
97. Schiano TD, Fiel MI. To B(iopsy) or not to B(iopsy) .... *Clin Gastroenterol Hepatol.* 2011;9:3-4.
98. Feld JJ, Dinh H, Arenovich T, Marcus VA, Wanless IR, Heathcote EJ. Autoimmune hepatitis: effect of symptoms and cirrhosis on natural history and outcome. *Hepatology.* 2005;42:53-62.
99. Al-Chalabi T, Boccato S, Portmann BC, McFarlane IG, Heneghan MA. Autoimmune hepatitis (AIH) in the elderly: a systematic retrospective analysis of a large group of consecutive patients with definite AIH followed at a tertiary referral centre. *J Hepatol.* 2006;45:575-83.
100. Czaja AJ, Muratori P, Muratori L, Carpenter HA, Bianchi FB. Diagnostic and therapeutic implications of bile duct injury in autoimmune hepatitis. *Liver Int.* 2004;24:322-9.
101. Dienes HP, Erberich H, Dries V, Schirmacher P, Lohse A. Autoimmune hepatitis and overlap syndromes. *Clin Liver Dis.* 2002;6:349-62, vi.
102. Benseler V, Warren A, Vo M, Holz LE, Tay SS, Le Couteur DG, et al. Hepatocyte entry leads to degradation of autoreactive CD8 T cells. *Proc Natl Acad Sci U S A.* 2011;108:16735-40.
103. Zen Y, Notsumata K, Tanaka N, Nakanuma Y. Hepatic centrilobular zonal necrosis with positive antinuclear antibody: a unique subtype or early disease of autoimmune hepatitis? *Hum Pathol.* 2007;38:1669-75.
104. Hofer H, Oesterreicher C, Wrba F, Ferenci P, Penner E. Centrilobular necrosis in autoimmune hepatitis: a histological feature associated with acute clinical presentation. *J Clin Pathol.* 2006;59:246-9.



105. Zachou K, Muratori P, Koukoulis GK, Granito A, Gatselis N, Fabbri A, et al. Review article: autoimmune hepatitis -- current management and challenges. *Aliment Pharmacol Ther.* 2013;38:887-913.
106. Te HS, Koukoulis G, Ganger DR. Autoimmune hepatitis: a histological variant associated with prominent centrilobular necrosis. *Gut.* 1997;41:269-71.
107. Fujiwara K, Fukuda Y, Yokosuka O. Precise histological evaluation of liver biopsy specimen is indispensable for diagnosis and treatment of acute-onset autoimmune hepatitis. *J Gastroenterol.* 2008;43:951-8.
108. Yasui S, Fujiwara K, Yonemitsu Y, Oda S, Nakano M, Yokosuka O. Clinicopathological features of severe and fulminant forms of autoimmune hepatitis. *J Gastroenterol.* 2011;46:378-90.
109. Guindi M. Histology of autoimmune hepatitis and its variants. *Clin Liver Dis.* 2010;14:577-90.
110. Tang J, Zhou C, Zhang Z-J, Zheng S-S. Association of polymorphisms in non-classic MHC genes with susceptibility to autoimmune hepatitis. *Hepatobiliary Pancreat Dis Int.* 2012;11:125-31.
111. Oo YH, Hubscher SG, Adams DH. Autoimmune hepatitis: new paradigms in the pathogenesis, diagnosis, and management. *Hepatol Int.* 2010;4:475-93.
112. Donaldson PT, Doherty DG, Hayllar KM, McFarlane IG, Johnson PJ, Williams R. Susceptibility to autoimmune chronic active hepatitis: human leukocyte antigens DR4 and A1-B8-DR3 are independent risk factors. *Hepatology.* 1991;13:701-6.
113. Czaja AJ, Souto EO, Bittencourt PL, Cancado ELR, Porta G, Goldberg AC, et al. Clinical distinctions and pathogenic implications of type 1 autoimmune hepatitis in Brazil and the United States. *J Hepatol.* 2002;37:302-8.
114. Fortes M del P, Machado I V, Gil G, Fernández-Mestre M, Dagher L, León R V, et al. Genetic contribution of major histocompatibility complex class II region to type 1 autoimmune hepatitis susceptibility in Venezuela. *Liver Int.* 2007;27:1409-16.
115. Czaja AJ, Carpenter HA, Santrach PJ, Moore SB. Significance of HLA DR4 in type 1 autoimmune hepatitis. *Gastroenterology.* 1993;105:1502-7.

116. Ota M, Seki T, Kiyosawa K, Furuta S, Hino K, Kondo T, et al. A possible association between basic amino acids of position 13 of DRB1 chains and autoimmune hepatitis. *Immunogenetics*. 1992;36:49-55.
117. Vázquez-García MN, Aláez C, Olivo A, Debaz H, Pérez-Luque E, Burguete A, et al. MHC class II sequences of susceptibility and protection in Mexicans with autoimmune hepatitis. *J Hepatol*. 1998;28:985-90.
118. Ma Y, Bogdanos DP, Hussain MJ, Underhill J, Bansal S, Longhi MS, et al. Polyclonal T-cell responses to cytochrome P450IID6 are associated with disease activity in autoimmune hepatitis type 2. *Gastroenterology*. 2006;130:868-82.
119. Djilali-Saiah I, Renous R, Caillat-Zucman S, Debray D, Alvarez F. Linkage disequilibrium between HLA class II region and autoimmune hepatitis in pediatric patients. *J Hepatol*. 2004;40:904-9.
120. Fainboim L, Marcos Y, Pando M, Capucchio M, Reyes GB, Galoppo C, et al. Chronic active autoimmune hepatitis in children. Strong association with a particular HLA-DR6 (DRB1\*1301) haplotype. *Hum Immunol*. 1994;41:146-50.
121. Pando M, Larriba J, Fernandez GC, Fainboim H, Ciocca M, Ramonet M, et al. Pediatric and adult forms of type I autoimmune hepatitis in Argentina: evidence for differential genetic predisposition. *Hepatology*. 1999;30:1374-80.
122. Bittencourt PL, Goldberg AC, Caçado EL, Porta G, Carrilho FJ, Farias AQ, et al. Genetic heterogeneity in susceptibility to autoimmune hepatitis types 1 and 2. *Am J Gastroenterol*. 1999;94:1906-13.
123. Czaja AJ, Carpenter HA, Moore SB. HLA DRB1\*13 as a risk factor for type 1 autoimmune hepatitis in North American patients. *Dig Dis Sci*. 2008;53:522-8.
124. de Boer YS, van Gerven NMF, Zwiers A, Verwer BJ, van Hoek B, van Erpecum KJ, et al. Genome-wide association study identifies variants associated with autoimmune hepatitis type 1. *Gastroenterology*. 2014;147:443-52.e5.
125. Manns MP, Lohse AW, Vergani D. Autoimmune hepatitis--Update 2015. *J Hepatol*. 2015;62:S100-11.

126. Chandok N, Silveira MG, Lindor KD. Comparing the simplified and international autoimmune hepatitis group criteria in primary sclerosing cholangitis. *Gastroenterol Hepatol (N Y)*. 2010;6:108-12.
127. Lapierre P, Béland K, Alvarez F. Pathogenesis of autoimmune hepatitis: from break of tolerance to immune-mediated hepatocyte apoptosis. *Transl Res*. 2007;149:107-13.
128. Longhi MS, Hussain MJ, Mitry RR, Arora SK, Mieli-Vergani G, Vergani D, et al. Functional study of CD4+CD25+ regulatory T cells in health and autoimmune hepatitis. *J Immunol*. 2006;176:4484-91.
129. Shevach EM, McHugh RS, Piccirillo CA, Thornton AM. Control of T-cell activation by CD4+CD25+ suppressor T cells. *Immunol Rev*. 2001;182:58-67.
130. Ng WF, Duggan PJ, Ponchel F, Matarese G, Lombardi G, Edwards AD, et al. Human CD4(+)CD25(+) cells: a naturally occurring population of regulatory T cells. *Blood*. 2001;98:2736-44.
131. Baecher-Allan C, Brown JA, Freeman GJ, Hafler DA. CD4+CD25high regulatory cells in human peripheral blood. *J Immunol*. 2001;167:1245-53.
132. Ferri S, Longhi MS, De Molo C, Lalanne C, Muratori P, Granito A, et al. A multifaceted imbalance of T cells with regulatory function characterizes type 1 autoimmune hepatitis. *Hepatology*. 2010;52:999-1007.
133. Vento S, Hegarty JE, Bottazzo G, Macchia E, Williams R, Eddleston AL. Antigen specific suppressor cell function in autoimmune chronic active hepatitis. *Lancet*. 1984;1:1200-4.
134. Longhi MS, Ma Y, Mitry RR, Bogdanos DP, Heneghan M, Cheeseman P, et al. Effect of CD4+CD25+ regulatory T-cells on CD8 T-cell function in patients with autoimmune hepatitis. *J Autoimmun*. 2005;25:63-71.
135. Invernizzi P, Mackay I-R. Autoimmune liver diseases. *World J Gastroenterol*. 2008;14:3290-1.
136. Shevach EM, Piccirillo CA, Thornton AM, McHugh RS. Control of T cell activation by CD4+CD25+ suppressor T cells. *Novartis Found Symp*. 2003;252:24-36; discussion 36-44, 106-14.

137. Vergani D, Mieli-Vergani G. Aetiopathogenesis of autoimmune hepatitis. *World J Gastroenterol*. 2008;14:3306-12.
138. Longhi MS, Meda F, Wang P, Samyn M, Mieli-Vergani G, Vergani D, et al. Expansion and de novo generation of potentially therapeutic regulatory T cells in patients with autoimmune hepatitis. *Hepatology*. 2008;47:581-91.
139. Longhi MS, Liberal R, Holder B, Robson SC, Ma Y, Mieli-Vergani G, et al. Inhibition of interleukin-17 promotes differentiation of CD25<sup>-</sup> cells into stable T regulatory cells in patients with autoimmune hepatitis. *Gastroenterology*. 2012;142:1526-35.e6.
140. La Cava A, Van Kaer L, Fu-Dong-Shi. CD4+CD25+ Tregs and NKT cells: regulators regulating regulators. *Trends Immunol*. 2006;27:322-7.
141. Kurokohchi K, Masaki T, Himoto T, Deguchi A, Nakai S, Morishita A, et al. Usefulness of liver infiltrating CD86-positive mononuclear cells for diagnosis of autoimmune hepatitis. *World J Gastroenterol*. 2006;12:2523-9.
142. Loetscher P, Ugucioni M, Bordoli L, Baggiolini M, Moser B, Chizzolini C, et al. CCR5 is characteristic of Th1 lymphocytes. *Nature*. 1998;391:344-5.
143. Ajuebor MN, Hogaboam CM, Le T, Proudfoot AEI, Swain MG. CCL3/MIP-1alpha is pro-inflammatory in murine T cell-mediated hepatitis by recruiting CCR1-expressing CD4(+) T cells to the liver. *Eur J Immunol*. 2004;34:2907-18.
144. Wen L, Ma Y, Bogdanos DP, Wong FS, Demaine A, Mieli-Vergani G, et al. Pediatric autoimmune liver diseases: the molecular basis of humoral and cellular immunity. *Curr Mol Med*. 2001;1:379-89.
145. Bittencourt PL, Palácios SA, Cañado ELR, Porta G, Carrilho FJ, Laudanna AA, et al. Cytotoxic T lymphocyte antigen-4 gene polymorphisms do not confer susceptibility to autoimmune hepatitis types 1 and 2 in Brazil. *Am J Gastroenterol*. 2003;98:1616-20.
146. Agarwal K, Czaja AJ, Jones DE, Donaldson PT. Cytotoxic T lymphocyte antigen-4 (CTLA-4) gene polymorphisms and susceptibility to type 1 autoimmune hepatitis. *Hepatology*. 2000;31:49-53.

147. Umemura T, Ota M, Yoshizawa K, Katsuyama Y, Ichijo T, Tanaka E, et al. Association of cytotoxic T-lymphocyte antigen 4 gene polymorphisms with type 1 autoimmune hepatitis in Japanese. *Hepatol Res.* 2008;38:689-95.
148. Djilali-Saiah I, Ouellette P, Caillat-Zucman S, Debray D, Kohn JI, Alvarez F. CTLA-4/CD 28 region polymorphisms in children from families with autoimmune hepatitis. *Hum Immunol.* 2001;62:1356-62.
149. Fan L-Y, Tu X-Q, Cheng Q-B, Zhu Y, Feltens R, Pfeiffer T, et al. Cytotoxic T lymphocyte associated antigen-4 gene polymorphisms confer susceptibility to primary biliary cirrhosis and autoimmune hepatitis in Chinese population. *World J Gastroenterol.* 2004;10:3056-9.
150. Brizzolara R, Montagna P, Soldano S, Cutolo M. Rapid interaction between CTLA4-Ig (abatacept) and synovial macrophages from patients with rheumatoid arthritis. *J Rheumatol.* 2013;40:738-40.
151. Scalapino KJ, Daikh DI. CTLA-4: a key regulatory point in the control of autoimmune disease. *Immunol Rev.* 2008;223:143-55.
152. Anthony RS, Mckelvie ND, Craig JI, Parker AC. Fas antigen (CD95) expression in peripheral blood progenitor cells from patients with leukaemia and lymphoma. *Leuk Lymphoma.* 1998;30:449-58.
153. Ogawa S, Sakaguchi K, Takaki A, Shiraga K, Sawayama T, Mouri H, et al. Increase in CD95 (Fas/APO-1)-positive CD4+ and CD8+ T cells in peripheral blood derived from patients with autoimmune hepatitis or chronic hepatitis C with autoimmune phenomena. *J Gastroenterol Hepatol.* 2000;15:69-75.
154. Tsirikoni A, Kyriakou DS, Rigopoulou EI, Alexandrakis MG, Zachou K, Passam F, et al. Markers of cell activation and apoptosis in bone marrow mononuclear cells of patients with autoimmune hepatitis type 1 and primary biliary cirrhosis. *J Hepatol.* 2005;42:393-9.
155. Pittoni V, Valesini G. The clearance of apoptotic cells: implications for autoimmunity. *Autoimmun Rev.* 2002;1:154-61.
156. Takahashi T, Tanaka M, Brannan CI, Jenkins NA, Copeland NG, Suda T, et al. Generalized lymphoproliferative disease in mice, caused by a point mutation in the Fas ligand. *Cell.* 1994;76:969-76.

157. Hiraide A, Imazeki F, Yokosuka O, Kanda T, Kojima H, Fukai K, et al. Fas polymorphisms influence susceptibility to autoimmune hepatitis. *Am J Gastroenterol*. 2005;100:1322-9.
158. Agarwal K, Czaja AJ, Donaldson PT. A functional Fas promoter polymorphism is associated with a severe phenotype in type 1 autoimmune hepatitis characterized by early development of cirrhosis. *Tissue Antigens*. 2007;69:227-35.
159. Cookson S, Constantini PK, Clare M, Underhill JA, Bernal W, Czaja AJ, et al. Frequency and nature of cytokine gene polymorphisms in type 1 autoimmune hepatitis. *Hepatology*. 1999;30:851-6.
160. Czaja AJ, Cookson S, Constantini PK, Clare M, Underhill JA, Donaldson PT. Cytokine polymorphisms associated with clinical features and treatment outcome in type 1 autoimmune hepatitis. *Gastroenterology*. 1999;117:645-52.
161. Bittencourt PL, Palácios SA, Caçado EL, Porta G, Drigo S, Carrilho FJ, et al. Autoimmune hepatitis in Brazilian patients is not linked to tumor necrosis factor alpha polymorphisms at position -308. *J Hepatol*. 2001;35:24-8.
162. Yoshizawa K, Ota M, Katsuyama Y, Ichijo T, Matsumoto A, Tanaka E, et al. Genetic analysis of the HLA region of Japanese patients with type 1 autoimmune hepatitis. *J Hepatol*. 2005;42:578-84.
163. Paladino N, Flores AC, Fainboim H, Schroder T, Cuarterolo M, Lezama C, et al. The most severe forms of type I autoimmune hepatitis are associated with genetically determined levels of TGF-beta1. *Clin Immunol*. 2010;134:305-12.
164. Szabo SJ, Kim ST, Costa GL, Zhang X, Fathman CG, Glimcher LH. A novel transcription factor, Tbet, directs Th1 lineage commitment. *Cell*. 2000;100:655-69.
165. Chen S, Zhao W, Tan W, Luo X, Dan Y, You Z, et al. Association of TBX21 promoter polymorphisms with type 1 autoimmune hepatitis in a Chinese population. *Hum Immunol*. 2011;72:69-73.
166. Ferreira AR, Roquete MLV, Toppa NH, de Castro LPF, Fagundes EDT, Penna FJ. Effect of treatment of hepatic histopathology in children and adolescents with autoimmune hepatitis. *J Pediatr Gastroenterol Nutr*. 2008;46:65-70.

167. Sogo T, Fujisawa T, Inui A, Komatsu H, Etani Y, Tajiri H, et al. Intravenous methylprednisolone pulse therapy for children with autoimmune hepatitis. *Hepatol Res.* 2006;34:187-92.
168. Al-Chalabi T, Heneghan MA. Remission in autoimmune hepatitis: what is it, and can it ever be achieved? *Am J Gastroenterol.* 2007;102:1013-5.
169. Kerkar N, Annunziato RA, Foley L, Schmeidler J, Rumbo C, Emre S, et al. Prospective analysis of nonadherence in autoimmune hepatitis: a common problem. *J Pediatr Gastroenterol Nutr.* 2006;43:629-34.
170. Samaroo B, Samyn M, Buchanan C, Mieli-Vergani G. Long-term daily oral treatment with prednisolone in children with autoimmune liver disease does not affect final adult height. *Hepatology.* 2006;44:438A.
171. Alvarez F. Autoimmune hepatitis and primary sclerosing cholangitis. *Clin Liver Dis.* 2006;10:89-107, vi.
172. Chang M-H, Hadzic D, Rouassant SH, Jonas M, Kohn IJ, Negro F, et al. Acute and chronic hepatitis: Working Group report of the second World Congress of Pediatric Gastroenterology, Hepatology, and Nutrition. *J Pediatr Gastroenterol Nutr.* 2004;39 Suppl 2:S584-8.
173. Saadah OI, Smith AL, Hardikar W. Long-term outcome of autoimmune hepatitis in children. *J Gastroenterol Hepatol.* 2001;16:1297-302.
174. Ferreira AR, Roquete MLV, Penna FJ, Toppa NH, Castro LPF de. [Type 1 autoimmune hepatitis in children and adolescents: assessment of immunosuppressive treatment withdrawal]. *J Pediatr (Rio J).* 81:343-8.
175. Schramm C, Wahl I, Weiler-Normann C, Voigt K, Wiegand C, Glaubke C, et al. Health-related quality of life, depression, and anxiety in patients with autoimmune hepatitis. *J Hepatol.* 2014;60:618-24.
176. Della Corte C, Sartorelli MR, Sindoni CD, Girolami E, Giovannelli L, Comparcola D, et al. Autoimmune hepatitis in children: an overview of the disease focusing on current therapies. *Eur J Gastroenterol Hepatol.* 2012;24:739-46.
177. Mieli-Vergani G, Vergani D. Autoimmune hepatitis in children. *Clin Liver Dis.* 2002;6:623-34.

178. Czaja AJ, Freese DK, American Association for the Study of Liver Disease. Diagnosis and treatment of autoimmune hepatitis. *Hepatology*. 2002;36:479-97.
179. Czaja AJ, Menon KVN, Carpenter HA. Sustained remission after corticosteroid therapy for type 1 autoimmune hepatitis: a retrospective analysis. *Hepatology*. 2002;35:890-7.
180. Rumbo C, Emerick KM, Emre S, Shneider BL. Azathioprine metabolite measurements in the treatment of autoimmune hepatitis in pediatric patients: a preliminary report. *J Pediatr Gastroenterol Nutr*. 2002;35:391-8.
181. Alvarez F, Ciocca M, Cañero-Velasco C, Ramonet M, de Davila MT, Cuarterolo M, et al. Short-term cyclosporine induces a remission of autoimmune hepatitis in children. *J Hepatol*. 1999;30:222-7.
182. Cuarterolo M, Ciocca M, Velasco CC, Ramonet M, González T, López S, et al. Follow-up of children with autoimmune hepatitis treated with cyclosporine. *J Pediatr Gastroenterol Nutr*. 2006;43:635-9.
183. Liberal R, Vergani D, Mieli-Vergani G. Paediatric Autoimmune Liver Disease. *Dig Dis*. 2015;33 Suppl 2:36-46.
184. Marlaka JR, Papadogiannakis N, Fischler B, Casswall TH, Beijer E, Németh A. Tacrolimus without or with the addition of conventional immunosuppressive treatment in juvenile autoimmune hepatitis. *Acta Paediatr*. 2012;101:993-9.
185. Mieli-Vergani G, Bargiota K, Samyn M, Vergani D. Therapeutic aspects of autoimmune liver disease in children. En: Dienes HP, Leuschner U, Lohse AW, Manns MP, editores. Autoimmune Liver Diseases-Falk Symposium. Dordrecht: Springer; 2005. p. 278-82.
186. el-Shabrawi M, Wilkinson ML, Portmann B, Mieli-Vergani G, Chong SK, Williams R, et al. Primary sclerosing cholangitis in childhood. *Gastroenterology*. 1987;92:1226-35.
187. Wurbs D, Klein R, Terracciano LM, Berg PA, Bianchi L. A 28-year-old woman with a combined hepatic/cholestatic syndrome. *Hepatology*. 1995;22:1598-605.
188. Gohlke F, Lohse AW, Dienes HP, Löhr H, Märker-Hermann E, Gerken G, et al. Evidence for an overlap syndrome of autoimmune hepatitis and primary sclerosing cholangitis. *J Hepatol*. 1996;24:699-705.



189. McNair AN, Moloney M, Portmann BC, Williams R, McFarlane IG. Autoimmune hepatitis overlapping with primary sclerosing cholangitis in five cases. *Am J Gastroenterol*. 1998;93:777-84.
190. Boberg KM, Chapman RW, Hirschfield GM, Lohse AW, Manns MP, Schrupf E, et al. Overlap syndromes: the International Autoimmune Hepatitis Group (IAIHG) position statement on a controversial issue. *J Hepatol*. 2011;54:374-85.
191. Hirschfield GM, Karlsen TH, Lindor KD, Adams DH. Primary sclerosing cholangitis. *Lancet (London, England)*. 2013;382:1587-99.
192. Hayee B, Samyn M, Shawcross D, Heneghan M, Bjarnason I. Autoimmune sclerosing cholangitis is associated with small bowel ulceration on capsule enteroscopy. *J Crohns Colitis*. 2014;8:S46.
193. Czaja A, Carpenter HA. Validation of scoring system for diagnosis of autoimmune hepatitis. *Dig Dis Sci*. 1996;41:305-14.
194. Bianchi FB, Cassani F, Lenzi M, Ballardini G, Muratori L, Giostra F, et al. Impact of international autoimmune hepatitis group scoring system in definition of autoimmune hepatitis. An Italian experience. *Dig Dis Sci*. 1996;41:166-71.
195. Toda G, Zeniya M, Watanabe F, Imawari M, Kiyosawa K, Nishioka M, et al. Present status of autoimmune hepatitis in Japan--correlating the characteristics with international criteria in an area with a high rate of HCV infection. Japanese National Study Group of Autoimmune Hepatitis. *J Hepatol*. 1997;26:1207-12.
196. Dickson RC, Gaffey MJ, Ishitani MB, Roarty TP, Driscoll CJ, Caldwell SH. The international autoimmune hepatitis score in chronic hepatitis C. *J Viral Hepat*. 1997;4:121-8.
197. van Buuren HR, van Hoogstraten HJE, Terkivatan T, Schalm SW, Vleggaar FP. High prevalence of autoimmune hepatitis among patients with primary sclerosing cholangitis. *J Hepatol*. 2000;33:543-8.
198. Omagari K, Masuda J, Kato Y, Nakata K, Kanematsu T, Kusumoto Y, et al. Re-analysis of clinical features of 89 patients with autoimmune hepatitis using the revised scoring system proposed by the International Autoimmune Hepatitis Group. *Intern Med*. 2000;39:1008-12.

199. Kaya M, Angulo P, Lindor KD. Overlap of autoimmune hepatitis and primary sclerosing cholangitis: an evaluation of a modified scoring system. *J Hepatol*. 2000;33:537-42.
200. Qiu D, Wang Q, Wang H, Xie Q, Zang G, Jiang H, et al. Validation of the simplified criteria for diagnosis of autoimmune hepatitis in Chinese patients. *J Hepatol*. 2011;54:340-7.
201. Czaja AJ. Comparability of probable and definite autoimmune hepatitis by international diagnostic scoring criteria. *Gastroenterology*. 2011;140:1472-80.
202. Ebbeson RL, Schreiber RA. Diagnosing autoimmune hepatitis in children: is the International Autoimmune Hepatitis Group scoring system useful? *Clin Gastroenterol Hepatol*. 2004;2:935-40.
203. Czaja AJ. Autoimmune hepatitis. 5.<sup>a</sup> ed. Pathology of the liver. New York: Churchill Livingstone; 2007.
204. Tobi H, van den Berg PB, de Jong-van den Berg LTW. Small proportions: what to report for confidence intervals? *Pharmacoepidemiol Drug Saf*. 2005;14:239-47.
205. Thulin M. The cost of using exact confidence intervals for a binomial proportion. *Electron J Stat*. 2014;8:817-40.
206. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44:837-45.
207. Ferri PM, Ferreira AR, Miranda DM, Simões E Silva AC. Diagnostic criteria for autoimmune hepatitis in children: a challenge for pediatric hepatologists. *World J Gastroenterol*. 2012;18:4470-3.
208. Jones T. The diagnosis of rheumatic fever. *JAMA*. 1944;126:481-4.
209. Ferrieri P, Jones Criteria Working Group. Proceedings of the Jones Criteria workshop. *Circulation*. 2002;106:2521-3.
210. Mileti E, Rosenthal P, Peters MG. Validation and modification of simplified diagnostic criteria for autoimmune hepatitis in children. *Clin Gastroenterol Hepatol*. 2012;10:417-21.e1-2.

211. Hiejima E, Komatsu H, Sogo T, Inui A, Fujisawa T. Utility of simplified criteria for the diagnosis of autoimmune hepatitis in children. *J Pediatr Gastroenterol Nutr.* 2011;52:470-3.
212. Mieli-Vergani G, Vergani D. Autoimmune liver diseases in children - what is different from adulthood? *Best Pract Res Clin Gastroenterol.* 2011;25:783-95.
213. Yilmaz B, Unlu O, Evcen R, Ugurluoglu C. Acute onset seronegative autoimmune hepatitis: are simplified diagnostic criteria sufficient? *Eur J Gastroenterol Hepatol.* 2016;28:607-8.
214. Ovchinsky N, Moreira RK, Lefkowitz JH, Lavine JE. Liver biopsy in modern clinical practice: a pediatric point-of-view. *Adv Anat Pathol.* 2012;19:250-62.
215. Gassert DJ, Garcia H, Tanaka K, Reinus JF. Corticosteroid-responsive cryptogenic chronic hepatitis: evidence for seronegative autoimmune hepatitis. *Dig Dis Sci.* 2007;52:2433-7.
216. Ranucci G, Socha P, Iorio R. Wilson disease: what is still unclear in pediatric patients? *Clin Res Hepatol Gastroenterol.* 2014;38:268-72.
217. Kaler SG. Inborn errors of copper metabolism. *Handb Clin Neurol.* 2013;113:1745-54.
218. Squires RH, Shneider BL, Bucuvalas J, Alonso E, Sokol RJ, Narkewicz MR, et al. Acute liver failure in children: the first 348 patients in the pediatric acute liver failure study group. *J Pediatr.* 2006;148:652-8.
219. Doménech JM, Navarro Pastor JB. Regresión logística binaria, multinomial, de Poisson y binomial negativa. 6.ª ed. Barcelona: Signo; 2011. 145-147.
220. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol.* 1996;49:1373-9.
221. Marín Reina P, Pereda Pérez A, Moreno A, Esteban MJ. Biopsia hepática por PAAF con control ecográfico: una técnica segura en nuestro medio. *Bol Soc Val Pediatr.* 2008;28:142.
222. Hajian-Tilaki K. Sample size estimation in diagnostic test studies of biomedical informatics. *J Biomed Inform.* 2014;48:193-204.
223. Jones SR, Carley S, Harrison M. An introduction to power and sample size estimation. *Emerg Med J.* 2003;20:453-8.

224. Buderer NM. Statistical methodology: I. Incorporating the prevalence of disease into the sample size calculation for sensitivity and specificity. *Acad Emerg Med*. 1996;3:895-900.
225. Echeverry J, Ardila E. Pruebas diagnósticas y proceso diagnóstico. En: Ardila E, Sánchez R, Echeverry J, editores. *Estrategias de Investigación en Medicina Clínica*. Bogotá: Manual Moderno; 2001. p. 135-68.
226. Kramer HC. *Medical test: objective and quantitative guidelines*. California: SAGE Publications Inc; 1992.
227. Arcos Machancoses J V., Molera Busoms C, Julio Tatis E, Bovo M V., Quintero Bernabeu J, Juampérez Goñi J, et al. Exactitud de los criterios simplificados de 2008 para el diagnóstico de la hepatitis autoinmune en población pediátrica. *Rev española pediatría clínica e Investig*. 2016;72:123-4.
228. Santiago Pérez MI, Hervada Vidal X, Naveira Barbeito G, Silva LC, Fariñas H, Vázquez E. El programa Epidat: usos y perspectivas. *Rev Panam Salud Pública*. 2010;27:80-2.
229. Malhotra RK, Indrayan A. A simple nomogram for sample size for estimating sensitivity and specificity of medical tests. *Indian J Ophthalmol*. 58:519-22.
230. Carley S, Dosman S, Jones SR, Harrison M. Simple nomograms to calculate sample size in diagnostic studies. *Emerg Med J*. 2005;22:180-1.
231. Schuetz GM, Schlattmann P, Dewey M. Use of 3x2 tables with an intention to diagnose approach to assess clinical performance of diagnostic tests: meta-analytical evaluation of coronary CT angiography studies. *BMJ*. 2012;345:e6717.
232. Fujiwara K, Yasui S, Tawada A, Fukuda Y, Nakano M, Yokosuka O. Diagnostic value and utility of the simplified International Autoimmune Hepatitis Group criteria in acute-onset autoimmune hepatitis. *Liver Int*. 2011;31:1013-20.
233. Qiu D, Wang Q, Wang H, Xie Q, Zang G, Jiang H, et al. Validation of the simplified criteria for diagnosis of autoimmune hepatitis in Chinese patients. *J Hepatol*. 2011;54:340-7.
234. Liu F, Pan ZG, Ye J, Xu D, Guo H, Li GP, et al. Primary biliary cirrhosis-autoimmune hepatitis overlap syndrome: simplified criteria may be effective in the diagnosis in Chinese patients. *J Dig Dis*. 2014;15:660-8.

235. Muratori P, Granito A, Pappas G, Muratori L. Validation of simplified diagnostic criteria for autoimmune hepatitis in Italian patients. *Hepatology*. 2009;49:1782-3; author reply 1783.
236. Kochar R, Fallon M. Diagnostic criteria for autoimmune hepatitis: what is the gold standard? *Hepatology*. 2010;51:350-1; author reply 351.
237. Czaja AJ. Performance parameters of the diagnostic scoring systems for autoimmune hepatitis. *Hepatology*. 2008;48:1540-8.
238. Delgado M, Llorca J, Doménech JM. Estudios para pruebas diagnósticas y factores pronósticos. Barcelona: Signo; 2005.
239. Phelps CE, Hutson A. Estimating diagnostic test accuracy using a «fuzzy gold standard». *Med Decis Making*. 1995;15:44-57.
240. Knottnerus JA, van Weel C, Muris JWM. Evaluation of diagnostic procedures. *BMJ*. 2002;324:477-80.
241. Stalpers LJA, Nelemans PJ, Geurts SME, Jansen E, de Boer P, Verbeek ALM. The indication area of a diagnostic test. Part II--the impact of test dependence, physician's decision strategy, and patient's utility. *J Clin Epidemiol*. 2015;68:1129-37.
242. Stalpers LJA, Nelemans PJ, Geurts SME, Jansen E, de Boer P, Verbeek ALM. The indication area of a diagnostic test. Part I--discounting gain and loss in diagnostic certainty. *J Clin Epidemiol*. 2015;68:1120-8.
243. Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Stat Med*. 1998;17:857-72.
244. Gardner MJ, Altman DG. Statistics with confidence: Confidence intervals and statistical guidelines. 1.<sup>a</sup> ed. London: British Medical Journal; 1989.
245. Dallal GE. An analytic approximation to the distribution of Lilliefors's test statistic for normality. *Am Stat*. 1986;40:294-6.
246. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin Chem*. 1993;39:561-77.

247. Doménech JM, Granero R. Macro !ROC for SPSS Statistics v2010.02.04. Bellaterra: Universitat Autònoma de Barcelona; 2010.
248. Doménech JM. Macro !DT for SPSS Statistics c2009.06.26. Bellaterra: Universitat Autònoma de Barcelona; 2009.
249. Doménech JM, Granero R. Macro !CIP for SPSS Statistics v2012.01.02. Bellaterra: Universitat Autònoma de Barcelona; 2012.
250. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011;155:529-36.
251. European Network for Health Technology Assessment. Guideline. Meta-analysis of diagnostic test accuracy studies. HIQA - Ireland, editor. EUnetHTA Joint Action 2; 2014. 19-34.
252. Chappell FM, Raab GM, Wardlaw JM. When are summary ROC curves appropriate for diagnostic meta-analyses? *Stat Med*. 2009;28:2653-68.
253. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *J Clin Epidemiol*. 2005;58:882-93.
254. Chen Y, Liu Y, Ning J, Cormier J, Chu H. Supplementary Materials for «A hybrid model for combining case-control and cohort studies in systematic reviews of diagnostic tests». *J R Stat Soc Ser C Appl Stat*. 2015;64:469-89.
255. Zamora J, Abaira V, Muriel A, Khan K, Coomarasamy A. Meta-DiSc: a software for meta-analysis of test accuracy data. *BMC Med Res Methodol*. 2006;6:31.
256. Harbord R. METANDI: Stata module to perform meta-analysis of diagnostic accuracy. Boston College Department of Economics; 2008.
257. Sterne J. METAFUNNEL: Stata module to produce funnel plots for meta-analysis. Boston College Department of Economics; 2003.

258. Gonçalves C, Ferreira M, Ferreira S, Nobre S, Gonçalves I. Use of diagnostic scores in children's autoimmune liver disease - IAIHG scores can rule out the diagnosis in clinical practice. Oral Communication. Oporto: LASPGHAN Annual Meeting. Oporto; 2017.
259. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull.* 1979;86:420-8.
260. Doménech JM, Granero R. Macro !KAPPA for SPSS Statistics. Weighted Kappa. V2009.07.31. Bellaterra: Universitat Autònoma de Barcelona; 2009.
261. Doménech JM, Granero R. Macro !AGREE for SPSS Statistics. Passing-Bablock & Bland-Altman methods. V2009.06.30. Bellaterra: Universitat Autònoma de Barcelona; 2009.
262. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33:159-74.
263. Fleiss JL, Levin B, Paik MC. Statistical methods for rates and proportions. 3ª edición. Hoboken NJ, editor. New York: John Wiley & Sons; 2003. 604.
264. Parés A. [Primary sclerosing cholangitis: diagnosis, prognosis and treatment]. *Gastroenterol Hepatol.* 2011;34:41-52.
265. Stoop JW, Zegers BJ, Sander PC, Ballieux RE. Serum immunoglobulin levels in healthy children and adults. *Clin Exp Immunol.* 1969;4:101-12.
266. Kleinbaum DG, Klein M. Logistic regression. A self-learning text. 2ª. New York: Springer-Verlag; 2002.
267. Hosmer D, Lemeshow S. Applied logistic regression. 2ª. New York: John Wiley & Sons; 2000.
268. Rubin DB, Schenker N. Multiple imputation in health-care databases: an overview and some applications. *Stat Med.* 1991;10:585-98.
269. Sesma R. Extension Command UAB AllSetsReg v0.0.7 (c) JM Doménech & JB Navarro. Barcelona: Laboratori d'Estadística Aplicada - Universitat Autònoma de Barcelona; 2012.
270. Nagelkerke NJD. A note of general definition of the coefficient of determination. *Biometrika.* 1991;78:691-2.

271. Younossi ZM, Koenig AB, Abdelatif D, Fazel Y, Henry L, Wymer M. Global epidemiology of nonalcoholic fatty liver disease-Meta-analytic assessment of prevalence, incidence, and outcomes. *Hepatology*. 2016;64:73-84.
272. Mameli C, Zuccotti GV, Carnovale C, Galli E, Nannini P, Cervia D, et al. An update on the assessment and management of metabolic syndrome, a growing medical emergency in paediatric populations. *Pharmacol Res*. 2017;119:99-117.
273. Dezsófi A, Baumann U, Dhawan A, Durmaz O, Fischler B, Hadzic N, et al. Liver biopsy in children: position paper of the ESPGHAN Hepatology Committee. *J Pediatr Gastroenterol Nutr*. 2015;60:408-20.
274. Piccinino F, Sagnelli E, Pasquale G, Giusti G. Complications following percutaneous liver biopsy. A multicentre retrospective study on 68,276 biopsies. *J Hepatol*. 1986;2:165-73.
275. Djulbegovic B, Desoky AH. Equation and nomogram for calculation of testing and treatment thresholds. *Med Decis Making*. 1996;16:198-9.
276. Djulbegovic B, van den Ende J, Hamm RM, Mayrhofer T, Hozo I, Pauker SG, et al. When is rational to order a diagnostic test, or prescribe treatment: the threshold model as an explanation of practice variation. *Eur J Clin Invest*. 2015;45:485-93.
277. Czaja AJ. Safety issues in the management of autoimmune hepatitis. *Expert Opin Drug Saf*. 2008;7:319-33.
278. Camarena Grande MC. Hepatitis autoinmune. En: Sociedad Española de Gastroenterología Hepatología y Nutrición Pediátrica, editor. Tratamiento en gastroenterología, hepatología y nutrición pediátrica. 3.ª ed. Majadahonda (Madrid): Ergon; 2012. p. 427-38.
279. Pauker SG, Kassirer JP. The threshold approach to clinical decision making. *N Engl J Med*. 1980;302:1109-17.
280. Latour J. Análisis de decisiones. Quaderns de salut pública i administració de serveis de salut, 12. València: Institut Valencià d'Estudis en Salut Pública; 2003.
281. Steyerberg EW, Van Calster B, Pencina MJ. Medidas del rendimiento de modelos de predicción y marcadores pronósticos: evaluación de las predicciones y clasificaciones. *Rev Española Cardiol*. Elsevier; 2011;64:788-94.



