ORIGINAL

Rediseño del triaje telefónico para dolor torácico no traumático mediante reglas deterministas derivadas por aprendizaje automático: desarrollo y validación temporal

Carlos Romero-Olóriz[1], María José Luque-Hernández[1], Álvaro Ritoré-Hidalgo[1-3], Cristina López-López[1], Ángel Serrano-Alarcón[1-3], Miguel Ángel Armengol de La Hoz[3]

Objetivos. Desarrollar y validar un sistema de triaje telefónico con reglas deterministas y preguntas dicotómicas para dolor torácico no traumático, orientado a reducir el infratriaje y la brecha diagnóstica por sexo.

Métodos. Estudio observacional retrospectivo en el Centro de Emergencias Sanitarias 061 de Andalucía. Se incluyeron 66.384 demandas con triaje estructurado mediante la guía vigente (GTTA23) y diagnóstico clínico extrahospitalario como referencia. Sus preguntas de opción múltiple se recodificaron en 36 variables dicotómicas. Mediante técnica de gradiente reforzado *(gradient boosting)* se derivaron 15 reglas de emergencia (P1, soporte vital avanzado inmediato) y 8 de no emergencia (NO-P1, demorables); el resto quedó como escucha del coordinador (ESC). Se utilizaron 53.323 episodios para el desarrollo (2020-2023) y 13.061 (año 2024) para la validación. Se siguieron las recomendaciones TRIPOD+AI y SAGER.

Resultados. En la validación la sensibilidad del sistema (P1 y ESC frente a NO-P1) fue del 97,8% (IC 95%: 96,96–98,42) frente al 74,0% del vigente (Δ = +23,82 puntos porcentuales; p < 0,0001). El infratriaje bajó del 6,7% al 4,0% y los falsos negativos, de 415 a 35. La brecha entre sexos disminuyó de 16,09 a 6,24 puntos porcentalaes (ΔΔ = –9,85; p = 0,0005). El sistema redujo el despacho inmediato de soporte vital avanzado del 52,8% al 28,5%; el 35,2% se resolvió deterministamente y el 64,8% pasó al coordinador.

Conclusiones. El sistema reduce el infratriaje por debajo del 5% y atenúa la brecha diagnóstica por sexo, preservando transparencia e interpretabilidad. Su impacto sobre carga asistencial, tiempos de coordinación y uso de recursos requiere una validación prospectiva.

Palabras clave: Triaje telefónico. Dolor torácico. Servicios médicos de urgencia. Reglas de decisión clínica. Aprendizaje automático. Centro coordinador de urgencias y emergencias.

Redesign of telephone triage for non-traumatic chest pain using deterministic rules derived through machine learning: development and temporal validation

Objective. To develop and validate a telephone triage system based on deterministic rules and dichotomous questions for non-traumatic chest pain, aimed at reducing undertriage and the sex-related diagnostic gap.

Method. Retrospective observational study conducted at the Centro de Emergencias Sanitarias 061 de Andalucía (Andalusia, Spain). A total of 66,384 calls triaged using the current structured guideline (GTTA23) were included, with the prehospital clinical diagnosis used as the reference standard. The multiple-choice questions were recoded into 36 dichotomous variables. Using gradient boosting, we derived 15 emergency rules (P1, immediate advanced life support) and 8 non-emergency rules (NO-P1, deferrable cases); the remaining cases were referred for coordinator review (ESC). The development cohort comprised 53,323 episodes recorded between 2020 and 2023, and the validation cohort comprised 13,061 episodes recorded in 2024. We followed the TRIPOD+AI and SAGER guidelines.

Results. In the validation cohort, system sensitivity (P1 ∪ ESC vs NO-P1) was 97.8% (95% CI, 96.96–98.42), compared with 74.0% for the current system (Δ = +23.82 percentage points; *P* < .0001). Undertriage dropped from 6.7% down to 4.0%, and the number of false negatives decreased from 415 to 35. The gap between the sexes decreased from 16.09 to 6.24 percentage points (ΔΔ = –9.85 percentage points; *P* = .0005). The system reduced immediate advanced life support dispatch from 52.8% to 28.5%; 35.2% of cases were resolved deterministically, while 64.8% were referred to the coordinator.

Conclusions. The system reduces undertriage to < 5% and attenuates the sex-related diagnostic gap while preserving transparency and interpretability. Its impact on workload, coordination times, and resource use requires prospective validation.

Keywords: Telephone triage. Chest pain. Emergency medical services. Clinical decision rules. Machine learning. Emergency coordination centre.

DOI: 10.55633/s3me/111.2026

Filiación de los autores:
[1]Centro de Emergencias Sanitarias 061 (CES-061) de Andalucía, Campanillas, Málaga, España.
[2]Ibima Plataforma Bionand, Instituto de Investigación Biomédica de Málaga, Málaga, España.
[3]Laboratorio de Ciencia de Datos, Fundación Progreso y Salud, Sevilla, España.

Contribución de los autores:
Todos los autores han confirmado su autoría en el documento de responsabilidades del autor, acuerdo de publicación y cesión de derechos a EMERGENCIAS.

Autor para correspondencia:
María José Luque Hernández.
Centro de Emergencias Sanitarias 061 (CES-061) de Andalucía.
Sede Central.
Parque Tecnológico de Andalucía.
C/ Severo Ochoa, 28.
29590 Campanillas, Málaga, España.

Correo electrónico:
mariajose.luque.sspa@juntadeandalucia.es

Información del artículo:
Recibido: 29-4-2026
Aceptado: 5-7-2026
Online: 4-9-2026

Editor responsable:
Aitor Alquézar Arbé

DOI:
10.55633/s3me/111.2026

Introducción

El dolor torácico no traumático es uno de los motivos más frecuentes de solicitud de asistencia urgente, con tendencia creciente en las últimas décadas[1,2]. Solo una minoría presenta síndrome coronario agudo (SCA) u otra condición tiempo-dependiente[3], con prevalencias de emergencia real del 11% al 17% según el entorno[2,4]. Esta desproporción entre volumen y gravedad exige equilibrar seguridad y eficiencia. Además, la presentación clínica del SCA difiere por sexo, con mayor frecuencia en mujeres de manifestaciones no clásicas o menos ajustadas al patrón tradicional de dolor coronario, lo que puede introducir inequidades si el protocolo no captura adecuadamente estos patrones[4,5].

Los centros coordinadores de urgencias y emergencias (CCUE) utilizan protocolos de triaje telefónico estructurado para asignar prioridades y movilizar recursos[6,7]. La evidencia sobre su rendimiento diagnóstico para dolor torácico es heterogénea, con sensibilidades del 76% al 93% y especificidades del 25% al 48%[7], lo cual releja la dificultad intrínseca de la evaluación remota. El infratriaje y el sobretriaje influyen en el balance seguridad-eficiencia como indicadores complementarios del rendimiento, con umbrales aceptables propuestos del 5-10% y del 30-50, respectivamente[8].

El proyecto trIAje se diseñó para evaluar y optimizar el triaje telefónico del dolor torácico no traumático en el CES 061 de Andalucía mediante una estrategia escalonada. En su Fase 1, nuestro grupo evaluó el rendimiento del sistema vigente (GTTA23) en el periodo 2020-2024 y confirmó limitaciones relevantes: infratriaje del 6,6%, sobretriaje del 82,6%, brecha de sensibilidad por sexo de 16,2 puntos porcentuales y baja cumplimentación completa del cuestionario estructurado[9]. En concreto, la mediana de preguntas completadas fue de 3 de 6 posibles y solo el 1,5% de las demandas recibieron el protocolo completo[9].

Sobre esta base, la presente Fase 2 analiza si el rendimiento puede mejorarse conservando el conjunto de variables ya recogidas por la GTTA23 vigente, pero reorganizándolas en dos niveles complementarios: por un lado, una arquitectura determinista e interpretable de reglas de decisión; por otro, una reformulación del cuestionario multicategoría original en preguntas dicotómicas de respuesta Sí/No. Este enfoque permite aislar una pregunta operativa concreta: ¿hasta qué punto puede optimizarse el algoritmo de decisión y explorarse un formato potencialmente más estandarizable del interrogatorio sin modificar todavía la información clínica recogida durante la llamada? Antes de introducir nuevas variables o modelos más complejos, resulta necesario estimar el rendimiento alcanzable con la información ya disponible[10-12]. En entornos de alto riesgo como los CCUE, la interpretabilidad de las reglas de decisión sigue siendo un requisito operativo relevante[13].

El objetivo primario de este estudio fue rediseñar el triaje telefónico estructurado para dolor torácico no traumático conservando las variables de la GTTA23 vigente y una arquitectura determinista interpretable, mediante una nueva selección y combinación de variables guiada por aprendizaje automático, para cuantificar el margen de mejora alcanzable bajo las restricciones del sistema actual. Como objetivos secundarios se definieron: (i) reducir el infratriaje sin incrementar de modo inaceptable la activación inmediata de soporte vital avanzado; (ii) reducir la brecha diagnóstica por sexo; y (iii) explorar la viabilidad operativa de un formato dicotómico como hipótesis a contrastar prospectivamente.