282. Hlatky MA, Greenland P, Arnett DK, Ballantyne CM, Criqui MH, Elkind MS V, et al. Criteria for evaluation of novel markers of cardiovascular risk: a scientific statement from the American Heart Association. *Circulation*. 2009;119:2408-16.
283. Pencina MJ, D'Agostino RB, D'Agostino RB, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med*. 2008;27:157-72; discussion 207-12.
284. Gulati A, Jabbour A, Ismail TF, Guha K, Khwaja J, Raza S, et al. Association of fibrosis with mortality and sudden cardiac death in patients with nonischemic dilated cardiomyopathy. *JAMA*. 2013;309:896-908.
285. Leening MJG, Steyerberg EW. Fibrosis and mortality in patients with dilated cardiomyopathy. *JAMA*. 2013;309:2547-8.
286. Leening MJG, Vedder MM, Wittteman JCM, Pencina MJ, Steyerberg EW. Net reclassification improvement: computation, interpretation, and controversies: a literature review and clinician's guide. *Ann Intern Med*. 2014;160:122-31.
287. Van Calster B, Vickers AJ, Pencina MJ, Baker SG, Timmerman D, Steyerberg EW. Evaluation of markers and risk prediction models: overview of relationships between NRI and decision-analytic measures. *Med Decis Making*. 2013;33:490-501.
288. Peirce CS. The numerical measure of the success of predictions. *Science*. 1884;4:453-4.
289. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 26:565-74.
290. Kerr KF, Brown MD, Zhu K, Janes H. Assessing the Clinical Impact of Risk Prediction Models With Decision Curves: Guidance for Correct Interpretation and Appropriate Use. *J Clin Oncol*. 2016;34:2534-40.
291. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ*. 2016;352:i6.
292. Primo J. Calculadora para diagnóstico de la red CASPe [Internet]. Critical Appraisal Skills Programme (en español). 2015 [citado 26 de julio de 2016]. Recuperado a partir de: <http://www.redcaspe.org/herramientas/calculadoras>

293. Dwamena B. MIDAS: Stata module for meta-analytical integration of diagnostic test accuracy studies. *Stat Softw Components*. Boston College Department of Economics; 2009;
294. Harbord RM, Whiting P. metandi: Meta-analysis of diagnostic accuracy using hierarchical logistic regression. *Stata J*. StataCorp LP; 2009;9:211-29.
295. Dwamena B. FAGAN: Stata module for Fagan's Bayesian nomogram. *Stat Softw Components*. Boston College Department of Economics; 2009;
296. Barkhordari M, Padyab M, Hadaegh F, Azizi F, Bozorgmanesh M. Stata Modules for Calculating Novel Predictive Performance Indices for Logistic Models. *Int J Endocrinol Metab*. 2016;14:e26707.
297. Rodríguez Artalejo F, Banegas Banegas JR, González Enríquez J, Martín Moreno JM, Vilar Álvarez F. Análisis de decisiones clínicas. *Med Clin*. 1990;94:348-54.
298. Latour J. Análisis de decisiones. Quaderns de salut pública i administració de serveis de salut, 12. València: Institut Valencià d'Estudis en Salut Pública; 1997.
299. Pita Fernández S, Pértegas Díaz S. Pruebas diagnósticas: Sensibilidad y especificidad. *Cad Aten Primaria*. 2003;10:120-4.
300. Sackett DL, Haynes RB. The architecture of diagnostic research. *BMJ*. 2002;324:539-41.
301. Fescina RH, Simini F, Belitzky R. Evaluación de los procedimientos diagnósticos. Aspectos metodológicos. *Salud Perinat PP*. 1985;2:39-43.
302. Almazán C, Espallargues M. La evaluación de pruebas diagnósticas: aplicación al diagnóstico por la imagen. Conceptos básicos. Informatiu AATM (Agència d'Avaluació de Tecnologia i Recerca Mèdiques de Catalunya) Número 23. 2001.
303. Popper K. Conjeturas y refutaciones: El desarrollo del conocimiento científico. 1.<sup>a</sup> ed. Barcelona: Paidós; 1987. 152.
304. Page MJ, Shamseer L, Tricco AC. Registration of systematic reviews in PROSPERO: 30,000 records and counting. *Syst Rev*. 2018;7:32.
305. Ruano J, Gómez-García F, Gay-Mimbrera J, Aguilar-Luque M, Fernández-Rueda JL, Fernández-Chaichio J, et al. Evaluating characteristics of PROSPERO records as predictors of eventual

- publication of non-Cochrane systematic reviews: a meta-epidemiological study protocol. *Syst Rev*. 2018;7:43.
306. Kohn MA, Carpenter CR, Newman TB. Understanding the direction of bias in studies of diagnostic test accuracy. *Acad Emerg Med*. 2013;20:1194-206.
307. Arcos-Machancoses JV, Molera Busoms C, Julio Tatis E, Bovo M V., Quintero Bernabeu J, Juampérez Goñi J, et al. Accuracy of the 2008 simplified criteria for the diagnosis of autoimmune hepatitis in children. *Pediatr Gastroenterol Hepatol Nutr*. 2018;21:118-26.
308. Ward AC. The role of causal criteria in causal inferences: Bradford Hill's «aspects of association». *Epidemiol Perspect Innov*. 2009;6:2.
309. Phillips C V, Goodman KJ. The missed lessons of Sir Austin Bradford Hill. *Epidemiol Perspect Innov*. 2004;1:3.
310. Schmidt RL, Factor RE. Understanding sources of bias in diagnostic accuracy studies. *Arch Pathol Lab Med*. 2013;137:558-65.
311. Mieli-Vergani G, Vergani D, Baumann U, Czubkowski P, Debray D, Dezsofi A, et al. Diagnosis and Management of Pediatric Autoimmune Liver Disease: ESPGHAN Hepatology Committee Position Statement. *J Pediatr Gastroenterol Nutr*. 2018;66:345-60.
312. Cho JM, Kim KM, Oh SH, Lee YJ, Rhee KW, Yu E. De novo autoimmune hepatitis in Korean children after liver transplantation: a single institution's experience. *Transplant Proc*. 2011;43:2394-6.
313. Lamia S, Sana K, Rachid J, Hajer A, Leila M, Nabil T, et al. Autoimmune hepatitis-primary sclerosing cholangitis overlap syndrome complicated by inflammatory bowel disease. *La Tunisie médicale*. 2012;90:899-900.
314. Mieli-Vergani G, Vergani D, Czaja AJ, Manns MP, Krawitt EL, Vierling JM, et al. Autoimmune hepatitis. *Nat Rev Dis Prim*. 2018;4:18017.
315. Maggiore G, Socie G, Sciveres M, Roque-Afonso A-M, Nastasio S, Johanet C, et al. Seronegative autoimmune hepatitis in children: Spectrum of disorders. *Dig Liver Dis*. 2016;48:785-91.

316. Pauker SG, Kassirer JP. Therapeutic decision making: a cost-benefit analysis. *N Engl J Med.* 1975;293:229-34.
317. Kerr KF, Wang Z, Janes H, McClelland RL, Psaty BM, Pepe MS. Net reclassification indices for evaluating risk prediction instruments: a critical review. *Epidemiology.* 2014;25:114-21.
318. Kassirer JP. Our stubborn quest for diagnostic certainty. A cause of excessive testing. *N Engl J Med.* 1989;320:1489-91.
319. Gaarder KR. Diagnosis. *South Med J.* 1989;82:1153-4.
320. Kassirer JP, Kopelman RI. The luxuriant language of diagnosis. *Hosp Pract (Off Ed).* 1989;24:36-8, 40, 42, 49.
321. Hui SL, Walter SD. Estimating the error rates of diagnostic tests. *Biometrics.* 1980;36:167-71.
322. Kassirer JP. Diagnostic reasoning. *Ann Intern Med.* 1989;110:893-900.
323. Sackett D, Haynes R, Guyatt G, Tugwell P. Interpretación de los datos diagnósticos. En: Sackett D, Haynes R, Guyatt G, Tugwell P, editores. *Epidemiología clínica Ciencia básica para la medicina clínica.* Buenos Aires: Editorial Médica Panamericana; 1994. p. 34-61.
324. Sandler G. The importance of the history in the medical clinic and the cost of unnecessary tests. *Am Heart J.* 1980;100:928-31.
325. Ibañez-Pradas V, Modesto i Alapont V. MBE en cirugía pediátrica. Lectura crítica de artículos. Pruebas diagnósticas (II). *Cir Pediatr.* 2006;19:130-5.
326. Okeh UM, Ugwu AC. Bayes' theorem: a paradigm research tool in biomedical sciences. *East Afr J Public Health.* 2009;6 Suppl:11-9.
327. Grenier B. *Décision médicale.* Paris: Masson; 1993.
328. Jiménez-Rivera C, Ling SC, Ahmed N, Yap J, Aglipay M, Barrowman N, et al. Incidence and Characteristics of Autoimmune Hepatitis. *Pediatrics.* 2015;136:e1237-48.
329. Yoshizawa K, Matsumoto A, Ichijo T, Umemura T, Joshita S, Komatsu M, et al. Long-term outcome of Japanese patients with type 1 autoimmune hepatitis. *Hepatology.* 2012;56:668-76.

330. Redondo Alvaro F. La lógica en la interpretación de pruebas diagnósticas. Madrid: Editorial Garsi; 1989.
331. Monsour M, Evans A, Kupper L. Confidence intervals for post-test probability. *Stat Med.* 1991;10:443-56.
332. Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic Research: Principles and Quantitative Methods. New York: John Wiley & Sons; 1982.
333. Grimes DA, Schulz KF. Refining clinical diagnosis with likelihood ratios. *Lancet (London, England).* 365:1500-5.
334. Brenner H, Gefeller O. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. *Stat Med.* 1997;16:981-91.
335. Fagan TJ. Letter: Nomogram for Bayes theorem. *N Engl J Med.* 1975;293:257.
336. Turing AM. On computable numbers, with an application to the Entscheidungsproblem. *Proc London Math Soc.* 1937;47:230-65.
337. Dick PK. ¿Sueñan los androides con ovejas eléctricas? 1.ª ed. Madrid: Ediciones Cátedra; 2015.
338. Turing AM. Computing machinery and intelligence. *Mind.* 1950;59:433-60.
339. Larrañaga P, Bielza C. Alan Turing and Bayesian statistics. *Mathw Soft Comput Mag.* 2012;19:23-4.
340. McGrayne SB. The theory that would not die: How Bayes' rule cracked the enigma code, hunted down Russian submarines, and emerged triumphant from two centuries of controversy. 1.ª ed. Londres: Yale University Press; 2011.
341. Good I. Studies in the history of probability and statistics. XXXVII A. M. Turing's statistical work in the World War II. *Biometrika.* 1979;66:393-6.
342. Mardia K V., Cooper SB. Alan Turing and enigmatic statistics. *Bol Of do Capitulo Bras da Int Soc Bayesian Anal.* 2012;5:2-7.
343. Miller RA. The cryptographic mathematics of Enigma. *Cryptologia.* 1995;19:65-80.

344. Jeffreys H. Theory of probability. Oxford: Clarendon; 1939.
345. Peirce CS. The probability of induction (Reimpresión de la versión original de 1878). En: Newman JR, editor. The World of Mathematics (Vol 2). New York: Simon and Schuster; 1956. p. 1341-54.
346. Gillies D. The Turing—Good weight of evidence function and Popper's measure of the severity of a test. *Br J Philos Sci.* 1990;41:143-6.
347. Youden WJ. Index for rating diagnostic tests. *Cancer.* 1950;3:32-5.
348. Escrig-Sos J, Martínez-Ramos D, Miralles-Tena JM. [Diagnostic tests: basic concepts for their correct interpretation and use]. *Cirugía española.* 2006;79:267-73.
349. Hulley SB, Cummings SR. Diseño de la investigación clínica. Madrid: Ediciones Doyma; 1993.
350. Latour J. El diagnóstico. Quaderns de salut pública i administració de serveis de salut, 21. València: Escola Valenciana d'Estudis per a la Salut; 2003.
351. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology.* 1982;143:29-36.
352. Hajian-Tilaki KO, Hanley JA. Comparison of three methods for estimating the standard error of the area under the curve in ROC analysis of quantitative data. *Acad Radiol.* 2002;9:1278-85.
353. Demler O V, Pencina MJ, D'Agostino RB. Misuse of DeLong test to compare AUCs for nested models. *Stat Med.* 2012;31:2577-87.
354. Burgueño MJ, García-Bastos JL, González-Buitrago JM. [ROC curves in the evaluation of diagnostic tests]. *Med clínica.* 1995;104:661-70.
355. López de Ullibarri Galparsoro I, Pita Fernández S. Curvas ROC. *Cad Aten Primaria.* 1998;5:229-35.
356. Lee WC. Probabilistic analysis of global performances of diagnostic tests: interpreting the Lorenz curve-based summary measures. *Stat Med.* 1999;18:455-71.
357. Pita S. Determinación del tamaño muestral. *Cad Aten Primaria.* 1996;3:138-44.

358. Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. *J Clin Epidemiol*. 1991;44:763-70.
359. Fleiss JL. Statistical methods for rates and proportions. 2.<sup>a</sup> ed. John Wiley & Sons; 1981. 38-48.
360. Robledo J. Diseños de muestreo (II). *NURE Investig*. 2005;12.
361. Epidat. Dirección Xeral de Innovación e Xestión da Saúde Pública de la Consellería de Sanidade (Xunta de Galicia);
362. Irwig L, Bossuyt P, Glasziou P, Gatsonis C, Lijmer J. Designing studies to ensure that estimates of test accuracy are transferable. *BMJ*. 2002;324:669-71.
363. Knottnerus JA, Muris JW. Assessment of the accuracy of diagnostic tests: the cross-sectional study. *J Clin Epidemiol*. 2003;56:1118-28.
364. Eggin TK, Feinstein AR. Context bias. A problem in diagnostic radiology. *JAMA*. 1996;276:1752-5.
365. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med*. 1978;299:926-30.
366. Begg CB. Biases in the assessment of diagnostic tests. *Stat Med*. 1987;6:411-23.
367. Lachs MS, Nachamkin I, Edelstein PH, Goldman J, Feinstein AR, Schwartz JS. Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infection. *Ann Intern Med*. 1992;117:135-40.
368. Knottnerus JA, Leffers P. The influence of referral patterns on the characteristics of diagnostic tests. *J Clin Epidemiol*. 1992;45:1143-54.
369. van der Schouw YT, Van Dijk R, Verbeek AL. Problems in selecting the adequate patient population from existing data files for assessment studies of new diagnostic tests. *J Clin Epidemiol*. 1995;48:417-22.
370. Joseph L, Gyorkos TW. Inferences for likelihood ratios in the absence of a «gold standard». *Med Decis Making*. 1996;16:412-7.

371. Feinstein A. Diagnostic and spectral markers. En: Feinstein A, editor. *Clinical Epidemiology The architecture of clinical research*. Philadelphia: Saunders Co; 1985. p. 597-631.
372. Choi BC. Sensitivity and specificity of a single diagnostic test in the presence of work-up bias. *J Clin Epidemiol*. 1992;45:581-6.
373. Cecil MP, Kosinski AS, Jones MT, Taylor A, Alazraki NP, Pettigrew RI, et al. The importance of work-up (verification) bias correction in assessing the accuracy of SPECT thallium-201 testing for the diagnosis of coronary artery disease. *J Clin Epidemiol*. 1996;49:735-42.
374. Begg CB, Greenes RA, Iglewicz B. The influence of uninterpretability on the assessment of diagnostic tests. *J Chronic Dis*. 1986;39:575-84.
375. Fleiss JL. The measurement of interrater agreement. En: Fleiss JL, editor. *Statistical methods for rates and proportions*. Toronto: John Wiley & Sons; 1981. p. 212-36.
376. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull*. 1968;70:213-20.
377. Bland JM, Altman DG. Measurement error. *BMJ*. 1996;313:744.
378. Altman DG, Bland JM. Comparing several groups using analysis of variance. *BMJ*. 1996;312:1472-3.
379. Fleiss JL. *The design and analysis of clinical experiments*. New York: John Wiley & Sons; 1986. 1-32.
380. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet (London, England)*. 1986;1:307-10.
381. Shapiro SS, Wilk MB. Analysis of variance test for normality (complete samples). *Biometrika*. 1965;52:591-611.
382. Sackett DL, Richardson WS, Rosenberg W, Haynes RB. *Medicina basada en la evidencia. Cómo ejercer y enseñar la MBE*. Madrid: Churchill Livingstone; 1997.
383. Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, et al. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *JAMA*. 1996;276:637-9.



384. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *BMJ*. 2003;326:41-4.
385. 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.
386. Pauker SG. Clinical decision making: handling and analyzing clinical data. En: Goldman L, Bennet JC, editores. *Cecil's Textbook of Medicine*. WB Saunders; 2000. p. 76-82.
387. Ebel MH. Evidence-based diagnosis: A handbook of clinical prediction rules. 1.<sup>a</sup> ed. New York: Springer-Verlag; 2001.
388. Wasson JH, Sox HC, Neff RK, Goldman L. Clinical prediction rules. Applications and methodological standards. *N Engl J Med*. 1985;313:793-9.
389. López Piñero JM. El papiro de Edwin smith. Medicina, historia, sociedad Antología de clásicos médicos. Esplugues de Llobregat (Barcelona): Editorial Ariel; 1973.
390. López Piñero JM. La colección hipocrática. Antología de clásicos médicos. Madrid: Editorial Triacastela; 1998.
391. McGinn T, Guyatt G, Wyer P, Naylor CD, Stiell IG. Diagnosis. Clinical prediction rules. En: Guyatt G, Rennie D, editores. *User's guides to the medical literature A manual for evidence-based clinical practice*. Chicago: AMA Press; 2002. p. 471-83.
392. Laupacis A, Sekar N, Stiell IG. Clinical prediction rules. A review and suggested modifications of methodological standards. *JAMA*. 1997;277:488-94.
393. Stiell IG, Greenberg GH, McKnight RD, Nair RC, McDowell I, Reardon M, et al. Decision rules for the use of radiography in acute ankle injuries. Refinement and prospective validation. *JAMA*. 1993;269:1127-32.
394. Stiell IG, McKnight RD, Greenberg GH, McDowell I, Nair RC, Wells GA, et al. Implementation of the Ottawa ankle rules. *JAMA*. 1994;271:827-32.

395. Auleley GR, Ravaud P, Giraudeau B, Kerboull L, Nizard R, Massin P, et al. Implementation of the Ottawa ankle rules in France. A multicenter randomized controlled trial. *JAMA*. 1997;277:1935-9.
396. Mark DB, Shaw L, Harrell FE, Hlatky MA, Lee KL, Bengtson JR, et al. Prognostic value of a treadmill exercise score in outpatients with suspected coronary artery disease. *N Engl J Med*. 1991;325:849-53.
397. Pauker SG, Kassirer JP. Decision analysis. *N Engl J Med*. 1987;316:250-8.
398. Kassirer JP, Pauker SG. Should diagnostic testing be regulated? *N Engl J Med*. 1978;299:947-9.
399. Richards RJ. Using threshold analysis to improve medical decisions. *Hosp Pract (1995)*. 1997;32:15-6, 19-21, 25-6.
400. Nease RF, Bonduelle Y. Solid recommendations from soft numbers: the test/treatment decision. *Med Decis Making*. 7:220-33.
401. Glasziou P. Threshold analysis via the Bayes' nomogram. *Med Decis Making*. 11:61-2.
402. Djulbegovic B, Hozo I, Lyman GH. Linking evidence-based medicine therapeutic summary measures to clinical decision analysis. *MedGenMed*. 2000;2:E6.
403. Halligan S. Systematic reviews and meta-analysis of diagnostic tests. *Clin Radiol*. 2005;60:977-9.
404. Sacks HS, Reitman D, Pagano D, Kupelnick B. Meta-analysis: an update. *Mt Sinai J Med*. 63:216-24.
405. Sousa MR de, Ribeiro ALP. Systematic review and meta-analysis of diagnostic and prognostic studies: a tutorial. *Arq Bras Cardiol*. 2009;92:229-38, 235-45.
406. Irwig L, Tosteson AN, Gatsonis C, Lau J, Colditz G, Chalmers TC, et al. Guidelines for meta-analyses evaluating diagnostic tests. *Ann Intern Med*. 1994;120:667-76.
407. Velanovich V. Meta-analysis for combining Bayesian probabilities. *Med Hypotheses*. 1991;35:192-5.

408. Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. *Med Decis Making*. 13:313-21.
409. Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Stat Med*. 1993;12:1293-316.
410. Pepe MS. The statistical evaluation of medical tests for classification and prediction. New York: Oxford University Press; 2003.
411. Update Software Ltd. La Biblioteca Cochrane Plus [Internet]. 2016 [citado 3 de agosto de 2016]. Recuperado a partir de: <http://www.biblioteca-cochrane.com/>
412. Zhou A, Obuchowski N, McClish D. Issues in meta-analysis for diagnostic tests. En: Zhou A, Obuchowski N, McClish D, editores. *Statistical methods in diagnostic medicine*. New York: Wiley & Sons; 2002. p. 222-40.
413. Haynes RB, Kastner M, Wilczynski NL, Hedges Team. Developing optimal search strategies for detecting clinically sound and relevant causation studies in EMBASE. *BMC Med Inform Decis Mak*. 2005;5:8.
414. Wilczynski NL, Haynes RB, Hedges Team. Developing optimal search strategies for detecting clinically sound prognostic studies in MEDLINE: an analytic survey. *BMC Med*. 2004;2:23.
415. Moayyedi P. Meta-analysis: Can we mix apples and oranges? *Am J Gastroenterol*. 2004;99:2297-301.
416. Egger M, Smith GD. Bias in location and selection of studies. *BMJ*. 1998;316:61-6.
417. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Quality of Reporting of Meta-analyses. *Lancet (London, England)*. 1999;354:1896-900.
418. Zamora J, Abaira V, Muriel A, Khan KS, Coomarasamy A. Meta-DiSc: a software for meta-analysis of test accuracy data. *BMC Med Res Methodol*. 2006;6:31.

419. Devillé WL, Buntinx F, Bouter LM, Montori VM, de Vet HCW, van der Windt DAWM, et al. Conducting systematic reviews of diagnostic studies: didactic guidelines. *BMC Med Res Methodol.* 2002;2:9.
420. Agresti A. Analysis of ordinal categorical data. New York: John Wileys & Sons; 1994.
421. Deeks JJ. Systematic reviews of evaluations of diagnostic and screening tests. En: Egger M, Smith GD, Altman DG, editores. Systematic reviews in health care Meta-analysis in context. Londres: BMJ Books; 2001. p. 248-82.
422. Dinnes J, Deeks J, Kirby J, Roderick P. A methodological review of how heterogeneity has been examined in systematic reviews of diagnostic test accuracy. *Health Technol Assess.* 2005;9:1-113, iii.
423. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ.* 2003;327:557-60.
424. Leemis LM, Trivedi KS. A comparison of approximate interval estimators for the Bernoulli parameter. *Am Stat.* 1996;50:63-8.
425. Zhou A, Obuchowski N, McClish D. Statistical methods for meta-analysis. En: Zhou A, Obuchowski N, McClish D, editores. Statistical methods in diagnostic medicine. New York: John Wiley & Sons; 2002. p. 396-417.
426. Hasselblad V, Hedges L V. Meta-analysis of screening and diagnostic tests. *Psychol Bull.* 1995;117:167-78.
427. Glas AS, Lijmer JG, Prins MH, Bossel GJ, Bossuyt PMM. The diagnostic odds ratio: a single indicator of test performance. *J Clin Epidemiol.* 2003;56:1129-35.
428. The Numerical Algorithms Group. NAG C Library [Internet]. 2004 [citado 3 de agosto de 2016]. Recuperado a partir de: <http://www.nag.co.uk>
429. Zamora Romero J, Plana MN, Abaira Santos V. Estudios de evaluación de la validez de una prueba diagnóstica: revisión sistemática y metanálisis. *Nefrología.* 2009;29:15-20.
430. Walter SD. Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. *Stat Med.* 2002;21:1237-56.

431. Mitchell MD. Validation of the summary ROC for diagnostic test meta-analysis: a Monte Carlo simulation. *Acad Radiol*. 2003;10:25-31.
432. Whiting P, Rutjes AWS, Dinnes J, Reitsma J, Bossuyt PMM, Kleijnen J. Development and validation of methods for assessing the quality of diagnostic accuracy studies. *Health Technol Assess*. 2004;8:iii, 1-234.
433. Whiting P, Rutjes AWS, Reitsma JB, Bossuyt PMM, 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*. 2003;3:25.
434. González Rodríguez MP, Velarde Mayol P. Listas guía de comprobación de estudios sobre pruebas diagnósticas incluidos en las revisiones sistemáticas: declaración QUADAS. *Evidencias en Pediatría*. 2012;8:20.
435. Leeflang MMG, Deeks JJ, Takwoingi Y, Macaskill P. Cochrane diagnostic test accuracy reviews. *Syst Rev*. 2013;2:82.
436. University of Bristol. QUADAS. A quality assessment tool for diagnostic accuracy studies. [Internet]. [citado 26 de julio de 2016]. Recuperado a partir de: <http://www.bris.ac.uk/quadas>
437. Whiting P, Rutjes AWS, Dinnes J, Reitsma JB, Bossuyt PMM, Kleijnen J. A systematic review finds that diagnostic reviews fail to incorporate quality despite available tools. *J Clin Epidemiol*. 2005;58:1-12.



13

Anexos



## **13.1. Los estudios de evaluación de sistemas diagnósticos**

---

### **13.1.1. El proceso diagnóstico**

Según la definición enunciada por Kassirer en 1989 el diagnóstico puede concebirse como un proceso de contraste de una hipótesis acerca de la naturaleza de la enfermedad de un paciente que se deriva de observaciones a través del uso de la inferencia. A su vez, se entiende por inferencia la propiedad de combinar los hechos observados en el paciente con la base de conocimiento del que se dispone, seleccionado los datos y pasos apropiados para presentar como resultado la solución al contraste de dicha hipótesis [318]. El diagnóstico es un resultado de alta significación tanto para el médico como para el paciente, aunque ambos protagonistas del proceso pueden no converger en la interpretación que dan del mismo [318,319]. La pretensión holística de esta definición entra en conflicto con la realidad reconocida tanto por los teóricos de la filosofía de la ciencia como por los médicos asistenciales, que es el hecho de que la enfermedad de un paciente no puede determinarse con certeza absoluta prácticamente nunca, independientemente de cuánta información se obtenga, cuántas observaciones se hagan o cuántas pruebas diagnósticas se realicen. Así pues, el objetivo del médico no puede ser alcanzar la certeza de un diagnóstico, sino reducir el nivel de incertidumbre lo suficiente como para tomar la decisión terapéutica [318,320,321].

La complejidad del proceso diagnóstico viene dada por el grado de incertidumbre de partida sobre la situación clínica problema, para la que se reconocen una serie de fuentes o causas: En primer lugar, que el conjunto de síntomas y signos en un paciente puede ser compatible con más de una enfermedad. Además, existen variaciones biológicas a menudo relevantes entre un enfermo y otro. A esto se suma que los pacientes son inexactos para recordar sucesos pasados. Y finalmente, que tanto los médicos como los instrumentos de los



que se sirven, suelen ser imprecisos para recoger, ponderar e integrar la información clínica [301,322].

Durante el proceso diagnóstico el médico se vale de distintas fuentes de información. Por un lado, tenemos la información epidemiológica y los datos obtenidos de la anamnesis y la exploración clínica. Por otro lado, tenemos los resultados de las llamadas pruebas complementarias u otras pruebas diagnósticas (*tests*, en la voz inglesa). Estas últimas se definen como toda herramienta (instrumento, pregunta, síntoma, signo, medición de laboratorio o prueba de imagen) utilizada para disminuir el grado de incertidumbre que se tiene del estado de un paciente en relación a un problema de salud real, subjetivo o potencial [323,324].

Atendiendo a este concepto, el término *sistema diagnóstico* podría incluso ser más idóneo (más intuitivo) que *prueba diagnóstica*, dado que dentro de la definición no solo se incluye la información proporcionada por las exploraciones complementarias (también llamadas *pruebas* complementarias, lo que puede ser origen de confusión), sino también la de los criterios de clasificación tal como se han descrito en el bloque anterior. Por este motivo, en esta tesis se preferirá hablar de sistemas diagnósticos, aunque ambas nomenclaturas se emplearán con la misma denotación.

Con todo, el diagnóstico es una de las habilidades del médico más apreciadas por los pacientes y por los mismos clínicos. Sin embargo, durante el grado el médico recibe escasa formación reglada en este terreno. Se ha planteado el temor que esto, unido a una deficiente calidad metodológica de algunos de los trabajos sobre evaluación de pruebas diagnósticas, puedan contribuir al fenómeno creciente de incorporar estrategias de diagnóstico (bien sean criterios de clasificación o exploraciones complementarias) de forma acrítica a la práctica clínica habitual con los consiguientes errores diagnósticos y encarecimiento sin rendimiento de los procesos médicos [325].

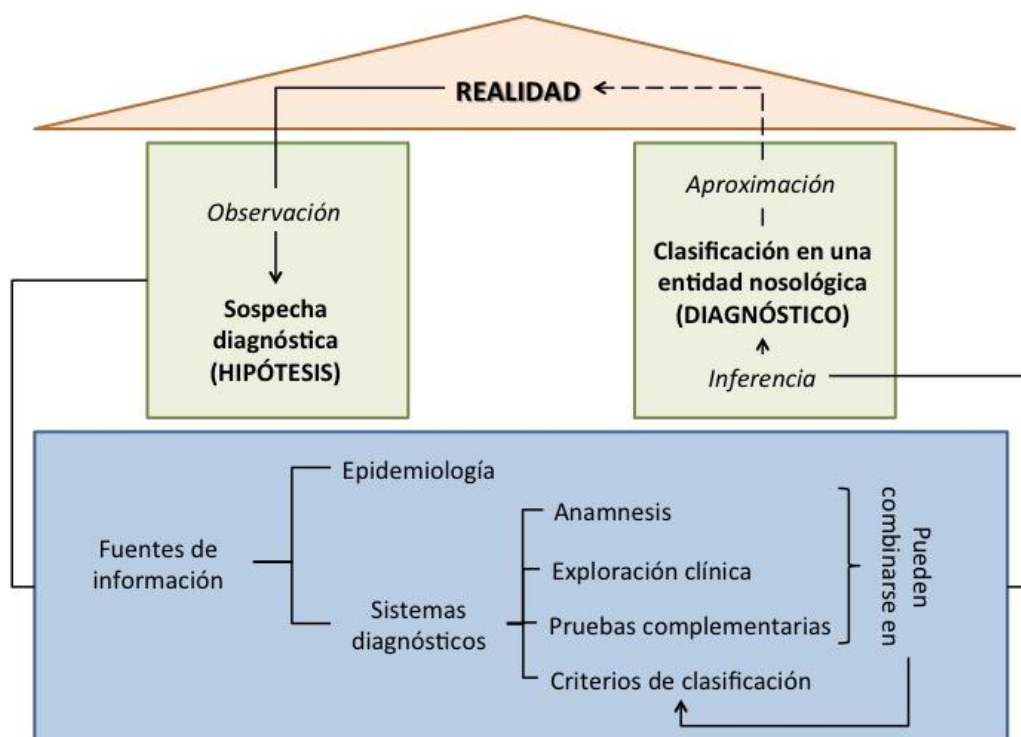


Figura 77: Componentes del proceso diagnóstico.

El uso eficiente de una prueba diagnóstica requiere entender previamente una serie de requisitos básicos. El primero es que una prueba diagnóstica es útil, desde el punto de vista clínico, solo si induce a tomar las decisiones (fundamentalmente terapéuticas) adecuadas. El segundo es que el diagnóstico es una actividad que el médico solo debe desarrollar en un ambiente de incertidumbre. Es decir, que el uso de los *tests* diagnósticos tiene sentido únicamente cuando la anamnesis, la exploración física y otras pruebas de diagnóstico básico no han proporcionado la suficiente certeza como para llevar a cabo una actitud terapéutica. Y el tercero es que solo tiene sentido plantear el uso de nuevas pruebas diagnósticas si sabemos que sus resultados van a ser capaces de disminuir nuestra situación de incertidumbre.

Así, el uso racional de un *test* requiere que el clínico: a) conozca la probabilidad de que el paciente presente la enfermedad antes de hacer el *test*

(probabilidad *a priori* o preprueba); b) conozca la capacidad que tiene el *test* de modificar esa probabilidad (probabilidad *a posteriori* o postprueba), y c) establezca el nivel de certeza que necesita tener para tomar una decisión terapéutica (umbral de acción).

El grado de incertidumbre o certeza que tenemos de que ocurra un evento (por ejemplo, que un determinado paciente tenga una hepatitis autoinmune), puede expresarse de dos maneras equivalentes: mediante una probabilidad o mediante una *odds*. Una probabilidad (un riesgo) es una cantidad entre 0 y 1 que coincide con la frecuencia de aparición del evento, expresada como el número de casos favorables partido por el total de casos. La *odds*, usada ampliamente por sus ventajas para el cálculo, es el mismo concepto pero expresado de una manera menos intuitiva: con una cantidad que oscila entre 0 e  $\infty$ , calculada con el número casos favorables partido por el de desfavorables (una probabilidad dividida por su complementaria).

Para conocer la capacidad que tiene un *test* diagnóstico de cambiar esa incertidumbre, se utiliza la sensibilidad, la especificidad y los valores predictivos. La sensibilidad y la especificidad se consideran los parámetros que mejor evalúan el rendimiento diagnóstico (validez interna) de una prueba: la sensibilidad representa la capacidad que tiene el *test* para detectar a los casos, y la especificidad, la capacidad que tiene el *test* para detectar a los sanos (no casos). Matemáticamente ambas son probabilidades condicionales, y se expresarían de la siguiente forma: sensibilidad =  $p(+|E)$ , es decir, probabilidad de que el *test* sea positivo en el caso de que sujeto esté enfermo; especificidad =  $p(-|\bar{E})$ , probabilidad de que el *test* sea negativo si el sujeto está sano (no enfermo). Ambos valores se obtienen tras aplicar la prueba a poblaciones en las que se conoce con certeza su estatus de enfermedad. Una prueba extremadamente sensible se utiliza para descartar la presencia de enfermedad. Por el contrario, una prueba extremadamente específica se utiliza para asegurar la presencia de enfermedad. Aunque idealmente las pruebas diagnósticas

deberían tener una alta sensibilidad y especificidad, por regla general ambos parámetros guardan una relación inversa, que viene representada por la curva COR.

Hay que señalar que la dificultad capital consiste en que en la práctica clínica diaria no se necesitan estos parámetros: la incertidumbre radica exactamente en que se desconoce el estado de salud del paciente, y lo que se conoce con certeza es el resultado del *test* o de los criterios de clasificación. La pregunta que el médico se hace a pie de cama es si el resultado positivo o negativo de la prueba es correcto o no. La respuesta son otras probabilidades condicionales: los valores predictivos del *test*. El valor predictivo positivo (VPP) es  $p(E|+)$ , probabilidad de que el sujeto esté enfermo dado que el *test* sea positivo (un signo patognomónico tendrá un VPP del 100%), y el negativo (VPN) es  $p(\bar{E}|-)$ , probabilidad de que el sujeto esté sano dado que el *test* sea negativo. En efecto, la expresión  $p(E|+)$  (VPP) es absolutamente diferente de  $p(+|E)$  (Sensibilidad). No es una sutileza baladí, sino la conocida falacia de transposición de los condicionales [325]. El valor predictivo global se define como la probabilidad que tiene una prueba de acertar.

El problema que plantea el uso de los valores predictivos es que su cálculo no es directo desde la sensibilidad y la especificidad, sino que dependen de la prevalencia de enfermedad: la probabilidad de enfermedad previa a hacer el *test*. Este inconveniente será discutido el bloque correspondiente con el ejemplo del problema diagnóstico de la hepatitis autoinmune en pediatría con los criterios simplificados, dado que los trabajos existentes al respecto muestran un diseño que no permite el cálculo de los valores predictivos porque no se estima la prevalencia de la enfermedad en el grupo de pacientes con situación clínica y analítica compatible.

Las probabilidades condicionales se rigen mediante el teorema de Bayes, y si se expresa el valor predictivo a partir de este teorema se comprueba que el resultado final depende de la prevalencia. Así, siendo A y B sucesos dependientes cualesquiera (y  $p$ , probabilidad), el teorema de Bayes establece que [326]:

$$p(A|B) = \frac{p(A) \times p(B|A)}{p(A) \times p(B|A) + p(\bar{A}) \times p(B|\bar{A})}$$

Que aplicado al VPP resulta en:

$$p(E|+) = \frac{p(E) \times p(+|E)}{p(E) \times p(+|E) + p(\bar{E}) \times p(+|\bar{E})}$$

Donde:

$p(E)$  = Probabilidad de estar enfermo = Prevalencia

$p(+|E)$  = Probabilidad de *test* positivo si se está enfermo = Sensibilidad

$p(\bar{E})$  = Probabilidad de no estar enfermo = 1 – Prevalencia

$p(+|\bar{E})$  = Probabilidad de *test* positivo si no se está enfermo (falso positivo) =  
= 1 – Especificidad

Así pues, se obtiene que:

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

Si se aplica al VPN, el teorema de Bayes establece que:

$$p(\bar{E}|-) = \frac{p(\bar{E}) \times p(-|\bar{E})}{p(E) \times p(-|E) + p(\bar{E}) \times p(-|\bar{E})}$$

Donde:

$p(\bar{E})$  = Probabilidad de no estar enfermo = 1 – Prevalencia

$p(-|\bar{E})$  = Probabilidad de *test* negativo si no se está enfermo = Especificidad

$p(E)$  = Probabilidad de estar enfermo = Prevalencia

$p(-|E)$  = Probabilidad de *test* negativo si se está enfermo (falso negativo) =  
= 1 – Sensibilidad

Y en este caso, se obtiene que:

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

Se observa que el VPP de un *test* o criterio diagnóstico aumenta si se aplica en un escenario con elevada prevalencia de la enfermedad de interés. Esto explica por qué los médicos de los niveles terciarios aciertan más con las mismas herramientas que los de atención primaria, ya que la primaria actúa como un filtro y aumenta la prevalencia de determinados procesos en la asistencia especializada.

De acuerdo con el impacto de la clasificación que se le asigna a un caso en función del resultado de una prueba complementaria o la aplicación de unos criterios diagnósticos, se obtiene el siguiente esquema de tipos básicos de enfermedades [327]:

**Tipo I:** Son todas aquellas enfermedades donde la peor equivocación que se puede cometer en el diagnóstico es un falso negativo.

**Tipo II:** Son todas aquellas enfermedades donde la peor equivocación que se puede cometer en el diagnóstico es un falso positivo.

**Tipo III:** Son las restantes, donde no se puede clasificar claramente como una de las anteriores.

Las enfermedades que se pueden clasificar como de tipo I son enfermedades curables si se detectan a tiempo, enfermedades graves que no pueden dejarse pasar inadvertidas y aquellas cuyo falso positivo no tiene un impacto psicológico ni económico importante para el paciente pero los diagnósticos falsos negativos sí. Son ejemplos el infarto de miocardio, enfermedades infecciosas curables y errores innatos del metabolismo con buen pronóstico si se instaura un tratamiento precoz como la fenilcetonuria. La hepatitis autoinmune también sería una enfermedad de tipo I dado que responde excelentemente al tratamiento y en estadios precoces se puede evitar su evolución a cirrosis. Además es una enfermedad potencialmente grave a largo plazo, con un tratamiento sencillo en la mayor parte de los casos y con una mortalidad poco relevante en el primer mundo y en la época actual [80,328,329]. El hecho de considerar peor un falso negativo en esta situación

justificará posteriormente algunos aspectos metodológicos de la tesis, como el método de asignación de los puntos de corte óptimos para los modelos diagnósticos a evaluar.

En cambio, una enfermedad donde para el paciente, un diagnóstico falso positivo es más peligroso que un falso negativo, se considera como del tipo II. Es el caso de enfermedades graves pero difícilmente curables o sin remisión, en las que es muy importante para el paciente o para la población el saberse un verdadero negativo, en la que los falsos positivos traumatizan seriamente al paciente o aquellas en las que el tratamiento de los falsos positivos ocasiona serios perjuicios al paciente. A modo de ejemplos se cuentan enfermedades en fase terminal, cáncer oculto, tributarias de tratamientos quimioterápicos agresivos, cirugías innecesarias o lobectomía.

Cuando la enfermedad no puede ser encasillada en ninguno de los dos casos anteriores, entonces se la considera como del tipo III, como por ejemplo: lupus eritematoso, ciertas formas de leucemia o linfoma, diabetes. En el caso del sida un falso positivo puede ser muy dañino para el paciente, pero un falso negativo sería muy peligroso para la sociedad y como ambos peligros son graves, lo mejor es clasificarla como del tipo III desde un punto de vista ético. Cabe destacar que una misma enfermedad puede ser clasificada de maneras diferentes, de acuerdo con el estadio en el que se encuentre el paciente.

### **13.1.2. Evaluación de la validez de un sistema diagnóstico**

La validez es el grado en el que los resultados de una medición corresponden al fenómeno real que se mide. La medición de la prueba o sistema diagnóstico problema se compara con un estándar aceptado. Fruto de esta comparación son los parámetros de validez interna que se describirán a continuación [330].

Para valorar una prueba se recurre normalmente a la construcción de una tabla de contingencia con cuatro casillas mutuamente excluyentes, en la que las

columnas representan el resultado según el criterio de verdad y las filas el resultado según la prueba o los criterios a evaluar.

Tabla 53: Distribución de casos para valorar la validez de una prueba diagnóstica.

Criterio de verdad →	Enfermedad	Ausencia de enfermedad
Prueba +	Verdadero positivo (VP)	Falso positivo (FP)
Prueba –	Falso negativo (FN)	Verdadero negativo (VN)

En los bloques sucesivos se expondrán las definiciones de los principales indicadores de validez y de fiabilidad de pruebas diagnósticas para una mejor comprensión posterior de los resultados de la tesis. Estos indicadores, al igual que los datos de estudios analíticos o experimentales, son susceptibles de agregarse mediante técnicas de meta-análisis, que en el caso de los estudios de pruebas diagnósticas tienen unas particularidades que se discutirán en el anexo correspondiente.

### 13.1.2.1. Sensibilidad

Es el porcentaje de resultados positivos en pacientes con una determinada enfermedad y por lo tanto mide el porcentaje de individuos enfermos correctamente diagnosticados. Es indicador de un bajo número de falsos negativos, motivo por el que las pruebas con una sensibilidad muy elevada son muy útiles para descartar la presencia de enfermedad. Una elevada sensibilidad es la propiedad que interesa para aplicar a enfermedades graves que no pueden permanecer ignoradas y son tratables como, por ejemplo, en los programas de cribado de cáncer de mama. Siguiendo el esquema de enfermedades ya mencionado, aquí encajarían las tipo I.

$$Se = \frac{VP}{VP + FN}$$

El intervalo de confianza al 95% para la sensibilidad, según la ley normal, tiene la siguiente expresión:



$$IC95\% (Se) = Se \pm 1,96 \times \sqrt{\frac{Se \times (1 - Se)}{\text{número de enfermos}}}$$

### 13.1.2.2. Especificidad

Es el porcentaje de resultados negativos en pacientes que no padecen esa enfermedad. Valora la capacidad de una prueba para detectar correctamente individuos sanos. Es indicador de una baja frecuencia de falsos positivos. La especificidad busca confirmar al que no tiene un proceso. Se desea una prueba con elevada especificidad en enfermedades importantes de difícil tratamiento efectivo y cuando el hecho de saber que no se tiene la enfermedad posee cierta importancia sanitaria o psicológica para el paciente, es decir, las enfermedades de tipo II.

$$Sp = \frac{VN}{VN + FP}$$

El intervalo de confianza al 95% para la especificidad, de forma análoga al de la sensibilidad, tiene la siguiente expresión:

$$IC95\% (Sp) = Sp \pm 1,96 \times \sqrt{\frac{Sp \times (1 - Sp)}{\text{número de sanos}}}$$

### 13.1.2.3. Valor predictivo positivo

El VPP es el porcentaje de pacientes enfermos entre todos los pacientes con resultados positivos. Valora la probabilidad de que una prueba positiva diagnostique correctamente a un individuo enfermo. Sería por tanto el porcentaje de pacientes enfermos con resultados positivos con respecto al total de resultados positivos.

Un VPP del 90% indica que de cada 100 pacientes que dan la prueba positiva solo 90 padecen la enfermedad, o lo que es lo mismo: si la prueba da positiva, la probabilidad de padecer la enfermedad es de un 90%. En este caso habrá un 10 % de individuos sanos diagnosticados incorrectamente como enfermos, sin que esto implique que haya un 10% de falsos positivos.

$$VPP = \frac{VP}{VP + FP}$$

De forma general, según la propuesta de Monsour, Evans y Kupper, se puede calcular su intervalo de confianza  $1-\alpha$  (siendo  $\alpha$  el error que se admite) a partir de la prevalencia ( $P$ ), la sensibilidad ( $Se$ ) y especificidad ( $Sp$ ) [331]. Emplearemos este método porque en el bloque correspondiente del trabajo se necesitará estimarlos con la fórmula de Bayes a partir de los datos obtenidos de otros estudios de la bibliografía utilizando nuestra prevalencia para mejorar la validez externa. En la expresión (error  $\alpha$  del 5% en el ejemplo),  $m_0$  hace referencia el número total de no casos (sanos);  $m_1$ , al número de casos (enfermos) y  $n$ , al tamaño muestral total.

$$IC95\% VPP = \frac{1}{1 + \frac{(1-Sp) \times (1-P)}{Se \times P} \times e^{\pm 1,96 \times \sqrt{\frac{1-Se}{m_1 \times Se} + \frac{Sp}{m_0(1-Sp)} + \frac{1}{n \times P(1-P)}}}}$$

#### 13.1.2.4. Valor predictivo negativo

El valor predictivo negativo Indica la frecuencia de pacientes no enfermos entre todos los pacientes con resultado negativo. Valora la probabilidad de que una prueba negativa diagnostique correctamente a un individuo sano. Es el porcentaje de individuos sanos con resultados negativos con relación al total de resultados negativos.

Un valor predictivo negativo del 95% indica que, de cada 100 pruebas negativas, 95 pertenecerán a individuos sanos, o lo que es igual, si la prueba da negativa, la probabilidad de no padecer la enfermedad es del 95%. En este caso hay un 5% de individuos enfermos que serán diagnosticados erróneamente como no enfermos, que tampoco significa que haya un 5% de falsos negativos.

$$VPN = \frac{VN}{VN + FN}$$

Puesto que esos dos índices VPP y VPN son los que interesan en la práctica clínica, parecería natural utilizarlos como índices de comparación a la hora de evaluar dos métodos diagnósticos diferentes. Sin embargo, presentan un grave inconveniente, ya que si se calculan a partir de la tabla dependen de la proporción de enfermos en la muestra estudiada. Por ello, para una determinada prueba,

resulta necesario determinar unos índices de valoración que, respondiendo a las necesidades reales en cuanto a la clasificación de pacientes, no dependan de esa proporción de enfermos en la muestra, tales como las razones de verosimilitud, que se describirán a continuación.

El intervalo de confianza al 95% para el VPN se calculará según la siguiente expresión, siguiendo el mismo principio que para el VPP:

$$IC95\% VPN = \frac{1}{1 + \frac{(1-Se) \times P}{Sp \times (1-P)} \times e^{\pm 1,96 \times \sqrt{\frac{Se}{m_1 \times (1-Se)} + \frac{1-Sp}{m_0 \times Sp} + \frac{1}{n \times P(1-P)}}}}$$

### 13.1.2.5. Razones de verosimilitud

Las razones de verosimilitud son una serie de parámetros para intentar resumir en sí el significado conjunto de la sensibilidad y la especificidad. Existen para un resultado positivo (razón de verosimilitud positiva, RV+ o coeficiente de probabilidad positivo) y para un resultado negativo (razón de verosimilitud negativa, RV- o coeficiente de probabilidad negativo). Tienen gran arraigo en el mundo anglosajón porque se basan en el concepto de *odds* (el cociente de una probabilidad dividida por su complemento).

$$RV+ = \frac{Se}{1 - Sp} \qquad RV- = \frac{1 - Se}{Sp}$$

Por un lado, la RV+ relaciona la *odds* preprueba de diagnosticar la enfermedad con la *odds* postprueba de un resultado positivo. La *odds* preprueba es la *odds* de prevalencia y la *odds* postprueba es la *odds* del valor predictivo positivo:

$$\frac{VPP}{1 - VPP} = \frac{P}{1 - P} \times \frac{Se}{1 - Sp} = \frac{P}{1 - P} \times RV+$$

Cuanto mayor sea la RV+ respecto a 1, mayor es la contribución de un resultado positivo de la prueba en el diagnóstico de la enfermedad. Una RV de 9 indica que el resultado positivo es proporcionalmente nueve veces más frecuente en los enfermos que en los sanos. La interpretación de la RV+ es como sigue: valores

por encima de 10 ofrecen una RV+ excelente; entre 5 y 10, buena; entre 2 y 5, regular y entre 1 y 2, pobre.

Por otro lado, la RV- se define como el cociente entre el complementario de la sensibilidad y la especificidad. Según este concepto y aplicando el teorema de Bayes, la RV- es la conexión entre la *odds* preprueba de enfermedad y el recíproco de la *odds* postprueba del resultado negativo:

$$\frac{1 - VPN}{VPN} = \frac{P}{1 - P} \times \frac{1 - Se}{Sp} = \frac{P}{1 - P} \times RV -$$

Definida de esta manera la RV- valora la contribución que realiza un resultado negativo en la no confirmación de la enfermedad (cuanto menor sea mejor). Así definida resulta difícil de entender. Se mueve en la escala del 1 al 0, recíproca de la escala de la RV+, siendo tanto más importante cuanto más se aproxima a 0. Al moverse en una escala diferente no se puede comparar directamente con la RV+.

La ventaja de las RV+ y RV- frente a los VPP y VPP de la prueba radica en que, a diferencia de éstos, no dependen de la proporción de enfermos en la muestra, sino tan solo de la sensibilidad y especificidad de ésta, de ahí su utilidad a la hora de comparar pruebas diagnósticas.

El cálculo del intervalo de confianza es complejo y existen dos métodos aproximados para muestras grandes [332]. Si la RV es próxima a 1 se emplea la fórmula de Miettinen (donde  $\chi^2$  es el estadístico de asociación de la tabla de contingencia):

$$IC95\% RV = RV^{\pm \frac{1,96}{\sqrt{\chi^2}}}$$

Si la RV no es próxima a 1 se lleva a cabo una aproximación de primer orden del desarrollo de Taylor (donde a, b, c y d son los valores de las casillas de la tabla de contingencia ordenadas desde arriba a la izquierda y en el sentido de la lectura):

$$IC95\% RV+ = e^{\ln(RV+) \pm 1,96 \times \sqrt{\frac{1-Se}{d} + \frac{Sp}{c}}}$$

$$IC95\% RV- = e^{\ln(RV-) \pm 1,96 \times \sqrt{\frac{Se}{b} + \frac{S1-Sp}{a}}}$$

### 13.1.2.5.1. Comportamiento e interpretación de las razones de verosimilitud

La incidencia de las razones de verosimilitud sobre los valores predictivos y sobre la utilización diagnóstica de los signos y de las pruebas es tal que resulta útil analizar el respectivo papel de sus dos componentes, sensibilidad y especificidad.

Si la especificidad es constante, la RV+ es función lineal de la sensibilidad. Cuando la sensibilidad aumenta, la RV+ aumenta en la misma proporción.

La razón de verosimilitud negativa es igualmente una función lineal de la sensibilidad, pero negativa: cuando aumenta la sensibilidad, RV- disminuye en la misma proporción.

Cuando  $Se = 1 - Sp = 0,1$ , las dos razones de verosimilitud son iguales a 1 la prueba no tiene ningún valor diagnóstico. El valor diagnóstico del resultado (positivo o negativo) de la prueba crece linealmente cuando la sensibilidad aumenta de 0,10 a 1,00.

Cuando  $Sp = 1 - Se = 0,8$ , las dos razones de verosimilitud son iguales a 1 y por lo tanto en esta situación la prueba carece de utilidad diagnóstica. Se nota que la progresión de RV+ es muy rápida cuando la especificidad es superior a 0,7. La RV- se modifica poco cuando la especificidad crece de 0,50 a 1,00, mientras que la RV+ se afecta mucho por variaciones mínimas de la especificidad cuando ésta es muy elevada; así, cuando la especificidad crece de 0,90 a 0,95, RV+ pasa de 8 a 16.

Si la sensibilidad es constante, las dos razones de verosimilitud son funciones hiperbólicas de la especificidad. La RV+, en particular, crece muy rápidamente y tiende hacia el infinito cuando la especificidad de una prueba o criterio diagnóstico tiende hacia 1.

Se dice que la eficacia diagnóstica de una prueba o unos criterios de clasificación está estrechamente ligada al crecimiento de su especificidad, cualidad que, precisamente, es estable en el caso de los sujetos no afectados por la enfermedad.

Por lo que respecta a su interpretación comparativa, si tenemos dos pruebas diagnósticas A y B, y calculamos sus  $RV_{+}$ , y vemos que  $RV_{+A} > RV_{+B}$ , diremos que la prueba A es mejor que la B para confirmar la presencia de enfermedad. Referente a los  $RV_{-}$ , si vemos que  $RV_{-A} < RV_{-B}$ , diremos que la prueba A es mejor que la B para confirmar la ausencia de enfermedad.

Comparando los cocientes de probabilidad de dos pruebas diagnósticas existen cuatro posibles resultados que se pueden reflejar en una gráfica. En el eje de las Y se representa la sensibilidad, en el eje de las X se representa (1-Especificidad). El punto representa los valores de Sensibilidad y (1-Especificidad) observados para una prueba A. Trazamos una línea desde el origen (0,0) que pase por el punto A y otra desde el valor (1,1) que también pase por ese punto, quedando dividido el gráfico en cuatro zonas, de tal manera que cualquier otra prueba cuyo resultado representemos en esa gráfica estará en alguna de las cuatro zonas.

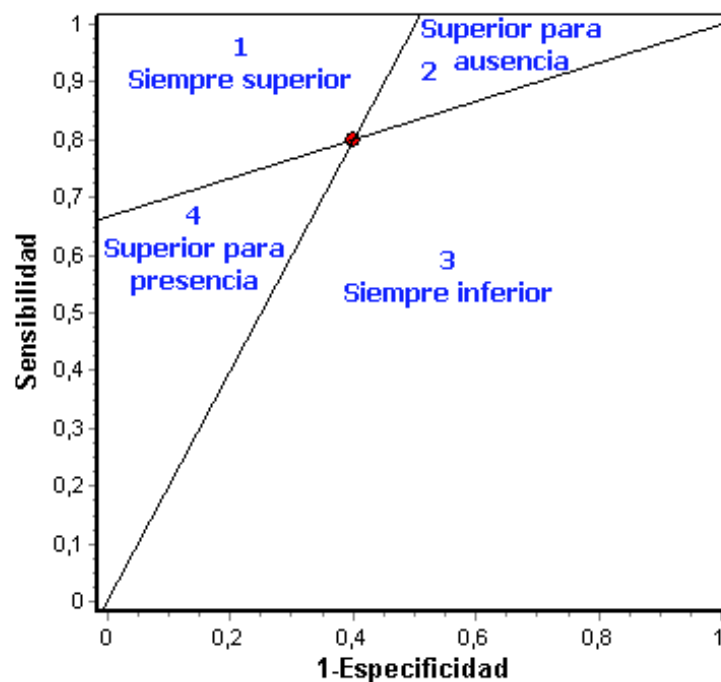


Figura 78: Ejemplo de representación de la validez de un sistema diagnóstico en una gráfica Sensibilidad / 1-Especificidad para establecer comparaciones con otros sistemas o pruebas diagnósticas.

Las zonas corresponden a las siguientes situaciones:

1. En esta región una hipotética prueba B es siempre superior a la A, tanto para confirmar presencia como ausencia de enfermedad.
2. Aquí la prueba B solo es superior a la A para confirmar ausencia de enfermedad.
3. En esta zona B siempre es inferior a A.
4. Si el resultado de B está en esta área, será superior a A para confirmar la presencia de enfermedad.

Todas estas consideraciones se refieren únicamente a la capacidad discriminatoria de una prueba. La decisión sobre qué prueba diagnóstica es más adecuada, es siempre algo más complicado, puesto que intervienen otros aspectos, como son el coste de la prueba, los riesgos que supone para el paciente y la valoración de las consecuencias que conlleva un falso positivo o un falso negativo.

Para comparar los parámetros de dos pruebas diagnósticas habrá que considerar que los valores obtenidos son solo estimaciones y están sometidos por tanto a posibles errores de muestreo, por lo que habrá que efectuar el correspondiente contraste estadístico para determinar si las diferencias encontradas son suficientemente importantes como para no poder ser atribuidas al azar.

A modo de resumen, los resultados de las razones de verosimilitud de un diagnóstico dado por un sistema en estudio se interpretan de la siguiente forma [333]:

$RV+ >10$  o  $RV- <0,1$  → Generan cambios amplios y a menudo concluyentes desde una probabilidad preprueba hasta una probabilidad postprueba.

$RV+ 5$  a  $10$  y  $RV- 0,1$  a  $0,2$  → Generan cambios moderados desde la probabilidad preprueba hasta la probabilidad postprueba.

$RV+ 2$  a  $5$  y  $RV- 0,2$  a  $0,5$  → Generan cambios pequeños, aunque en ocasiones pueden ser importantes, de la probabilidad postprueba.

$RV+ 1$  a  $2$  y  $RV- 0,5$  a  $1$  → Alteran la probabilidad postprueba en un grado normalmente insignificante.

Además si se conoce o se puede hacer una estimación de la probabilidad preprueba de que un sujeto padezca la enfermedad, utilizando las razones de verosimilitud, al realizar la prueba se podrá corregir ese valor de acuerdo con el resultado, de tal manera que la probabilidad aumenta o disminuye según que el resultado sea positivo o negativo, aplicando la siguiente fórmula donde  $P$  es la probabilidad preprueba y  $VP$  y  $RV$  deben de ser del mismo signo [334]:

$$VP = \frac{P \times RV}{1 + P \times (RV - 1)}$$

Finalmente cabe mencionar una herramienta clásica que describe esta relación entre las razones de verosimilitud y la probabilidad preprueba (prevalencia) para estimar la probabilidad postprueba y, por lo tanto, los valores predictivos. Fue descrita por Fagan en el año 1975 y se inspira en el teorema de Bayes para convertirlo en una función lineal simple. Se emplea trazando una línea que conecte la prevalencia (la línea de la izquierda) con la  $RV+$  o  $RV-$  (línea central) y leyendo la probabilidad postprueba (línea de la derecha) cuya interpretación, positiva o negativa, depende del signo de la  $RV$  empleada [335].

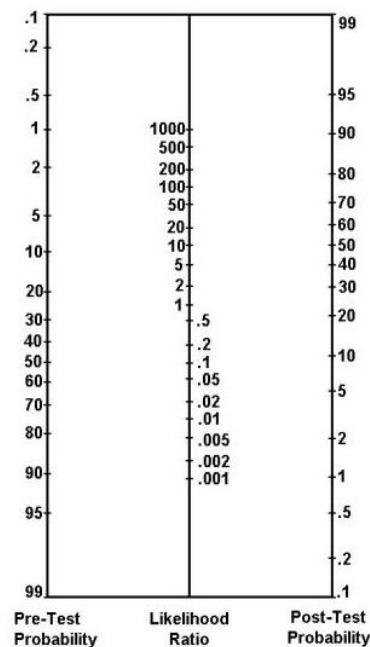


Figura 79: Nomograma de Fagan.



### 13.1.2.5.1.1. La contribución de Turing

El trabajo pionero de Alan Mathison Turing (Londres 1912 – Chesire 1954) no solo se circunscribe a las máquinas de computación lógica y su relación con el problema de decisión propuesto por David Hilbert (*Entscheidungsproblem*, es decir, la cuestión sobre si las matemáticas permiten definir un método que pueda aplicarse a cualquier sentencia matemática y que nos diga si esa sentencia es cierta o no), que resolvió en un sentido negativo [336]. Es sencillo constatar que esta es la faceta de su trabajo más conocida. De hecho ha llegado a la cultura popular por la analogía del ficticio *test* de Voight-Kampff [337] con el *test* de Turing [338], un intento de determinar un proceso (*juego*, tal como fue descrito inicialmente) para demostrar si una máquina autónoma está pensando. Sin embargo, A. M. Turing también hizo contribuciones significativas a la estadística basadas en el análisis bayesiano, de las que se han beneficiado disciplinas tales como la biología, la lingüística y la criptografía, esta última de forma fundamental durante la Segunda Guerra Mundial.

La estadística bayesiana, a diferencia de la estadística frecuentista que se fundamenta en la idea de cuantificar la probabilidad de un suceso a partir de la frecuencia relativa de aparición, parte de la noción de que la probabilidad representa el grado de creencia que otorgamos al suceso en cuestión. La aproximación bayesiana explica la manera en que cada persona revisa su creencia en el suceso una vez que recibe nueva información. Es decir, a medida que recabamos nueva evidencia, decidimos si dicha nueva información apoya la hipótesis de partida o si, por el contrario, la nueva evidencia favorece una nueva hipótesis alternativa. De acuerdo a este esquema de actualización secuencial de las hipótesis, no solo cuentan las evidencias recibidas a favor de una hipótesis, sino también la perspectiva personal que otorgue el observador a la experiencia previa [339].

En la época de Turing esto se consideraba demasiado subjetivo y poco científico. Sin embargo, los gobiernos británico y americano lo utilizaban para resolver problemas bélicos de alto secreto. Afortunadamente los excelentes

resultados que ha proporcionado esta aproximación bayesiana en innumerables problemas del mundo real han servido para constatar su superioridad sobre la aproximación frecuentista [340].

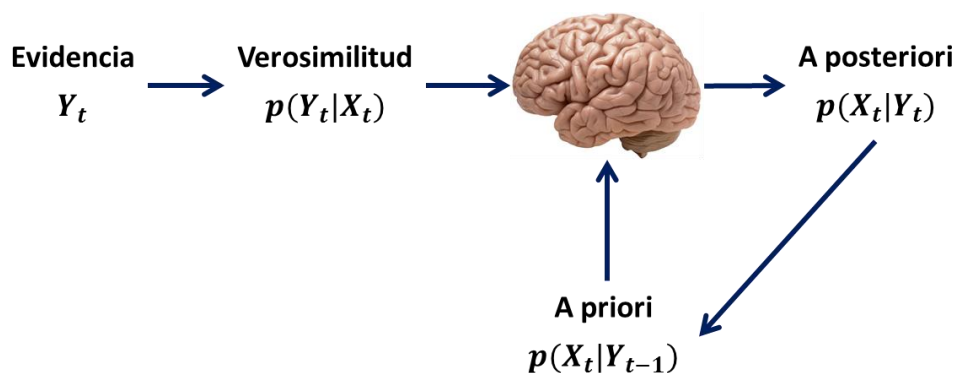


Figura 80: Esquema del razonamiento bayesiano para la solución del contraste de la hipótesis Y en base a la información X en un universo  $t$ .

Esta faceta estadística del trabajo de Turing fue registrada por Irwing J. Good, quien a principios de los años cuarenta le asistió en el trabajo que llevó al descifrado de la máquina Enigma. En efecto, las aportaciones de Turing a la estadística se desarrollaron en relación con el algoritmo *bamburismus* que sirvió para descifrar los mensajes enviados por la armada naval germana, durante la Segunda Guerra Mundial. Dichos mensajes eran de capital importancia para la población británica cuyo abastecimiento dependía de manera crítica de la supervivencia de los convoyes marítimos aliados. Good explica en un artículo publicado en *Biometrika* en 1979 las aportaciones metodológicas de Turing a la teoría bayesiana, tanto al denominado peso de la evidencia como a la introducción de un *test* de hipótesis basado en la razón de verosimilitudes con el que confrontar hipótesis nulas y alternativas [341]. Gracias a recientes desclasificaciones de documentos relacionados con el *bamburismus* por parte del gobierno americano, conocemos que dicho algoritmo se fundamenta en el *test* de hipótesis diseñado por Turing [342].

La máquina Enigma constaba de un teclado, un panel donde las letras se iluminaban y varios rotores. Para cifrar un mensaje se comenzaba colocando los

rotores en una determinada posición (lo que se denominaba configuración inicial) y se escribía el mensaje, obteniendo el mensaje cifrado en el panel. Para descifrar un mensaje cifrado, el proceso era simétrico. Simplemente había que colocar los rotores en la configuración inicial y teclear el mensaje cifrado, que iba apareciendo decodificado en el panel. Las configuraciones iniciales se distribuían a los usuarios de las máquinas, mensualmente al principio y con mayor frecuencia según avanzaba la guerra.

El mecanismo de cifrado de la máquina Enigma se basaba en los rotores, los cuales permitían cambiar la letra del alfabeto en la que comenzaba la asignación de la letra A. Así, por ejemplo, si la letra A se transformaba en una F, entonces la letra B se transformaba en una G, y así sucesivamente. Cualquier letra del alfabeto podía tomarse como inicio en cualquiera de las ruedas, lo que da una idea de la explosión combinatoria a que da origen el cómputo de todas las combinaciones posibles de encriptación [343].

A medida que se recibían mensajes codificados, la creencia sobre la configuración hipotética de la máquina iba actualizándose de acuerdo con el esquema de razonamiento bayesiano. Cuando el peso de la evidencia a favor de una configuración determinada de la máquina Enigma era lo suficientemente alto, es decir estadísticamente significativo atendiendo al *test* de la razón de verosimilitudes, dicha configuración se consideraba como probable. Se probaban de forma exhaustiva todas las configuraciones probables con los mensajes recibidos. Como resultado de estas pruebas se consiguió descifrar el código [339].

Desde un punto de vista pragmático es útil emplear en una discusión términos que denoten su significado de forma intuitiva. Según Good, Turing introdujo la expresión “factor de Bayes a favor de una hipótesis” como elemento de partida para el razonamiento que llevó al desarrollo del análisis secuencial de las razones de verosimilitud. El factor de Bayes a favor una hipótesis  $H$ , proporcionado por la evidencia  $E$  es  $Odds(H|E)/Odds(H)$ , o lo que es lo mismo, el factor por el que se debe multiplicar la *odds* inicial de  $H$  para obtener la *odds* final. Constituye un

teorema de fácil demostración el que el factor de Bayes es igual a  $P(E|H)/P(E|\bar{H})$ , donde  $P$  es probabilidad y  $\bar{H}$  es la negación de la hipótesis  $H$ . Posiblemente el mismo Bayes se quedó a mitad camino de la descripción del factor de Bayes, tal como lo denominó Turing. Un año antes, el matemático H. Jeffreys lo describió de forma independiente pero sin ponerle nombre, lo que sí hizo con acierto el propio Turing [341,344].

En realidad la invención de análisis secuencial con el empleo de la transformación logarítmica del factor de Bayes por parte de Turing también ocurrió de forma independiente a una propuesta anterior, en 1878, por el filósofo Charles S. Peirce, que la publicó bajo la denominación de peso de la evidencia (*weight of evidence* o WoE) [345]. Irwing Good identificó la homología y propuso recuperar el término anterior, de modo que actualmente se habla de peso de la evidencia de Good-Turing [346].

El peso de la evidencia a favor de  $H$  proporcionado por  $E$  se expresa  $W(H:E)$ , donde el signo “:” debe diferenciarse de “|”, que se lee como “dado” en vez de como “proporcionado por”. De forma general se define el peso de la evidencia a favor de  $H$  y en contra de  $H'$  (hipótesis alternativa), proporcionado por la evidencia  $E$  del siguiente modo:

$$W\left(\frac{H}{H'}:E\right) = \log \frac{\text{Odds}\left(\frac{H}{H'}|E\right)}{\text{Odds}\left(\frac{H}{H'}\right)} = \log \frac{P(E|H)}{P(E|H')} = W(H:E|H \text{ o } H')$$

Se aprecia que el peso de la evidencia está relacionado con la cantidad de la información al respecto de  $H$  proporcionado por  $E$ ,  $I(H:E)$ , definida como  $\log[P(E|H)]/P(E)$ . De hecho:

$$W\left(\frac{H}{H'}:E\right) = I(H:E) - I(H':E)$$

Esta cantidad de información es un caso especial de peso de la evidencia  $W(H|H':E)$  en el que  $H'$  es reemplazado por una tautología. Si se utiliza el peso de la evidencia para el problema de la resolución de un diagnóstico en el que caben dos posibilidades (ausencia y presencia, mutuamente excluyentes) y la evidencia es el

resultado de una prueba complementaria, se puede intercambiar  $H'$  por  $\bar{H}$ . Es decir, que considerando la hipótesis a contrastar como la presencia de un evento binario tal como el diagnóstico de una enfermedad, la hipótesis alternativa se comporta como equiparable al complementario de la hipótesis  $H$ . Al efectuar la sustitución pertinente, la expresión del factor de Bayes a favor del diagnóstico  $H$  (o peso de la evidencia a favor del diagnóstico  $H$ ) resulta:

$$W(H/\bar{H} : E) = W(H : E) = \log \frac{P(E|H)}{P(E|\bar{H})}$$

Una tabla de contingencia como la tabla 2, substituyendo los elementos de cada casilla por su equivalente en la expresión anterior, queda de la siguiente forma:

**Tabla 54:** Tabla de contingencia donde la evidencia es la prueba a validar y la hipótesis es el criterio de verdad.

Criterio de verdad →	Enfermedad ( $H$ )	Ausencia de enfermedad ( $\bar{H}$ )
Prueba + ( $E$ )	$E \cap H$	$E \cap \bar{H}$
Prueba - ( $\bar{E}$ )	$\bar{E} \cap H$	$\bar{E} \cap \bar{H}$

Se aprecia que  $Se = P(E|H)$  y que  $1 - Sp = P(E|\bar{H})$ . Por consiguiente:

$$W(H : E) = WoE = \log \frac{Se}{1 - Sp} = \log RV +$$

Análogamente, se puede expresar así el peso de la evidencia de ausencia de enfermedad proporcionada por un resultado negativo de la prueba diagnóstica:

$$W(\bar{H} : \bar{E}) = Wo\bar{E} = \log \frac{P(\bar{E}|H)}{P(\bar{E}|\bar{H})} = \log \frac{1 - Se}{Sp} = \log RV -$$

Turing fue también el primero en reconocer el valor de denominar las unidades en cuyos términos se mide el peso de la evidencia. Para una base  $e$  del logaritmo de la fórmula nombró la unidad como un ban natural y para una base 10, simplemente ban. No ha sido hasta años más tarde que una unidad de información de base 2 se ha denominado bit, que de hecho, también puede emplearse como unidad de medida de la información del peso de la evidencia. Por analogía con el decibel, el mismo Turing introdujo también el concepto de deciban en el sentido de

una décima parte de un ban. Un deciban se interpreta como la mínima unidad de cambio en el peso de la evidencia discernible por el ser humano. En el manuscrito de Good en el que se explica el fundamento de esta técnica estadística ya apunta que: “*presiento que esto supone una ayuda importante al razonamiento humano y podría mejorar los juicios de los médicos, abogados y otros miembros de la sociedad*” [341]. El motivo de adoptar esta nomenclatura viene de la época en la que Turing, Good y otros descifradores de códigos se dedicaban al descifrado de Enigma. Este trabajo se desarrolló en las instalaciones militares de Bletchley Park, Inglaterra. Cada nuevo día los alemanes cambiaban el código de encriptación. La tarea de los aliados suponía un enorme problema de inferencia: deducir, en base al texto cifrado que los aliados interceptaban a los alemanes cada día, el orden de los tres rotores de la máquina que se debía de emplear, cuáles eran sus posiciones iniciales (configuración inicial) y cuáles de los mensajes originales en alemán eran los cifrados. La evidencia a favor de cada hipótesis particular se anotó en hojas de papel que fueron especialmente impresas en Banbury, una ciudad a unas 30 millas de Bletchley Park. La tarea de inferencia fue bautizada como *banburismus* y sus unidades se denominaron ban, en referencia a esta ciudad [346].

Tabla 55: Relación entre la razón de verosimilitud y el peso de la evidencia (WoE).

	Razón de verosimilitud	Fórmula	WoE (deciban)
Positiva	10	$10 \times \log_{10} 10$	10
	5	$10 \times \log_{10} 5$	6,9897
	2	$10 \times \log_{10} 2$	3,0103
	1	$10 \times \log_{10} 1$	0
Negativa	$1/2 = 0,5$	$10 \times \log_{10} 0,5$	-3,0103
	$1/5 = 0,2$	$10 \times \log_{10} 0,2$	-6,9897
	$1/10 = 0,1$	$10 \times \log_{10} 0,1$	-10

Tal como se observa en la tabla anterior, el WoE es un parámetro más intuitivo que la razón de verosimilitud y su ancho de distribución de posibles valores va de  $-\infty$  a  $+\infty$  (de -10 a +10 en deciban, de forma simétrica para *RV* habituales). Dada su posibilidad de aplicarse en el estudio de la validez de pruebas diagnósticas,

refleja mejor y de una forma más entendible, respecto a las razones de verosimilitud, el peso que un resultado positivo de un sistema diagnóstico tiene sobre la posibilidad real de ese diagnóstico. Sorprendentemente, los WoE se han empleado escasamente en la literatura médica pese a tratarse de una medición con buenas raíces históricas. Como novedad metodológica, en el bloque correspondiente de esta tesis, se hará uso de los WoE a modo exploratorio como parte de la evaluación de la validez de los sistemas diagnósticos para el caso de la hepatitis autoinmune en población pediátrica.