Método

Diseño del estudio y aspectos éticos

Estudio observacional retrospectivo de desarrollo y validación temporal de un sistema de reglas clínicas para el triaje telefónico del dolor torácico no traumático, siguiendo las recomendaciones TRIPOD+AI[14] sobre los datos del trabajo previo del mismo grupo[9]. El proyecto trIAje (PI-0179-2024) fue aprobado por el Comité Coordinador de Ética de la Investigación Biomédica de Andalucía (SICEIA-2024-002592). Se eximió del consentimiento informado al ser retrospectivo con datos anonimizados.

Ámbito y población de estudio

El CES-061 proporciona cobertura a 8,5 millones de habitantes mediante ocho CCUE provinciales. Se incluyeron las demandas por dolor torácico no traumático (código A23) gestionadas entre enero de 2020 y diciembre de 2024 con triaje estructurado mediante la Guía de Triaje Telefónico A23 (GTTA23)[15] y diagnóstico clínico del equipo extrahospitalario disponible. De las 248.116 demandas iniciales registradas en el periodo, se excluyeron secuencialmente 50.522 sin asistencia efectiva (anulación de la demanda, autoactivaciones, traslados primarios sin triaje), 99.689 sin diagnóstico clínico documentado al cierre, 31.418 con asignación directa de prioridad sin cumplimentación del cuestionario GTTA23 estructurado y 103 con datos faltantes en variables clave del modelo predictivo. La muestra analítica final fue de 66.384 demandas (Figura 1).

Partición temporal

La validación fue temporal, y se usó para el desarrollo los datos de los años 2020-2023 (n = 53.323; prevalencia de emergencia real [P1REAL] 13,1%) y para la validación los del 2024 (n = 13.061; prevalencia 12,2%).

Variables y patrón de referencia

Las variables predictoras procedieron de la información recogida durante la coordinación telefónica. Las opciones del GTTA23 se recodificaron como variables dicotómicas (Tabla 1) y se añadieron el sexo y tres um-

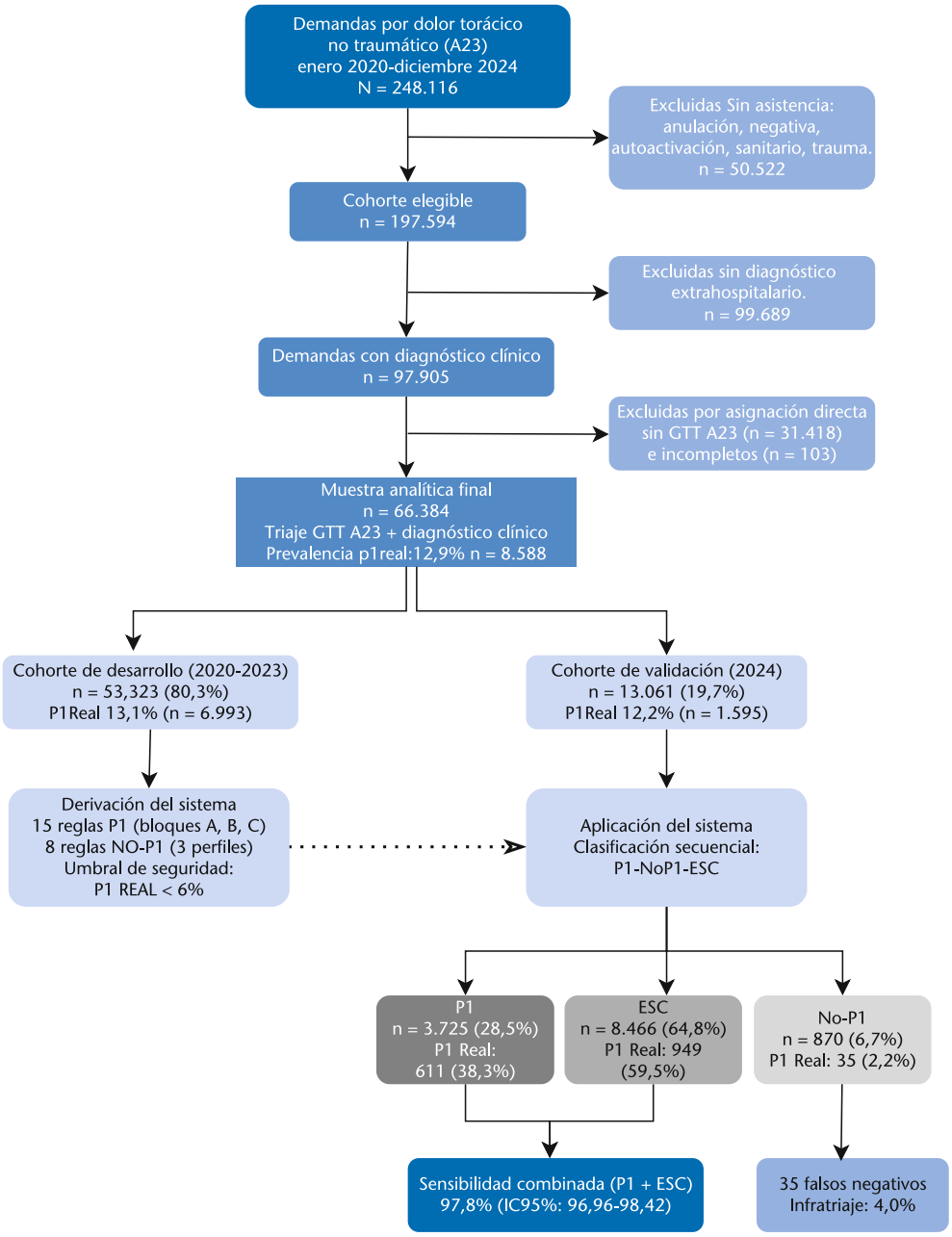


Figura 1. Diagrama de flujo del estudio según TRIPOD + AI.
CES-061: Centro de Emergencias Sanitarias 061; ESC: escucha y valoración por médico coordinador; GTTA23: Guía de Triaje Telefónico A23; IC: intervalo de confianza; NO-P1: no emergencia; P1: emergencia; P1REAL: emergencia verdadera (referencia diagnóstica); TRIPOD+AI: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis – Artificial Intelligence.

brales de edad (55, 65 y 75 años), resultando 36 variables. El patrón de referencia fue el diagnóstico clínico final del equipo extrahospitalario, dicotomizado como emergencia verdadera (P1REAL) según el consenso Delphi previamente descrito[9]. Se definieron infratriaje (1 – VPN) y sobretriaje (1 – VPP).

Desarrollo del sistema de reglas

El desarrollo siguió tres etapas. Primero, se entrenó un modelo de gradiente reforzado (*gradient boosting*, HGB_d4, HistGradientBoosting profundidad 4), seleccionado entre 14 arquitecturas por validación cruzada estratificada, para cuantificar la importancia variable por permutación sobre el AUC-ROC. No es el clasificador final, sino que aporta el techo discriminativo y su probabilidad calibrada por Platt se usó para el DCA.

Segundo, se derivaron 15 reglas de activación (P1, emergencia vital tiempo-dependiente) por combinación estructurada de patrones clínicos consolidados por las guías de SCA, confirmados por el orden de importancia. Las reglas P1 se organizaron en tres bloques pres-

Tabla 1. Cuestionario dicotómico y sistema de reglas propuesto para la clasificación del dolor torácico no traumático (cohorte de validación, N = 13.061)

Panel A — Cuestionario dicotómico: recodificación del protocolo GTTA23 en preguntas de respuesta Sí/No	
Pregunta original (GTTA23)	**Preguntas dicotómicas para el operador (Sí/No)**
1. Antecedentes cardiacos (4 opciones)	¿Ha tenido ingresos por angina o infarto? ¿Tiene tarjeta de corazón? ¿Tiene algún otro antecedente de corazón?
2. Localización del dolor (2 opciones)	¿Le duele en el centro del pecho? ¿Le duele en otra zona (no en el centro)?
3. Tipo de dolor (4 opciones)	¿Es un dolor que le aprieta o le pesa? ¿Es un dolor muy fuerte, muy intenso? ¿Es un dolor que le quema o le pincha?
4. Relación con el movimiento (3 opciones)	¿Empeora con el movimiento, la tos o al respirar hondo?
5. Síntomas acompañantes (6 opciones)	¿Está sudando mucho? ¿Está pálido/a? ¿Tiene náuseas o vómitos? ¿Le falta el aire o tiene dificultad para respirar? ¿Tiene alteración del nivel de conciencia?
6. Respuesta a fármacos (4 opciones)	¿Ha mejorado al tomar nitroglicerina (Cafinitrina/Vernies/...)?
Datos demográficos	Sexo: ¿Es hombre o mujer? Edad: ¿Cuántos años tiene? (umbrales: < 55, ≥ 65, ≥ 75 años)
El protocolo original tiene 6 preguntas con 23 opciones de respuesta (algunas de selección múltiple). El cuestionario propuesto las transforma en 15 preguntas Sí/No + sexo + edad → entrada para las reglas del Panel B.	

Panel B — Sistema de reglas: combinaciones de las variables del Panel A que determinan la clasificación				
Paso 1 — Reglas de activación (P1): si alguna se cumple → activación inmediata de soporte vital avanzado				
ID	**Condiciones clínicas**	**n**	**P1REAL (%)**	**VP**
Bloque A — Dolor típico + síntomas vegetativos				
A1	Dolor opresivo/peso + sudoración profusa	71	21,1	15
A2	Dolor opresivo/peso + palidez	48	6,2	3
A3	Dolor opresivo/peso + náuseas/vómitos	85	12,9	11
A4	Dolor opresivo/peso + disnea	92	7,6	7
A5	Dolor opresivo/peso + alteración de conciencia	0	–	0
A6	Dolor muy intenso + sudoración profusa	26	23,1	6
Bloque B — Antecedentes cardiacos + dolor típico o respuesta a NTG				
B1	Ingreso previo angina/infarto + dolor opresivo/peso	1.498	15,9	238
B2	Ingreso previo angina/infarto + dolor muy intenso	342	19,3	66
B3	Tarjeta corazón + dolor opresivo/peso	39	15,4	6
B4	Tarjeta corazón + dolor muy intenso	9	11,1	1
B5	Mejoría con NTG (criterio único)	28	14,3	4
Bloque C — Sexo masculino + edad + características del dolor				
C1	Hombre + edad ≥ 65 años + centrotorácico + opresivo	893	18,1	162
C2	Hombre + centrotorácico + dolor muy intenso	299	24,1	72
C3	Hombre + edad ≥ 75 años + centrotorácico	1.303	16,0	209
C4	Hombre + edad ≥ 65 años + opresivo	1.491	17,0	253
Total P1		**3.725**	**16,4**	**611**

(Continúa)

pecificados: dolor típico con síntomas vegetativos (bloque A, 6 reglas), antecedentes cardiacos con dolor típico o respuesta a nitroglicerina (bloque B, 5 reglas) y sexo masculino combinado con edad y características del dolor (bloque C, 4 reglas).

Tercero, las reglas de no emergencia (NO-P1) se generaron *de novo* mediante combinaciones de variables clínicas con características menos sugestivas de emergencia tiempo-dependiente, junto con moduladores demográficos de menor riesgo. Las candidatas se evaluaron mediante dos filtros secuenciales de seguridad (*standalone* –evaluados regla a regla sobre la cohorte completa–): un filtro principal sobre desarrollo, que exige prevalencia de P1REAL < 6% (umbral fijado como criterio de no inferioridad respecto al infratriaje del GTTA23), y un filtro de robustez sobre la validación del año 2024. De 12 candidatas, 2 se excluyeron en el filtro principal y 2 adicionales en el filtro de robustez por deriva temporal, por lo que quedaron 8 reglas NO-P1 (Tabla 1). Todas incorporan el modulador MUJER = 1 por el patrón epidemiológico detectado. Las demandas sin activación de ninguna regla se asignaron a escucha por el médico coordinador (ESC). La clasificación es secuencial y excluyente: P1 → NO-P1 → ESC.

Análisis estadístico

El rendimiento se evaluó en validación bajo dos formulaciones binarias. La formulación A (P1 Δ ESC frente a NO-P1), principal, evalúa la seguridad operativa global del sistema –la capacidad de no clasificar como au-

Tabla 1. Cuestionario dicotómico y sistema de reglas propuesto para la clasificación del dolor torácico no traumático (cohorte de validación, N = 13.061) (Continuación)

Panel B — Sistema de reglas: combinaciones de las variables del Panel A que determinan la clasificación				
Paso 2 — Reglas de exclusión (NO-P1): si alguna se cumple → demorable (sin activación inmediata de SVA)				
ID	**Condiciones clínicas**	**n**	**P1REAL (%)**	**FN**
Perfil I — Mujer joven (< 55 años) con características menos sugestivas de emergencia tiempo-dependiente				
N01	Mujer + edad < 55 años + sin antecedentes + tipo quemazón/pincha	117	2,6	3
N03	Mujer + edad < 55 años + localización no centrotorácica + tipo quemazón/pincha	101	3,0	3
N04	Mujer + edad < 55 años + localización no centrotorácica	376	4,0	15
Perfil II — Mujer con patrón osteomuscular				
N05	Mujer + localización no centrotorácica + empeora con movimiento	365	5,2	19
N06	Mujer + tipo quemazón/pincha + empeora con movimiento	145	2,8	4
N07	Mujer + localización no centrotorácica + tipo quemazón/pincha + empeora con movimiento	104	3,8	4
N08	Mujer + sin antecedentes + localización no centrotorácica + tipo quemazón/pincha	261	5,0	13
Perfil III — Mujer con perfil respiratorio				
N12	Mujer + tos / fiebre	46	4,3	2
Reglas candidatas retiradas — descartadas en los filtros sucesivos de seguridad standalone (P1REAL < 6%)				
ID	**Condiciones clínicas**	**n (val)**	**P1REAL dev (%)**	**P1REAL val (%)**
Filtro principal (sobre cohorte de desarrollo): excluidas por P1REAL ≥ 6% en desarrollo				
N09	Mujer + < edad 55 años + tipo quemazón/pincha + empeora con movimiento	70	6,4	5,7
N10	Mujer + < edad 55 años + localización no centrotorácica + empeora con movimiento	106	7,2	4,7
Filtro de robustez (sobre cohorte de validación 2024): retiradas por deriva temporal (P1REAL < 6% en desarrollo pero ≥ 6% en validación)				
N02	Mujer + < edad 55 años + sin antecedentes + empeora con movimiento	56	4,8	7,1
N11	Localización no centrotorácica + tipo quemazón/pincha + empeora con movimiento	184	5,6	6,5
Total NO-P1		**870**	**4,0**	**35**
Paso 3 — ESC: si ninguna regla se activa → escucha y valoración por médico coordinador (red de seguridad). En validación: n = 8.466 (64,8%); P1REAL: 11,2% (n = 949).				

Panel A: las 6 preguntas multicategoría del protocolo vigente (GTTA23) se recodificaron como preguntas dicotómicas individuales (Sí/No). El sistema no incorpora preguntas nuevas; reutiliza exclusivamente la información ya recogida durante la coordinación telefónica. La condición «sin antecedentes cardiacos» que aparece en las reglas NO-P1 se cumple cuando las tres preguntas sobre antecedentes (pregunta 1) son «No».
Panel B: la clasificación es secuencial y excluyente (paso 1 → paso 2 → paso 3). La activación de cualquier regla P1 clasifica la demanda como emergencia e interrumpe la evaluación. Las reglas NO-P1 solo se evalúan si ninguna P1 activó. Si ninguna regla se activa, la demanda se deriva al médico coordinador (ESC). Cuando una pregunta no fue formulada, las reglas que la incluyen no se activan (sesgo conservador hacia ESC). La regla A5 registró 0 activaciones en validación porque la evaluación secuencial asigna esos casos a reglas previas.
Las columnas n, P1REAL, VP y FN corresponden a la cohorte de validación (2024). De 12 candidatas iniciales NO-P1, 4 se eliminaron en filtros sucesivos de seguridad standalone con umbral P1REAL < 6%: 2 en el filtro principal sobre la cohorte de desarrollo (N09 y N10) y 2 en el filtro de robustez por deriva temporal sobre la cohorte de validación (N02 y N11). Las cuatro reglas descartadas se muestran en gris al final del Paso 2 con su prevalencia en ambas cohortes (dev: desarrollo / val: validación). Las cifras de cada regla individual reflejan sus activaciones independientes; las filas «Total» reportan demandas únicas tras la clasificación secuencial, por lo que la suma de las filas individuales es superior al total al existir solapamiento entre reglas.
ESC: escucha/valoración por médico coordinador; FN: falsos negativos (emergencias verdaderas clasificadas como NO-P1); GTTA23: Guía de Triaje Telefónico A23 (dolor torácico del adulto); NTG: nitroglicerina; P1REAL: prevalencia de emergencia verdadera en el subgrupo (%); SVA: soporte vital avanzado; VP: verdaderos positivos (emergencias verdaderas clasificadas como P1).

tónomamente resoluble un caso que requiere atención–; la categoría ESC no equivale a sobrepriorización, dado que el caso se deriva al coordinador con el cuestionario ya cumplimentado, sin movilización de soporte vital avanzado. La formulación B (P1 frente a ESC y NO-P1), complementaria, evalúa el componente autónomo del algoritmo. Se calcularon sensibilidad, especificidad, valores predictivos, infratriaje, sobretriaje, F1 y falsos negativos, con IC 95% por remuestreo *(bootstrap)* de 2.000 réplicas[17]. La comparación con el GTTA23 se realizó mediante test de McNemar, intervalos de Newcombe[18] para diferencias de proporciones y test de DeLong pareado[19] para diferencias de las áreas bajo la curva (ABC). Se reportó el índice de reclasificación neto (NRI) con descomposición en eventos y no-eventos[20,21]; el índice de inflación (IDI) se omitió por la asimetría de escalas entre el GTTA23 binario y la probabilidad calibrada del HGB_d4[22-24]. Se realizó análisis de curvas de decisión (DCA)[25,26]. El infratriaje se reporta como descriptivo con IC *bootstrap.*

Análisis por subgrupos, sensibilidad y perspectiva de sexo/género

Los análisis se estratificaron por sexo, con cruce sexo × edad en cinco tramos (< 45, 45-54, 55-64, 65-74, ≥ 75 años). La subcohorte SAGER excluyó 19 registros con sexo no informado (0,1%) y 4 adicionales con edad no informada (0,0%) para el cruce etario. Las brechas hombre-mujer se calcularon con IC de Newcombe y el contrastes entre brechas mediante *bootstrap* pareado. Se realizaron análisis de sensibilidad de las reglas y al umbral en el rango 4%-9%. La perspectiva de sexo y género siguió las recomendaciones de las guías SAGER[27] y PROGÉNEROS[28].