### **13.1.2.6. Valor predictivo global**

El valor predictivo global (*efficiency* en inglés,  $VP_G$ ) representa la probabilidad que tiene una prueba diagnóstica en acertar, es la proporción de resultados válidos entre la totalidad de las pruebas efectuadas. En realidad, no es un parámetro de validez interna de un *test* porque su resultado depende, además de la prevalencia de la enfermedad como los valores predictivos positivo y negativo, de la relación entre sensibilidad y especificidad. Es posible que aumentos en la prevalencia de la condición a estudio, produzcan descensos en el  $VP_G$  difíciles de predecir. Su importante sensibilidad a cambios en los verdaderos indicadores de validez y su dificultad de interpretación han conducido a que no se utilice habitualmente en los estudios de validez de pruebas diagnósticos a pesar de ser un parámetro clásico de los textos y manuales sobre evaluación de pruebas diagnósticas.

$$VP_G = \frac{VP + VN}{VP + VN + FP + FN}$$

Esta relación entre la prevalencia ( $P$ ), la sensibilidad ( $Se$ ) y la especificidad ( $Sp$ ) está predicha por el teorema de Bayes como se ha expuesto previamente, afecta a los valores predictivos y está fundamentada en probabilidades condicionales. Así, se demuestra que:

$$VP+ = \frac{P \times Se}{P \times Se + (1 - P) \times (1 - Sp)}$$

$$VP^- = \frac{(1 - P) \times Sp}{(1 - P) \times Sp + P \times (1 - Se)}$$

$$VP_G = P \times Se + (1 - P) \times Sp$$

Estas fórmulas permiten calcular fácilmente los valores predictivos cuando la prueba diagnóstica se aplica en comunidades con diferentes prevalencias. Tal como se aprecia en la figura, los cambios en el  $VP^+$  son bruscos cuando la prevalencia de una condición es inferior al 50% (que suele ser lo habitual en la mayoría de los contextos clínicos), sin ser tan importantes sobre el  $VP^-$ . Cuando la frecuencia de la enfermedad es elevada (en ciertos ámbitos especialistas con poca diversidad de enfermedades), habría que dudar si los resultados de una prueba con resultado negativo es correcta o no, debido a la caída del  $VP^-$  [238].

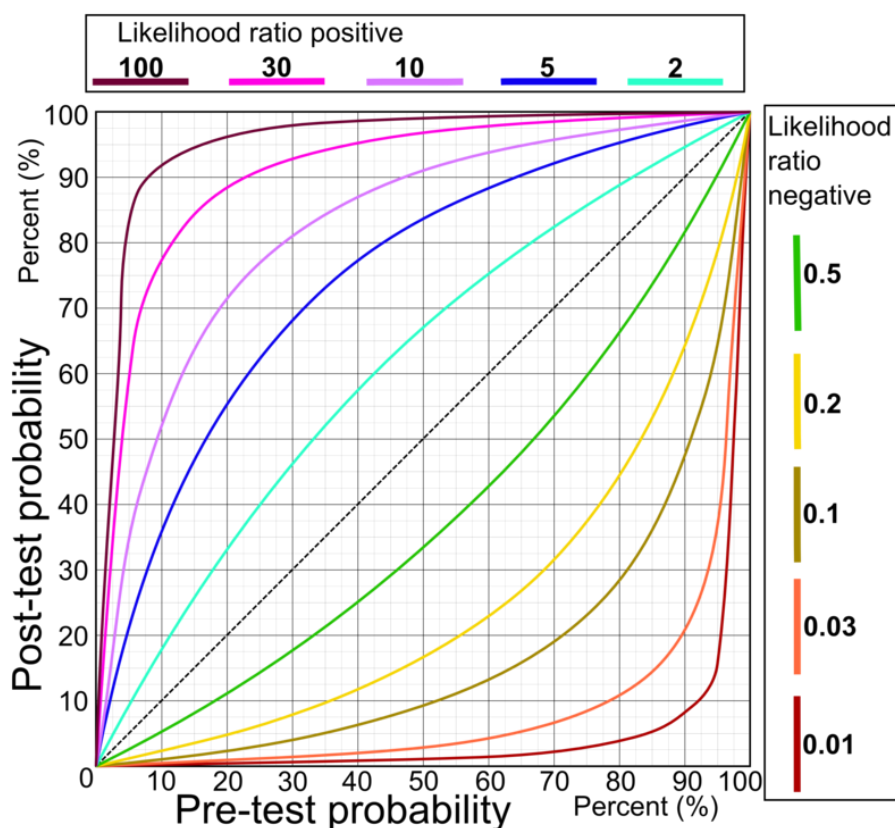


Figura 81: Valores predictivos en función de la razón de verosimilitud y la prevalencia. Creado por Mikael Häggström y no sujeto a derechos de autor (*Public Domain Dedication*, CC0 1.0). Las razones de verosimilitud positiva de 100, 30, 10, 5 y 2 son equivalentes a unos pesos de evidencia positivos de 20, 15, 10, 7 y 3, respectivamente. Las razones de verosimilitud negativa de 0,5, 0,2, 0,1, 0,03 y 0,01, por su parte, son equivalentes a unos pesos de la evidencia negativos de -3, -7, -10, -15 y -20, también respectivamente.



**13.1.2.7. Odds ratio diagnóstica**

La *odds ratio* diagnóstica (ORD) es conocida como un índice estadístico en los estudios epidemiológicos de casos y controles representando la fuerza de asociación entre el factor de riesgo y la enfermedad. Por analogía tiene utilidad para mostrar la misma asociación pero entre un sistema de clasificación o diagnóstico y la enfermedad. Este índice traduce las prestaciones de una prueba con un solo valor que no está influenciado por la prevalencia.

Es la razón de la *odds* de estar enfermo si la prueba da positivo y la *odds* de no estar enfermo si la prueba da negativo.

$$ORD = \frac{\frac{VP}{FN}}{\frac{VN}{FP}} = \frac{RV +}{RV -} = \frac{\frac{VP+}{1-VN+}}{\frac{1-VN-}{VP-}}$$

Los valores de la ORD varían de cero a infinito (cuantos más altos son los valores, mejor es la prueba). El valor  $ORD = 1$  significa que la prueba no es discriminante, es decir, es una prueba inútil. Los valores mayores de 1 significan que es más probable que la prueba dé positivo en el caso de enfermos que en sanos.

**13.1.2.8. Efectividad de una prueba**

La efectividad de la prueba ( $\delta$ ), cuya distribución es aproximadamente normal, se define de esta manera:

$$\delta = \sqrt{\frac{3}{\pi}} \times \left( \ln \frac{Se}{1-Se} + \ln \frac{Sp}{1-Sp} \right) = \sqrt{\frac{3}{\pi}} \times \ln \frac{RV +}{RV -}$$

Puede interpretarse como la diferencia entre las medias de los resultados entre una población de enfermos y otra de sanos en una escala normalizada. Si  $\delta = 1$  la prueba no es efectiva y si  $\delta > 3$  es altamente efectiva.

### 13.1.2.9. El índice de Youden

Este índice clínico fue propuesto por Youden (1950) para analizar la capacidad del método de diagnóstico, usando un único valor en reemplazo de la forma dual de hacerlo (sensibilidad y especificidad). Se define como [347]:

$$J = Se + Sp - 1$$

Varía de  $-1$  a  $+1$ . Si es inferior o igual a  $0$ , la prueba no tiene ningún valor informativo. La prueba es tanto de mejor cuanto el índice de Youden se acerca a  $+1$ . No es recomendable presentarlo en lugar de la sensibilidad y la especificidad porque enmascara la presencia de valores deficientes en estos indicadores. Por ejemplo, una prueba con una  $Se = 0,51$  y una  $Sp = 0,99$  ( $J = 0,51 + 0,99 - 1 = 0,5$ ) obtiene el mismo índice que otra con  $Se = 0,75$  y  $Sp = 0,75$  ( $J = 0,5$ ).

### 13.1.2.10. Ganancia diagnóstica

Se denomina ganancia diagnóstica,  $G(+)$ , a la diferencia entre la probabilidad preprueba de la enfermedad y la probabilidad postprueba obtenida por el resultado del *test*. Si el resultado de la prueba es positivo, la  $G(+)$  se mide según la siguiente expresión, donde  $\hat{p}$  es probabilidad,  $E$  es enfermedad,  $T$  es *test* y  $P$  es prevalencia [348]:

$$G(+) = \hat{p}(E + | T+) - \hat{p}(E +) = VP(+) - P$$

Si el resultado de la prueba es negativo, la reducción de la probabilidad de la enfermedad,  $G(-)$ , se mide en valor absoluto de esta diferencia:

$$G(-) = \hat{p}(E +) - \hat{p}(E + | T-) = P - (1 - VP(-))$$

El conjunto de la ganancia diagnóstica alcanzada por la aplicación de la prueba, o contenido diagnóstico  $\Gamma$  (gamma mayúscula) de la prueba es igual a la suma de estas dos ganancias diagnósticas:

$$\begin{aligned} \Gamma &= G(+) + G(-) = \\ &= [\hat{p}(E + | T+) - \hat{p}(E +)] + [\hat{p}(E +) - \hat{p}(E + | T-)] = \\ &= \hat{p}(E + | T+) - \hat{p}(E + | T-) \end{aligned}$$

Se aprecia que el contenido diagnóstico  $\Gamma$  alcanzado por la aplicación de una prueba se mide por la diferencia entre las dos probabilidades postprueba de la enfermedad, que son obtenidas respectivamente por el resultado positivo y negativo del *test* o sistema diagnóstico.

Por ejemplo, una prueba con respuesta cualitativa binaria cuyas cualidades diagnósticas son  $Se = 0,8$ ,  $Sp = 0,9$ ,  $RV(+)= 8$  y  $RV(-) = 0,222$  se aplica a la investigación de una enfermedad cuya probabilidad preprueba o prevalencia es del 30%, es decir, 0,30.

Si el resultado de la prueba va a favor de la presencia de la enfermedad, la probabilidad postprueba de la enfermedad se eleva a:

$$\hat{p}(E + | T+) = VP(+) = \frac{P \times RV(+)}{P \times [RV(+)-1] + 1} = \frac{0,30 \times 8}{(0,30 \times 7) + 1} = 0,77$$

La ganancia diagnóstica aportada por la respuesta positiva de la prueba es:

$$G(+)= VP(+)- P = 0,77 - 0,30 = 0,47$$

Si el resultado de la prueba va en contra de la presencia de la enfermedad, la probabilidad de la enfermedad baja a:

$$\hat{p}(E + | T-) = 1 - VP- = \frac{P \times RV(-)}{P \times [RV(-)-1] + 1} = \frac{0,30 \times 0,222}{[0,30 \times (0,222 - 1)] + 1} = 0,09$$

La ganancia diagnóstica obtenida por la respuesta negativa de la prueba, en valor absoluto es:

$$G(-)= P - (1 - VP-) = 0,3 - 0,09 = 0,21$$

Las ganancias diagnósticas se pueden visualizar en la figura. En función de la probabilidad preprueba en el eje de las abscisas, las dos curvas dan los valores de las probabilidades postprueba de la enfermedad según que el resultado de la prueba sea positivo (curva superior) o negativo (curva inferior). La ganancia diagnóstica  $G(+)$  aportada por el resultado positivo es igual a la distancia OA; la ganancia  $G(-)$  aportada por un resultado negativo es igual en valor absoluto a la distancia OB; la ganancia total o contenido diagnóstico  $\Gamma$  de la prueba cuando la probabilidad preprueba es 0,30, es igual a la distancia AB entre las dos curvas:

$$\Gamma = 0,47 + 0,21 = 0,68$$

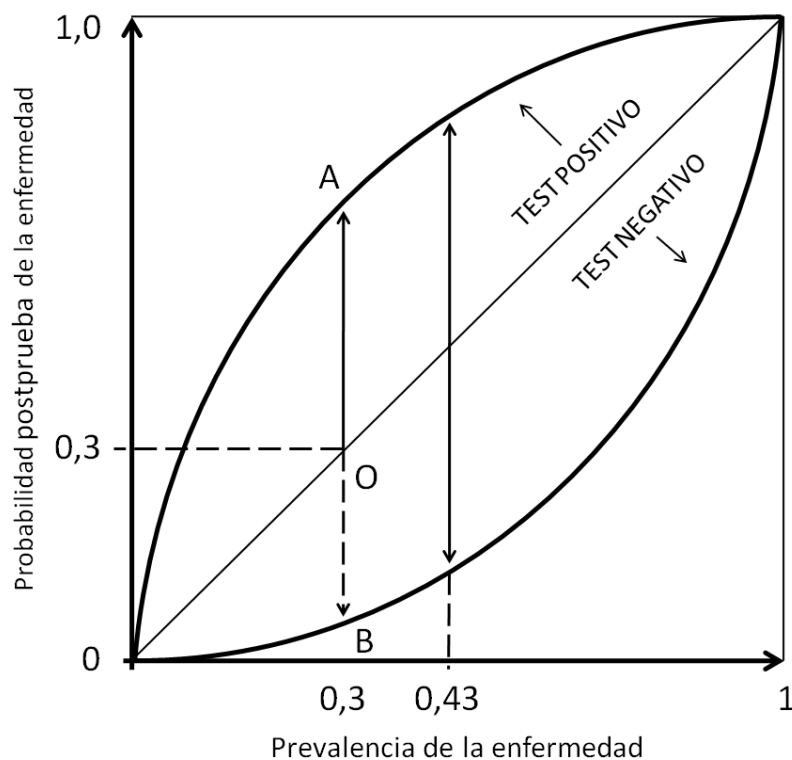


Figura 82: Ganancia diagnóstica obtenida por la respuesta de una prueba en función de la probabilidad preprueba o prevalencia. Se observa que la ganancia diagnóstica obtenida por la prueba tiende a anularse cuando la probabilidad primaria de la enfermedad se aproxima a los valores extremos 0 y 1.

El cálculo muestra que la ganancia diagnóstica aportada por la respuesta positiva del *test* es máxima cuando la probabilidad preprueba es igual a:

$$P = \frac{1}{1 + \sqrt{RV +}}$$

En el ejemplo,  $P = 1/(1 + \sqrt{8}) = 0,26$ .

En caso de respuesta negativa, la ganancia diagnóstica es máxima cuando la prevalencia es:

$$P = \frac{1}{1 + \sqrt{RV -}}$$

En el ejemplo,  $P = 1/(1 + \sqrt{0,222}) = 0,68$ .

Por otro lado, el contenido diagnóstico de la prueba es máximo cuando la prevalencia es:

$$P = \frac{1}{1 + \sqrt{RV(+)} \times RV(-)}$$

En el ejemplo,  $P = 1/(1 + \sqrt{8 \times 0,222}) = 0,43$ .

Esta probabilidad es aquella en que los dos valores predictivos son iguales. En el ejemplo, cuando  $P$  es igual a  $P_{max} = 0,43$ , los dos valores predictivos son  $VP(+)=VP(-)=0,86$ . Es cuando la probabilidad preprueba está próxima a valores medianos (próxima al 50%) cuando la respuesta de la prueba aporta la información diagnóstica más elevada. Este valor  $P_{max}$  de la probabilidad preprueba es 0,50 cuando el producto  $RV(+)\times RV(-)=1$ , es decir cuando las dos razones de verosimilitud son inversas la una de la otra.

Estos cálculos permiten obtener, además de un indicador de la utilidad global de un sistema diagnóstico, la prevalencia de la enfermedad bajo la cual dicho método diagnóstico es óptimo. De este modo el investigador puede deducir algunos elementos de las condiciones bajo las cuales una prueba maximiza su eficacia, aunque el estudio en el que se llevó a cabo la evaluación de la validez de esa prueba tuviera lugar en otras circunstancias diferentes.

### **13.1.2.11. Curva de características operativas del receptor**

Hasta el momento se han expuesto las características de pruebas o criterios diagnósticos cuando son aplicadas a dos grupos de la población: el grupo con la enfermedad y el grupo sin ella. Los resultados de tales pruebas son citados como positivos o negativos según señale o no hacia la presencia de la enfermedad en cuestión. Pero la realidad suele ser más compleja. En algunas instancias, pueden ser necesarias más de dos categorías para enmarcar la condición de cada paciente, el resultado de una prueba, o de ambos.

Uno de estos casos es cuando los resultados de una prueba son de naturaleza cuantitativa u ordinal. Es decir, cuando el resultado de una prueba diagnóstica es un número, un rango o un nivel (que puede adoptar grados distintos de definición, como, por ejemplo, de más a menos, 250 UI/mL de concentración plasmática de

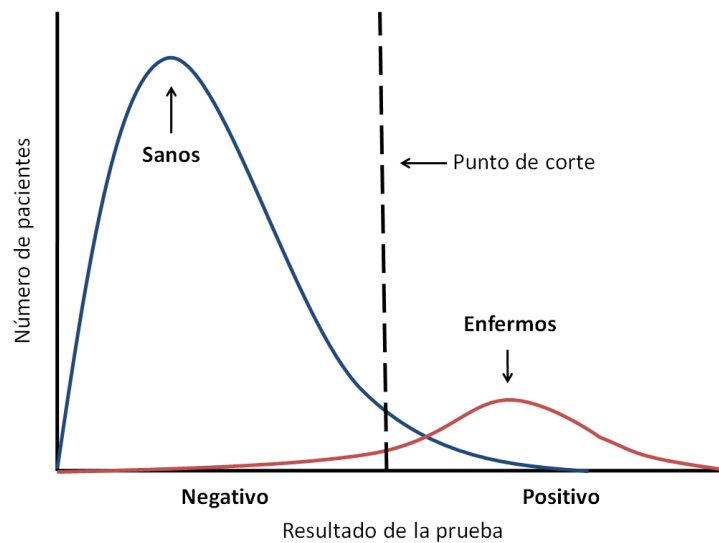
alanina aminotransferasa, título de 1:120 de anticuerpos antinucleares o grado de fibrosis moderado). En la mayoría de trabajos sobre pruebas diagnósticas se establece un solo punto de corte dentro del espectro cuantitativo (o semicuantitativo, como en el caso del número de puntos de un sistema de clasificación para una enfermedad) que separa a los enfermos de los no enfermos. Esto es equivalente a señalar un punto en el rango de resultados posibles que divide a los pacientes en “probablemente enfermos” y “probablemente no enfermos”.

La mayoría de estudios sobre validez de pruebas diagnósticas convierten las variables cuantitativas en variables cualitativas dicotómicas, en el sentido de resultado positivo frente a resultado negativo o enfermo frente a sano.

La curva de características operativas del receptor (COR) facilita la elección de los mejores puntos de corte que proporcionen una mayor sensibilidad y especificidad a la prueba (mayor área bajo la curva). Por lo general, estos puntos se corresponden con los puntos de inflexión de la curva.

Clásicamente, para conocer la eficacia de las pruebas con resultados cuantitativos, se elige un solo valor de entre los resultados posibles de una prueba, que permita declarar a los pacientes “con resultado positivo” o “con resultado negativo”, y estimar entonces los indicadores de eficacia, transformándose así los resultados en variables dicotómicas [349].

Así, las estimaciones de sensibilidad y especificidad de la prueba dependen de un solo punto de corte seleccionado que el médico debe escoger según sus necesidades. Sin embargo, la noción de “punto de corte óptimo” no es única ya que, por un lado, son casi inexistentes las pruebas diagnósticas con sensibilidad y especificidad ambas muy altas (próximas a 1) y, por otro lado, la práctica clínica es versátil en sus necesidades de sensibilidad y especificidad elevadas. Así pues, fijar un punto de corte, un valor determinado de la prueba que marque el límite entre sano y enfermo, no suele ser una tarea sencilla.



**Figura 83:** Habitualmente existe un solapamiento en los resultados de las pruebas diagnósticas cuantitativas entre enfermos y sanos que hace imposible encontrar un punto de corte perfecto.

Existe una zona de posibles resultados de la prueba para la que las distribuciones de sujetos sanos y enfermos se solapan. El desplazamiento del punto de corte en el sentido de incluir a todos los enfermos en la zona de valores positivos conllevará asimismo la inclusión de una mayor parte de los sanos, con lo cual aumentará la proporción de falsos positivos. Lo mismo ocurre en sentido contrario, de modo que normalmente una estrategia de aumento de la sensibilidad acarrea un empobrecimiento de la especificidad y viceversa.

Una herramienta útil para evaluar la capacidad diagnóstica de una prueba cuantitativa para todos los posibles puntos de corte es la denominada curva COR, cuya morfología ayudará a definir aquellos puntos de corte que optimicen tanto la sensibilidad como la especificidad. Esta curva también servirá para comparar diferentes pruebas diagnósticas.

Latour explica el origen de las curvas COR como la solución al problema de la comparación de la eficacia de los radares de los submarinos. La idea se desarrolló en la década de los 50 en plena Guerra Fría, cuando se necesitaba evaluar la capacidad de un radar de distinguir entre las verdaderas señales electromagnéticas reflejadas

por un objeto de interés y el “ruido” generado por otras fuentes. Igual que una prueba diagnóstica, el radar puede equivocarse de dos formas: fallando en la detección de la señal (falso negativo) o interpretando como verdadera una señal de fondo (falso positivo). La posibilidad de ajustar el umbral de detección de señales del radar origina cambios en las distintas tasas de errores relacionados entre sí: a medida que el umbral disminuye, la proporción de falsos negativos desciende (aumenta la sensibilidad) y aumenta la de falsos positivos (disminuyendo la especificidad) [350].

En esencia la curva COR es una representación gráfica de la sensibilidad y el complementario de la especificidad calculados para todos los puntos de corte posibles de los resultados de un sistema diagnóstico cuantitativo o categórico ordinal [302].

La curva COR es no paramétrica cuando el trazo que une cada uno de los puntos está configurado por líneas horizontales y verticales formando un ángulo recto. Así, la curva pasa por todos los puntos de corte y los ángulos corresponden a los resultados de la aplicación real de la prueba en el ejercicio de validación. Se trata de la representación más estricta de la realidad porque evita inferir sensibilidad y especificidad de resultados no obtenidos en el experimento. Trazar una curva COR permite además elegir los puntos de corte óptimos por métodos geométricos. Por lo general, estos puntos de corte que maximizan los indicadores de validez se corresponden con los puntos de inflexión de la curva [246]. Esto permite agrupar los datos en categorías, pudiéndose utilizar también modelos paramétricos para el ajuste de la curva COR. Las curvas COR paramétricas tienen la ventaja de proporcionar un línea de visualización más clara, sin el gráfico abigarrado de las curvas no paramétricas. El inconveniente lógico es que, al agrupar datos, la curva no pasa necesariamente por los puntos que reflejan información verdadera y el cálculo del área bajo la curva puede diferir con la obtenida para la trazado no paramétrico [351].



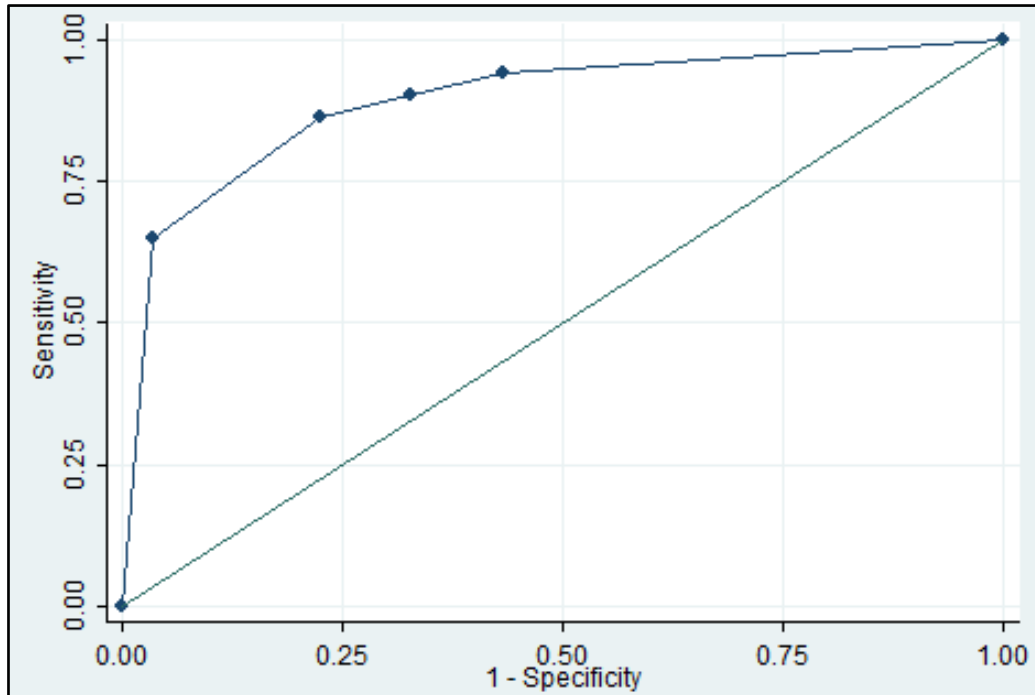


Figura 84: Curva COR paramétrica para datos discretos.

En la situación ideal, una prueba que discrimina perfectamente entre enfermos y sanos quedaría representada en la gráfica como una línea que coincidiría con los lados izquierdo y superior del cuadrado; mientras que una prueba que no discrimine en absoluto correspondería a la línea diagonal que aparece en la figura. Cuanto más desplazada esté la curva COR hacia el vértice superior izquierdo, mejor es la capacidad discriminatoria de la prueba. Una forma de evaluar de manera cuantitativa y global esa capacidad de discriminación consiste en calcular el área del polígono que queda debajo de la curva COR, parámetro que permite efectuar comparaciones de la calidad de los *tests*. A mayor área bajo la curva, mejor capacidad diagnóstica.

El valor del área bajo la curva oscila entre 0 y +1. El intervalo de confianza del área bajo la curva se puede calcular por varios métodos entre los que destacan el propuesto por Delong y el de Hanley y McNeil [352]. Para el presente trabajo se utilizará el método de Delong, basado en asunción de normalidad, que solo ha demostrado presentar un rendimiento subóptimo en modelos anidados [353].

La curva COR de una prueba perfecta (sensibilidad = 1 y especificidad = 1) será aquella en que la curva coincide con el eje de ordenadas y la paralela al eje de abscisas que pasa por el punto  $S = 1$ , que daría un área bajo la curva igual a 1.

Si el área es igual a 0,5 la curva coincidirá con la diagonal que divide el recuadro en un triángulo superior-izquierdo y otro inferior-derecho. Es el caso de emplear un *test* equiparable al de asignar la categoría de enfermo o sano en función del resultado de lanzar una moneda al aire.

En resumen, el valor del área que queda debajo de la curva COR equivale a la probabilidad de clasificar correctamente en función del resultado de la prueba diagnóstica [354]. El área bajo dicha curva se convierte así en un buen indicador de la capacidad de un elemento diagnóstico de clasificar en enfermos y sanos, independiente de la prevalencia de la enfermedad en la población de referencia y en base a la cual se podrán establecer comparaciones entre diferentes pruebas diagnósticas [355].

#### **13.1.2.12. Curva de Lorenz**

La curva de Lorenz es un instrumento gráfico desarrollado dentro del área de la economía por Mark O. Lorenz en 1905 para representar las desigualdades en los ingresos de los hogares en una región empleando proporciones acumuladas.

Gráficamente representa la distribución relativa de una variable en un dominio determinado. El dominio para el que se empleó originalmente fue el conjunto de hogares de un país; y la variable de estudio, el ingreso de dichas unidades familiares. El trazado de la curva considera en el eje de las abscisas el porcentaje acumulado de hogares del dominio en cuestión y en el eje de las ordenadas el porcentaje acumulado del ingreso.

De esta manera, cada punto de la curva se lee como porcentaje acumulado dentro del dominio. La curva parte del origen (0,0) y termina en el punto (1,1). Cuando se utiliza en el marco de la evaluación de pruebas diagnósticas con resultado cuantitativo o categórico ordinal, la curva de Lorenz es un gráfico de coordenadas

cuyo eje de abscisas representa las diversas  $X_i$  (proporción acumulada de no enfermos para el punto de corte  $i$ ), mientras que el eje de ordenadas representan las diversas  $Y_i$  (proporción acumulada de enfermos para ese punto de corte  $i$ ) [356]. Del mismo modo que en la curva COR, se representa la diagonal, el origen y el punto (1,1). La curva, igual que la COR, también da una idea de la desigualdad entre sensibilidad y el complementario de la especificidad.

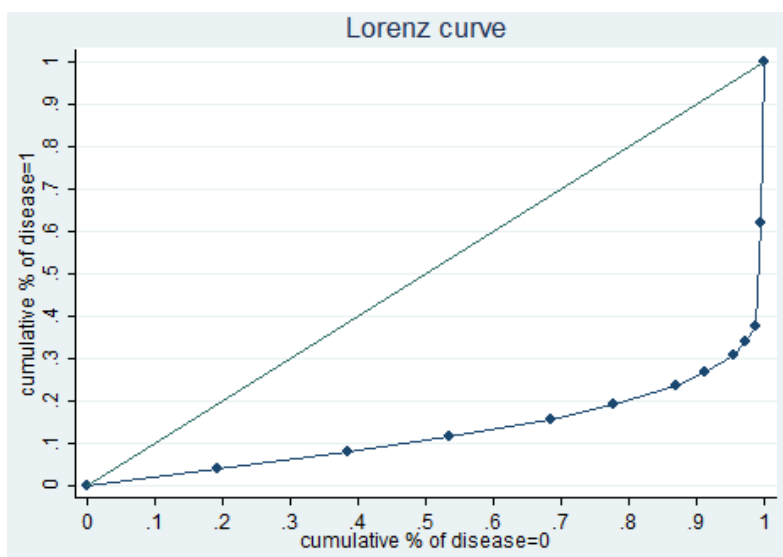


Figura 85: Curva de Lorenz. *Disease=0*, en el eje horizontal, representa el porcentaje acumulado de no enfermos. *Disease=1*, en el eje vertical, el de los sanos. Gráfico generado con Stata 14®.

Si en cada punto de corte la proporción acumulada de enfermos fuera igual a la de no enfermos, la sensibilidad y 1–especificidad serían iguales en todos los puntos al igualarse sus denominadores. En ese caso, la curva sería la diagonal del cuadrado y la razón de verosimilitud sería igual a 1 en todos los puntos, lo que denotaría que la prueba diagnóstica no tiene valor alguno como medio diagnóstico. Lo contrario sería que la curva tuviera una concavidad máxima, entonces la prueba sería perfecta.

En resumen, cuando se aplica en el área de la evaluación de un sistema diagnóstico, a mayor concavidad de la curva de Lorenz, mejor capacidad resolutive del sistema.

### 13.1.2.12.1. Índices de Gini y Pietra

Los índices de Gini y de Pietra representan un intento de convertir la concavidad de la curva de Lorenz en un parámetro útil para efectuar comparaciones entre diferentes pruebas diagnósticas.

El índice de Gini es la proporción que representa el área comprendida entre la curva de Lorenz y la curva de igualdad perfecta (la diagonal del cuadrado) con respecto al área total debajo de la diagonal. En la figura vendría dado por la expresión:

$$\text{Índice de Gini} = A/(A + B)$$

Donde  $A$  y  $B$  son las áreas de las superficies homónimas de la figura. Igual que la curva de Lorenz, normalmente se utiliza para medir la desigualdad en los ingresos dentro de un territorio, pero puede utilizarse para medir cualquier forma de distribución desigual.

El índice de Pietra es la distancia vertical máxima entre la diagonal y la curva de Lorenz. Sería el módulo del vector  $P$  de la figura.

Ambos índices admiten valores entre 0 y +1, donde 1 representa la bondad máxima de la prueba diagnóstica [356].

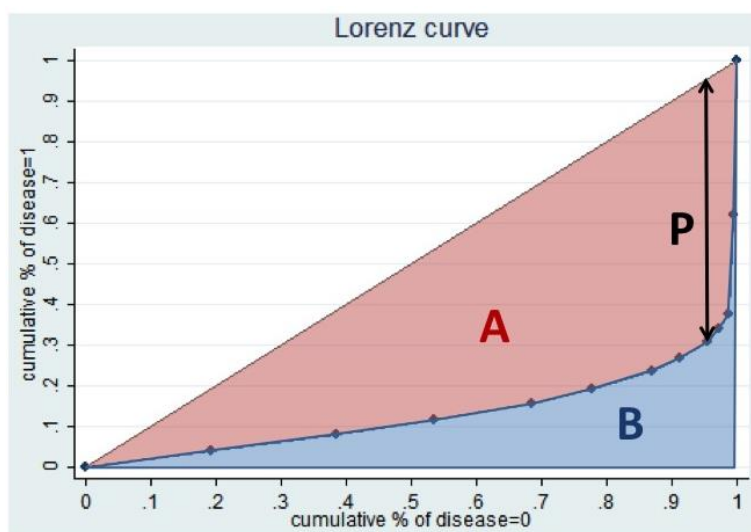


Figura 86: Elementos de los índices de Gini y Pietra sobre una curva de Lorenz. Imagen modificada de la obtenida con Stata 14®.

### **13.1.3. Cálculo del tamaño muestral para estudios de evaluación de pruebas o sistemas diagnósticos**

No existe un consenso sobre la mejor metodología para el cálculo del tamaño muestral en estudios de evaluación de pruebas diagnósticas. Muestra de ello es la diversidad de fórmulas y recomendaciones para efectuar la estimación. A diferencia de para el contraste de hipótesis con comparación de proporciones, medias o varianzas con métodos estadísticos paramétricos, para los fines perseguidos en los estudios de sistemas diagnósticos no se requiere una precisión extrema en el cálculo del tamaño muestral [222,223].

A continuación, se enunciarán los métodos más empleados en la literatura médica.

En primer lugar se ha postulado que en la muestra no debe haber un número inferior a 30 pacientes en el grupo de “no casos” y otros 30 que presenten positivamente la condición de interés en base al patrón oro [225].

También se ha sugerido, como tamaño de muestra adecuado, una cifra global de un mínimo de 200 pacientes, aunque otros autores señalan que 50 es el tamaño muestral mínimo aconsejable para evitar estimaciones demasiado imprecisas. Para el estudio piloto con el que se empezará el proyecto global de esta tesis se tendrá en cuenta este último dintel, con el que se pretende recoger información suficiente para poder efectuar una estimación del tamaño muestral más adecuada.

Una recomendación adicional se basa en el número mínimo de sujetos en las casillas de la tabla tetracórica o de contingencia 2x2. De acuerdo con ésta aproximación, un tamaño de muestra suficiente sería aquel que permitiese un mínimo de 10 pacientes en cada casilla marginal de la tabla [225,226].

Desde la perspectiva del contraste de hipótesis también se puede aproximar el mínimo tamaño de muestra necesario. Se puede estudiar la hipótesis de presencia de diferentes indicadores de validez entre el *test* bajo estudio y la prueba de referencia y también se puede evaluar una hipótesis de concordancia entre las dos

pruebas. Para ello es recomendable disponer de un *gold standard* robusto, elemento del que no es posible disponer en algunas ocasiones [225].

Existen también fórmulas matemáticas para el cálculo del tamaño muestral. La lógica de todos estos métodos se basa en la asunción de un margen de error (o precisión absoluta), que no siempre es fácil de determinar a priori y que, en cualquier caso, se trata de una decisión esencialmente arbitraria del investigador. Normalmente se establece en 10%, 5% o 1%. Un menor margen de error significará un mayor tamaño de la muestra [357]. Como se expondrá en el bloque de la metodología, para este trabajo se fijará en 5%. El intervalo de confianza se debe de situar entre el 90 y 99% para obtener estimaciones adecuadas, siendo 95% el nivel más ampliamente empleado. Se mencionarán algunas fórmulas basadas en los siguientes parámetros:

- Estimación del tamaño de la muestra en base al intervalo de confianza deseado de las razones de verosimilitud [358].

- Cálculo basado en una razón de verosimilitud positiva deseada. Se basa en fijar un error alfa y un error beta y en una estimación previa de la especificidad, la razón de verosimilitud que se desea detectar, el intervalo de confianza deseado para la sensibilidad y la prevalencia esperada de la enfermedad en estudio [359].

- Cálculo basado en las fórmulas de Buderer [224]:

Tamaño muestral basado en la sensibilidad ( $Se$ )=

$$= \frac{Z_{1-\alpha/2}^2 \times Se \times (1 - Se)}{L^2 \times Prevalencia}$$

Tamaño muestral basado en la especificidad ( $Sp$ )=

$$= \frac{Z_{1-\alpha/2}^2 \times Sp \times (1 - Sp)}{L^2 \times (1 - Prevalencia)}$$

Donde  $1 - \alpha/2$  es la desviación estándar normal correspondiente al tamaño especificado de la región crítica ( $\alpha$ ) y  $L$  es la precisión absoluta deseada, de modo que  $100 \times L$  representa, en %, la mitad del intervalo de confianza de indicador.

Para obtener una prevalencia esperada y una previsión de sensibilidad y especificidad, resulta interesante realizar un estudio piloto. Se tendrá en cuenta esta recomendación para el desarrollo de la tesis. Cuanto más se aleje la prevalencia de la población objetivo del 50%, en cualquiera de los dos sentidos, mayor tamaño muestral se necesitará [360].

Recientemente se han desarrollado tanto nomogramas que facilitan el cálculo como programas informáticos que incorporan las fórmulas necesarias para la estimación.

El programa Epidat incluyen funciones de análisis epidemiológico y estadístico y ha sido desarrollado por la *Dirección Xeral de Innovación e Xestión da Saúde Pública de la Consellería de Sanidade (Xunta de Galicia)* con el apoyo de la Organización Panamericana de la Salud y la Universidad CES de Colombia. Realiza el cálculo del tamaño muestral en función de la prevalencia estimada de la enfermedad en la población de referencia, el valor esperado de sensibilidad o especificidad, el nivel de confianza aceptable para el par sensibilidad/especificidad y el margen de error muestral aceptado del estudio [228,361].

Finalmente, se han propuesto varias traducciones de estas fórmulas a nomogramas para facilitar la estimación del tamaño muestral sin necesidad de cálculos.