Uso de inteligencia artificial generativa

Se emplearon modelos de lenguaje comerciales como apoyo para la generación y depuración de código Python y la revisión lingüística del manuscrito, bajo

Tabla 2. Comparación formal del sistema propuesto frente al protocolo GTTA23 vigente sobre la cohorte de validación temporal de 2024 (N = 13.061)

Panel A — Seguridad global del dispositivo · GTTA23: P1 vs no-P1 / Sistema: P1 y ESC vs NO-P1				
Métrica	**GTTA23 vigente**	**Sistema propuesto**	**Δ (Sistema – GTTA23)**	**Valor de p**
Sensibilidad	74,0 [71,8; 76,1]	97,8 [97,0; 98,4]	+23,8 [+21,6; +26,1]	< 0,0001
Especificidad	50,1 [49,2; 51,1]	7,3 [6,8; 7,8]	–42,9 [–43,8; –41,9]	< 0,0001
VPP	17,1 [16,2; 18,0]	12,8 [12,2; 13,4]	–4,3	–
VPN	93,3 [92,6; 93,9]	96,0 [94,5; 98,0]	+2,7	–
Infratriaje (1 – VPN)	6,7 [6,1; 7,4]	4,0 [2,9; 5,5]	–2,7	–
Sobretriaje (1 – VPP)	82,9 [82,0; 83,8]	87,2 [86,6; 87,8]	+4,3	–
F1	27,8	22,6	–5,2	–
AUC binario	0,621	0,525	–0,095	–
Panel B — Activación autónoma de SVA · GTTA23: P1 vs no-P1 / Sistema: solo P1 vs ESC ∪ NO-P1				
Sensibilidad	74,0 [71,8; 76,1]	38,3 [35,9; 40,7]	–35,7 [–38,6; –32,6]	< 0,0001
Especificidad	50,1 [49,2; 51,1]	72,8 [72,0; 73,6]	+22,7 [+21,6; +23,8]	< 0,0001
VPP	17,1 [16,2; 18,0]	16,4 [15,2; 17,6]	–0,7	–
VPN	93,3 [92,6; 93,9]	89,5 [88,8; 90,1]	–3,8	–
Infratriaje (1 – VPN)	6,7 [6,1; 7,4]	10,5 [9,9; 11,2]	+3,8	–
Sobretriaje (1 – VPP)	82,9 [82,0; 83,8]	83,6 [82,4; 84,7]	+0,7	–
F1	27,8	23,0	–4,8	–
ABC binario	0,621	0,556	–0,065	–
Panel C — Reclasificación y discriminación · NRI y wNRI miden el cambio de clasificación de pacientes; ΔAUC mide el cambio en discriminación probabilística				
Métrica	**Valor (IC 95%)**	**Test**	**p-valor**	
Reclasificación · Formulación A (P1 ∪ ESC vs NO-P1)				
Componente eventos (P1REAL = 1)	+23,8 [+21,7; +26,1]	*Bootstrap* 2.000	–	
Componente no-eventos (P1REAL = 0)	–42,9 [–43,9; –41,9]	*Bootstrap* 2.000	–	
NRI global	–19,0 [–21,5; –16,6]	*Bootstrap* 2.000	–	
wNRI (ponderado por prevalencia)	–34,7 [–35,7; –33,8]	*Bootstrap* 2.000	–	
Reclasificación · Formulación B (P1 ∪ ESC vs NO-P1)				
Componente eventos (P1REAL = 1)	–35,7 [–38,6; –32,7]	*Bootstrap* 2.000	–	
Componente no-eventos (P1REAL = 0)	+22,7 [+21,6; +23,8]	*Bootstrap* 2.000	–	
NRI global	–13,0 [–16,1; –9,7]	*Bootstrap* 2.000	–	
wNRI (ponderado por prevalencia)	+15,6 [+14,4; +16,6]	*Bootstrap* 2.000	–	
Discriminación probabilística				
ABC HGB_d4 (modelo continuo, referencia técnica)	0,661 [0,647; 0,675]	–	–	
ΔABC (HGB_d4 – GTTA23)	+0,041 [+0,025; +0,056]	DeLong	< 0,0001	

Paneles A y B: métricas como porcentaje (%) con IC 95% entre corchetes (*bootstrap* pareado, 2.000 réplicas; Newcombe para porcentajes; McNemar para contraste de sensibilidad y especificidad; DeLong para ABC). El GTTA23 distingue solo P1 frente a no-P1, por lo que sus datos son idénticos en ambos paneles. El sistema propuesto adopta dos formulaciones binarias distintas: Panel A (P1 ∪ ESC vs NO-P1, seguridad operativa global) y Panel B (P1 ∪ ESC vs NO-P1, activación autónoma de SVA).
Panel C: el NRI cuantifica la reclasificación neta entre clasificadores y se descompone en componentes de eventos y no-eventos; el wNRI los pondera por prevalencia. Los valores negativos del NRI global y del wNRI son un artefacto previsible del colapso de un sistema tricategórico (P1/ESC/NO-P1) a un clasificador binario y no indican inferioridad clínica (Pencina 2014; Kerr 2014). El ΔAUC representa la ganancia discriminativa del modelo continuo HGB_d4 (referencia técnica del techo discriminativo) frente al GTTA23.
Definiciones del proyecto trIAje: infratriaje = 1 – VPN; sobretriaje = 1 – VPP; F1 = media armónica de sensibilidad y VPP; AUC binario = capacidad discriminativa de la clasificación binaria (0,5-1).
ABC: área bajo la curva de la característica operativa del receptor; ESC: escucha y valoración por médico coordinador; GTTA23: Guía de Triaje Telefónico A23; HGB_d4: HistGradientBoosting de profundidad 4 (modelo de aprendizaje automático del que derivan las reglas; sirve como referencia del techo discriminativo); IC: intervalo de confianza; NRI: índice de reclasificación neta; P1: emergencia; P1REAL: emergencia verdadera (referencia); SVA: soporte vital avanzado; VPN: valor predictivo negativo; VPP: valor predictivo positivo; wNRI: NRI ponderado por prevalencia.

supervisión directa de los investigadores. Ninguna herramienta de inteligencia artificial figura como autora.

Resultados

Se analizaron 66.384 demandas (Figura 1). La mediana de edad fue de 76 años (rango intercuartil 62-84) y el 53,0% correspondió a mujeres. La prevalencia de P1REAL fue del 12,9%, con diferencia marcada por sexo: 17,6% en hombres frente al 8,8% en mujeres (p < 0,001). En validación (n = 13.061), el modelo de referencia HGB_d4 alcanzó un ABC de la característica operativa del receptor (COR) de 0,6615 (IC 95%: 0,65-0,67), dentro del rango estrecho de las 14 arquitecturas evaluadas (0,66-0,68). El análisis de importancia por permutación identificó el sexo femenino como predictor dominante (Δ entre ABC = 0,079), más de cuatro veces superior al segundo predictor (edad ≥ 75 años, ΔAUC = 0,019).

El sistema derivado comprende 15 reglas P1 en tres bloques clínicos y 8 reglas NO-P1 (Tabla 1). En validación, el sistema clasificó el 28,5% de las demandas como P1 (n = 3.725), el 6,7% como NO-P1 (n = 870) y el 64,8% como ESC (n = 8.466).