Por un lado, tenemos la propuesta de Malhotra e Indrayan, que incluye cuatro parámetros: sensibilidad o especificidad esperadas, nivel de precisión absoluta, prevalencia estimada y número de pacientes necesarios. Está basado en la fórmula de Buderer. Teniendo tres de estos indicadores o variables, se puede obtener el cuarto. Como limitación presenta el hecho de que no incorpora el error de tipo II asumible y por lo tanto no sería útil para el objetivo de un contraste de hipótesis de igualdad de sensibilidad o especificidad. Otro inconveniente es la precisión irregular de la lectura de la línea del número de sujetos, aunque las desviaciones por este motivo pueden no ser relevantes en la práctica [229].

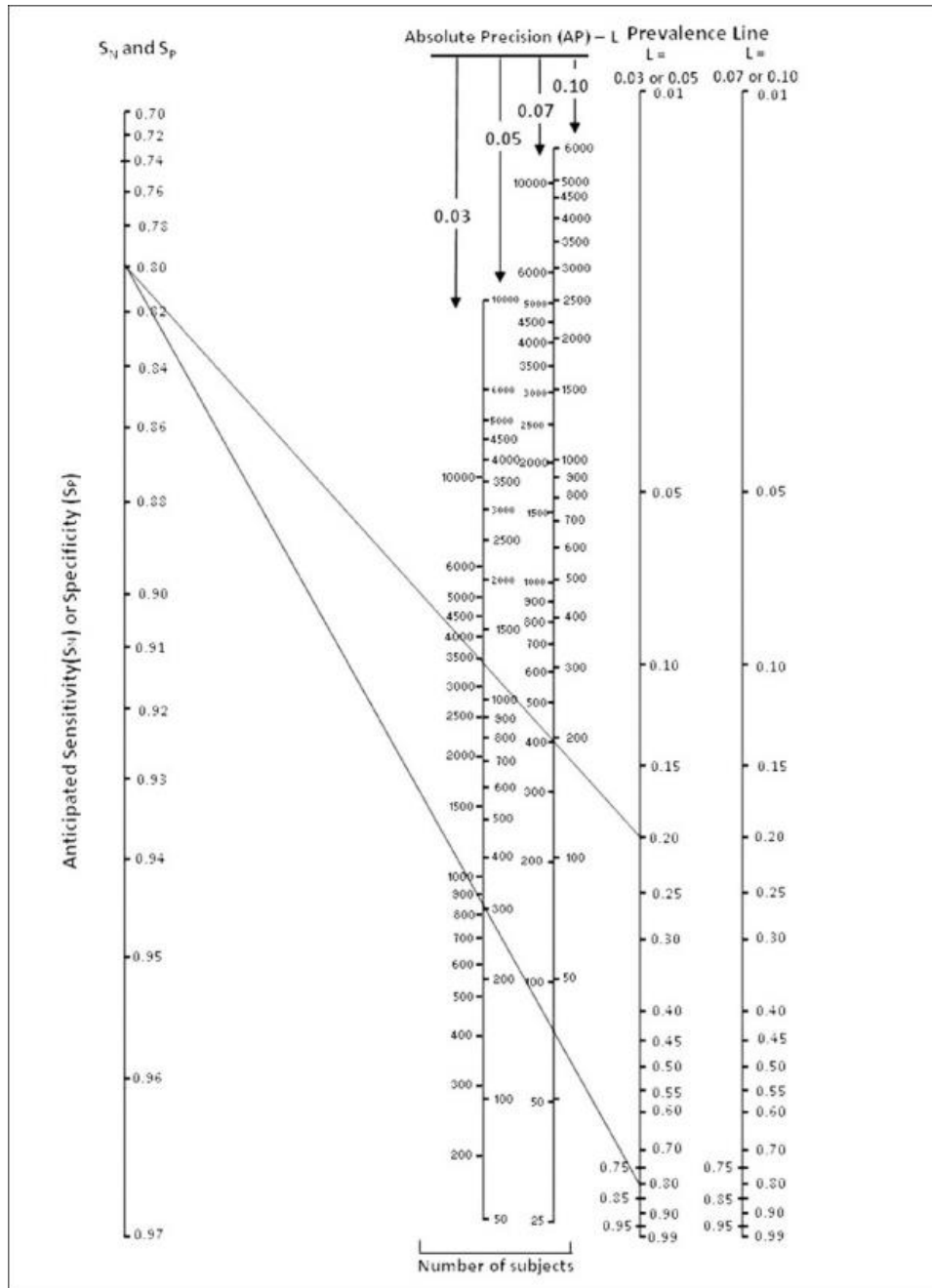


Figura 87: Nomograma de Malhotra e Indrayan (basado en la ecuación de Buderer) para el tamaño muestral de un estudio de validación de pruebas diagnósticas a partir de la prevalencia y la sensibilidad/especificidad ( $S_e/S_p$ ) esperadas, y el margen de error deseado. Dos trazos de ejemplo en la figura partiendo de  $S_e$  o  $S_p = 0,8$ . Reproducido de Malhotra e Indrayan. Indian J Ophthalmol. 2010;58:520. Con licencia *Creative Commons*.

Por último, existe otro nomograma que integra la especificidad o sensibilidad estimadas, la prevalencia esperada de la enfermedad y el intervalo de confianza



deseado para una precisión del 5%. Se trata de los nomogramas de Carley, ideados para variables respuesta de carácter binario y, por tanto (igual que la propuesta de Malhotra e Indrayan), también adecuados para el problema de la validación de criterios de clasificación enfermo/no enfermo. Existe una versión para el parámetro sensibilidad y otro para la especificidad, de modo que la elección entre uno u otro se pueda basar en el indicador de validez cuyo valor esperado sea más fiable o más interesante en función de la utilidad de la prueba. También es razonable elegir el número de pacientes indicado por la versión que arroje un valor superior. Funciona trazando una línea horizontal desde la prevalencia esperada hasta el intervalo de confianza requerido. En la intersección se dibuja una recta perpendicular hasta el cruce con la sensibilidad o la especificidad esperadas. Desde este último punto se proyecta una línea horizontal hasta la derecha del gráfico que señala el tamaño muestral total requerido (incluyendo casos y no casos) [230].

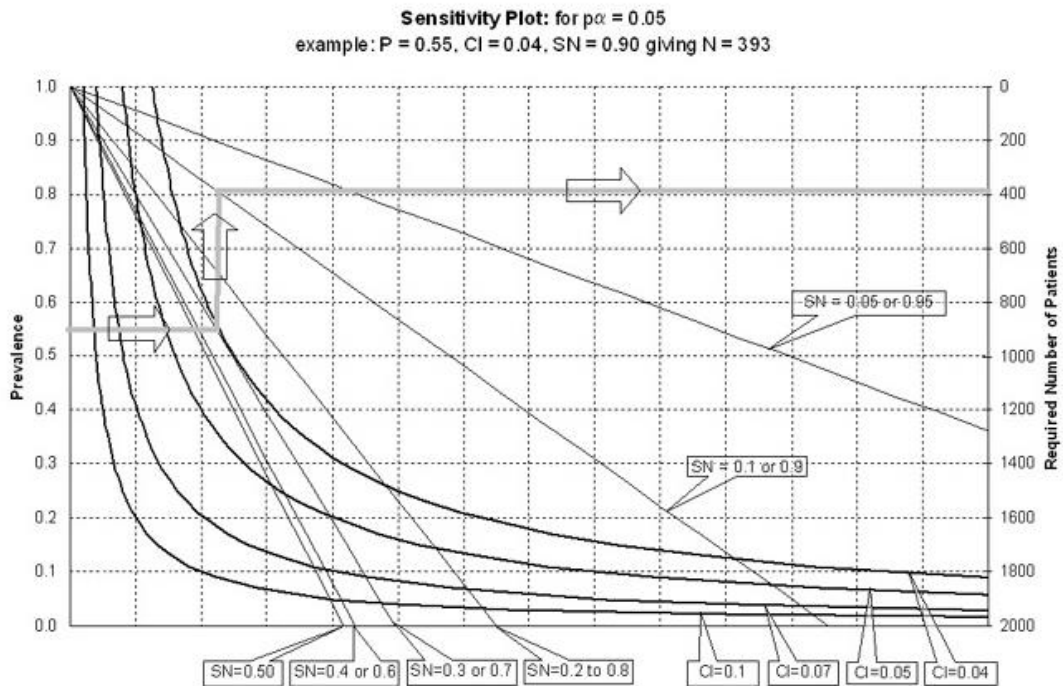


Figura 88: Nomograma de Carley para la sensibilidad con  $\alpha = 0,05$ . Ejemplo: Prevalencia = 0,55. Intervalo de confianza = 0,04. Sensibilidad = 0,90. Resultado de tamaño muestral de 393.

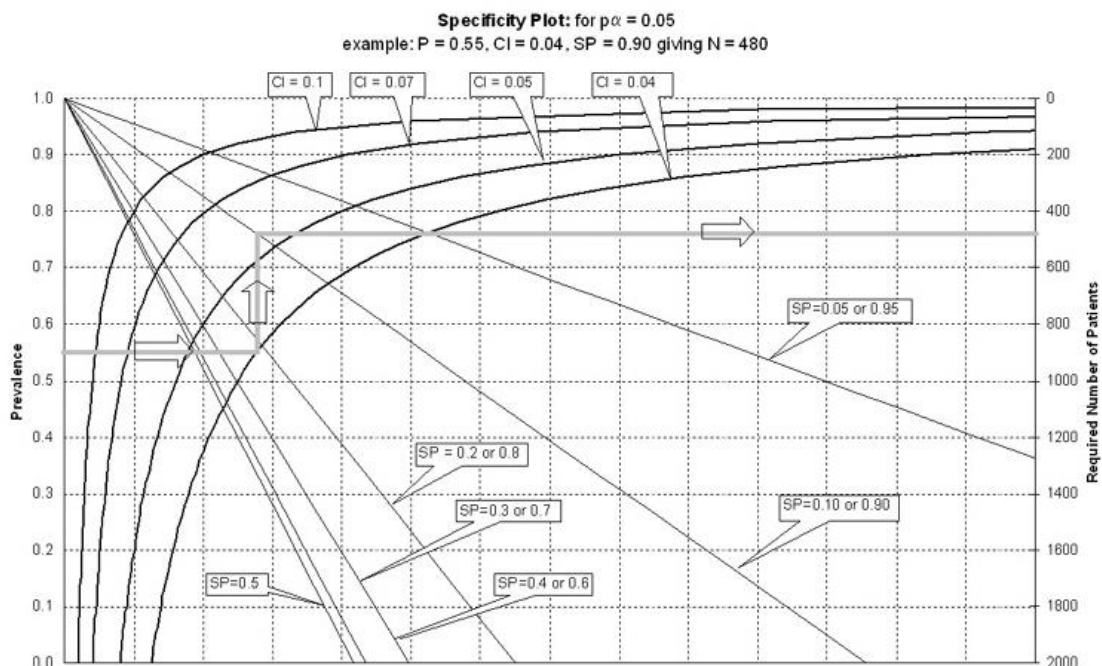


Figura 89: Nomograma de Carley para la especificidad con  $\alpha = 0.05$ . Ejemplo: Prevalencia = 0,55. Intervalo de confianza = 0,04. Especificidad = 0,90. Resultado de tamaño muestral de 480.

### 13.1.4. Diseño de estudios para valorar la validez de sistemas diagnósticos

#### 13.1.4.1. Consideraciones iniciales para la gestión de la arquitectura de un estudio de validación de unos criterios de clasificación

Para el diseño de un estudio para valorar la validez de un sistema, unos criterios o una prueba diagnóstica conviene tener en cuenta las cuestiones que formalizaron Irwing *et al.* en el 2002 [238,362]:

1) ¿Cuál es la enfermedad problema y el estándar de referencia? Es decir, definir como adquiere el diagnóstico de certeza. En ausencia de este, tan solo se puede analizar la reproducibilidad.

2) ¿El objetivo es valorar la ejecución de la prueba mediante una medida global o una medida que permita estimar la probabilidad de enfermedad en los individuos? Esto define los parámetros a estimar. Las medidas globales de calidad

diagnóstica de una prueba son la curva COR y la *odds ratio*. La sensibilidad, la especificidad, valores predictivos, razones de verosimilitud y los pesos de la evidencia permiten estudiar probabilidades de acierto de la prueba.

3) ¿Cuál es la población y el problema clínico? Esta pregunta permite conocer el ámbito de utilización futuro de la prueba.

4) ¿La prueba en estudio será sustituta de otras o se añadirá a las existentes? Si es una prueba sustituta, la hipótesis de trabajo es su superioridad frente a las existentes.

5) ¿En qué medida se quieren estudiar las razones de variabilidad de los resultados dentro de la población? Esto plantea un análisis pormenorizado de todos los factores (de la prueba o los criterios diagnósticos, del paciente...) que pueden influir en el resultado.

6) ¿En qué medida se quieren estudiar la transferencia de los resultados a otros ámbitos? La aplicabilidad de unos criterios diagnósticos en un entorno diferente requiere estas condiciones:

- Definición de la enfermedad constate.
- Empleo de la misma prueba diagnóstica.
- Constancia de los umbrales entre categorías de los resultados de la prueba o los criterios (entre negativo y positivo, por ejemplo).

- Que la distribución de los resultados de la prueba en el grupo de enfermos sea constante en la media y en la forma de distribución: el espectro de la enfermedad no cambia.

- Que la distribución de los resultados de la prueba en el grupo de no enfermos sea constante en la media y en la forma de la distribución: los procesos que dan falsos positivos son los mismos.

- Que la prevalencia sea constante.

El que se cumplan estas condiciones es lo que asegura la validez externa de una prueba o criterios diagnósticos para un contexto diferente de aquél en el que se

obtuvieron sus indicadores de validez. Será analizado para el caso concreto y discutido en el cuerpo de la tesis.

#### **13.1.4.2. Fases en el estudio de un sistema diagnóstico**

Se diferencian cuatro partes en la cronología de un estudio completo sobre la bondad de una prueba diagnóstico o unos criterios diagnósticos [300]:

- **Fase I:** Se comprueba que la prueba da valores diferentes en los enfermos y en los que no lo están, comparando enfermos verdaderos con personas teóricamente sanas. El grupo de sanos no tiene por qué estar constituido por pacientes en los que se realiza un descarte efectivo de la enfermedad a través una prueba *gold standard*.

- **Fase II:** Se comprueba si los sujetos con ciertos valores tienen más riesgo que los otros de estar enfermos. Se puede usar la misma serie de datos que para la fase I, pero con diferente aproximación. Se trata de comprobar si la prueba discrimina bajo circunstancias ideales y a partir de qué resultados se puede considerar un paciente enfermo, o qué riesgo de enfermedad existe para resultados de diferente magnitud. Se admite que se excluyan pacientes por pérdidas de resultados o diagnósticos no bien determinados.

- **Fase III:** Se realiza el análisis de la fase II en condiciones reales, comprobando si la prueba discrimina los pacientes con la enfermedad entre aquellos en los que se sospecha (no frente a sujetos claramente sanos). En esta fase se calcula la sensibilidad, la especificidad y los valores predictivos. La diferencia esencial con la fase II es que la prueba se estudia en una serie consecutiva de pacientes en los que se sospecha la enfermedad. En ocasiones se utiliza el término “intención de diagnosticar” para referirse a esta situación: la fase III se aplica en pacientes en los que realmente sería necesaria la prueba en la clínica real, es decir, los pacientes en los que no está claro si tienen o no la enfermedad de interés. Idealmente no se deberían de excluir pacientes por ningún motivo para efectuar el análisis, a diferencia de la fase II. A menudo el efecto (la presencia de enfermedad si

se considera un efecto binario) no se define de según las mismas referencias para los pacientes con y sin la enfermedad, incluso se puede considerar el buen pronóstico sin tratamiento como prueba de ausencia de enfermedad o la curación con el tratamiento como confirmación de la presencia de la enfermedad. Esta última diferencia con los estudios de fase II es particularmente importante para la metodología que se utilizará en la tesis, porque constituye una característica que en parte propone una solución para el problema de la ausencia de patrón de referencia para la confirmación de la enfermedad.

· **Fase IV:** Por último, se puede valorar si los que se diagnostican tienen un mejor pronóstico. Esta fase no siempre se realiza.

#### **13.1.4.3. Descripción de los diseños genéricos para estudios sobre pruebas o sistemas diagnósticos**

Para los estudios de fase IV se pueden usar diseños de cohortes o de casos y controles y también diseños experimentales [240]. Si se está estudiando una prueba no aceptada se realizará un estudio experimental, generalmente sin seguimiento.

Para los estudios de fase I a III se disponen de varias opciones, pero se ha de tener siempre en cuenta que el diagnóstico se refiere a una situación en un punto concreto dentro de la historia natural de la enfermedad. A continuación se describen los principales diseños observacionales utilizados [238].

##### **13.1.4.3.1. Estudio de cohortes**

La prueba se aplica a una muestra representativa de sujetos de la población de referencia. Hay dos cohortes, los que dan un resultado positivo en la prueba y los que lo dan negativo. El seguimiento se usa para confirmar los negativos porque por razones éticas o económicas no se puede aplicar el criterio de verdad a estos sujetos.

Por ejemplo, para evaluar la detección de sangre oculta en heces en el diagnóstico precoz de cáncer de colon en población sin factores predisponentes de

este cáncer (colitis ulcerosa, poliposis), se aplica la prueba a una muestra de sujetos. Los que dan resultado positivo pueden ser evaluados mediante colonoscopia y posterior biopsia. Esta valoración no está exenta de complicaciones, por lo que su uso está más difícilmente justificado en los que dan un resultado negativo. El seguimiento se puede utilizar para verificar el resultado negativo, si al cabo de un cierto tiempo es estos sujetos no se diagnostica el cáncer, se considerará que el resultado fue un verdadero negativo; si, por el contrario, en el intervalo aparece un cáncer de colon, éste tenía que estar presente cuando se le realizó la determinación de sangre oculta en heces, y es por lo tanto un resultado falso negativo. El intervalo de tiempo que define el investigador es crucial. En este ejemplo los investigadores podrían adoptar como criterio un lapso de dos años. Estos diseños se utilizan con poca frecuencia [238].

#### **13.1.4.3.2. Estudio transversal**

Es la opción más frecuente y causa menos problemas que los estudios de casos y controles. A una muestra representativa de los sujetos en los que luego se utilizará la prueba se les aplica la misma y todos los resultados, positivos y negativos, se confirman mediante la prueba de referencia.

Generalmente estos estudios se hacen sobre una serie consecutiva de pacientes que entran dentro del diagnóstico diferencial de la enfermedad a diagnosticar. Todos los parámetros de la prueba (sensibilidad, especificidad, valores predictivos y otros posibles índices de validez) se estiman sin inconvenientes, ya que este diseño estima la prevalencia de la condición en estudio [238]. Por ejemplo, si propone unos nuevos criterios para el diagnóstico en niños de la hepatitis autoinmune, la muestra deberá incorporar una serie representativa de todos los sujetos con diagnóstico diferencial de hepatitis autoinmune. Esta es una condición cardinal de las que debe reunir todo estudio de corte o transversal para valorar un sistema diagnóstico según la propuesta y revisión de Knottnerus en 2003. De entre las otras circunstancias recomendables destaca la de llevar a cabo una recogida

prospectiva de los datos, aunque también son aceptables las aproximaciones retrospectivas o ambispectivas. El *standard* de referencia debe de ser independiente de los resultados de la prueba diagnóstica a estudio. Dado que establecer un patrón de referencia puede ser difícil (o incluso imposible en ausencia de un concepto fisiopatológico claro) en algunos casos, algunas posibles soluciones son establecer un panel de expertos independiente que justifique con claridad el diagnóstico o llevar a cabo un reclutamiento transversal de los casos con posterior seguimiento (estudio transversal diferido). Para que los resultados sean relevantes en la práctica clínica real el análisis de los indicadores de validez debe de realizarse con “intención de diagnosticar”, es decir, considerando también los casos excluidos por cualquier motivo entre el total sobre el que se referencian los eventos del indicador [363].

#### **13.1.4.3.3. Estudio de casos y controles**

Es el diseño que origina más problemas. Se selecciona un grupo de sujetos con la enfermedad diana y otro que no la tienen, según los resultados de la prueba de referencia. Los casos deben representar el espectro de la enfermedad y no ceñirse solo a los más graves como ocurre con frecuencia, que suelen ser los casos en los que más evidente es el diagnóstico cuando en la práctica clínica real lo más probable es que la prueba bajo evaluación tenga que aplicarse a los pacientes en los que el diagnóstico no está claro. Los controles deben representar el conjunto de diagnósticos diferenciales de la enfermedad. En los enfermos se estima la sensibilidad y en los no enfermos la especificidad.

Si los enfermos no mantienen la prevalencia respecto a los no enfermos, los valores predictivos no se pueden estimar. Dar unos valores para los mismos en base a la tabla de contingencia que genera el estudio es un error frecuente dado que la prevalencia de la enfermedad se desconoce. Es el investigador el que decide el número de controles por cada caso, no la epidemiología de la enfermedad la que genera dicha relación. Es esta situación hay que obtener la prevalencia de otras

fuentes o realizar un análisis de sensibilidad, variando la prevalencia dentro de un rango razonable.

Si el proceso se realiza prospectivamente no suele haber otras fuentes de error sistemático. No sucede así si se hace retrospectivamente porque hay que reconstruir cuál fue la secuencia de pruebas diagnósticas y si se realizaron a todos los individuos por igual [238].

#### **13.1.4.4. *Sesgos en la validez de sistemas o tests diagnósticos***

##### **13.1.4.4.1. Sesgo de selección**

Asumir que los valores de una prueba diagnóstica son constantes en diferentes subpoblaciones de pacientes puede no ser cierto. Este es un aserto que con frecuencia se realiza en la toma de decisiones en la clínica. Por ejemplo, en el diagnóstico de la hepatitis autoinmune, los valores de sensibilidad y especificidad de los criterios simplificados no son iguales en población adulta que en menores de 16 años [61,210,211]. Esto pone de manifiesto la importancia de no cometer errores en la selección de los pacientes, para que los valores promedio que se obtengan correspondan a la realidad.

Otro de los aspectos importantes es que las diferentes sensibilidades y especificidades encontradas en distintos estudios pueden justificarse en base a diferentes tipos de enfermos. Esto implica también que en la valoración de pruebas diagnósticas se tiene que investigar cuáles son los factores que influyen en el diagnóstico, en unas mejores o peores sensibilidad y especificidad, qué es lo que determina que un sistema o criterios diagnósticos den un resultado positivo o negativo.

En un diseño de casos (prevalentes) y controles, estos últimos deben representar la población de la que se originan los casos y debe incluir el espectro apropiado de todos los no enfermos en los que se aplicará la prueba, que entran dentro del diagnóstico diferencial de la patología objetivo. Quizá el error más cometido es la comparación de sujetos claramente enfermos (en los que no cabe la



menor duda diagnóstica) con sujetos sin ninguna patología o con escasas comorbilidades, claramente diferenciables de la patología índice por otros criterios que no son la prueba diagnóstica problema. Esto justifica el que inicialmente muchos criterios o pruebas diagnósticas parezcan prometedoras y luego decaiga el entusiasmo que motivaron al aplicarse dentro de contexto real, con pacientes limítrofes. Ésta es la historia del antígeno carcinoembrionario con el cáncer de colon. En una primera etapa se compararon pacientes con estadios III de Dukes (cáncer invasivo) con donantes sanguíneos. Al aplicarse es estadios menos avanzados y en pacientes con otra patología digestiva, perdió capacidad discriminativa y hoy ha quedado como marcador pronóstico. Este error justifica el que muchas veces la expectativa inicialmente despertada por una prueba diagnóstica no se confirme en estudios posteriores [238].

Si la prueba diagnóstica se analiza retrospectivamente se debe garantizar que no influya en la posterior elección del criterio de referencia. En este tipo de diseños se puede presentar con mayor facilidad el sesgo de verificación (*work-up bias*), que consiste en que a los pacientes que dan positivo en la prueba se les practica con más frecuencia la prueba de referencia que a los que dan negativo. En los estudios prospectivos, el sesgo de verificación puede corregirse mediante el seguimiento, para comprobar que la enfermedad no existe en los que dieron un resultado negativo a la prueba que se valora.

La comprobación de que los resultados de una prueba cambian en función de las características de presentación de la enfermedad anterior agrava además el problema del sesgo de verificación si la probabilidad de un resultado anormal cambia con la sintomatología y ésta influye en la remisión del paciente para confirmación diagnóstica. Esto se produce con la clínica de la angina y de ciertos diagnósticos diferenciales (como la hiperventilación) y la probabilidad de un resultado positivo a la prueba de ejercicio. Si esta relación es directa, se remitirán para confirmación diagnóstica (angiografía en el caso de que la condición problema sea la enfermedad trombotica coronaria) sobre todo los pacientes más graves, en

los que se encontrarán más resultados positivos, tanto enfermos como no enfermos, La sensibilidad aumentará y la especificidad tenderá a descender (ya que se detectarán más falsos positivos). La sensibilidad seguirá aumentando incluso si la sintomatología de los individuos no enfermos no se relaciona con la prueba (en cuyo caso la especificidad, naturalmente, no cambia).

Si la verificación diagnóstica no solo se realiza por razones de la sintomatología, sino también en función de los resultados de una prueba diagnóstica: la sensibilidad aumenta ligeramente, mientras que la especificidad desciende drásticamente. Lo anterior se justifica con facilidad: si pocos sujetos con la prueba negativa se envían para la prueba de confirmación, los individuos falsos negativos serán pocos tras la confirmación diagnóstica y la sensibilidad aumenta. Por la misma razón, los sujetos verdaderos negativos estarán en escaso número en relación con los falsos positivos y la especificidad desciende.

En presencia del sesgo de verificación aumenta la prevalencia en todos los casos. Esto subraya la precaución que hay que tener con los datos retrospectivos en los que se sospecha la existencia de un sesgo de verificación, que puede afectar a todos los parámetros de valoración diagnóstica. La realización de la prueba de referencia basada exclusivamente en los resultados de la prueba diagnóstica que se valora no altera los valores predictivos de una prueba diagnóstica, aunque sí la sensibilidad y la especificidad. La pregunta que hay que plantearse es si la referencia de los pacientes para confirmación diagnóstica se hace solo basándose en los resultados de la prueba, sin que influya ninguna otra característica. El único escenario donde no puede tener lugar este fenómeno es cuando el criterio diagnóstico de referencia se aplica a todos los pacientes, casos y no casos [238].

Sea cual sea el mecanismo por el que se produce el sesgo de verificación, la consecuencia suele ser la misma: la muestra pierde representatividad, ya que el número de resultados positivos (verdaderos y falsos) aumenta respecto a los negativos y no se puede calcular directamente la sensibilidad y la especificidad.

Tampoco se conoce la prevalencia del proceso en estudio porque solo una parte de los sujetos tienen confirmación diagnóstica.

En cambio, si el criterio para aplicar el sistema diagnóstico de referencia se realiza exclusivamente según la prueba diagnóstica (se ignora el resto de la información clínica del paciente), sí se pueden estimar los valores predictivos ya que los pacientes son remitidos en función de los resultados positivos o negativos. A partir de estos valores predictivos se puede calcular la sensibilidad y la especificidad de la prueba si se conoce la prevalencia de la enfermedad en la población diana a partir de las fórmulas dadas por el teorema de Bayes.

Si se sabe que no solo influye el resultado de la prueba a estudio para llevar a cabo la confirmación diagnóstica (por desgracia es lo más habitual) la corrección anterior no sirve. Si se conocen y se han cuantificado los elementos que gobiernan la referencia, hay ecuaciones que permiten la corrección del sesgo de verificación [238].

Otro error frecuente al valorar una prueba es la exclusión de los casos dudosos. Los resultados de las pruebas diagnósticas no son siempre positivos y negativos claros, también los hay dudosos, y éstos, con frecuencia, ni siquiera figuran en la metodología de un estudio por lo que no se sabe si se han producido y que es lo que se ha hecho con ellos. La exclusión de los casos dudosos tiene como consecuencia inmediata aumentar la sensibilidad y la especificidad de una prueba diagnóstica (y su inclusión las reducen, aproximándolas a la realidad) [238].

#### **13.1.4.4.2. Sesgo de información**

Al valorar una prueba diagnóstica debe garantizarse que su lectura sea independiente de la del criterio de referencia. Este es un criterio de calidad aceptado para que no haya influencias entre las pruebas. Con ello se evita el sesgo de revisión.

Una radiografía de tórax no tiene por qué verse igual si el clínico conoce la existencia de una prueba de Mantoux de 20 mm de diámetro y una lesión dudosa

puede ser adscrita más fácilmente a una probable tuberculosis basándose en el conocimiento adicional que se tiene. De ahí el desconocimiento que debe existir de los resultados de la prueba de referencia a la hora de evaluar la prueba diagnóstica.

El control del sesgo de revisión ha mejorado con el tiempo y cada vez más investigadores se preocupan de que las evaluaciones diagnósticas se hagan con independencia.

Es conocido que la experiencia puede interferir en el diagnóstico, pero su influencia pocas veces se ha documentado numéricamente. La experiencia en el campo diagnóstico se podría definir en parte como la integración subjetiva de la prevalencia de la enfermedad dentro de la experiencia clínica, la forma de presentación de la enfermedad.

Egglin y Feinstein (1996) lo comprobaron en la radiología y lo llamaron sesgo del contexto. Para estudiarlo alteraron la prevalencia de embolismo pulmonar de una manera muy original sin notificarlo a los clínicos encargados del diagnóstico. Suministraron a un grupo A de radiólogos 40 arteriografías pulmonares con una frecuencia de embolismo pulmonar del 60%, y a otro grupo B les suministraron 40 arteriografías con una frecuencia de embolismo pulmonar del 20%. Las revisaron y evaluaron y al cabo de ocho semanas el grupo A evaluó la serie B y viceversa (un diseño cruzado). Se halló una mayor sensibilidad y una menor especificidad en el grupo A, aunque no llegaron a ser significativas. La habituación de las primeras radiografías condujo a que el clínico diera con mucha mayor proporción un resultado positivo (de embolismo pulmonar) y en menor proporción negativo. Esto condujo a un aumento (proporcional) de falsos positivos y a un descenso proporcional de falsos negativos [364].

En realidad, el clínico cambia su criterio porque intuye que la mayor probabilidad preprueba de embolismo pulmonar condiciona ya un alto valor de la probabilidad postprueba (valor predictivo positivo). Las alteraciones de la sensibilidad y especificidad surgen como consecuencia de la búsqueda de un alto valor predictivo positivo.

Los sujetos con mayor probabilidad de ser erróneamente clasificados son los que tienen un valor real del parámetro diagnóstico próximo al punto de corte (cuanto más alejado, menor es la probabilidad de cometer un error). Los valores reales del parámetro diagnóstico dependen del espectro de enfermedad que existe en una colectividad.

El espectro de enfermedad existente en una comunidad es un determinante de la prevalencia de enfermedad diagnosticada y además influyen la mala clasificación de los individuos, si como se ha mencionado, abundan los casos con valores del parámetro diagnóstico en la frontera del punto de corte. Esto establece una relación entre la prevalencia de enfermedad y los parámetros de validez interna de una prueba diagnóstica: sensibilidad especificidad y razones de verosimilitud, y no solo con los valores predictivos. En la mayor parte de los casos se comprueba que hay una fuerte relación de dependencia entre la prevalencia con los parámetros de validez. Lo más habitual es que un aumento de la frecuencia de la enfermedad produce un incremento de la sensibilidad y un descenso en la especificidad.

Lo anterior supone que la clásica dependencia de los valores predictivos de la prevalencia se mitiga en parte. Por ejemplo, si la prevalencia desciende, el valor predictivo positivo disminuye; pero si la especificidad aumenta, desciende el número de falsos positivos.

Las razones de verosimilitud también se influyen por el error de mala clasificación: ambas (definiendo la negativa como el inverso de la misma) se asocian negativamente con el error de medición. La prevalencia presenta relación negativa con la razón de verosimilitud positiva (a mayor prevalencia menor razón de verosimilitud) y relación positiva con la razón de verosimilitud negativa.

Los hechos anteriores permiten anticipar el fenómeno por el cual una prueba diagnóstica investigada en un medio clínico tendrá una peor sensibilidad y mayor especificidad en el cribado de una población general asintomática.

### 13.1.4.4.3. Otros sesgos

En la siguiente tabla se resumen otros sesgos habituales (y su impacto y forma de corrección) en los estudios sobre pruebas o criterios diagnósticos.

Tabla 56: Sesgos en estudios de evaluación de sistemas diagnósticos. Reproducido de Cabello López et al. Rev Esp Cardiol. 1997;150;515. Con permiso de Elsevier.

Tipo de sesgo	Modo de producción	Consecuencias	Modos de control
Sesgo por inadecuado espectro de enfermedad o sesgo de selección de casos [365–369]	No se tiene en cuenta el espectro clínico, patológico o de comorbilidad	Sobreestima Se y Sp si se representa a los casos graves. Si se trata de casos leves infraestima Se y Sp	1) Representar el espectro completo en la muestra. 2) Describir el espectro en el análisis. 3) Análisis de la prueba en los subgrupos.
Sesgo del <i>gold standard</i> imperfecto [239,321,370]	No se dispone de un buen <i>gold standard</i> y se usa el disponible (aunque no clasifique óptimamente)	Generalmente sobrestima Se y Sp (a veces infraestima)	1) Seguimiento clínico de los pacientes para ver si son enfermos o no. 2) Correcciones matemáticas si se dispone de un subconjunto de pacientes con adscripción definitiva.
Sesgo de incorporación [365,366,371]	Elementos del <i>test</i> forman parte del <i>gold standard</i> (están incorporados)	Sobrestima Se y Sp	Conceptualización adecuada del <i>gold standard</i> y el <i>test</i>
Sesgo por revisión del diagnóstico o del <i>test</i> [365,366,371]	La interpretación del <i>test</i> o del estándar se realiza conociendo el otro resultado, es decir, de modo no ciego	Sobrestima Se y Sp	Cegado de las personas que interpretan (y realizan) la prueba a validar y la de referencia
Sesgo de verificación diagnóstica [365,366,372,373]	El resultado del <i>test</i> condiciona la realización del <i>gold standard</i>	Sobrestima Se e infraestima Sp	1) Realizar el estándar en todos los pacientes del estudio. 2) Seguimiento de resultados negativos. 3) Correcciones matemáticas.
Resultados no interpretables [365,366,374]	Es una eventualidad que se produce en cualquier prueba o estándar	Sobrestima Se y Sp	1) Repetición del <i>test</i> , si es posible. 2) Inclusión en el análisis de los casos no interpretables.
Sesgo por variabilidad en la interpretación de resultados [365,366]	Diversos observadores que actúan dentro del estudio tienen diferente Se y Sp. El mismo observador cambia su Se y Sp dentro del estudio debido a entrenamiento.	Generalmente infraestima Se y Sp	1) Estudios previos (piloto) de consistencia interobservadores. 2) Correcciones matemáticas.

Se: Sensibilidad. Sp: Especificidad.

#### **13.1.4.4.4. Errores más comunes en el razonamiento clínico probabilístico**

A medida que se avanza por el desarrollo de esta introducción, se pone de manifiesto que toda exploración de casos clínicos mediante sistemas diagnósticos se basa en criterios de probabilidad. Una interpretación incorrecta de estos fundamentos lleva a que, en ocasiones, exista más preocupación en la práctica asistencial por la reunión de la información que de su integración. Según Delgado, Llorca y Doménech, los errores más frecuentes que afectan al correcto razonamiento clínico probabilístico son:

1) Con frecuencia se desechan las evidencias que no confirman el diagnóstico pensado, mientras que se realizan los que confirman lo deseado.

2) Se suelen ignorar los resultados negativos (o ausencia de un resultado positivo). En este sentido es conveniente recordar que un signo puede ser patognomónico (valor predictivo positivo del 100%), pero esto no dice nada de la frecuencia con que se presenta en la enfermedad. Siempre que se señale este tipo de signos se debería de citar también la frecuencia de su presentación en el total de enfermos (lo que representaría su sensibilidad).

3) Los médicos, como todos los humanos, tienen limitaciones como observadores y cometen errores. Algunos factores contribuyen a la presencia de estos errores: la ambigüedad de presentación de un signo; las condiciones en las que se realiza la observación (por ejemplo, la ictericia se ve más fácilmente a la luz del sol), el estado físico y emocional del observador, las expectativas del médico que realiza la observación y la influencia de compañeros que también están involucrados en la experiencia de dicha observación.

4) Las personas somos malos estimadores intuitivos de la probabilidad de un evento. Además, se ha demostrado que los médicos son conservadores en los ajustes de la misma [238].

El diseño empleado para llevar a cabo un estudio sobre criterios diagnósticos debe de contemplar que estos errores se van a producir sin adecuados métodos de compensación que deben de explicarse suficientemente para facilitar la

reproductibilidad del estudio de una forma objetiva y una adecuada discusión de sus resultados.

### **13.1.5. Evaluación de la fiabilidad de un sistema diagnóstico**

Hasta el momento se ha abordado el análisis de la validez de los sistemas diagnósticos pero su calidad no depende exclusivamente de la validez, también de la fiabilidad.

La fiabilidad (también llamada consistencia) de una prueba es su capacidad para producir los mismos resultados cada vez que se aplica en las mismas condiciones. Unos criterios diagnósticos son fiables si son reproducibles sin exceso de variabilidad. Sin embargo, las mediciones realizadas por los distintos sistemas o pruebas diagnósticas están sujetas a múltiples fuentes de variabilidad. El origen de esta variabilidad puede hallarse en el propio sujeto objeto de la medición (variabilidad biológica), en el instrumento de medida que se emplee o en el observador que la ejecuta o interpreta. A la hora de estudiar la fiabilidad diagnóstica de un método tiene especial interés evaluar la variabilidad encontrada entre las mediciones realizadas por múltiples observadores o instrumentos, y la variabilidad encontrada entre mediciones repetidas realizadas por el mismo observador o instrumento [3].