Bajo la formulación principal (P1 y ESC frente a NO-

P1), la sensibilidad del sistema fue del 97,8% (IC 95%: 97,0-98,4) frente al 74,0% (IC 95%: 71,8-76,1) del GTTA23 (Δ = +23,8 puntos porcentuales [pp]; IC 95%: +21,6 a +26,1; p < 0,0001; Tabla 2, Panel A). El infratriaje se redujo del 6,7% al 4,0% (IC 95%: 2,9-5,5; Δ descriptivo = –2,7 pp), con disminución de los falsos negativos de 415 a 35. La especificidad fue del 7,3% frente al 50,1% del vigente (Δ = –42,9 pp; p < 0,0001). El NRI mostró descomposición asimétrica (+23,8 pp en eventos, –42,9 pp en no-eventos; NRI global –19,0%, IC 95% –21,5 a –16,6; Tabla 2, Panel C), artefacto previsible del colapso de un sistema tricategórico a un clasificador binario (discusión). En términos clínicos, la reducción de los falsos negativos de 415 a 35 sobre la cohorte de validación equivale a 380 emergencias adicionales correctamente identificadas como necesitadas de atención médica, manteniendo el infratriaje por debajo del umbral aceptado en la literatura.

Bajo la formulación complementaria (P1 frente a ESC y NO-P1), que aísla el componente autónomo del algoritmo, la sensibilidad fue del 38,3%, la especificidad del 72,8% y el infratriaje del 10,5% (Tabla 2, Panel B).

Al estratificar por sexo (Tabla 3, Panel A), la sensibilidad bajo la formulación A fue del 100% en hombres (0 falsos negativos) y del 93,8% (IC 95%: 91,4-95,5) en mujeres; la clasificación NO-P1 está confinada al subgrupo femenino: los 870 casos clasificados como NO-P1 son todos mujeres y los 35 falsos negativos residuales (4,02% del NO-P1) recaen íntegramente sobre la cohorte femenina, dado que el modulador MUJER = 1 forma parte de todas las reglas NO-P1. La brecha de sensibilidad hombre-mujer se redujo de +16,1 pp con el GTTA23 a +6,2 pp (IC 95%: +4,5 a +8,5) con el sistema propuesto; el contraste formal entre brechas (ΔΔ brechas = –9,8 pp; IC 95%: –15,0 a –4,7; p = 0,0005) constituye el hallazgo central del análisis SAGER. El cruce sexo × edad (Tabla 3, Panel B) identificó a las mujeres de 45-54 años como el subgrupo con menor sensibilidad (75,7%), con aumento significativo de brecha (Δ = +22,5 pp; p = 0,010) y reducción significativa en los tramos ≥ 55 años, donde se concentra el 86% de la patología P1REAL. Clínicamente, esto representa diez puntos porcentuales menos de brecha en los tramos donde reside la mayor parte de la patología emergente (≥ 55 años), con vigilancia explícita en mujeres de 45-54 años.

Los análisis de sensibilidad (Tabla 4) confirmaron la robustez del sistema: el umbral 4%-9% mantuvo el principio de no inferioridad y el volumen de P1 fue invariante, modulando solo el reparto entre ESC y NO-P1. El sistema intermedio aplicado a validación (10 reglas NO-P1 supervivientes únicamente del filtro principal sobre desarrollo) arrojó sensibilidad del 96,7% e infratriaje del 5,1%, frente al 97,8% y 4,0% del sistema final, cuantificando el optimismo introducido por el filtro de robustez en < 1,1 puntos porcentuales de sensibilidad. En términos clínicos, las decisiones fueron estables frente a umbrales de seguridad razonables: cualquier valor entre el 4% y el 9% conservó el infratriaje por debajo del 6,7% del protocolo vigente.

Respecto al dispositivo operativo, el despacho inmediato de soporte vital avanzado se redujo del 52,8% al 28,5% (–46% relativo); el sistema resolvió deterministamente el 35,2% de las demandas y derivó el 64,8% al coordinador. Aunque el sobretriaje estadístico aumentó del 82,9% al 87,2% bajo la formulación principal, esta métrica incluye P1 y ESC como categoría positiva y debe interpretarse como activación inmediata o derivación a valoración por el médico coordinador, no como activación automática de soporte vital avanzado en todos los casos. El análisis de curvas de decisión (Figura 2) confirmó beneficio neto superior al GTTA23 en el rango clínicamente pertinente de umbrales.

Discusión

El sistema propuesto redujo sustancialmente el infratriaje del triaje telefónico para dolor torácico no traumático, situándolo por debajo del umbral del 5% propuesto como objetivo de seguridad y alcanzando una sensibilidad combinada superior al rango descrito en la revisión sistemática más citada sobre este proceso asistencial[7], sin añadir complejidad a su estructura determinista.

La reducción de la brecha de sensibilidad entre sexos (ΔΔ = –9,8 pp; p = 0,0005) es el hallazgo central del análisis SAGER, concentrada en los tramos donde reside el 86% de la patología P1REAL (≥ 55 años). La dominancia del sexo femenino es coherente con la menor probabilidad pre-test de SCA en mujeres premenopáusicas y la mayor frecuencia de manifestaciones no clásicas[5,30]. No obstante, el sistema redistribuye el riesgo residual hacia subgrupos concretos de mujeres: el subgrupo de 45-54 años presenta una sensibilidad del 75,7% (9 falsos negativos, ≈ 25,7% del total residual del sistema), consecuencia estructural del modulador MUJER = 1. Cualquier implementación prospectiva deberá monitorizar específicamente este subgrupo y evaluar si la ganancia de sensibilidad global compensa éticamente la concentración del riesgo residual.

Los valores negativos del NRI global no contradicen la ganancia de seguridad, sino que reflejan la penalización por reclasificar no-eventos hacia ESC dentro de la formulación A, fenómeno esperable al reducir un sistema tricategórico a una comparación binaria[20,21,29]. Su interpretación debe centrarse en los componentes: más emergencias verdaderas identificadas a costa de derivar más no-eventos a valoración médica.

El aumento del sobretriaje estadístico de 82,9% a 87,2% no debe interpretarse automáticamente como mayor activación de soporte vital avanzado, porque la categoría positiva agrupa P1 y ESC. P1 implica activación inmediata, mientras que ESC supone valoración por el médico coordinador. Aunque el despacho inmediato de soporte vital avanzado se redujo del 52,8% al 28,5%, la redistribución del 64,8% hacia ESC puede incrementar carga médica, tiempos de coordinación o decisiones conservadoras; su impacto real sobre recursos y resultados clínicos deberá evaluarse prospectivamente.

Tabla 3. Análisis de equidad por sexo y por grupo etario sobre la cohorte de validación temporal de 2024 (N = 13.061)

Panel A — Rendimiento global por sexo · Formulación A (P1 ∪ ESC vs NO-P1); cohorte de validación temporal 2024 (N = 13.061)			
Sistema	**Hombres**	**Mujeres**	**Brecha (H – M)**
Sensibilidad · proporción de emergencias verdaderas correctamente identificadas			
GTTA23 vigente	79,73 [77,2; 82,1]	63,6 [59,6; 67,5]	+16,1 [+11,4; +20,8]
Sistema propuesto	100,00 [99,6; 100,0]	93,8 [91,4; 95,5]	+6,2 [+4,5; +8,5]
Δ brechas (Sistema – GTTA23)	–	–	–9,8 [–15,00; –4,7]; p = 0,0005
Especificidad · proporción de no-emergencias correctamente identificadas como no-P1 (o NO-P1)			
GTTA23 vigente	43,8 [41,8; 44,5]	55,5 [54,3; 56,7]	–12,3 [–14,1; –10,5]
Sistema propuesto	0,0 [0,0; 0,1]	12,9 [12,1; 13,7]	–12,9 [–13,7; –12,1]
Δ brechas (Sistema – GTTA23)	–	–	–0,6 [–2,5; +1,4]; p = 0,589
Infratriaje (1 – VPN) · proporción de emergencias verdaderas que el sistema clasificó como negativas			
GTTA23 vigente	8,8 [7,8; 10,1]	5,4 [4,7; 6,1]	+3,5 [+2,2; +4,9]
Sistema propuesto	n/a (a)	4,0 [2,9; 5,5]	n/a (a)
Δ brechas (Sistema – GTTA23)	–	–	n/a (a)
Panel B — Sensibilidad por sexo y grupo etario · Formulación A; cinco tramos de edad (cohorte de validación 2024)			
Sistema	**Hombres**	**Mujeres**	**Brecha (H – M)**
< 45 años · n = 726; n P1REAL = 35; prevalencia = 4,8%			
GTTA23 vigente	93,3 [78,7; 98,1]	100,0 [56,5; 100,0]	–6,7 [–21,3; +37,0]
Sistema propuesto	100,0 [88,6; 100,0]	80,0 [37,5; 96,4]	+20,0 [+0,1; +62,4]
Δ brechas (Sistema – GTTA23)	–	–	+26,7 [0,0; +73,3]; p = 0,106 (b)
45-54 años · n = 1.099; n P1REAL = 174; prevalencia = 15,8%			
GTTA23 vigente	88,3 [81,9; 92,7]	86,5 [72,0; 94,1]	+1,8 [–8,1; +16,9]
Sistema propuesto	100,0 [97,3; 100,0]	75,7 [59,9; 86,6]	+24,3 [+13,0; +40,1]
Δ brechas (Sistema – GTTA23)	–	–	+22,5 [+5,1; +40,2]; p = 0,010 (b)
55-64 años · n = 1.671; n P1REAL = 306; prevalencia = 18,3%			
GTTA23 vigente	85,7 [80,7; 89,6]	72,1 [60,4; 81,3]	+13,7 [+3,1; +25,9]
Sistema propuesto	100,0 [98,4; 100,0]	97,1 [89,9; 99,2]	+2,9 [+0,3; +10,1]
Δ brechas (Sistema – GTTA23)	–	–	–10,7 [–21,7; –0,7]; p = 0,034
65-74 años · n = 2.280; n P1REAL = 364; prevalencia = 16,0%			
GTTA23 vigente	83,1 [77,9; 87,3]	78,5 [70,4; 84,9]	+4,6 [–3,6; +13,8]
Sistema propuesto	100,0 [98,4; 100,0]	95,9 [90,7; 98,2]	+4,1 [+1,3; +9,3]
Δ brechas (Sistema – GTTA23)	–	–	–0,5 [–9,6; +8,2]; p = 0,935
≥ 75 años · n = 7.262; n P1REAL = 713; prevalencia = 9,8%			
GTTA23 vigente	69,7 [64,9; 74,1]	53,3 [47,9; 58,6]	+16,4 [+9,2; +23,3]
Sistema propuesto	100,0 [99,0; 100,0]	94,5 [91,5; 96,5]	+5,4 [+3,2; +8,5]
Δ brechas (Sistema – GTTA23)	–	–	–10,9 [–18,4; –3,9]; p = 0,004

Marco general: los valores se expresan como porcentaje (%) con intervalo de confianza al 95% entre corchetes (*bootstrap,* 2.000 réplicas; Newcombe para porcentajes individuales). La brecha (H – M) es la diferencia, en puntos porcentuales, entre el valor en hombres y el valor en mujeres dentro de un mismo sistema; cuantifica la disparidad de género del sistema (un valor próximo a 0 indica equidad). El Δ brechas (Sistema – GTTA23) es la diferencia entre las dos brechas anteriores; cuantifica cuánto se reduce o se amplía la disparidad de género al pasar del protocolo vigente al sistema propuesto. Un Δ brechas negativo indica reducción de la disparidad (más equidad); un Δ brechas positivo indica ampliación. La p del Δ brechas se calculó por *bootstrap* pareado.

Panel A: la métrica clínicamente prioritaria es la sensibilidad (capacidad de detectar emergencias verdaderas). La brecha de sensibilidad por sexo del GTTA23 (+16,09 pp en favor de hombres) se reduce a +6,24 pp con el sistema propuesto: una reducción absoluta de 9,85 pp en la disparidad de género (p = 0,0005). En especificidad, la brecha por sexo se mantiene similar (Δ no significativo, p = 0,589) y refleja un patrón conocido: las mujeres siguen recibiendo más clasificaciones positivas que los hombres en ambos sistemas, en parte por diferencias en la expresión clínica del dolor torácico y por la mayor frecuencia de manifestaciones no clásicas. En infratriaje (proporción de falsos negativos sobre el total de clasificados como negativos), el sistema propuesto no clasifica como negativos a ningún hombre bajo formulación A –todos los varones son derivados a P1 o ESC–, por lo que la métrica es matemáticamente indefinida en ese subgrupo (anotación a).

Panel B: cinco tramos etarios sobre la cohorte SAGER (n = 13.038). En los tramos donde reside el 86% de la patología P1REAL (≥ 55 años) el sistema reduce significativamente la brecha de sensibilidad por sexo: 55-64 años (p = 0,034) y ≥ 75 años (p = 0,004); en 65-74 años no hay cambio significativo (p = 0,935). En los tramos jóvenes la brecha aumenta nominalmente: < 45 años no alcanza significación (p = 0,106; IC amplio por escasez de eventos) y 45-54 años sí (p = 0,010), constituyendo una limitación reconocida (ver nota b). IC y p-valores por *bootstrap* pareado (2.000 réplicas).

Notas: (a) bajo el sistema propuesto en formulación A todos los hombres se clasifican como P1 o ESC y ninguno va a NO-P1, por lo que VPN e infratriaje no son matemáticamente definibles para el subgrupo masculino (división por cero). Es consecuencia mecánica del diseño. (b) los Δ brechas positivos en tramos jóvenes tienen interpretación distinta: en < 45 años la amplitud del IC refleja escasez de eventos (5 mujeres P1REAL; p = 0,106); en 45-54 años el aumento es estadísticamente significativo (p = 0,010) y constituye una limitación reconocida del sistema en mujeres jóvenes con manifestaciones menos sugestivas de emergencia tiempo-dependiente según las variables disponibles en la llamada.

Definiciones del proyecto trIAje: infratriaje = 1 – VPN; sobretriaje = 1 – VPP.

ESC: escucha y valoración por médico coordinador; GTTA23: Guía de Triaje Telefónico A23; H: hombres; IC: intervalo de confianza; M: mujeres; n/a: no aplicable; P1: emergencia; P1REAL: emergencia verdadera (referencia); pp: puntos porcentuales; VPN: valor predictivo negativo; VPP: valor predictivo positivo.

El formato dicotómico podría reducir la carga cognitiva del operador al reutilizar las variables del GTTA23 reformuladas como preguntas Sí/No, hipótesis a contrastar prospectivamente dado que solo el 1,5% de las demandas recibieron el protocolo completo[9]. Asimismo, la lógica secuencial podría disminuir el número efectivo de preguntas al detener el interrogatorio tras una regla P1; este efecto sobre duración,

Tabla 4. Análisis de robustez del sistema propuesto: sensibilidad al umbral, Leave-One-Out (LOO) y análisis contrafactual sobre la cohorte de validación temporal de 2024 (N = 13.061)

Panel A — Análisis de sensibilidad al umbral · rendimiento del sistema variando el umbral máximo de prevalencia P1REAL admisible para una regla NO-P1 (cohorte de validación 2024)

Umbral	Reglas NO-P1 activas	Sensibilidad (%)	Especificidad (%)	Infratriaje (%)
< 4%	3 reglas (N01, N06, N07)	99,62 [99,18; 99,83]	2,01 [1,77; 2,29]	2,53 [1,17; 5,41]
< 5%	7 reglas (N01, N03, N04, N06, N07, N08, N12)	98,56 [97,85; 99,04]	5,76 [5,34; 6,20]	3,37 [2,25; 5,00]
< 6% (a)	8 reglas (N01, N03, N04, N05, N06, N07, N08, N12)	97,81 [96,96; 98,42]	7,28 [6,82; 7,77]	4,02 [2,91; 5,54]
< 7%	10 reglas (+ N09, N11)	97,30 [96,39; 97,99]	7,94 [7,46; 8,45]	4,51 [3,37; 6,02]
< 8%	12 reglas (+ N02, N10) (b)	97,18 [96,25; 97,88]	8,12 [7,63; 8,63]	4,61 [3,46; 6,11]
< 9%	12 reglas (saturación) (b)	97,18 [96,25; 97,88]	8,12 [7,63; 8,63]	4,61 [3,46; 6,11]

Panel B — Análisis Leave-One-Out (LOO) de las 8 reglas NO-P1 · impacto de retirar cada regla individualmente, reclasificando sus demandas a ESC (cohorte de validación 2024; baseline = sistema completo)

Regla retirada	n reclasificadas a ESC	n P1REAL recuperados	Δ Sensibilidad (pp)	Δ Infratriaje (pp)
N05 · Mujer · osteomuscular (mov. + loc.)	187	12	+0,75	–0,66
N08 · Mujer · sin antecedentes + osteomuscular	154	9	+0,56	–0,39
N04 · Mujer joven · loc. no centrotorácica simple	211	7	+0,44	+0,23
N12 · Respiratorio · mujer con tos/fiebre	19	1	+0,06	–0,03
N01 · Mujer joven · tipo quemazón/pincha	35	0	0,00	+0,17
N06 · Mujer · osteomuscular (mov. + tipo)	33	0	0,00	+0,16
N03 · Mujer joven · localización no centrotorácica	0	0	0,00	0,00
N07 · Mujer · osteomuscular completo	0	0	0,00	0,00
Sistema completo (baseline)	–	–	–	4,02% (referencia)

Panel C — Análisis contrafactual: ¿y si las reglas NO-P1 no existieran? · comparación entre el GTTA23 vigente, el sistema propuesto y el escenario contrafactual donde las 870 demandas clasificadas como NO-P1 se reasignan al médico coordinador (ESC)

Métrica	GTTA23 vigente (%)	Sistema propuesto (%)	Contrafactual (%) (c)	Lectura clínica
Sensibilidad	73,98 [71,77; 76,07]	97,81 [96,96; 98,42]	100,00	el contrafactual tiene 0 falsos negativos por construcción
Especificidad	50,15 [49,23; 51,06]	7,28 [6,82; 7,77]	0,00	el contrafactual no clasifica nada como negativo
Infratriaje (1 – VPN)	6,73 [6,13; 7,38]	4,02 [2,91; 5,54]	n/a (d)	matemáticamente indefinido (denominador = 0)
Sobretriaje (1 – VPP)	82,89 [81,98; 83,76]	87,20 [86,60; 87,78]	87,79	el sobretriaje del sistema es prácticamente equivalente al contrafactual

Marco general: los análisis de robustez someten el sistema a tres pruebas independientes orientadas a verificar la solidez del diseño y a caracterizar las propiedades del sistema desde ángulos complementarios. Todos los análisis utilizan la formulación A (P1 ∪ ESC vs NO-P1), que es la directamente comparable con el GTTA23 vigente. Los IC al 95% se calcularon por bootstrap (Newcombe para porcentajes; 2.000 réplicas; semilla 42).