Los métodos para la valoración de la fiabilidad de las mediciones clínicas están en función del tipo de variable a medir.

#### **13.1.5.1. Variables categóricas nominales**

A las pruebas cuyos resultados admiten varias categorías posibles sin un orden jerárquico entre ellas se les aplica el índice *kappa*.

Se puede construir una tabla de contingencia que refleje los recuentos de casos en los que hay acuerdo (casillas a y d) y desacuerdo (casillas b y c).



Tabla 57: Tabla de contingencia entre los resultados dados por dos observadores diferentes sobre una misma muestra de casos.

Observador 1 →	<b>Enfermedad</b>	<b>Ausencia de enfermedad</b>
Observador 2 ↓	<b>Enfermedad</b>	<b>Ausencia de enfermedad</b>
<b>Enfermedad</b>	Acuerdo positivo (A)	Desacuerdo (B)
<b>Ausencia de enfermedad</b>	Desacuerdo (C)	Acuerdo negativo (D)

La forma más sencilla de expresar la concordancia entre las dos evaluaciones es mediante el porcentaje o proporción de acuerdo o concordancia simple observada ( $P_O$ ), que corresponde a la proporción de observaciones concordantes:

$$P_O = \frac{A + D}{Total}$$

Un indicador de concordancia útil solo se puede interpretar si se tiene en cuenta que parte del acuerdo encontrado puede ser debido al azar. Las observaciones esperadas ( $O_E$ ) por azar en cada casilla de la tabla de contingencia se pueden calcular a partir del producto de los marginales de la fila y columna correspondientes, dividido por el total. Así, las  $O_E$  para cada casilla tienen las siguientes expresiones:

$$O_E^A = \frac{(A + B) \times (A + C)}{Total}$$

$$O_E^B = \frac{(B + A) \times (B + D)}{Total}$$

$$O_E^C = \frac{(C + A) \times (C + D)}{Total}$$

$$O_E^D = \frac{(D + B) \times (D + C)}{Total}$$

La proporción de acuerdo esperada ( $P_E$ ), de forma análoga a la observada, corresponde a la proporción de observaciones iguales esperadas:

$$P_E = \frac{O_E^A + O_E^D}{Total}$$

El índice *kappa* ( $\kappa$ ) es el resultado de formular el cálculo de la proporción (en vez del recuento) de observaciones no atribuibles al azar. Dicho de otra forma, ofrece una estimación del grado de acuerdo no debido al azar a partir de la proporción de acuerdo observado y la proporción de acuerdo esperado:

$$\kappa = \frac{P_O - P_E}{1 - P_E}$$

El índice *kappa* puede adoptar valores entre  $-1$  y  $1$ . Es  $1$  si existe un acuerdo total,  $0$  si el acuerdo observado es igual al esperado y menor de  $0$  si el acuerdo observado es inferior al esperado por azar. La interpretación más aceptada de los rangos de valores situados entre  $0$  y  $1$  es la propuesta por Landis y Koch en 1977 (tabla) [262]. Al igual que con otros estimadores poblacionales expuestos en esta tesis, los índices *kappa* se deben calcular con sus intervalos de confianza [375]. Para el caso de un estudio de la concordancia de la capacidad diagnóstica de dos criterios diferentes, en los que la respuesta sea presencia o ausencia de la enfermedad de interés, cabría utilizar este índice.

Tabla 58: Interpretación de los valores del índice *kappa*.

Valor de <i>kappa</i>	Grado de concordancia
0,81 – 1,00	Excelente
0,61 – 0,80	Buena
0,41 – 0,60	Moderada
0,21 – 0,40	Ligera
<0,20	Mala

El índice *kappa* también puede ser aplicado a pruebas cuyos resultados tengan más de dos categorías nominales utilizando el mismo cálculo que el del acuerdo esperado por azar.

### 13.1.5.2. Variables categóricas ordinales

Cuando el resultado de la prueba analizada adopte más de dos categorías entre las que existe cierto orden jerárquico (resultados discretos ordinales) se debe de aplicar el índice *kappa* ponderado como estimador del grado de concordancia. En esta situación, pueden existir distintos grados de acuerdo o desacuerdo entre las evaluaciones repetidas. Es el caso de, por ejemplo, una variable que represente un riesgo categorizado en los términos “bajo”, “medio” y “alto”. Es evidente que no

puede considerarse igual una discrepancia entre riesgo bajo y medio, que entre bajo y alto [3].

El índice *kappa* ponderado nos permite estimar el grado de acuerdo, considerando de forma diferente esas discrepancias. Para ello, debemos asignar diferentes pesos a cada nivel de concordancia. Habitualmente se asignará un peso 1 al acuerdo total (100% de acuerdo) y un peso 0 al desacuerdo extremo. A los desacuerdos intermedios se les asignarán pesos intermedios, en función del significado que tengan las distintas discordancias en el atributo estudiado.

El índice *kappa* ponderado se calcula de forma similar al índice *kappa*, con la diferencia de que, en las fórmulas de las proporciones de acuerdo observado y esperado, las frecuencias de las distintas casillas se deben multiplicar por sus pesos respectivos [376]. El problema en este caso es cómo asignar el esquema de ponderación. Dos son los más aceptados: la solución cuadrática y la ponderación por los errores absolutos. Cuando no se disponen de ponderaciones objetivas de las diferentes discordancias, se prefieren los pesos cuadráticos, que corresponden al cuadrado de las distancias de cada casilla respecto a la diagonal principal en la tabla de contingencia (donde se sitúan las casillas de acuerdo perfecto). La ponderación por los errores absolutos asigna a cada casilla un peso equivalente (lineal) a las unidades de distancia entre la misma y la diagonal principal [238].

El uso del índice *kappa* presenta algunos problemas clínicos. El valor de *kappa* tiende a disminuir al aumentar el número de categorías. Esto tiene más relación con la forma en que se definen las categorías que con la propia reproducibilidad de los procedimientos. Por este motivo se debe de justificar que la categorización del evento que se pretende diagnosticar esté justificada desde un punto de vista práctico. Si se emplea para comparar la fiabilidad de criterios diagnósticos para la hepatitis autoinmune, parece claro que el máximo grado de categorías que se debe de admitir son las tres originales con las que se validaron los criterios clásicos: ausencia de hepatitis autoinmune, diagnóstico probable y diagnóstico definitivo [61]. Además, la influencia de los desequilibrios entre

resultados positivo y negativos está en cierto modo gobernada por la prevalencia subyacente de la condición en estudio, lo que indica que la prevalencia afecta el resultado de *kappa*. Se entiende que esto tiene derivaciones clínicas importantes si se tiene en cuenta que prevalencias superiores al 50% suelen ser poco frecuentes, un *kappa* realizado en un área de frecuencia más elevada de la enfermedad será mayor que el de un área de baja prevalencia. Esto quiere decir que dos pruebas de fiabilidad serán comparables si las circunstancias subyacentes de frecuencia de la enfermedad son similares, con independencia de otros detalles metodológicos [238].

### **13.1.5.3. Variables cuantitativas continuas**

#### **13.1.5.3.1. Desviación estándar intrasujetos**

Cuando el resultado de una prueba se mide en una escala continua, podemos estimar el error de medición calculando la variabilidad existente entre medidas repetidas en los mismos sujetos. El parámetro que mejor refleja dicha variabilidad es la desviación estándar intrasujetos (excluyendo la observada entre sujetos). Para calcularlo necesitamos una serie de sujetos a los que se les realice al menos dos mediciones. La desviación estándar intrasujetos puede calcularse fácilmente a través del análisis de la varianza (ANOVA), que permite obtener el parámetro CMr (cuadrados medios de los residuos), que es la varianza intrasujetos. Si realizamos la raíz cuadrada de CMr obtendremos la desviación estándar intrasujetos ( $s_i$ ). La  $s_i$  puede calcularse igualmente a partir del ANOVA para estudios con más de dos mediciones por sujeto.

Utilizando la  $s_i$  podemos cuantificar el margen de error de las mediciones. Así, se puede estar seguro de que la diferencia entre una medición determinada y el verdadero valor no será mayor de 1,96 veces la  $s_i$  en el 95% de las observaciones. También que la diferencia entre dos mediciones repetidas en un mismo sujeto no superarán 2,77 veces la  $s_i$  en el 95% de las observaciones [377,378].

### 13.1.5.3.2. Coeficiente de correlación intraclase

Si solo se realizan dos mediciones por sujeto, la forma más intuitiva de compararlas es representarlas en un diagrama de puntos, examinar si existe relación lineal entre ambas y calcular su coeficiente de correlación de Pearson ( $r$ ).

Sin embargo, la existencia de una fuerte relación lineal con un alto coeficiente de correlación no indica que haya una buena concordancia entre las mediciones, solamente que los puntos en el diagrama se ajustan a una recta. El coeficiente de correlación depende, en gran manera, de la variabilidad entre sujetos, por ello, varía sensiblemente en función de las características de la muestra donde se estima, afectándole especialmente la presencia de valores extremos. Si una de las mediciones es sistemáticamente mayor que otra, el coeficiente de correlación será muy alto, a pesar de que las mediciones nunca concuerden. Estos problemas son evitados utilizando el coeficiente de correlación intraclase (CCI).

El CCI estima la concordancia entre dos o más medidas repetidas. El cálculo del CCI se basa en un modelo de ANOVA con medidas repetidas, aplicándose distintas fórmulas en función del diseño y los objetivos del estudio [379]. El escenario más simple es aquél en el que se estima la variabilidad de las medidas, sin tener en cuenta la variabilidad aportada por los distintos observadores (diseño de una vía con factor aleatorio) o aquél en el que varios observadores realizan una sola medida para cada caso. Considerando este diseño, y utilizando los resultados del ANOVA, podemos calcular el CCI con la siguiente fórmula:

$$CCI = \frac{CMp - CMr}{CMp + (k - 1) \times CMr}$$

Donde  $k$  es el número de observaciones por caso,  $CMp$  son los cuadrados medios entre pacientes y  $CMr$  los cuadrados medios de los residuos.

Si el CCI fuera mucho menor que  $r$ , habría que pensar que existe un cambio sistemático entre una medida y otra, lo que podría estar causado por un efecto de aprendizaje [3].

### 13.1.5.3.3. Método de Bland-Altman

Un método alternativo para analizar la concordancia entre dos observaciones repetidas que se miden en una escala continua es el método gráfico descrito por Bland y Altman [380]. Consiste en representar en un diagrama de puntos la diferencia entre los pares de mediciones contra su media. Ello permite examinar el ancho de las diferencias y su relación con la magnitud de la medición. Además, se puede estimar la desviación estándar de las diferencias y los intervalos entre los que cabe esperar que se encuentre el 95% de las diferencias.

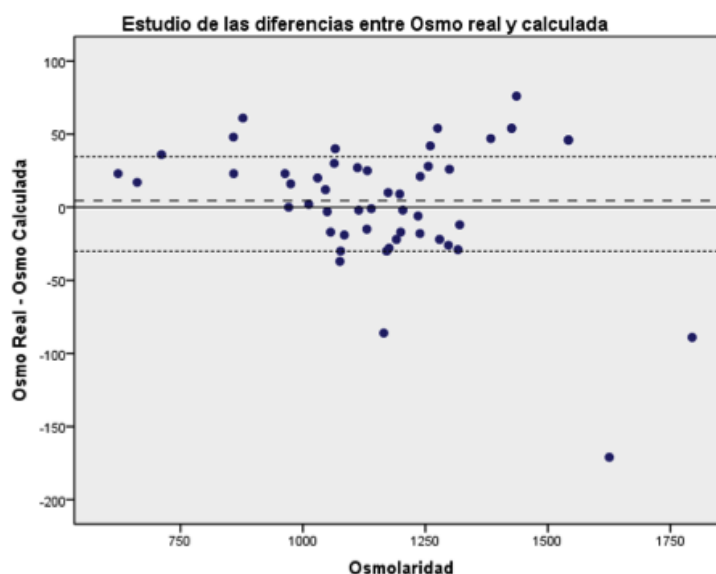


Figura 90: Ejemplo de diagrama de Bland-Altman que representa las diferencias entre dos métodos distintos de estimar la osmolaridad de las soluciones de nutrición parenteral en función de la magnitud del valor real. Obtenido a partir de datos propios con el paquete estadístico SPSS 21.0®.

El procedimiento de Bland y Altman supone normalidad de la variable diferencia e independencia entre la magnitud de la diferencia y la magnitud de los promedios. A efectos prácticos, para demostrar que se cumplen las condiciones para emplear este método, se puede contrastar la normalidad de la variable diferencia con la prueba de Shapiro-Wilk [381]. Cuando la variabilidad en las medidas no es constante, sino que cambia al aumentar o disminuir la magnitud de la medida

(diferencias de tipo proporcional), el cálculo se complica. Si existe correlación significativa entre las diferencias y las medias, la variabilidad no será constante. En ese caso, puede intentarse realizar transformaciones logarítmicas de los datos o analizar la variabilidad por separado para varios intervalos de valores [3].

### **13.1.6. Lectura crítica de estudios para pruebas diagnósticas**

Los requisitos para valorar críticamente los artículos sobre pruebas diagnósticas equivalen a las preguntas que debe responder el artículo en cuestión. Para Sackett las preguntas fundamentales a las que tiene que hacer referencia un estudio sobre una prueba diagnóstica son [382]:

1) ¿Se ha hecho una comparación independiente y enmascarada con el criterio de referencia? Se refiere a la existencia de un sesgo de mala clasificación.

2) ¿La prueba se ha evaluado en una muestra con el espectro apropiado de pacientes, análogo al que se utilizaría en la práctica? Se plantea un posible sesgo de selección.

3) ¿Se aplicó el criterio de referencia con independencia del resultado de la prueba? Evita el sesgo de verificación.

Cuando se analiza un artículo sobre pruebas diagnósticas hay que plantearse su evaluación en tres ejes básicos: validez de los resultados, su forma de presentación y la aplicación de los mismos.

#### **13.1.6.1. Validez de los resultados**

Ante a lectura del material y métodos de un artículo de este tipo hay que reflexionar acerca de las siguientes cuestiones:

1) ¿Se hizo una comparación independiente y ciega con un criterio de referencia? La lectura de cada prueba se hace sin conocer el resultado de la otra. Intenta evitar un sesgo de mala clasificación. Es más fácil que este criterio se cumpla en los estudios prospectivos o ambispectivos.

2) ¿La muestra incluye un espectro apropiado de pacientes en los que se aplicará la prueba diagnóstica?

3) ¿Los resultados de la prueba en evaluación influyeron en la decisión de realizar el criterio de referencia?

4) ¿Se describieron los métodos de la prueba con suficiente detalle como para permitir su replicación? Esto es una característica general de cualquier investigación.

#### **13.1.6.2. Exposición de los resultados**

Ha de haber datos sobre la probabilidad preprueba y si el diseño no permite su cálculo es conveniente indicarlo, así se evitan malas interpretaciones por parte de los lectores. En la literatura científica anglosajona se insiste sobre la prioridad que tiene la presentación de las razones de verosimilitud, como los mejores parámetros en la interpretación de la bondad de una prueba diagnóstica. Por este motivo también puede ser muy conveniente presentar su transformación en los pesos de evidencia de Good-Turing. Al menos deben figurar la sensibilidad y especificidad, que permiten su cálculo. Los valores de sensibilidad, especificidad y los otros indicadores de la validez no deben limitarse a las estimaciones puntuales: deben aportarse también sus intervalos de confianza u otra medida de variabilidad.

#### **13.1.6.3. Aplicabilidad de los resultados**

La aplicabilidad de los resultados depende de la validez externa de la investigación. En la valoración de ésta intervienen criterios subjetivos. A modo de guía se sugiere que para valorar la validez externa se analice la reproducibilidad de la técnica en otro medio y el espectro de enfermos en los que se ha aplicado (si es o no similar a los que se presentan en el entorno del lector).

Con independencia de lo anterior se debe valorar la utilidad de los resultados. Ésta depende de si la prueba cambia de decisión sobre los pacientes. Una forma de aproximarse es calculando una proporción de sujetos que tienen



valores muy altos de las razones de verosimilitud (alejadas de la unidad), que son los individuos en los que se alcanza una razonable exactitud diagnóstica. La utilidad de los resultados está también unida a un aspecto que no debe olvidarse nunca y es que una prueba debe aplicarse para mejorar el estado de los pacientes, esto es, va ligada a una intervención preventiva o terapéutica. Una forma de analizar formalmente la utilidad de un sistema diagnóstico desde este punto de vista se expondrá más adelante.

En la presentación de los resultados de un estudio de validez diagnóstica, ante la falta de normalización y la baja calidad de muchos estudios, se ha emprendido una labor de homogeneización para la comunicación de los resultados similar a la declaración CONSORT para los ensayos clínicos [383]. Esta iniciativa se llama STARD (*Standards for Reporting of Diagnostic Accuracy*) y ha sido actualizada en el año 2015 [384,385]. Los puntos principales de la información que se deben conocer están resumidos en los anexos. Se aconseja que en los resultados se presente un flujo de pacientes similar al de la figura.

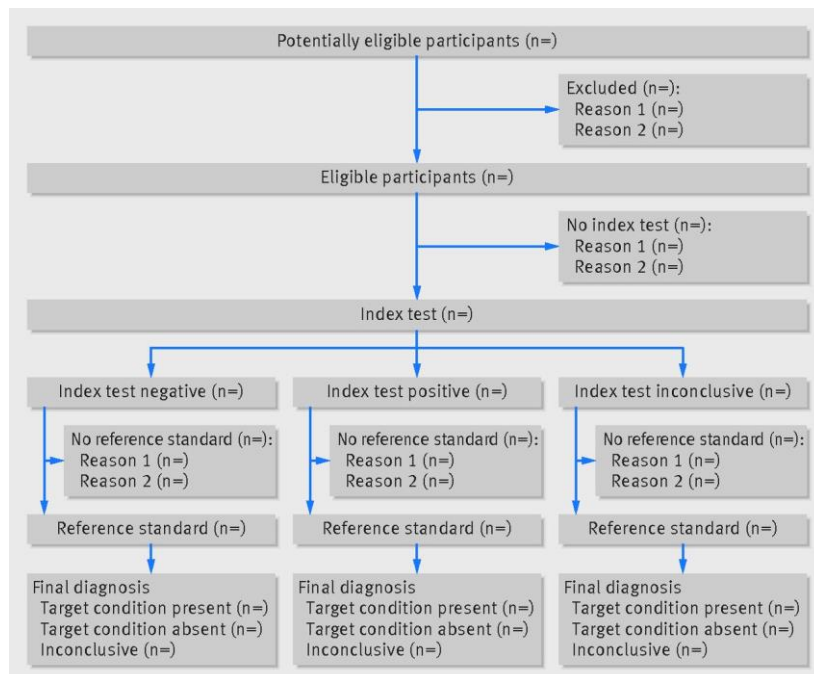


Figura 91: Diagrama prototípico de la declaración STARD para comunicar el flujo de pacientes a lo largo del estudio. Reproducido de Bossuyt et al. *BMJ*. 2015;351:h5527. Con permiso de BMJ Publishing Group Ltd.

### **13.1.7. Índices de predicción clínica como fundamento de la toma de decisiones en Medicina asistencial**

Determinar el diagnóstico y establecer el pronóstico de un paciente son dos actividades íntimamente relacionadas que constituyen el centro de la práctica de cualquier médico dedicado a la asistencia. Condicionan seriamente las decisiones que éste toma sobre las recomendaciones que realiza a las personas que trata: las pruebas que va a ordenar, el pronóstico que va a predecir o el tratamiento que se debe aplicar. Es claro que el principal papel del médico es tomar decisiones [386]. Pero ya se ha visto como todas las decisiones médicas -como las de cualquier científico- se toman en un ambiente de incertidumbre e inseguridad. Partiendo de este marco, y de la necesidad de formalizar la adopción de actitudes diagnósticas y terapéuticas, es cada vez más evidente como aumenta el grado de instrucción en las técnicas de decisión basadas en enfoques matemáticos adaptados de los mundos militar y mercantil [386].

La experiencia en la práctica clínica dota de una sensación intuitiva (el juicio u “ojo” clínico) de qué hallazgos de la anamnesis, la exploración física o los exámenes complementarios son cruciales para hacer un diagnóstico certero y predecir adecuadamente un pronóstico. Pero, desgraciadamente, la experiencia también enseña que esta misma intuición resulta a veces muy engañosa: como humanos estamos inclinados a dirigir nuestra atención hacia lo nuevo, lo raro, lo interesante o lo emocionalmente atractivo y no somos muy buenos haciendo observaciones sistemáticas, no sesgadas y consistentes en el tiempo [387].

Así, es habitual observar cómo una experiencia negativa con un paciente individual puede perjudicar la correcta práctica futura de un médico, incluso si esa experiencia fue una excepción rara. O cómo, por el contrario, si un enfermo responde muy bien a una medicación, se es propenso a creer que también funcionará con el próximo paciente. Nuestra vivencia durante el periodo de formación, a menudo en grandes hospitales terciarios, pudo predisponer a convivir

con la patología rara o muy abigarrada, y ello suele traer como resultado que se sobreestime la probabilidad de encontrarnos de nuevo ante esas formas de enfermedad.

La disponibilidad de información, como la lectura reciente de un artículo en una revista médica, aumenta la conciencia que se tiene de un determinado proceso o condición y hace también más probable que (correcta o incorrectamente) se diagnostique en los próximos pacientes. Además, la opinión de colegas o la literatura médica no aprendida críticamente puede estar más basada en anécdotas que en datos objetivos sobre prevalencia y/o evolución de una enfermedad. Estos y otros sesgos, que no son sino la traslación a la Medicina del problema de la inducción en ciencia, privan de capacidad para acertar siempre con nuestras predicciones clínicas.

Para ayudar a establecer objetivamente un pronóstico, basándose en la experiencia acumulada durante años por clínicos de gran prestigio, los médicos de la antigüedad desarrollaron reglas de predicción que expresaban en forma de aforismo [388]. Así, por ejemplo, los médicos del Egipto arcaico (tercer milenio antes de nuestra era) dejaron escrito:

*“Si examinas a un hombre que tiene una rotura en la cámara de la nariz y encuentras su nariz torcida, su cara desfigurada y una hinchazón por encima que sobresale, dirás acerca de él: una enfermedad que voy a tratar.*

*Si examinas un hombre que tiene una herida abierta en la cabeza, que penetra en el hueso, fractura el cráneo y deja el cerebro al descubierto, deberás palpar su herida. Comprobarás si la fractura que tiene en su cráneo es semejante a los pliegues que se forman en el cobre fundido, y si palpita y cede bajo tus dedos como la parte débil de la coronilla de un niño antes de que se suelde. Cuando suceda que no palpites ni ceda bajo tus dedos mientras que el cerebro está al descubierto y el paciente arroja sangre por ambas fosas nasales y tiene rigidez en el cuello, dirás acerca de él: una enfermedad que no es posible tratar” [389].*

De una época más cercana a la nuestra (siglos V-IV antes de nuestra era) data la descripción clásica de la facies hipocrática:

*“Conviene investigar así en las enfermedades agudas. Primeramente, observar la cara del enfermo, si es semejante a la de los sanos, sobre todo a la del mismo enfermo cuando tenía salud; porque esto sería lo mejor, y cuanto más diste de lo semejante tanto será más temible. Por ejemplo: la nariz afilada, hundidos los ojos, caídas las sienas, frías y encogidas las orejas y sus pulpejos retorcidos, y dura la cutis del rostro y tirante y árida, y la color de todo el semblante amarilla y amoratada. Si tal se presenta el semblante al comienzo de la enfermedad, sin que todavía por las demás señales puedan hacerse conjeturas, conviene preguntar (desde luego) si el enfermo estuvo desvelado, si padeció abundantes cámaras o si tiene por ventura mucha hambre. Cuando confesare alguna de estas cosas debe tenerse por menos de cuidado, juzgándose de todos modos en un día y una noche si por aquellas causas tiene tal apariencia el semblante. Pero si nada de esto confiesa ni en el tiempo dicho se compone su rostro, entiéndase que es señal de muerte segura” [390].*

Tales muestras de experiencia destilada son memorables. Sin embargo, simplifican en exceso la compleja realidad de los problemas individuales de la práctica clínica diaria, y evitan que se valore adecuadamente toda la riqueza de detalles o matizaciones con que se plantean las situaciones reales [388]. Ello les resta eficacia como herramientas que guíen el proceso de toma de decisiones en las condiciones de incertidumbre que constituyen la esencia de la actividad médica cotidiana.

En la Medicina científica actual, se llama regla o índice de predicción clínica (en inglés *clinical prediction rule* o *predictive index*) a una herramienta matemática de uso clínico capaz de ponderar cuantitativamente la contribución individual que varios componentes de la anamnesis, la exploración física y los resultados de análisis básicos de laboratorio tienen sobre el diagnóstico, el pronóstico y/o la respuesta

probable al tratamiento de un paciente individual [391]. En general, las modernas reglas o índices de predicción clínica (IPC) han sido desarrolladas en estudios realizados con bases de datos derivados de miles de pacientes y utilizando métodos matemáticos de análisis multivariable de gran sofisticación [392]. En la literatura internacional también se conocen como *clinical prediction guides*, o *clinical decision rules*. *Predicción* indica su poder de ayudar al médico a avanzar o adivinar con antelación la aparición de un evento clínico futuro. *Decisión* implica la capacidad para ayudar al médico a elegir entre una o varias alternativas de acción diagnóstica o terapéutica.

Los estudios sobre una IPC son estudios para establecer un pronóstico, pero desde el punto de vista epidemiológico son estudios superponibles a los que se realizan sobre métodos diagnósticos. La naturaleza de la relación entre la variable predictora (el *test* diagnóstico o el factor pronóstico) y la de resultado (presencia o ausencia de enfermedad, aparición o no del resultado) raramente es causal. El resultado del *test* o la presencia del factor pronóstico suele ser la consecuencia del proceso patológico, no su causa, o una situación intermedia en la cadena causal entre la enfermedad y su resultado, por ejemplo, la muerte. Por eso, en la mayoría de las situaciones clínicas reales, el límite entre diagnóstico y pronóstico se borra completamente hasta desaparecer. La aplicación de un IPC a veces condiciona una decisión terapéutica (por ejemplo, decidir a qué pacientes con lesiones traumáticas menores de tobillo debe practicarse una radiografía [393–395]), otras veces establece una predicción sobre un pronóstico (por ejemplo, cuál es la probabilidad de que un paciente con sospecha de enfermedad coronaria muera durante los próximos 4 años [396]) y otras muchas veces provee al médico de una probabilidad post-prueba o de una razón de verosimilitud para su aplicación en un problema de diagnóstico diferencial, que sería el caso del objeto de la tesis. Aunque el nombre correcto de estos últimos debería ser índice, regla o guía de diagnóstico clínico, se emplearán las siglas IPC independientemente de que su resultado sugiera una estrategia diagnóstico-terapéutica apropiada, dé la probabilidad estimada de un

evento clínico futuro o indique una variación en la verosimilitud de un determinado diagnóstico.

Cualquiera que sea el resultado generado, los actuales IPC demuestran su verdadero potencial cuando se utilizan en situaciones clínicas en las que el proceso de toma de decisiones es muy complejo, los riesgos clínicos potenciales son muy altos, o existe oportunidad de aumentar la eficiencia, ahorrando costes sin comprometer la eficacia del cuidado de los pacientes. Pero para ello es muy importante que estén bien desarrollados, validados, y se haya comprobado su impacto real en la excelencia de la atención sanitaria.

Los datos de validez de unos sistemas diagnósticos en formato de criterios de puntuación para el diagnóstico de la hepatitis autoinmune en niños tienen potencial de ser empleados como fundamento de un IPC, aspecto que será uno de los temas a explorar en este trabajo. En el bloque siguiente se detalla cómo se usa correctamente una prueba diagnóstica, desde el punto de vista clínico. Se trata del modelo de Pauker-Kassirer, que incorpora la utilización adecuada de un IPC al proceso de toma de decisiones en condiciones reales [397].

### **13.1.8. Análisis de decisiones clínicas**

Para aplicar eficientemente un sistema diagnóstico se necesita el nivel de certeza que se requiere para llevar a cabo una actitud terapéutica. Este grado de certeza traduce la probabilidad a partir de la cuál conviene proponer un tratamiento porque los beneficios esperables superan a los perjuicios potenciales. De forma análoga, también se puede estimar el grado de sospecha requerido solo para el hecho de aplicar la prueba diagnóstica. Racionalizar esta decisión es el objeto del llamado modelo de Pauker-Kassirer del análisis de decisiones clínicas [325].

En efecto, Pauker y Kassirer, en una serie de artículos publicados en los años 70 del siglo XX, desarrollaron un abordaje sistemático sencillo para ayudar a los clínicos a un uso correcto de las pruebas diagnósticas a pie de cama del enfermo

[279,316,398]. Está basado en la teoría matemática de la decisión y en el cálculo de probabilidades. La propuesta para el uso clínico correcto de las pruebas diagnósticas que se empleará en este trabajo se basa en la modificación que el profesor Jaime Latour ha hecho del modelo de Pauker-Kassirer [280]. Desde su perspectiva, el análisis decisional es una disciplina que, a través de una evaluación matemática de la validez de una prestación diagnóstica para una indicación patológica concreta, da una respuesta (emite un juicio diagnóstico y propone una actitud al respecto) coincidente con la que daría un médico dotado de excelente razonamiento. De este modo, aunque no asegure completamente la veracidad del resultado, expresa la mejor opción esperada, basándose en el beneficio de los indicadores de validez, dentro de un margen de seguridad y coste. Permite tomar la mejor decisión en términos de aproximación diagnóstica, tratamiento y pronóstico [280,350].

La premisa sobre la que se basa el modelo de Pauker-Kassirer para el análisis de decisiones clínicas es la posibilidad de estimar unos umbrales de probabilidad de enfermedad a los que denomina *umbrales de acción o decisión*. El cálculo de los umbrales de acción está basado en un procesado matemático del beneficio y el riesgo tanto de aplicar el sistema diagnóstico como del tratamiento. Para ello se precisa cuantificar la utilidad y la probabilidad de aparición de los posibles resultados de cada decisión clínica [297]. Más adelante se expondrán los métodos empleados para su cálculo.

Existen dos contextos clínicos en los que aplicar el modelo:

1) Escenario en el que existan pruebas diagnósticas válidas y reconocidas para la enfermedad de interés. En este caso, la decisión de aplicarlas o no, o de iniciar tratamiento o no, dependerá de dos umbrales de acción: el umbral diagnóstico y el umbral diagnóstico-terapéutico (conjuntamente denominados umbrales de acción para una prueba diagnóstica).

2) Escenario en el que no existen más pruebas diagnósticas válidas más allá de la estudiada para obtener sus indicadores de validez. En este caso solo tiene sentido calcular el umbral terapéutico para decidir si iniciar el tratamiento.

### **13.1.8.1. Valor de un procedimiento terapéutico o diagnóstico**

Los beneficios y los riesgos de una intervención, bien sea con fines diagnósticos o de tratamiento, hacen referencia a los resultados de dicha intervención, que pueden ser expresados en términos de utilidad. Cada proceder diagnóstico o terapéutico genera varias situaciones clínicas o resultados subsiguientes que tiene una probabilidad de ocurrencia. La suma de todas estas probabilidades es la unidad.

No siempre es posible presentar los resultados en la misma escala. Los resultados pueden describirse en términos de supervivencia, alivio de los síntomas, presencia de complicaciones importantes, o incluso unidades arbitrarias que midan diferentes valores relativos al paciente. Cuando no se usan las mismas unidades de medida, deberán ponderarse o ajustarse cada uno de los resultados en función de la importancia que se le asigne en relación con los otros resultados, de manera que presenten diferente peso específico a la hora de comparar resultados con otros. Por ejemplo, el riesgo/beneficio “supervivencia” debe de considerarse en mayor valía que al riesgo/beneficio “dolor”, multiplicando el parámetro “supervivencia” por un coeficiente mayor que el factor para “dolor” [279].

La utilidad de un procedimiento es una medida de la suma de las utilidades de cada uno de los resultados que conlleva dicha opción, ajustadas a la probabilidad de ocurrir cada resultado. La utilidad es pues un número con el que se mide el impacto del resultado. Este método permite la inclusión de las utilidades en el cálculo del beneficio y el riesgo, de un tratamiento y de una prueba diagnóstica.

#### **13.1.8.1.1. Beneficio neto de un tratamiento apropiado**

Son las consecuencias positivas que como media se producen entre los pacientes que tienen la enfermedad y son tratados correctamente [297].

Dado que se maneja en términos netos, se calcula restando a la utilidad ( $U$ ) de un tratamiento apropiado sobre los enfermos ( $T + |E +$ ), la utilidad de no tratar a un enfermo ( $T - |E +$ ). Si una enfermedad con tratamiento tiene una



supervivencia del 60% y sin tratamiento, del 30%, el beneficio neto de tratamiento apropiado ( $B_t$ ) puede expresarse como esta diferencia, es decir 0,3.

$$B_t = U(T + |E +) - U(T - |E +)$$

#### **13.1.8.1.2. Riesgo neto de un tratamiento inapropiado**

Son las consecuencias negativas que como media se producen entre los pacientes que no tienen la enfermedad y son tratados por equivocación [297].

Representa la diferencia en términos absolutos, es decir, sin signo negativo, entre la utilidad de no tratar a un sano y la de tratarlo.

Siguiendo el mismo ejemplo, si a un no enfermo se le somete a un tratamiento, se crea una situación que define la utilidad de tratarlo y que se puede expresar como la inversa de las complicaciones del tratamiento. Si la mortalidad asociada a dicho tratamiento es del 10%, la utilidad puede expresarse como supervivencia del 90%. Si a este mismo no enfermo (o no caso) no se le aplica el tratamiento, el resultado de utilidad suele ser, en las mismas unidades, supervivencia del 100%. Para esta situación el riesgo neto de tratamiento inapropiado ( $R_t$ ) es de 0,1.

$$R_t = U(T - |E -) - U(T + |E -)$$

#### **13.1.8.1.3. Riesgo neto de una prueba diagnóstica**

Son las consecuencias negativas que, como media, se derivan de complicaciones secundarias al uso del sistema diagnóstico [297].

Hace referencia a los efectos colaterales que pueda provocar. Al igual que los riesgos y beneficios del tratamiento, en su cálculo intervienen factores como la probabilidad de los posibles efectos colaterales ajustada a la utilidad que se le asigne a cada efecto colateral. A efectos de cálculo matemático, el riesgo neto de una prueba diagnóstica ( $R_d$ ) es la diferencia entre la utilidad del resultado esperado de aplicar una prueba diagnóstica ( $D$ ) y la utilidad del resultado esperado en pacientes a los que no se les aplica la prueba diagnóstica, que habitualmente puede considerarse nula.

$$R_d = U(D+) - U(D-) \cong U(D+)$$

### 13.1.8.2. Umbrales de acción para una prueba diagnóstica

Ante la disponibilidad de un sistema diagnóstico con suficiente validez como para aplicarlo a un caso concreto, el médico se encuentra ante tres posibles situaciones:

1) Descartar el diagnóstico sin aplicar al paciente ningún sistema diagnóstico por considerar que la probabilidad preprueba es lo suficientemente baja.

2) Diagnosticar el paciente sin llevar a cabo ninguna prueba diagnóstica por considerar que la probabilidad preprueba es lo suficientemente elevada.

3) Aplicar la prueba con la intención de que su resultado pueda contribuir a modificar el estado de diagnóstico del paciente por considerar que la probabilidad preprueba se encuentra en el ancho de valores de la zona de incertidumbre.

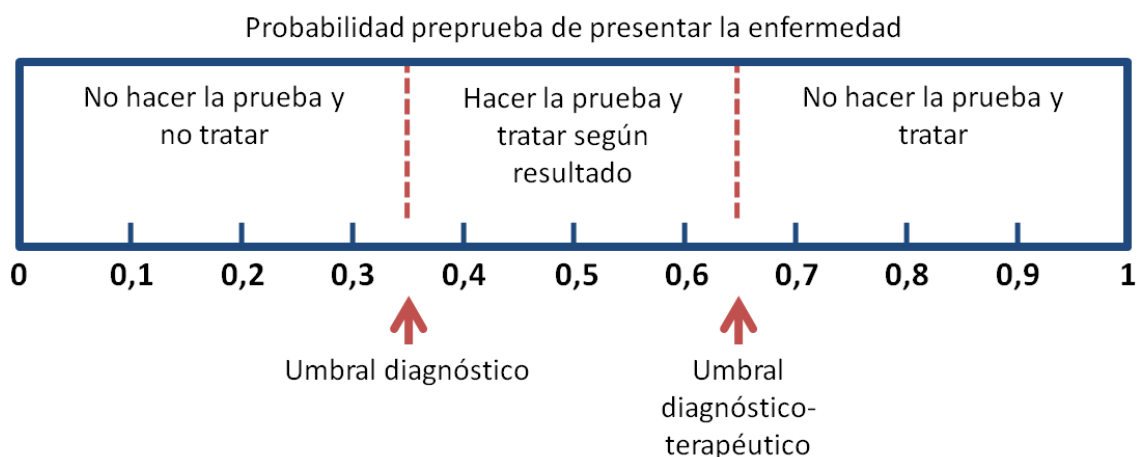


Figura 92: Análisis de umbrales de probabilidad y decisiones diagnósticas y terapéuticas acordes a la probabilidad preprueba de padecer una enfermedad concreta.

Estas tres situaciones vienen determinadas por dos umbrales para el valor de la probabilidad preprueba de la enfermedad que, como se ha mencionado previamente, son los umbrales de acción para una prueba diagnóstica.

· **El umbral diagnóstico.** Una probabilidad de estar enfermo inferior a este umbral obliga al clínico a no poder aceptar el diagnóstico y contraindica aplicar el sistema diagnóstico dado que su resultado no será capaz de modificar este criterio. Por consiguiente, no se justifica ninguna actitud terapéutica.