Panel A — análisis de sensibilidad al umbral: rendimiento del sistema variando el umbral máximo de prevalencia P1REAL admisible para una regla NO-P1. El umbral operativo del sistema final es < 6% (nota a), fijado por criterio de no inferioridad respecto al GTTA23. Umbrales más restrictivos (4-5%) reducen la auto-resolución; umbrales más permisivos (7-9%) la aumentan a costa de mayor exposición. La saturación a partir del 8% (nota b) refleja la admisión de las 12 candidatas iniciales.

Panel B — análisis Leave-One-Out (LOO): impacto de retirar cada regla NO-P1 individualmente (las otras 7 permanecen activas; baseline = sistema completo). Cada fila reporta cuántas demandas pasan a ESC, cuántos P1REAL se recuperan, y los Δ resultantes en sensibilidad e infratriaje. Las reglas más activas son N04, N05 y N08; ninguna retirada produce una ganancia > 0,75 pp de sensibilidad, lo que confirma la robustez del sistema frente a la pérdida de reglas individuales.

Panel C — análisis contrafactual: escenario simulado donde las 870 demandas clasificadas como NO-P1 por el sistema se reasignan a ESC (nota c). El contrafactual alcanza por construcción 100% de sensibilidad y 0% de especificidad; el infratriaje no es definible (nota d). El sobretriaje del sistema (87,20%) es prácticamente idéntico al contrafactual (87,79%). El sistema se aproxima al escenario contrafactual en términos de sensibilidad, pero conserva una fracción limitada de clasificación NO-P1, con el correspondiente balance entre seguridad, especificidad y carga de valoración médica.

Notas: (a) umbral operativo del sistema final (< 6%), fijado por criterio de no inferioridad respecto al infratriaje del GTTA23 vigente. (b) saturación: a partir del umbral del 8% se admiten las 12 reglas candidatas iniciales y no hay diferencias entre umbrales superiores. (c) contrafactual: construcción analítica, no métrica observada; reasigna las 870 demandas NO-P1 al coordinador. (d) denominador = 0: el escenario contrafactual no clasifica ninguna demanda como negativa, lo que hace que el infratriaje (1 – VPN) sea matemáticamente indefinido.

Definiciones del proyecto trIAje: infratriaje = 1 – VPN; sobretriaje = 1 – VPP.

ESC: escucha y valoración por médico coordinador; GTTA23: Guía de Triaje Telefónico A23; IC: intervalo de confianza; LOO: Leave-One-Out (eliminación de uno cada vez); n/a: no aplicable; NO-P1: regla de no-emergencia (auto-resuelve sin escucha por médico coordinador); P1: emergencia; P1REAL: emergencia verdadera (referencia); pp: puntos porcentuales; VPN: valor predictivo negativo; VPP: valor predictivo positivo.

cumplimentación y carga cognitiva requiere validación prospectiva.

Los análisis de curvas de decisión y contrafactual respaldaron la estabilidad del diseño en el rango clínicamente pertinente de umbrales bajo el principio de «no perder emergencias»[25,26], aunque su interpretación debe considerarse complementaria al resultado clínico principal: reducción de falsos negativos e infratriaje. El estrecho rango de ABC COR (0,658-0,683) refleja una limitación inherente al triaje telefónico basado en información verbal breve[7]; la ausencia de mejoras relevantes con modelos probabilísticos más complejos basados en las mismas variables respalda la elección de reglas interpretables[13] y proporciona una referencia para futuros estudios con nuevas fuentes de información o diseños prospectivos.

El estudio presenta varias limitaciones. El modulador MUJER = 1 confina la clasificación NO-P1 al subgrupo femenino y deja indefinido el infratriaje masculino bajo formulación A; las reglas con prevalencia residual más

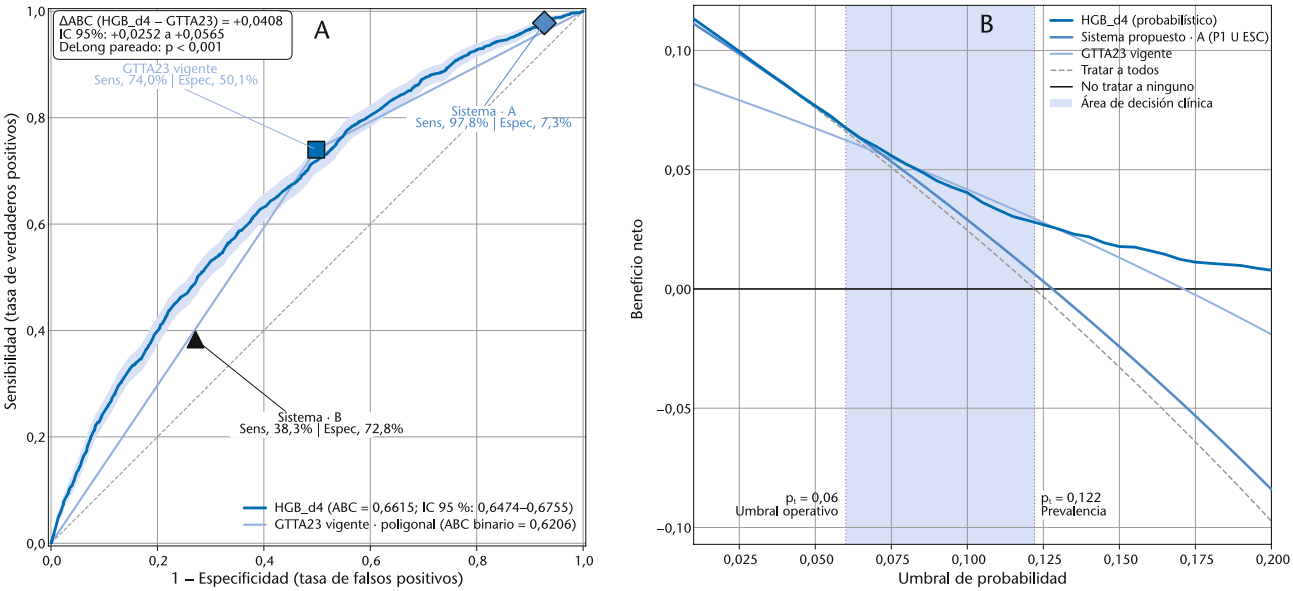


Figura 2. Discriminación y utilidad clínica sobre la cohorte de validación 2024 (n = 13.061).
Panel A · Curvas de la característica operativa del receptor. La curva COR representa la sensibilidad (verdaderos positivos) frente a 1 – especificidad (falsos positivos) para todos los umbrales posibles de decisión; el área bajo la curva (ACD) resume la capacidad discriminativa global (1 = perfecta, 0,5 = azar). La curva azul corresponde al modelo subyacente HGB_d4 (probabilidad continua), con ACD = 0,6615 (IC 95%: 0,647 – 0,675) y banda de incertidumbre estimada por remuestreo (2.000 réplicas). La poligonal naranja corresponde al protocolo GTTA23 vigente, que es un clasificador binario y opera en un único punto (cuadrado naranja); su ABC binario fue 0,621. La diferencia entre ambas áreas (ΔACD = +0,041; IC 95%: +0,025 a +0,056; test de DeLong: p < 0,001) cuantifica la mejora discriminativa del modelo respecto al protocolo vigente. Los marcadores indican los puntos operativos del estudio: GTTA23 (cuadrado naranja), sistema propuesto en formulación A (rombo verde: P1 y ESC vs. NO-P1) y formulación B (triángulo rojo: solo P1 vs. ESC y NO-P1).
El sistema propuesto (◆) prioriza la sensibilidad (97,8%; 35 falsos negativos) frente al vigente (■; 74,0%; 415 falsos negativos), reduce la activación inmediata de unidades de SVA (28,5% vs. 52,8%) y detecta más emergencias verdaderas, a costa de una mayor derivación al médico coordinador (que explica la baja especificidad aparente, 7,3%). El triángulo (▲) representa una lectura alternativa del mismo sistema centrada solo en la activación directa de unidades de emergencias (P1 vs. resto): sensibilidad 38,3%, especificidad 72,8%.
Panel B · Análisis de curva de decisión (ACD). El ACD traduce la discriminación en utilidad clínica mediante el beneficio neto, que combina aciertos y falsos positivos en una única escala según el umbral de probabilidad pt por encima del cual el clínico decide actuar (en este caso, derivar al médico coordinador). Un beneficio neto superior al de las estrategias triviales («tratar a todos» y «no tratar a ninguno») indica utilidad clínica añadida. La franja sombreada en color crema delimita el área de decisión clínicamente relevante, comprendida entre el umbral operativo del sistema (pt = 0,06) y la prevalencia observada (pt = 0,122). En el umbral operativo del sistema (pt = 0,06, borde izquierdo del sombreado), el sistema propuesto en formulación A (verde) aporta un beneficio neto prácticamente idéntico al del modelo probabilístico subyacente HGB_d4 (azul) y superior al del GTTA23 vigente (naranja). La pendiente decreciente de la línea verde refleja la propiedad de cualquier clasificador binario evaluado fuera de su umbral de calibración y no debe interpretarse como pérdida de rendimiento.
ABC: área bajo la curva; ACD: análisis de curva de decisión; ESC: escucha y valoración por médico coordinador; GTTA23: Guía de Triaje Telefónico A23; HGB_d4: modelo *Histogram-based Gradient Boosting* con profundidad 4; IC: intervalo de confianza; NO-P1: no emergencia; P1: emergencia; pt: umbral de probabilidad.

alta (N05, N08, N12) son candidatas a revaluación. El filtro de robustez aplicado sobre la cohorte del año 2024 constituye un ajuste asistido sobre la cohorte de validación, con optimismo cuantificado inferior a 1,1 puntos de sensibilidad, por lo que requiere validación externa independiente. Las respuestas dicotómicas proceden de la recodificación retrospectiva de las opciones del GTTA23 y su rendimiento prospectivo requiere de un estudio externo. La prevalencia de P1REAL en los casos excluidos no es estimable por construcción, al ser el patrón de referencia la variable ausente. Finalmente, dicho patrón –diagnóstico clínico extrahospitalario dicotomizado por consenso Delphi de cinco expertos[9]– no se enlazó sistemáticamente con datos hospitalarios, biomarcadores cardiacos ni eventos cardiovasculares a 30 días, aspecto que deberá abordarse en estudios posteriores.

En conclusión, el sistema propuesto redujo el infratriaje por debajo del 5%, disminuyó los falsos negativos, atenuó la brecha de sensibilidad entre sexos y clasificó deterministamente más de un tercio de las demandas, manteniendo una arquitectura interpretable basada en preguntas dicotómicas. Estas ganancias se obtuvieron a costa de mayor sobretriaje estadístico y de una redistribución relevante hacia la valoración del médico coordinador, por lo que su impacto sobre la carga asistencial, tiempos, uso de recursos y resultados clínicos deberá evaluarse prospectivamente antes de su implementación. La concentración del riesgo residual en subgrupos femeninos específicos, particularmente mujeres de 45-54 años, requerirá monitorización en futuras validaciones.

Conflicto de intereses: Los autores declaran no tener conflicto de intereses en relación con este artículo.

Financiación: Consejería de Salud y Consumo de la Junta de Andalucía, convocatoria CSyC 2024 – Proyectos de Investigación en Salud (referencia: PI-0179-2024, "trIAje"). El organismo financiador no ha participado en el diseño del estudio, la recogida o el análisis de los datos, la interpretación de los resultados ni la redacción del manuscrito.

Responsabilidades éticas: Todos los autores han confirmado el mantenimiento de la confidencialidad y respeto de los derechos de los pacientes en el documento de responsabilidades del autor, acuerdo de publicación y cesión de derechos a EMERGENCIAS. El proyecto trIAje (PI-0179-2024) fue aprobado por el Comité Coordinador de Ética de la Investigación Biomédica de Andalucía (SICEIA-2024-002592).

Artículo no encargado por el Comité Editorial y con revisión externa por pares.

Agradecimientos: Los autores agradecen al personal gestor y sanitario de los Centros Coordinadores de Urgencias y Emergencias del CES-061 su labor asistencial diaria. Asimismo, agradecen a la dirección del Centro de Emergencias Sanitarias 061-Servicio Andaluz de Salud y a la Consejería de Salud y Consumo de la Junta de Andalucía por facilitar los recursos necesarios, así como a los cinco expertos que participaron en el consenso Delphi para la clasificación de diagnósticos de emergencia.

Bibliografía

1 Gulati M, Levy PD, Mukherjee D, Amsterdam E, Bhatt DL, Birtcher KK, et al. 2021 AHA/ACC/ASE/CHEST/SAEM/SCCT/SCMR Guideline for the Evaluation and Diagnosis of Chest Pain. Circulation. 2021;144:e368-454.
2 Lindskou TA, Andersen PJ, Christensen EF, Søvsø MB. More emergency patients presenting with chest pain. PLoS One. 2023;18:e0283454.
3 Byrne RA, Rossello X, Coughlan JJ, Barbato E, Berry C, Chieffo A, et al. 2023 ESC Guidelines for the management of acute coronary syndromes. Eur Heart J. 2023;44:3720-826.
4 Fanaroff AC, Rymer JA, Goldstein SA, Simel DL, Newby LK. Does this patient with chest pain have acute coronary syndrome? The rational clinical examination systematic review. JAMA. 2015;314:1955-65.
5 Vogel B, Acevedo M, Appelman Y, Bairey Merz CN, Chieffo A, Figtree GA, et al. The Lancet women and cardiovascular disease Commission: reducing the global burden by 2030. Lancet. 2021;397:2385-438.
6 Bohm K, Kurland L. The accuracy of medical dispatch—a systematic review. Scand J Trauma Resusc Emerg Med. 2018;26:94.
7 Alotaibi A, Alghamdi A, Reynard C, Body R. Accuracy of emergency medical services (EMS) telephone triage in identifying acute coronary syndrome (ACS) for patients with chest pain: a systematic literature review. BMJ Open. 2021;11:e045815.
8 Graversen DS, Pedersen AF, Christensen MB, Folke F, Huibers L. Factors associated with undertriage and overtriage in telephone triage in Danish out-of-hours primary care: a natural quasi-experimental cross-sectional study of randomly selected and high-risk calls. BMJ Open. 2023;13:e064999.
9 Romero-Olóriz C, Luque-Hernández MJ, López-López C, Caballero-García A, Péculo Carrasco JA, Armengol de La Hoz MÁ. Rendimiento diagnóstico del triaje telefónico para dolor torácico no traumático en el centro coordinador de urgencias y emergencias: estudio poblacional en Andalucía. Emergencias. 2025;37:433-41.
10 Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med. 2022;28:31-8.
11 Michel J, Manns A, Boudersa S, Jaubert C, Dupic L, Vivien B, et al. Clinical decision support system in emergency telephone triage: a scoping review of technical design, implementation and evaluation. Int J Med Inform. 2024;184:105347.
12 Raff D, Stewart K, Yang MC, Shang J, Cressman S, Tam R, et al. Improving triage accuracy in prehospital emergency telemedicine: scoping review of machine learning-enhanced approaches. Interact J Med Res. 2024;13:e56729.
13 Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1:206-15.
14 Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378.
15 Empresa Pública de Emergencias Sanitarias. Guía Clínica de Triaje Telefónico A23: Dolor Torácico del Adulto. Sevilla: Consejería de Salud, Junta de Andalucía; 2018.
16 Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. 2nd ed. Cham: Springer; 2019.
17 Davison AC, Hinkley DV. Bootstrap Methods and Their Application. Cambridge: Cambridge University Press; 1997.
18 Newcombe RG. Interval estimation for the difference between independent proportions: comparison of eleven methods. Stat Med. 1998;17:873-90.
19 Sun X, Xu W. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. IEEE Signal Process Lett. 2014;21:1389-93.
20 Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, 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.
21 Pencina KM, Pencina MJ, D'Agostino RB Sr. What to expect from net reclassification improvement with three categories. Stat Med. 2014;33:4975-87.
22 Hilden J, Gerds TA. A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index. Stat Med. 2014;33:3405-14.
23 Pepe MS, Feng Z, Huang Y, Longton G, Prentice R, Thompson IM, et al. Integrating the predictiveness of a marker with its performance as a classifier. Am J Epidemiol. 2008;167:362-8.
24 Leening MJ, Vedder MM, Witteman JC, 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.
25 Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26:565-74.
26 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.
27 Heidari S, Babor TF, De Castro P, Tort S, Curno M. Sex and Gender Equity in Research: rationale for the SAGER guidelines and recommended use. Res Integr Peer Rev. 2016;1:2.
28 Sugranyes G, Sebastià MC, García-Delgar B, Forcadell E, Coll-Vinent B, en representación del Grup de Treball de Gènere en Salut del Hospital Clínic de Barcelona, et al. Consideraciones respecto al uso de la variable sexo/género en investigación para avanzar hacia una buena praxis: Decálogo PROGÉNEROS. Emergencias. 2023;35:303-5.
29 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.
30 Dawson LP, Nehme E, Nehme Z, Davis E, Bloom J, Cox S, et al. Sex differences in epidemiology, care, and outcomes in patients with acute chest pain. J Am Coll Cardiol. 2023;81:933-45.