· **El umbral diagnóstico-terapéutico.** Una probabilidad de estar enfermo superior a este umbral obliga al clínico a aceptar el diagnóstico de enfermedad sin necesidad de aplicar el sistema diagnóstico dado que tampoco será capaz de modificar este criterio. Así, justifica el inicio de la actitud terapéutica específica incluso sin resultados de pruebas diagnósticas.

Con la probabilidad de enfermedad entre los dos valores umbrales anteriores, sí estará indicado realizar la prueba diagnóstica, ya que su resultado sí puede variar la probabilidad de enfermedad más allá de los puntos señalados por los umbrales diagnóstico y diagnóstico-terapéutico. Este escenario supone un nicho de oportunidad para un sistema diagnóstico, cuyos indicadores de validez (para una prevalencia conocida de la enfermedad de interés) finalmente señalarán la utilidad real [279,348,382,399].

Aunque existen varias propuestas para efectuar el cálculo de los umbrales de acción para una prueba diagnóstica, todas se basan en fórmulas matemáticas que incorporan la sensibilidad ( $S$ ), la especificidad ( $E$ ) o las razones de verosimilitud ( $RV$ ) del sistema diagnóstico. Además también requieren una estimación de la prevalencia de la enfermedad en la población diana y el cálculo del beneficio neto del tratamiento apropiado ( $B_t$ ), del riesgo neto de un tratamiento inapropiado ( $R_t$ ) y del riesgo neto de la prueba diagnóstica ( $R_d$ ) [279].

$$\text{Umbral diagnóstico} = \frac{(1 - E) \times R_t + R_d}{(1 - E) \times R_t + S \times B_t}$$

$$\text{Umbral diagnóstico/terapéutico} = \frac{E \times R_t - R_d}{E \times R_t + (1 - S) \times B_t}$$

Djulgovic y Desoky, a partir de las fórmulas del modelo de Pauker-Kassirer, desarrollaron una ecuación para aquellas pruebas diagnósticas en las que el riesgo neto de la prueba es despreciable, haciendo uso de la razón de verosimilitud. En el caso de la

evaluación de unos criterios de clasificación diagnóstica que funcionen como un sistema de puntos y para el que no supone riesgos disponer de toda la información necesaria para aplicarlo, es razonable asumir que el riesgo de la prueba se aproxima a 0 y, por lo tanto, se puede utilizar esta ecuación [275]:

$$\text{Umbral diagnóstico} = \frac{1}{RV(+)\times\frac{B_t}{R_t} + 1}$$

$$\text{Umbral diagnóstico/terapéutico} = \frac{1}{RV(-)\times\frac{B_t}{R_t} + 1}$$

### 13.1.8.3. Umbral terapéutico

Como se ha visto, el abanico de posibilidades de actuación ideal en un acto médico abarca las decisiones de no tratar, tratar o realizar más pruebas diagnósticas para añadir más información que ayude a decidir por alguna de las dos primeras opciones.

Sin embargo, es posible encontrarse con que no se dispone de más sistemas diagnósticos que contribuyan a decidir si establecer un tratamiento. En estos momentos, la teoría del análisis de decisiones establece que el clínico debe optar por tratar o no tratar en base a la probabilidad de enfermedad del paciente.

El umbral terapéutico es la probabilidad de estar enfermo por encima de la cual el paciente recibirá un tratamiento, y por debajo de la que no lo recibirá. Es referido a un tratamiento concreto, para una enfermedad concreta y en un paciente concreto (cuando se le hace al paciente partícipe en la toma de decisiones clínicas, a través de la asignación de utilidades y sus preferencias). Representa la probabilidad que iguala la utilidad de tratar ( $U_t$ ) y la utilidad de no tratar ( $U_{\bar{t}}$ ) [280,350].

La utilidad de tratar es la utilidad de tratar a un enfermo ajustada por la probabilidad de enfermedad ( $P$ ) sumada a la utilidad de tratar a un no enfermo ajustada por el inverso de la probabilidad de enfermedad.

$$U_t = P \times U(T + |E +) + (1 - P) \times U(T + |E -)$$

La utilidad de no tratar es la utilidad de no tratar a un enfermo ajustada por la probabilidad de enfermedad sumada a la utilidad de no tratar a un no enfermo ajustada por el inverso de la probabilidad de enfermedad.

$$U_{\bar{t}} = P \times U(T - |E +) + (1 - P) \times U(T - |E -)$$

Estos conceptos están ligados a los de beneficio neto de tratamiento apropiado ( $B_t$ ) y riesgo neto de tratamiento inapropiado ( $R_t$ ), expresados anteriormente. Así, para el umbral terapéutico se cumple que:

$$U_t = U_{\bar{t}}$$

$$P \times U(T + |E +) + (1 - P) \times U(T + |E -) = P \times U(T - |E +) + (1 - P) \times U(T - |E -)$$

$$P \times U(T + |E +) - P \times U(T - |E +) = (1 - P) \times U(T - |E -) - (1 - P) \times U(T + |E -)$$

$$P \times [U(T + |E +) - U(T - |E +)] = (1 - P) \times [U(T - |E -) - U(T + |E -)]$$

$$P \times B_t = (1 - P) \times R_t$$

Expresado según la fórmula anterior, el umbral terapéutico es la probabilidad de estar enfermo en la que se cumple que el beneficio esperado de tratar enfermos ajustado por la probabilidad de estar enfermo se iguala con el riesgo de tratar no enfermos ajustado por la probabilidad de no estar enfermo.

Dado que el umbral terapéutico se define como una probabilidad, su expresión se obtiene aislando  $P$  de la fórmula anterior [275,280,316,400,401]:

$$\text{Umbral terapéutico} = P = \frac{R_t}{R_t + B_t}$$

Las ecuaciones incluidas en la calculadora de la red CASP (*critical appraisal skills programme*) para análisis decisional incluyen esta fórmula [292].

Se comprueba que, si el riesgo de tratar sanos es pequeño y/o el beneficio de tratar enfermos es elevado, el umbral terapéutico es bajo, con lo que la decisión de tratar será más fácilmente tomada incluso a partir de resultados positivos de sistemas diagnósticos con un peso de la evidencia o una razón de verosimilitud moderada.

A la inversa, el umbral terapéutico será alto cuando en el riesgo de tratar sanos sea elevado y/o el beneficio de tratar enfermos sea bajo. En esta situación se requerirán resultados positivos de pruebas diagnósticas con alta capacidad de generar probabilidades postprueba elevadas.

De forma similar a la propuesta de Djulbegovic y Desoky para el cálculo de los umbrales de acción para pruebas diagnósticas, se puede obtener un umbral terapéutico si se establece que  $RV = 1$ , con lo que quedaría la formulación siguiente [275,402]:

$$\text{Umbral terapéutico} = \frac{1}{\frac{B_t}{R_t} + 1}$$

Desde este punto de vista, un sistema diagnóstico es útil si permite desplazar la probabilidad preprueba hasta la probabilidad postprueba a través del umbral terapéutico en cualquiera de los sentidos.

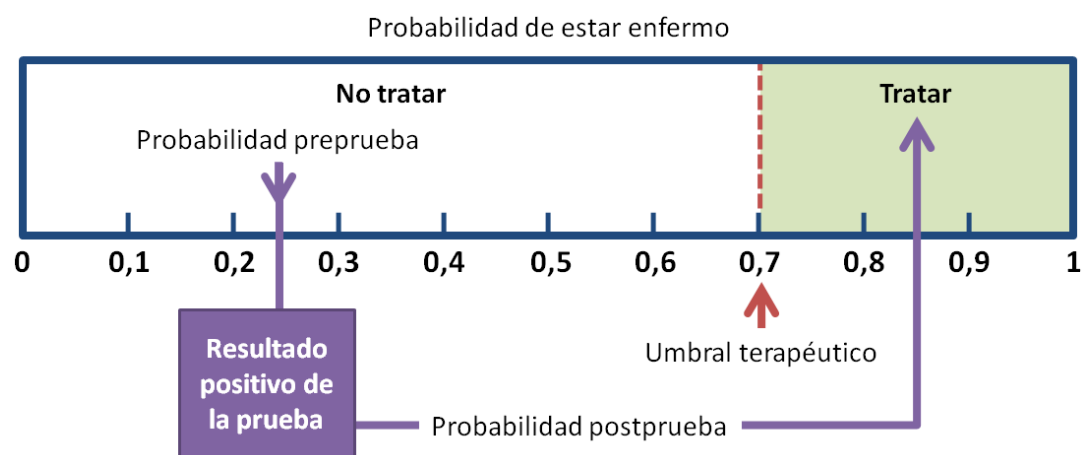


Figura 93: Prueba diagnóstica cuyo resultado se traduce en un cambio de actitud terapéutica a favor del tratamiento.

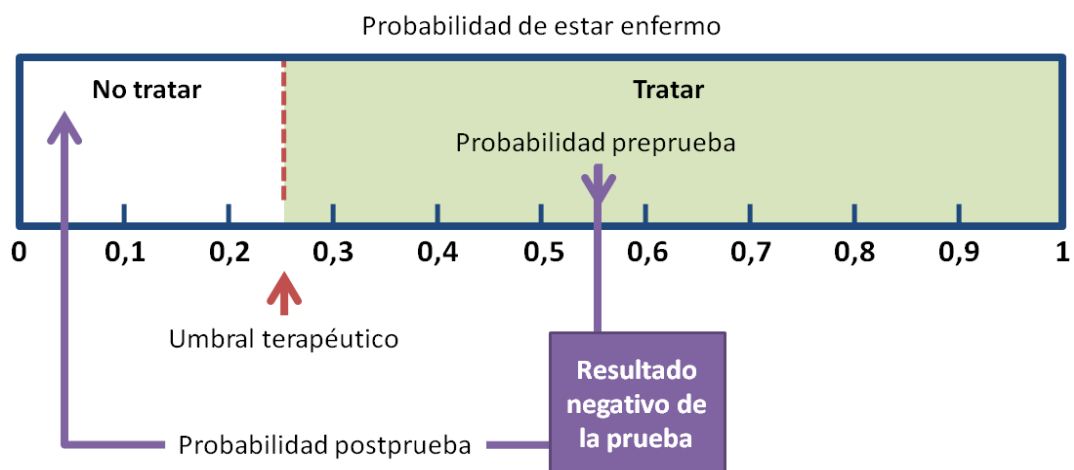


Figura 94: Prueba diagnóstica cuyo resultado se traduce en un cambio de actitud terapéutica en contra del tratamiento.



### 13.2. Criterios originales revisados para el diagnóstico de la HAI (1999)

Parámetro y discriminador	Puntuación
Sexo femenino	+2
Relación FA/AST (o ALT)	
<1,5	+2
1,5 – 3	0
>3	-2
Valor por encima de lo normal de inmunoglobulinas o IgG	
>2	+3
1,5 – 2	+2
1 – 1,5	+1
<1	0
ANA, anti-Sm o anti-LKM1	
>1:80	+3
1:80	+2
1:40	+1
<1:40 <sup>1</sup>	0
Anticuerpos anti-mitocondriales (AMA)	-4
Marcadores de hepatitis vírica	
Positivos <sup>2</sup>	-4
Negativos	+1
Ingesta media diaria de alcohol	
<25 g/día	+2
>60 g/día	-2
Histopatología hepática	
Hepatitis de interfase	+3
Infiltración con predominio linfoplasmocitario	+2
Formaciones de hepatocitos <i>en roseta</i>	+1
Nada de lo anterior	-5
Afectación biliar <sup>3</sup>	-3
Otros cambios que sugieran distinta etiología	-3
Otra(s) enfermedad(es) autoinmune(s) en el paciente o en familiar de primer grado	+2
Parámetros adicionales opcionales	
Otros anticuerpos (anti-SLA/LP, LC1, ASGPR, pANCA, anti-actina)	+2
HLA DR3 o DR4	+1
Respuesta al tratamiento	
Completa	+2
Con recaída	+3



<sup>1</sup>Títulos más bajos en niños se puede considerar positivos y se les debe asignar +1 punto.

<sup>2</sup>Se debe descartar infección aguda, o crónica activa si es posible por el tipo de virus, por VHA, VHB y VHC. Considerar también descartar infección por otros virus hepatotropos como VEB y CMV.

<sup>3</sup>Incluye colangitis granulomatosa, fibrosis concéntrica periductal, ductopenia, proliferación marginal de conductillos biliares y colangiolitias.

Significado de la puntuación total:

Pre-tratamiento

>15: HAI definitiva

10 – 15: HAI probable

Post-tratamiento

>17: HAI definitiva

12 – 17: HAI probable

Fuente: Alvarez F, Berg PA, Bianchi FB, Bianchi L, Burroughs AK, Cancado EL, et al. International Autoimmune Hepatitis Group Report: review of criteria for diagnosis of autoimmune hepatitis. J Hepatol. 1999;31:929-38.

### 13.3. Criterios simplificados para el diagnóstico de la HAI (2008)

Parámetro y discriminador	Puntuación
ANA o anti-Sm	
≥1:40	+1
≥1:80	+2
Anti-LKM1 ≥1:40	+2
Anti-SLA positivos <sup>1</sup>	+2
IgG	
Por encima del límite superior de la normalidad	+1
>1,1 veces por encima del límite superior de la normalidad	+2
Histopatología hepática	
Compatible con HAI <sup>2</sup>	+1
Típica de HAI <sup>3</sup>	+2
Ausencia de marcadores de hepatitis vírica	+2

<sup>1</sup>La suma máxima de puntos por presencia de autoanticuerpos es 2. La máxima total de todo el sistema, 8.

<sup>2</sup>Las características compatibles son aquellas de una hepatitis crónica con infiltración linfocítica en ausencia de los tres rasgos considerados típicos.

<sup>3</sup>Presencia simultánea de los tres rasgos siguientes: hepatitis de interfase, infiltración linfocítica o linfoplasmocitaria en tractos portales con extensión hacia el lóbulo y formaciones de hepatocitos en roseta.

Significado de la puntuación total:

≥7: HAI definitiva

6: HAI probable

Fuente: Hennes EM, Zeniya M, Czaja AJ, Pares A, Dalekos GN, Krawitt EL, et al. Simplified criteria for the diagnosis of autoimmune hepatitis. *Hepatology*. 2008;48:169-76.



### 13.4. Criterios diagnósticos de la HAI pediátrica propuestos por la ESPGHAN y la NASPGHAN (2009)

---

#### Parámetro y discriminador

---

Hipertransaminasemia

Presencia de autoanticuerpos:

ANA y/o anti-Sm a título  $\geq 1:20$

Anti-LKM1 a título  $\geq 1:10$

Anti-LC1

Anti-SLA

Hipergammaglobulinemia

Histopatología hepática:

Hepatitis de interfase

Colapso multilobular

Ausencia de marcadores de hepatitis vírica

Descarte de enfermedad de Wilson

Colangiograma (colangio-resonancia magnética o colangiografía retrógrada) normal

Fuente: Mieli-Vergani G, Heller S, Jara P, Vergani D, Chang M-H, Fujisawa T, et al. Autoimmune hepatitis. J Pediatr Gastroenterol Nutr. 2009;49:158–64.



### 13.5. Lista STARD (*Standards for Reporting of Diagnostic Accuracy*) para la comunicación de estudios de validez de pruebas diagnósticas

Sección	Nº	Ítem
<b>Título y palabras clave</b>	<b>1</b>	Identificar el artículo como un estudio de validez diagnóstica utilizando, en las palabras clave, al menos una medida de precisión según terminología MeSH (como sensibilidad, especificidad, valores predictivos o área bajo la curva)
<b>Resumen</b>	<b>2</b>	Resumen estructurado del estudio que incluya diseño general, métodos, resultados y conclusiones (existe una guía específica para los resúmenes en la página web de la red Equator)
<b>Introducción</b>	<b>3</b>	Fundamentos clínicos y científicos incluyendo el uso previsto y el papel clínico de la prueba a validar (prueba índice)
	<b>4</b>	Hipótesis de trabajo y objetivos del estudio
<b>Métodos</b>		
<i>Diseño del estudio</i>	<b>5</b>	Señalar si la obtención de los datos se planificó antes de que la prueba índice o la de referencia se realizaran (estudio prospectivo) o después (estudio retrospectivo)
<i>Participantes</i>	<b>6</b>	Describir los criterios de inclusión
	<b>7</b>	Describir el reclutamiento de los pacientes: si se basó en la presencia de síntomas, resultados de pruebas previas o en la realización a ellos de la prueba índice o la de referencia
	<b>8</b>	Señalar el ámbito y la cronología de la identificación de pacientes elegibles
<i>Métodos de la prueba</i>	<b>9</b>	Describir el muestreo de los pacientes: ¿fue una serie consecutiva de pacientes que cumplían los criterios de los ítems 6 y 7? En caso negativo, señalar si fue un muestreo aleatorio o de conveniencia
	<b>10a</b>	Describir la prueba índice en suficiente detalle como para permitir la replicación del estudio por otros investigadores
	<b>10b</b>	Describir la prueba de referencia en suficiente detalle como para permitir la replicación del estudio por otros investigadores
	<b>11</b>	Justificación de la elección de la prueba de referencia y señalar si existen alternativas
	<b>12a</b>	Definición y justificación de los puntos de corte establecidos para cada categoría de los resultados que arroja la prueba índice, distinguiendo los pre-especificados de los empleados a modo exploratorio
<b>12b</b>	Definición y justificación de los puntos de corte establecidos para cada categoría de los resultados que arroja la prueba de referencia, distinguiendo los pre-especificados de los empleados a modo exploratorio	

<i>Análisis y métodos estadísticos</i>	<b>13a</b>	Señalar si existe enmascaramiento de la información clínica del paciente y los resultados de la prueba de referencia a los lectores e intérpretes de la prueba índice
	<b>13b</b>	Señalar si existe enmascaramiento de la información clínica del paciente y los resultados de la prueba índice a los evaluadores del estándar de referencia
	<b>14</b>	Describir los métodos para estimar o comparar las medidas de validez diagnóstica
	<b>15</b>	Describir el tratamiento de los resultados indeterminados, tanto de la prueba índice como de la prueba de referencia
	<b>16</b>	Describir el tratamiento de los valores perdidos de la prueba índice y de la prueba de referencia
	<b>17</b>	Explicar si se llevó a cabo cualquier análisis de la variabilidad de la validez diagnóstica de las pruebas, especificando si se trata de análisis contemplados previamente en el diseño o de análisis exploratorios posteriores a la obtención de los datos
	<b>18</b>	Estimar el tamaño muestral necesario y señalar los criterios empleados para el cálculo
<b>RESULTADOS</b>		
<i>Participantes</i>	<b>19</b>	Describir mediante un diagrama el flujo de los participantes en el estudio
	<b>20</b>	Datos demográficos y clínicos basales de los participantes en el estudio
	<b>21a</b>	Perfil de gravedad de los pacientes incluidos con la enfermedad de interés
	<b>21b</b>	Distribución de los diagnósticos alternativos en la parte de la muestra sin la enfermedad de interés
	<b>22</b>	Especificar el intervalo de tiempo, y si se ha realizado cualquier tipo de intervención, entre la aplicación de la prueba índice y el estándar de referencia
<i>Resultados de la prueba</i>	<b>23</b>	Elaboración de una tabla de contingencia con los resultados de la prueba índice y los de la prueba de referencia (o una tabulación cruzada de ambas distribuciones)
	<b>24</b>	Aportar los resultados de los indicadores de validez con alguna medida de precisión como, por ejemplo, su intervalo de confianza al 95%
	<b>25</b>	Señalar cualquier evento adverso secundario al empleo de la prueba índice o la prueba de referencia
<b>DISCUSIÓN</b>	<b>26</b>	Señalar las limitaciones del estudio al respecto de fuentes potenciales de sesgos, falta de certeza estadística y de la capacidad de generalización o validez externa
	<b>27</b>	Propuesta de implicaciones para la práctica clínica de la prueba índice

<b>OTRA INFORMACIÓN</b>	<b>28</b>	Número y otros datos de registro del estudio
	<b>29</b>	Información de acceso al protocolo completo del estudio
	<b>30</b>	Señalar fuentes de financiación u otras ayudas y el papel de los mecenas

Fuente: Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, LijmerJG Moher D, Rennie D, de Vet HCW, Kressel HY, Rifai N, Golub RM, Altman DG, Hooft L, Korevaar DA, Cohen JF, For the STARD Group. STARD 2015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies. BMJ. 2015;351:h5.





### **13.6. Fundamentos de la revisión sistemática y el meta-análisis de estudios de sistemas diagnósticos**

---

Se conoce por revisión sistemática de la literatura a la identificación, selección y evaluación crítica de estudios primarios relevantes sobre una cuestión claramente planteada. El objetivo de la sistematización es reducir los posibles sesgos que pudieran ocurrir en una revisión no sistemática, tanto los observados en el método de selección de los artículos como aquellos detectados tras la evaluación crítica de cada estudio [403]. Las revisiones sistemáticas incluyen a los meta-análisis, que son el conjunto de métodos estadísticos que integran los resultados de los estudios recopilados para aumentar el poder estadístico de la investigación primaria [404].

La aplicación de metodología estadística a las pruebas diagnósticas se ha desarrollado con posterioridad a la de estudios de intervención o experimentales. Existen diferencias importantes entre el meta-análisis de estudios de intervención y el meta-análisis de sistemas diagnósticos, más recientes y menos estandarizados en comparación con los otros. Principalmente, los meta-análisis de estudios que comparan intervenciones generalmente integran estudios aleatorizados, con dos grupos semejantes, que evalúan la misma intervención, en general comparada con placebo o con tratamiento convencional. Por el contrario, los meta-análisis de estudios de pruebas diagnósticas enfrentan retos distintos, como puntos de corte diferentes para el resultado positivo o negativo de un examen o evaluación de pruebas que han sido realizadas en estudios prospectivos para estudio de intervenciones terapéuticas [405]. El uso de meta-análisis para exámenes diagnósticos está todavía en fase de desarrollo, pero viene ganando cada vez más importancia desde la década de 1990, cuando surgieron nuevas técnicas estadísticas de combinación de estudios de sistemas diagnósticos y se publicó la primera guía de recomendaciones al respecto [406–409].

En este anexo se describirán, con más detalle de lo que permite la exposición de la metodología y basado en las técnicas de las referencias anteriores, los pasos a seguir para la revisión sistemática y el meta-análisis de estudios de pruebas diagnósticas.

### **13.6.1. Definición de la pregunta diagnóstica de interés**

Consiste en especificar claramente la prueba diagnóstica en cuestión, la enfermedad en estudio, cómo se realiza el diagnóstico y con qué fin se planteó la pregunta a resolver a través de la revisión sistemática y el meta-análisis. De modo general, el sistema diagnóstico bajo estudio se compara a un patrón de referencia para el diagnóstico de la enfermedad. Se acepta, sin embargo, que los métodos estadísticos utilizados para el meta-análisis de exámenes diagnósticos pueden tener una aplicación más bien amplia. Además de ello, se debe también aclarar si se realizará una comparación de pruebas diagnósticas [410].

### **13.6.2. Búsqueda en diversas fuentes de todos los estudios confiables que tratan del tema**

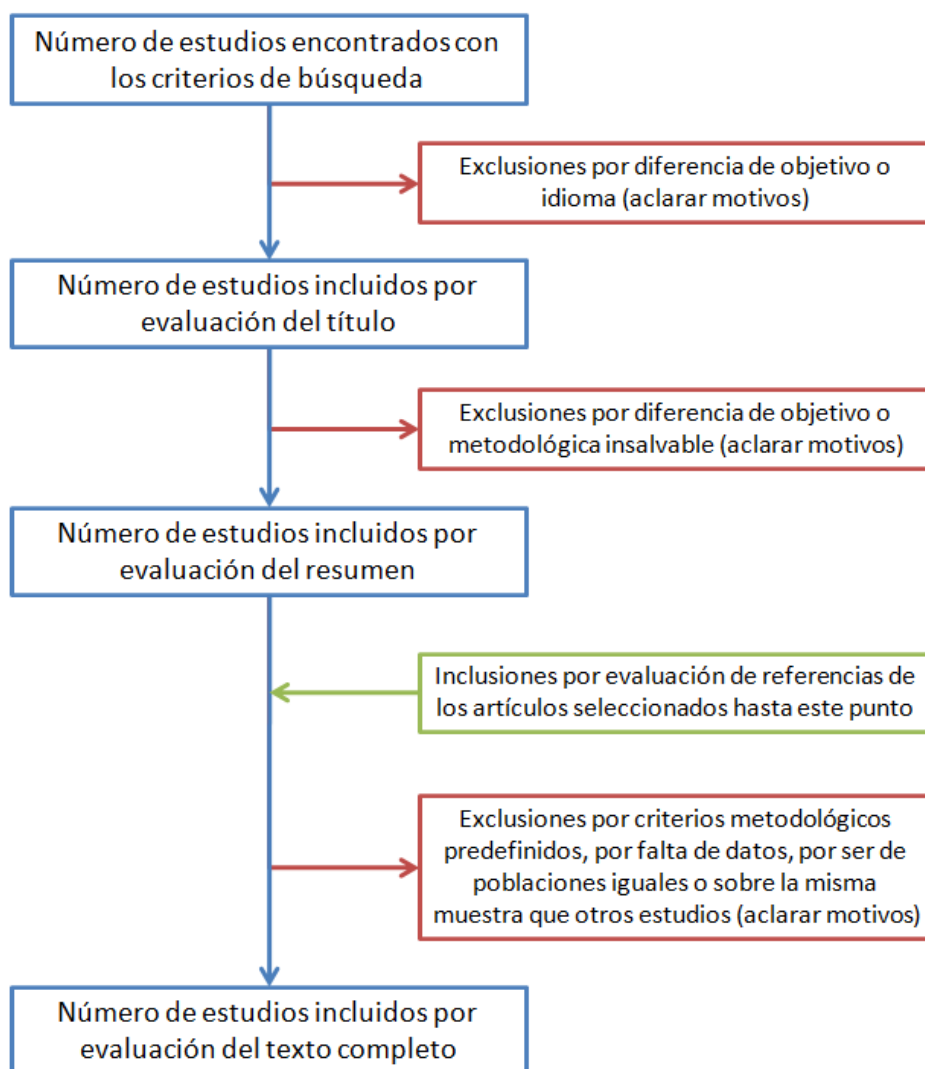
Se recomienda ampliar al máximo las fuentes de búsqueda, de modo que se incluya tanto literatura indexada en los principales directorios médicos (MEDLINE, EMBASE, LILACS) como literatura gris (publicaciones gubernamentales, comisiones de ética, resúmenes en anales de congresos, tesis). Además, es importante consultar la biblioteca de revisiones Cochrane para verificar si dicha revisión ya está hecha [411].

Aunque la intención no sea de utilizar datos no publicados, el contacto con investigadores de estudios en marcha o no publicados puede ser importante [412].

Para la averiguación en la base de datos MEDLINE, vale especificar claramente el procedimiento de búsqueda con términos MeSH (*medical subject*

*headings*), tomando en consideración criterios de inclusión y exclusión [406]. La forma de investigar con términos de búsqueda puede interferir en la sensibilidad de la revisión sistemática [413,414]. Se obtiene la mejor estrategia, en general, mediante la combinación de los términos MeSH utilizados con palabras textuales [414].

Es importante señalar cómo fue el proceso de revisión de la literatura, idealmente con ayuda de un diagrama de flujo que describa los pasos de búsqueda y selección de artículos.



**Figura 95:** Esquema del proceso de búsqueda y selección de artículos para una revisión sistemática de estudios de sistemas diagnósticos.

El sesgo de publicación es la distorsión de los resultados de un una revisión sistemática debida a la tendencia de publicar más frecuentemente los trabajos con resultados positivos sobre los de resultados negativos. Esto es tanto más probable cuanto menor sea la muestra, motivo por el que algunos autores preconizan la exclusión de estudios con un tamaño muestral reducido [415]. Para minimizar la posibilidad de sesgo de publicación, se debe ampliar al máximo las fuentes de búsqueda. Se puede evaluar la presencia de sesgo de publicación a través de gráficos de dispersión en embudo (*funnel plot*), que relacionen las *odds ratio* diagnósticas de cada estudio primario con su tamaño muestral. El nombre del gráfico viene del hecho de que, en ausencia de sesgo de publicación, a mayor tamaño muestral de los estudios habrá una menor dispersión de las *odds ratio* diagnósticas alrededor del valor real en la población de origen por el principio estadístico de regresión a la media. La apariencia asimétrica sugiere la presencia de dicho sesgo [416].

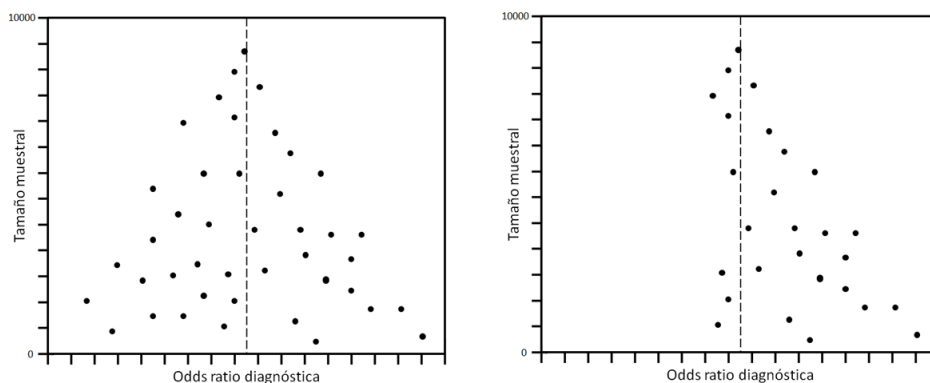


Figura 96: Gráfico de dispersión. Cada punto representa un estudio. A la izquierda, imagen simétrica que apunta a ausencia de sesgo de publicación. A la derecha, imagen asimétrica que señala que existe un sesgo en el sentido de infrapublicación de estudios con *odds ratio* diagnósticas pobres.

### 13.6.3. Selección de estudios por medio de criterios claros de inclusión y de exclusión

Idealmente, dos investigadores deben buscar y evaluar los estudios de forma independiente. Se puede utilizar la prueba estadística *kappa* para estudiar la

concordancia entre los dos investigadores [405]. Se debe de explicar también cómo se solucionan las discordancias entre ellos, lo que en general se hace mediante un acuerdo y en base a la opinión de un tercer investigador experimentado. Posteriormente se debe confeccionar una lista con las características de cada estudio primario, así como sus resultados.

Este paso se facilita si el estudio ha sido publicado de acuerdo con la estandarización STARD, formulada para garantizar más claridad, rigor metodológico y posibilidad de comparación de los estudios de sistemas diagnósticos [385]. Se debe evaluar también la calidad en base a los principios de la lista QUADAS [250]. El contenido de estos estándares se expone en sendos materiales anexos.

#### 13.6.4. Extracción y presentación de los datos de cada estudio

Acompañar la revisión de tablas de comparación de las diferencias clínicas y metodológicas de los estudios es una estrategia muy útil de presentación de los resultados. También es interesante evaluar la distribución por edad, sexo, forma de selección de pacientes, covariables relevantes, tiempo de seguimiento y tamaño de la muestra [405,417]. Para la obtención de los datos a combinar (a meta-analizar) hay que recopilar los valores originales de falsos y verdaderos positivos y de falsos y verdaderos negativos. Eventualmente, esos datos pueden estimarse a partir de valores de sensibilidad, especificidad y de los valores de ocurrencia del desenlace o prueba de referencia. Igual que en los estudios primarios, se pueden presentar en forma de tabla de contingencia.

Tabla 59: Nomenclatura para los elementos de cada estudio individual sobre un sistema diagnóstico.

Criterio de verdad →	Enfermedad	Ausencia de enfermedad	Total
<b>Prueba +</b>	Verdadero positivo (a)	Falso positivo (b)	P
<b>Prueba –</b>	Falso negativo (c)	Verdadero negativo (d)	N
<b>Total</b>	D	ND	T

### 13.6.5. Evaluación de la homogeneidad entre los estudios primarios

El grado de variabilidad entre los resultados de los estudios puede evaluarse gráficamente presentando la sensibilidad y especificidad de cada estudio en un *forest plot*. Alguna dispersión debería aparecer por el azar en la selección de las muestras en los estudios, pero otros factores la pueden aumentar. Con respecto a los meta-análisis sobre tratamiento, en los de validez de pruebas diagnósticas aparece una fuente extra de variabilidad entre estudios: los trabajos incluidos pueden haber usado, explícitamente o no, diferentes umbrales para definir los resultados positivo y negativo de la prueba. Para explorar esta fuente de variación es útil una gráfica de las sensibilidades y especificidades en el plano COR. Si existiera efecto umbral (que es como se denomina este fenómeno) los puntos mostrarían un patrón curvilíneo [418]. Este efecto umbral también puede evaluarse mediante el coeficiente de correlación de Spearman, ya que si el efecto existe aparece una correlación inversa entre sensibilidad y especificidad [419]. En este caso, combinar los resultados de los estudios exige, en lugar de promediar las sensibilidades y especificidades o las razones de verosimilitud, ajustar los puntos a una curva COR.

Además de gráficamente, la homogeneidad de las sensibilidades y especificidades puede contrastarse usando la prueba de la razón de verosimilitud [420]. Haciendo uso de la notación de la tabla anterior, e indicando con el subíndice  $i$ , un estudio individual, y con el subíndice  $T$ , el global de estudio:

$$G_{Se}^2 = 2 \times \sum_i \left( a_i \times \ln \frac{a_i}{\frac{a_T \times D_i}{D_T}} + c_i \times \ln \frac{c_i}{\frac{c_T \times D_i}{D_T}} \right)$$

$$G_{Sp}^2 = 2 \times \sum_i \left( d_i \times \ln \frac{d_i}{\frac{d_T \times ND_i}{ND_T}} + b_i \times \ln \frac{b_i}{\frac{b_T \times ND_i}{ND_T}} \right)$$

Donde  $a_T = \sum_i a_i$ ;  $c_T = \sum_i c_i$ ;  $D_T = \sum_i D_i$ ;  $b_T = \sum_i b_i$ ;  $d_T = \sum_i d_i$  y  $ND_T = \sum_i ND_i$ .

En la hipótesis de homogeneidad ambos se distribuyen asintóticamente como una  $\chi^2$  con  $k-1$  grados de libertad (siendo  $k$  el número de estudios).

La homogeneidad de las razones de verosimilitud y de las *odds ratio* diagnósticas se contrasta con la prueba  $Q$  de Cochran usando como pesos los inversos de las varianzas [421]. El estadístico  $Q$  también tiene una distribución  $\chi^2$  con  $k-1$  grados de libertad.

$$Q = \sum_i \frac{1}{EE(\ln\theta_i)^2} \times (\ln\theta_i - \ln\theta_T)^2$$

Siendo  $\theta$  la razón de verosimilitud positiva o negativa o la *odds ratio* diagnóstica, y  $EE$  el error estándar.

Otra medida de heterogeneidad que se puede obtener desde ese valor  $Q$  es el estadístico  $I^2$ , que se conoce como medida de inconsistencia y se expresa en tanto por cien, obtenida mediante la fórmula:

$$I^2 = \frac{[Q - (k - 1)]}{Q} \times 100$$

La medida de inconsistencia  $I^2$  describe el porcentaje de variabilidad del efecto que es consecuencia de la heterogeneidad y no del azar. Cuando  $I^2$  presenta un valor superior a un 50%, se considera que hay heterogeneidad substancial [422,423].

Los métodos de combinación, que se describirán a continuación, calculan los promedios ponderados de los resultados de los estudios. Dichos métodos se dividen usualmente en dos categorías: métodos con efectos fijos y métodos con efectos aleatorios. En la combinación que utiliza métodos con efectos fijos, se atribuye un valor a cada estudio que es el inverso de la varianza ( $1/v$ ) del estudio. Métodos de combinación con efectos aleatorios atribuyen un valor a cada estudio que es el inverso de la varianza sumada a la heterogeneidad ( $1/v + h$ ). Si el contraste de la hipótesis de homogeneidad da un resultado estadísticamente no significativo (por convención, valor  $p > 0,05$ ) se permite el empleo de un modelo de efectos fijos. De modo simplificado, es como si los métodos con efectos fijos consideraran que la



variabilidad entre los estudios ocurrió solo al azar e ignoraran la heterogeneidad entre ellos [415]. Por su parte, los modelos basados en métodos de efectos aleatorios incorporan parcialmente la heterogeneidad entre los estudios en los resultados. De esa manera, se generan resultados combinados con mayor intervalo de confianza. A pesar de tener esa ventaja y ser más recomendados, los métodos con efectos aleatorios son criticados por atribuir un mayor valor a estudios menores [415].

Existen otras fuentes de heterogeneidad más allá de la de los indicadores de validez, que tiene que ver con el diseño de los estudios primarios. En efecto, agrupar resultados de estudios que adolecen de sesgos diversos en un único indicador resumen no siempre es apropiado. Otro problema consiste en la combinación de estudios con diseño diferente, como por ejemplo el caso de mezclar estudios transversales o de cohortes con estudios de caso-control. Para esta situación particular se ha publicado recientemente una solución basada en un enfoque híbrido que modeliza conjuntamente la prevalencia de la enfermedad junto con la sensibilidad y la especificidad de la prueba diagnóstica en los estudios de cohortes (o transversales), y la sensibilidad y especificidad en los estudios de casos y controles. En esencia es un procedimiento matemáticamente complejo de inferencia basado en las razones de verosimilitud combinadas. Para una explicación más detallada se puede consultar el artículo de Chen et al de 2015 y su suplemento [254].

### **13.6.6. Métodos de combinación de resultados de estudios sobre sistemas diagnósticos**

#### **13.6.6.1. Combinación de sensibilidades y especificidades**

La sensibilidad ( $Se$ ) y la especificidad ( $Sp$ ) globales son:

$$Se_T = \frac{\sum_i a_i}{\sum_i D_i} \qquad Sp_T = \frac{\sum_i d_i}{\sum_i ND_i}$$

Estas fórmulas equivalen a las de promedios ponderados, en los que el peso de cada estudio es proporcional a su tamaño muestral.

Los intervalos de confianza de la sensibilidad y la especificidad global se pueden calcular usando el método exacto para las proporciones binomiales basado en la distribución F de Snedecor [424]. Opcionalmente también se pueden calcular los intervalos corregidos por sobredispersión, a través de una aproximación normal a la binomial, que es el método que utiliza el software de distribución libre Meta-DiSc, compilado por la Unidad de Bioestadística clínica del Hospital Ramón y Cajal [418]. De este modo, los errores estándar de estos indicadores combinados son:

$$EE(Se_T) = \sqrt{\frac{Se_T \times (1 - Se_T)}{\sum_i D_i}} \quad EE(Sp_T) = \sqrt{\frac{Sp_T \times (1 - Sp_T)}{\sum_i ND_i}}$$

Y los intervalos corregidos por sobredispersión:

$$Se_T \pm z_{\alpha/2\varphi_{Se}} \times EE(Se_T) \quad Sp_T \pm z_{\alpha/2\varphi_{Sp}} \times EE(Sp_T)$$

Siendo los factores de corrección:

$$\varphi_{Se} = \sqrt{\frac{\chi_{Se}^2}{k-1}}, \quad \text{con} \quad \chi_{Se}^2 = \sum_i \left[ \frac{\left(a_i - \frac{a_T \times D_i}{D_T}\right)^2}{\frac{a_T \times D_i}{D_T}} + \frac{\left(c_i - \frac{c_T \times D_i}{D_T}\right)^2}{\frac{c_T \times D_i}{D_T}} \right]$$

$$\varphi_{Sp} = \sqrt{\frac{\chi_{Sp}^2}{k-1}}, \quad \text{con} \quad \chi_{Sp}^2 = \sum_i \left[ \frac{\left(d_i - \frac{d_T \times ND_i}{ND_T}\right)^2}{\frac{d_T \times ND_i}{ND_T}} + \frac{\left(b_i - \frac{b_T \times ND_i}{ND_T}\right)^2}{\frac{b_T \times ND_i}{ND_T}} \right]$$

### 13.6.6.2. Combinación de razones de verosimilitud

Las razones de verosimilitud pueden ser agrupadas por el método de Mantel-Haenszel (modelo de efectos fijos) o, para incorporar la variación entre estudios, por el método de DerSimonian-Laird (modelo de efectos aleatorios). Ambos métodos calculan promedios ponderados, y la diferencia está justamente en los pesos usados y en qué se promedia. Con el método de Mantel-Haenszel (*MH*) se promedia directamente la razón de verosimilitud, mientras que en el de DerSimonian-Laird (*DL*) se promedia su logaritmo [421].

$$\theta_T^{MH} = \frac{\sum_i w_i^{MH} \times \theta_i}{\sum_i w_i^{MH}}$$

$$\ln \theta_T^{DL} = \frac{\sum_i w_i^{DL} \times \ln \theta_i}{\sum_i w_i^{DL}}$$

Los pesos de Mantel-Haenszel son:

Para la razón de verosimilitud positiva,

$$w_i^{MH} = \frac{b_i \times D_i}{T_i}$$

Para la razón de verosimilitud negativa,

$$w_i^{MH} = \frac{d_i \times D_i}{T_i}$$

Por su parte, el peso de DerSimonian-Laird es:

$$w_i^{DL} = \frac{1}{EE(\ln \theta_i)^2 + \tau^2}$$

Siendo, si  $Q > k - 1$ ,

$$\tau^2 = \frac{Q - (k - 1)}{\sum_i w_i - \left( \frac{\sum_i w_i^2}{\sum_i w_i} \right)}$$

O, si  $Q < k - 1$ ,  $\tau^2 = 0$ . Donde  $Q$  es el estadístico de homogeneidad de Cochran para el estimador global de Mantel-Haenszel y  $w_i$  los pesos del inverso de la varianza.

La razón de verosimilitud combinada tiene la ventaja de poder analizar sistemas diagnósticos cuyo resultado es una variable continua o con muchas categorías (tipo criterios de puntuación), evitándose pérdidas de información al dicotomizarse la variable. Otra ventaja es que la *odds* postprueba de la enfermedad, una vez que se sabe que el resultado de la prueba es positivo, puede calcularse por medio de la fórmula: *odds* postprueba = *odds* preprueba x razón de verosimilitud [406].

Dado que la relación entre *odds* y probabilidad viene dada por la siguiente expresión:

$$Probabilidad = \frac{Odds}{1 + Odds}$$

La probabilidad postprueba se calcularía como el cociente entre la *odds* postprueba y  $1 + odds$  postprueba.

La distribución de los logaritmos de las razones de verosimilitud combinadas o globales, estimados por el método de Mantel-Haenszel, es aproximadamente normal con los siguientes errores estándar:

$$EE(\ln RV_+^{MH}) = \sqrt{\frac{\sum \frac{[D_i \times ND_i \times (a_i + b_i) - a_i \times b_i \times T_i]}{T_i^2}}{\sum \frac{a_i \times ND_i}{T_i} \times \sum \frac{c_i \times D_i}{T_i}}}$$

$$EE(\ln RV_-^{MH}) = \sqrt{\frac{\sum \frac{[D_i \times ND_i \times (a_i + b_i) - a_i \times b_i \times T_i]}{T_i^2}}{\sum \frac{b_i \times ND_i}{T_i} \times \sum \frac{d_i \times D_i}{T_i}}}$$

Por lo tanto, el intervalo de confianza sigue la expresión clásica del método exacto para proporciones binomiales:

$$RV \times e^{\pm z_{\alpha/2} EE(\ln RV^{MH})}$$

La distribución de los logaritmos de las razones de verosimilitud globales estimados por el método de DerSimonian-Laird es también aproximadamente normal con errores estándar dados por:

$$EE(\ln RV^{DL}) = \frac{1}{\sqrt{\sum w_i^{DL}}}$$

Por ello, el intervalo de confianza también es:

$$RV \times e^{\pm z_{\alpha/2} EE(\ln RV^{DL})}$$

### 13.6.6.3. Combinación de odds ratio diagnósticas

Las *odds ratio* diagnósticas (*ORD*) también pueden ser agrupadas siguiendo los mismos métodos que para las razones de verosimilitud. La única diferencia es que los pesos en el método de Mantel-Haenszel son:

$$w_i^{MH} = \frac{b_i \times c_i}{T_i}$$

El resto de las expresiones anteriores son equivalentes para  $\theta_i = \text{odds ratio}$  diagnóstica.

Análogamente, el intervalo de confianza de la *odds ratio* diagnóstica combinada se estima igual que el de las razones de verosimilitud combinadas. Aquí también, la única salvedad es que, para el cálculo por el método de Mantel-Haenszel, el error estándar tiene la siguiente formulación:

$$EE(\ln ORD) = \sqrt{\frac{1}{2} \times \left[ \frac{\sum \frac{(a_i+d_i)a_i d_i}{T_i^2}}{\left(\sum \frac{a_i \times d_i}{T_i}\right)^2} + \frac{\sum \frac{(a_i+d_i)b_i c_i}{T_i^2} + \sum \frac{(b_i+c_i)a_i d_i}{T_i^2}}{\sum \frac{a_i \times d_i}{T_i} \times \sum \frac{b_i \times c_i}{T_i}} + \frac{\sum \frac{(b_i+c_i)b_i c_i}{T_i^2}}{\left(\sum \frac{b_i \times c_i}{T_i}\right)^2} \right]}$$

Aunque la *odds ratio* diagnóstica es difícil de aplicar clínicamente, resulta muy útil por diversos motivos:

- 1) Es una medida estadística de desempeño global de la prueba.
- 2) Se la puede obtener fácilmente mediante el producto cruzado de la tabla de contingencia 2x2.
- 3) Es frecuentemente constante independientemente del punto de corte utilizado en los diversos estudios sobre sistemas diagnósticos con resultados cuantitativos.
- 4) Es útil en la construcción del intervalo de confianza de la curva COR resumen.

#### 13.6.6.4. Scores de efectividad diagnóstica

El *score* de efectividad diagnóstica cuantifica el grado de superposición de resultados entre enfermos y no enfermos. Se puede interpretar como el número de desviaciones estándar, al separarse el promedio entre las dos curvas de distribución (enfermos y no enfermos) de resultados que se comportan como variables continuas. Se puede obtener por medio de una fórmula propia o a partir de la *odds ratio* diagnóstica [425,426]. Es la medida de la distancia estandarizada entre los

promedios de dos poblaciones, también denominada medida del tamaño del efecto o medida de efectividad, que se puede evaluar por medio de modelos de efectos fijos o aleatorios. Es una medida cuantitativa que se puede usar para comparar métodos diagnósticos o para resumir resultados de estudios en meta-análisis. Para más detalles de su obtención, se puede consultar el trabajo de Hasselblad y Hedges, que revisa el método [426]. Así como la curva COR resumen, el *score* de efectividad suministra una descripción de la separación de dos distribuciones de resultados de exámenes (entre enfermos y no enfermos), independientemente del modo de distribución de los resultados.

### **13.6.6.5. Curva resumen de características operativas del receptor**

Si el umbral diagnóstico, el punto de corte, varía entre los estudios de la revisión sistemática, el mejor resumen de los resultados es una curva COR, en lugar de solo un índice global. La forma de la curva depende de las distribuciones de probabilidad subyacentes de los resultados de la prueba diagnóstica en las personas con y sin la enfermedad [427].

Hay dos maneras de ajustar la curva COR. Las pruebas diagnósticas en las que la ORD es constante, independiente del umbral, tienen curvas COR simétricas alrededor de la línea “Se=Sp”. En este caso es posible combinar las ORD por los métodos de Mantel-Haenszel o DerSimonian-Laird, y a partir de la ORD global o combinada calcular la mejor curva COR: la ecuación de la curva está dada por [409]:

$$Se = \frac{1}{1 + \frac{1}{ORD_T \times \frac{1-Sp}{Sp}}}$$

Pero cuando las ORD cambian con el umbral, la curva COR es asimétrica. Para estudiar la variación de la ORD combinado con el umbral y ajustar, en función de ello, una curva simétrica o asimétrica se usa el método de Moses-Shapiro-Littenberg, que consiste en estudiar esta relación ajustando los datos de los estudios a la recta [409]:

$$\ln ORD_T = a + b \times \ln \left( \frac{Se}{1 - Se} \times \frac{1 - Sp}{Sp} \right)$$

El *software* Meta-DiSc lleva a cabo la estimación de los parámetros a y b, y sus errores estándar y covarianza, por el método de los mínimos cuadrados, ordinario o ponderado usando la librería NAG C. Los pesos de ponderación pueden ser la inversa de la varianza del logaritmo de la ORD o el tamaño muestral [428].

El contraste sobre si hay variación del rendimiento diagnóstico (medido por la ORD) con el umbral es equivalente al realizado sobre el parámetro b. Si  $b = 0$  no hay variación y el método da lugar a una curva COR simétrica y  $e^a$  es la estimación de la ORD combinada o global. Sin embargo, si  $b \neq 0$ , existe variación y la curva COR es asimétrica, formulada según la expresión [429]:

$$Se = \frac{1}{1 + \frac{1}{e^{\frac{a}{1-b} \times \left(\frac{1-Sp}{Sp}\right)^{\frac{1+b}{1-b}}}}}$$

Un estadístico útil cuando se agrupan estudios mediante la curva COR es el área bajo la curva, que resume el rendimiento diagnóstico en un solo número [351]. Las pruebas perfectas tienen un área bajo la curva cercano a 1, y las inútiles, cercano a 0,5. Se calcula integrando la ecuación de la curva numéricamente por el método trapezoidal.

El error estándar del área bajo la curva COR, si la curva es simétrica, viene dado por la fórmula [430]:

$$EE(ABC_{sim}) = \frac{ORD_T}{(ORD_T - 1)^3} [(ORD_T + 1) \ln ORD_T - 2(ORD_T - 1)] EE(\ln ORD_T)$$

Si, por el contrario, la curva es asimétrica, el error estándar es:

$$EE(ABC_{asim}) = \sqrt{A^2 var(a) + B^2 var(b) + 2ABcov(a, b)}$$

Donde A y B representan:

$$A = \left( \frac{1}{1 - b} \right) e^{\left(\frac{a}{1-b}\right)} \int_0^1 \frac{\left(\frac{x}{1-x}\right)^p}{\left[1 + \left(\frac{x}{1-x}\right)^p e^{\left(\frac{a}{1-b}\right)}\right]^2} dx$$

$$B = \left(\frac{1}{1-b}\right)^2 e^{\left(\frac{a}{1-b}\right)} \int_0^1 \frac{\left(\frac{x}{1-x}\right)^p \left[a + 2 \ln\left(\frac{x}{1-x}\right)\right]}{\left[1 + \left(\frac{x}{1-x}\right)^p e^{\left(\frac{a}{1-b}\right)}\right]^2} dx$$

Siendo  $p = (1 + b)/(1 - b)$ .

Cuando se restringe el rango de la curva COR al cuadrante superior izquierdo, el error estándar del área bajo la curva se calcula con la fórmula de la curva asimétrica substituyendo adecuadamente los límites de integración. De hecho, el programa Meta-DiSc, si conviene emplear restricciones porque el ancho de los intervalos de confianza incluye el 1, presenta el error estándar bajo la curva solo si la  $ORD_T$  se calcula a partir del modelo de Moses.

Introduciendo en la ecuación de la curva COR simétrica, en lugar de la ORD global, los límites superior e inferior de su intervalo de confianza, se obtienen un intervalo de confianza para la curva. En el caso de la curva asimétrica obtenida a partir del modelo de Moses–Shapiro-Littenberg, Mitchell sugiere construir un intervalo de confianza para la curva aplicando la transformación inversa a la banda de confianza del modelo lineal [431]. La transformación inversa está dada por:

$$Se = \frac{1}{1 + e^{-\frac{D+S}{2}}}$$

$$Sp = \frac{1}{1 + e^{-\frac{D-S}{2}}}$$

Siendo D y S

$$D = \ln ORD_T$$

$$S = \ln \left( \frac{Se}{1 - Se} \times \frac{1 - Sp}{Sp} \right)$$

Otro estadístico útil es el índice  $Q^*$  (no confundir con el  $Q$  de Cochran para el estudio de la homogeneidad entre indicadores de los estudios primarios), que se define como el punto en el que la sensibilidad y la especificidad son iguales, que es el punto de la curva más cercano al ideal extremo superior del plano COR [430]:

$$Q^* = \frac{ORD_T}{1 + \sqrt{ORD_T}}$$



El error estándar de  $Q^*$  es

$$EE(Q^*) = \frac{\sqrt{ORD_T}}{2(1 + \sqrt{ORD_T})^2} \times EE(\ln ORD_T)$$

Como alternativa a la curva COR resumen para evaluar globalmente el sistema diagnóstico, se sugiere la medida del índice  $Q^*$ , que no cambia de acuerdo con la heterogeneidad y presenta bastante robustez [430]. Admite valores entre 0,5 y 1,0 (cuanto mayor, mejor) y se considera una medida global de la eficacia de la prueba. Si se evalúan diez estudios como mínimo, la distribución de  $Q^*$  es gaussiana. Se puede utilizar el valor de  $Q^*$  para comparar métodos o verificar sesgos, separándose los estudios con problemas metodológicos en subgrupos y comparando su valor con el valor de los otros subgrupos de estudios [425]. El error estándar del área bajo la curva COR y el del índice  $Q^*$  son próximos numéricamente. Cuando el intervalo de confianza del valor de  $Q^*$  o del área bajo la curva COR pasa por 0,5, el examen no presenta desempeño diagnóstico significativo y, por lo tanto, no contribuye para el estudio de la enfermedad.

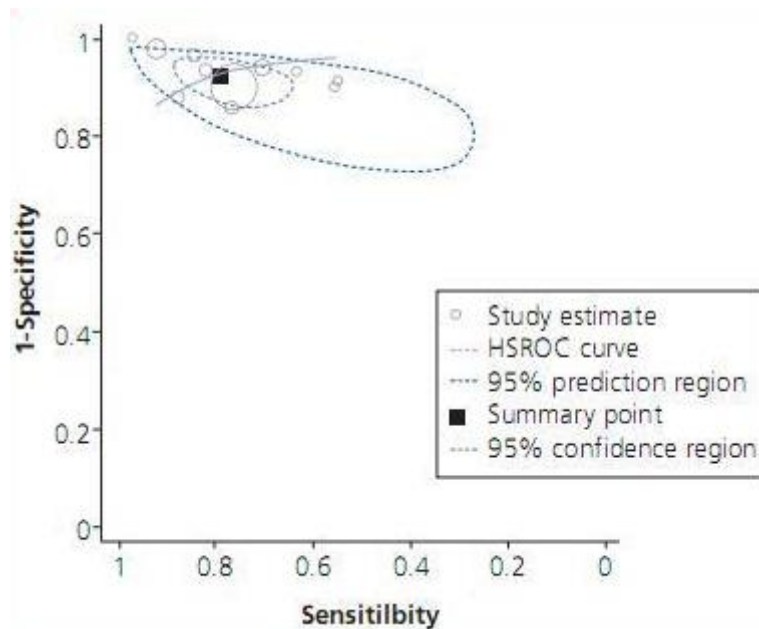


Figura 97: Ejemplo de plano de la curva resumen de características operativas del receptor. Incluye las regiones de confianza de la curva y el par sensibilidad/especificidad resumen. Reproducido de Zamora Romero et al. *Nefrología*. 2009;29;18. Con permiso de Elsevier.

### **13.6.7. Comentarios sobre el formato de presentación del meta-análisis**

Según la conferencia QUORUM (*Quality of Reporting of Metaanalysis*) la correcta publicación de un meta-análisis de estudios diagnósticos, se basa en una descripción detallada de la metodología, explicitando cada etapa del proceso [417].

El título debe identificar el trabajo como meta-análisis o como revisión sistemática.

Se debe estructurar el resumen con la descripción de los aspectos que siguen: la cuestión clínica, las fuentes y bases de datos, los métodos de revisión y selección de la literatura y de síntesis cuantitativa de los datos de forma reproducible, los resultados con sus estimaciones y sus intervalos de confianza, y la conclusión con los resultados principales.

La introducción debe contextualizar y fundamentar el objetivo.

La metodología debe detallar las fuentes y el modo de búsqueda, el período e idioma, los criterios de selección de los estudios, la manera de evaluación de sesgo de publicación, la evaluación de la calidad y validez metodológicas de los estudios, el modo de extracción de los datos idealmente por dos investigadores, las características de los estudios, la manera de evaluación de la heterogeneidad y el modo de sintetizar matemáticamente los datos.

Los resultados deben presentar el flujo de la revisión, las características de los estudios, una evaluación de la distribución por edad, sexo, modo de diagnóstico o selección de pacientes, covariables relevantes, tiempo de seguimiento, tamaño de la muestra, y los estimadores combinados de desempeño diagnóstico, con los debidos intervalos de confianza.

En la discusión, especificar los puntos clave, discutir las inferencias clínicas con base en la validez interna y externa, interpretar los resultados enfocando la totalidad de las evidencias, describir las limitaciones y los potenciales sesgos, específicamente el sesgo de publicación, y sugerir estudios futuros [405].



### **13.7. Lista QUADAS (Quality Assessment Diagnostic Accuracy Studies) para la comprobación de estudios sobre pruebas o sistemas diagnósticos incluidos en revisiones sistemáticas y meta-análisis**

---

La herramienta QUADAS (*Quality Assessment Diagnostic Accuracy Studies*) se desarrolló como un proyecto colaborativo entre el *Centre for Reviews and Dissemination*, de la Universidad de York, y la *Academic Medical Centre* de la Universidad de Amsterdam [432,433]. Fue financiado por el programa *Health Technology Assessment* y se publicó en el año 2003 [434].

Una versión modificada del QUADAS ha sido utilizada por la Colaboración Cochrane en las revisiones de la precisión de pruebas o sistemas diagnósticos, lo que la consolida como referencia a la hora de ponderar la calidad de estos artículos cuando se incluyen en revisiones sistemáticas [435].

Un grupo de expertos diseñaron una lista de ítems relevantes, recogidos en la literatura médica. Utilizando el método Delphi seleccionaron 14 ítems. Cada uno se puntuaba como “sí”, “no” o “dudoso”. El “sí” indicaba siempre una buena respuesta. El QUADAS incluye el riesgo de sesgo, aplicabilidad y calidad en la descripción del estudio. La versión Cochrane de la herramienta omitió los ítems relacionados con la calidad en la descripción del estudio [434,435].

A partir de la experiencia de los autores y las aportaciones de la Cochrane en cuanto a dificultades con la utilización de QUADAS, se procedió a revisar la primera versión y a desarrollar el QUADAS-2, que terminó presentándose en el año 2010. El QUADAS-2 está formado por cuatro áreas fundamentales que incluyen [250,436]:

- 1) La selección de los pacientes.
- 2) El *test* o prueba en estudio.
- 3) Los estándares de referencia.
- 4) El flujo de los pacientes y el cronograma.

En cada una de las áreas se evalúa el riesgo de sesgo y las dudas acerca de su aplicabilidad. Esta evaluación se realiza con una serie de preguntas orientadas a evaluar existencia de un sesgo.

El QUADAS-2 se aplica en cuatro fases [434]:

1) La fase 1 es un resumen de la pregunta de la revisión. En esta fase, los autores deben describir las características de la revisión sistemática: los pacientes, la prueba diagnóstica, la prueba de referencia y la enfermedad o situación que se estudia.

2) La fase 2 incluye adaptar al estudio en concreto que se analiza, las preguntas orientativas que ayudarán a interpretar la calidad del estudio.

3) La fase 3 se refiere a establecer un diagrama de trabajo (diagrama de flujo). Se revisa el diagrama de flujo de cada estudio primario y si no lo tiene, se realiza un diagrama de cada estudio. Con un diagrama de trabajo apropiado será más fácil evaluar el riesgo de sesgo. De esta manera se obtiene información acerca de método de inclusión de los pacientes (es decir si son pacientes consecutivos con síntomas que hacen sospechar que tienen la enfermedad, o si son casos y controles).

4) En la fase 4 se valora el riesgo de sesgo y la aplicabilidad:

· *Riesgo de sesgo*: La primera parte de cada área se valora el sesgo y está estructurada en tres secciones: 1) qué información se proporciona para poder evaluar el riesgo de sesgo; 2) las preguntas orientativas, y 3) la valoración del riesgo de sesgo. Las preguntas orientativas se responden con los términos: “sí”, “no”, o “dudoso”. El riesgo de sesgo se valora como “bajo”, “alto”, o “dudoso”. Si todas las preguntas orientativas son respondidas como sí, entonces el riesgo es bajo. Si alguna se responde como no, existe riesgo de sesgo. En ese caso los autores deben utilizar las guías desarrolladas en la fase dos para juzgar el riesgo de sesgo.

· *Aplicabilidad*: los autores de la revisión deben registrar la información en base a lo que se concluye acerca de la aplicabilidad.

Área	Selección de los pacientes	Prueba diagnóstica en estudio	Prueba de referencia	Flujo y cronograma
<b>Descripción</b>	Describe los métodos utilizados para seleccionar a los pacientes: pruebas previas, ámbito, uso previsto de la prueba en estudio	Describe la prueba, cómo se realizó y su interpretación	Describe la prueba de referencia, cómo se realizó y su interpretación	Describe a los pacientes que no van a recibir la prueba de estudio, la prueba de referencia o que se excluyen de la tabla 2 x 2: describe el intervalo y cualquier intervención entre la prueba en estudio y la de referencia
	¿Es una muestra consecutiva o aleatoria?	¿Se interpretaron los resultados de la prueba sin el conocimiento de los de la prueba de referencia? Lo correcto es realizar primero la prueba de estudio	¿La prueba de referencia clasifica correctamente la enfermedad en estudio?	¿Describe el intervalo de tiempo entre las dos pruebas? ¿El intervalo de tiempo es el adecuado?
<b>Preguntas clave (sí/no/dudoso)</b>	¿Se evitó un diseño de casos y controles?  ¿Se evitaron exclusiones inapropiadas?	Si se usó un punto de corte (umbral), ¿se especificó previamente?	¿Los resultados de la prueba de referencia se interpretaron independientemente de la prueba de estudio? ¿Hay algún elemento de la prueba en estudio que forme parte de la prueba de referencia?	¿Se aplicó a todos los pacientes el patrón de referencia? ¿Todos los pacientes recibieron la misma prueba de referencia independientemente del resultado de la prueba en estudio? ¿Se incluyeron todos los pacientes en el análisis?

<b>Riesgo de sesgo (alto/bajo/dudoso)</b>	¿Hay sesgo en la selección de los pacientes?	¿Podría haber sesgos en la realización e interpretación de la prueba?	¿Podría haber sesgos en la realización e interpretación de la prueba?	¿El flujo de seguimiento del paciente podría haber producido algún sesgo?
<b>Aplicabilidad (alta/baja/dudosa)</b>	¿Hay dudas de que los pacientes incluidos y su ámbito de estudio no se ajusten a la pregunta de la revisión? Es decir, que sean diferentes de la población diana	¿Hay dudas de que la prueba (realización e interpretación) difieran de la pregunta de revisión? Cualquier modificación de la tecnología, interpretación o realización merma su aplicabilidad	¿Hay dudas de que la condición de estudio (enfermedad) definida por la prueba de referencia (realización e interpretación) difiera o no se ajustara a la pregunta de revisión?	

Las iniciativas QUADAS y STARD coinciden en la búsqueda de un instrumento que detecte la variación y el sesgo de los estudios de sistemas diagnósticos utilizando la Medicina basada en la evidencia. Difieren entre sí en la intención del instrumento: STARD tiene como objetivo el proporcionar una lista que sirva de guía para la publicación de los estudios de precisión de sistemas diagnósticos. Es una herramienta que se utiliza de forma prospectiva para realizar un diseño adecuado de un estudio; por tanto, interesan a los investigadores en la fase de diseño del estudio y a los editores. QUADAS-2, sin embargo, es una herramienta para valorar la calidad de los estudios primarios en las revisiones sistemáticas y meta-análisis. Se utiliza de forma retrospectiva para realizar un análisis crítico del rigor metodológico de un estudio de sistemas diagnósticos [434].

Como consideraciones finales, QUADAS-2 no debe utilizarse para generar una escala de puntuación de la calidad. Si un estudio se considera como “bajo” en todas las áreas pues el estudio se describe como *bajo riesgo de sesgo*. Si “alto” o “dudoso”, se describe como *con riesgo de sesgo* o como *existencia de dudas acerca de la aplicabilidad*.

Los resultados se pueden describir en forma de resumen o de tabla, describiendo cuantos estudios tienen bajo, alto o dudoso riesgo de sesgo y aplicabilidad en cada área.

Los autores pueden elegir incluir únicamente los estudios de sistemas diagnósticos con bajo riesgo de sesgo en todas las áreas. También pueden realizar análisis de subgrupos y análisis de sensibilidad.

Se ha criticado al QUADAS que presenta limitaciones por lo que respecta a su reproductibilidad, sobre todo en los ítems relativos a los resultados indeterminados o no concluyentes, las pérdidas y las retiradas del estudio. Aun así, se recomienda utilizar tanto STARD como QUADAS a la hora de evaluar la calidad de los estudios de sistemas diagnósticos. A pesar de la existencia de estas herramientas, aproximadamente la mitad de las revisiones sistemáticas sobre sistemas diagnósticos no llevan a cabo una evaluación reglada de la calidad de los trabajos que incluyen. Esto debe condicionar la ponderación de los estudios y, si no se tiene en cuenta, puede imprimir un sesgo en las conclusiones, dado que es probable que se infraestime la heterogeneidad [437].

Fuente: Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, Leeflang MM, Sterne JAC, and Bossuyt PMM. QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. *Ann Intern Med.* 2011;155(8):529-536.







### 13.8. Formulario de entrada a la base de datos

Captura de pantalla de la interfaz de usuario empleada para la entrada de información a la base de datos creada con Microsoft Access 2010®.

**AlHDataEntryForm**

## Formulario de Entrada de Datos

Propuesta y validación de criterios simplificados para el diagnóstico de hepatitis autoinmune en pediatría

Número de Historia Clínica:	<input type="text"/>	Fecha de nacimiento:	<input type="text"/>
Hospital de origen:	<input type="text" value="HVH"/> <input type="text" value="HSJD"/>	Fecha de inicio del tratamiento: <small>(Fecha de diagnóstico)</small>	<input type="text"/>
Sexo:	<input type="text" value="Hombre"/> <input type="text" value="Mujer"/>	Diagnósticos previos:	<input type="text"/>

### Criterios originales revisados de 1999

<p><b>Sexo Femenino:</b></p> <p>Sí <input type="text" value="+2"/> No <input type="text" value="0"/></p> <p><b>Valor por encima de lo normal de inmunoglobulinas o IgG:</b></p> <p>&gt;2 <input type="text" value="+3"/> 1,5 - 2 <input type="text" value="+2"/> 1 - 1,5 <input type="text" value="+1"/> &lt;1 <input type="text" value="0"/></p> <p><b>Anticuerpos anti-mitocondriales (AMA):</b></p> <p>Sí <input type="text" value="-4"/> No <input type="text" value="+1"/></p> <p><b>Histopatología hepática:</b></p> <p>Hepatitis de interfase <input type="text" value="+3"/> Infiltración con predominio linfoplasmocitario <input type="text" value="+2"/> Formaciones de hepatocitos en roseta <input type="text" value="+1"/> Nada de lo anterior <input type="text" value="-5"/> Afectación biliar <input type="text" value="-3"/> Otros cambios que sugieran distinta etiología <input type="text" value="-3"/></p> <p><b>Ingesta media diaria de alcohol:</b></p> <p>&lt;25 g/día <input type="text" value="+2"/> &gt;60 g/día <input type="text" value="-2"/></p> <p><b>Parámetros adicionales opcionales:</b></p> <p>Otros anticuerpos (anti-SLA/LP, LC1, ASGPR, pANCA, antiactina) <input type="text" value="+2"/> HLA DR3 o DR4 <input type="text" value="+1"/></p> <p><b>Respuesta al tratamiento:</b></p> <p>Completa <input type="text" value="+2"/> Con recaída <input type="text" value="+3"/></p>	<p><b>Relación FA/AST (o ALT):</b></p> <p>&lt;1,5 <input type="text" value="+2"/> 1,5 - 3 <input type="text" value="0"/> &gt;3 <input type="text" value="-2"/></p> <p><b>ANA, anti-SM o anti-LKM1:</b></p> <p>&gt;1:80 <input type="text" value="+3"/> 1:80 <input type="text" value="+2"/> 1:40 <input type="text" value="+1"/> &lt;1:40 <input type="text" value="0"/></p> <p><b>Marcadores de hepatitis viral:</b></p> <p>Positivos <input type="text" value="-4"/> Negativos <input type="text" value="+1"/></p> <p><b>Otra(s) enfermedad(es) autoinmune(s) en el paciente o en familiar de primer grado:</b></p> <p>Sí <input type="text" value="+2"/> No <input type="text" value="0"/></p> <p><b>Puntuación total en los criterios originales revisados de 1999:</b></p> <input type="text"/>
---	--

**Diagnóstico definitivo de hepatitis autoinmune**

Diagnóstico definitivo (alternativo a HAI):

Notas sobre el diagnóstico:

### Criterios simplificados de 2008

<p><b>ANA o anti-SM:</b></p> <p>≥1:40    <input type="text" value="+1"/></p> <p>≥1:80    <input type="text" value="+2"/></p>	<p><b>Anti-LKM1 ≥1:40:</b></p> <p>Sí    <input type="text" value="+2"/></p> <p>No    <input type="text" value=""/></p>
<p><b>Anti-SLA positivos:</b></p> <p>Sí    <input type="text" value="+2"/></p> <p>No    <input type="text" value=""/></p>	<p><b>Histopatología hepática:</b></p> <p>Compatible con HAI    <input type="text" value="+1"/></p> <p>Típica de HAI    <input type="text" value="+2"/></p>
<p><b>Niveles de inmunoglobulina G:</b></p> <p>Por encima del límite superior de la normalidad    <input type="text" value="+1"/></p> <p>&gt;1,1 veces por encima del límite superior de la normalidad    <input type="text" value="+2"/></p>	
<p><b>Ausencia de marcadores de hepatitis viral:</b></p> <p>Sí    <input type="text" value="+2"/></p> <p>No    <input type="text" value=""/></p>	<p><b>Puntuación total en los criterios simplificados de 2008:</b></p> <input style="width: 100px;" type="text"/>

### Criterios diagnósticos de la HAI pediátrica propuestos por la ESPGHAN y la NASPGHAN (2009)

<p><input type="checkbox"/> <b>Hipertransaminasemia</b></p> <p><input type="checkbox"/> <b>Hipergammaglobulinemia</b></p> <p><input type="checkbox"/> <b>Ausencia de marcadores de hepatitis viral</b></p> <p><input type="checkbox"/> <b>Colangiograma (colangiorensonancia magnética o colangiografía retrógrada) normal</b></p>	<p><input type="checkbox"/> <b>Presencia de auto-anticuerpos</b> ANA y/o anti-SM a título ≥1:20 Anti-LKM1 a título ≥1:10 Anti-LC1 Anti-SLA</p> <p><input type="checkbox"/> <b>Histopatología hepática</b> Hepatitis de interfase Colapso multilobular</p> <p><input type="checkbox"/> <b>Descarte de enfermedad de Wilson</b></p>
--	---

### Resultados de exploraciones complementarias

Informe de A. P. de la biopsia hepática: <input style="width: 150px;" type="text"/>	Niveles de Ig G (mg/dL): <input style="width: 50px;" type="text"/>
Informe de la colangio-RM: <input style="width: 150px;" type="text"/>	Niveles de AST (U/L): <input style="width: 50px;" type="text"/>
Informe de la CPRE: <input style="width: 150px;" type="text"/>	Niveles de ALT (U/L): <input style="width: 50px;" type="text"/>
	Niveles de FA (U/L): <input style="width: 50px;" type="text"/>
	Niveles de GGT (U/L): <input style="width: 50px;" type="text"/>
Autoanticuerpo 1: ANA <input style="width: 30px;" type="text"/>	Título del autoanticuerpo 1 (1:X): <input style="width: 50px;" type="text"/>
Autoanticuerpo 2: ANA <input style="width: 30px;" type="text"/>	Título del autoanticuerpo 2 (1:X): <input style="width: 50px;" type="text"/>
Autoanticuerpo 3: ANA <input style="width: 30px;" type="text"/>	Título del autoanticuerpo 3 (1:X): <input style="width: 50px;" type="text"/>
Autoanticuerpo 4: <input style="width: 50px;" type="text"/>	Título del autoanticuerpo 4 (1:X): <input style="width: 50px;" type="text"/>

### 13.9. Hoja informativa y de consentimiento informado

---

Estamos realizando un estudio con el fin de validar los criterios diagnósticos de la hepatitis autoinmune en la población de pacientes con problemas del hígado que seguimos en el hospital. Aunque no padezca esta enfermedad, igualmente necesitamos saber cómo se comportan dichos criterios en otras enfermedades hepáticas. Ello implica conocer los datos básicos del paciente, los resultados de las pruebas que se le han hecho, el diagnóstico final al que se ha llegado y la respuesta al tratamiento que le han planteado.

Pedimos su colaboración para que podamos incluirle (o incluir a su hijo/tutelado) en el registro. Su participación, si acepta, consistirá únicamente en que utilizaremos los datos que figuren en su historia clínica. No se le someterá a ninguna prueba que pueda resultarle perjudicial.

Que usted acepte o rechace participar en el estudio no cambia la manera en que será tratado por el médico cuando acuda a consulta o cuando sea ingresado.

Le agradecemos de antemano su colaboración

*José Vicente Arcos Machancoses (Coordinador / Investigador)*

---

Yo, \_\_\_\_\_, en calidad de \_\_\_\_\_ (paciente o tutor de) he leído la información indicada en esta página y comprendo que :

- Puedo aceptar o rechazar libremente mi participación (o la participación de mi hijo o tutelado).
- La asistencia médica que recibiré (o recibirá mi hijo o tutelado) no depende de mi participación en el estudio.
- Puedo retirarme/retirarlo del estudio cuando lo desee sin que eso modifique la asistencia médica que reciba

Libremente acepto participar en el estudio que se me propone.

Firma y fecha



## Publicaciones en relación con la tesis

---

Arcos-Machancoses JV, Molera Busoms C, Julio Tatis E, Victoria Bovo M, Quintero Bernabeu J, Juampérez Goñi J, Crujeiras Martínez V, Martín de Carpi J. Accuracy of the 2008 Simplified Criteria for the Diagnosis of Autoimmune Hepatitis in Children. *Pediatric Gastroenterology Hepatology and Nutrition*. 2018.

**DOI: 10.5223/pghn.2018.21.2.118**

Arcos-Machancoses JV, Molera Busoms C, Julio Tatis E, Victoria Bovo M, Martín de Carpi J. Accuracy of the simplified criteria for autoimmune hepatitis in children: Systematic review and decision analysis. *Journal of Clinical and Experimental Hepatology*. 2018.

**DOI: 10.1016/j.jceh.2018.10.006**

Arcos-Machancoses JV, Molera Busoms C, Julio Tatis E, Victoria Bovo M, Quintero Bernabeu J, Juampérez Goñi J, Crujeiras Martínez V, Martín de Carpi J. Development and validation of a new simplified diagnostic scoring system for pediatric autoimmune hepatitis. *Digestive and Liver Disease*. 2019.

**DOI: 10.1016/j.dld.2019.02.018**