Ciencia y Educación
(L-ISSN: 2790-8402 E-ISSN: 2707-3378)
Vol. 7 No. 8.1
Edición Especial VIII 2026

SISTEMA DE PUNTUACIÓN PARA SELECCIONAR PACIENTES EN TRATAMIENTO PERSONALIZADO DEL CÁNCER MAMARIO CON INTELIGENCIA ARTIFICIAL

SCORING SYSTEM FOR PATIENT SELECTION FOR PERSONALIZED BREAST CANCER TREATMENT USING ARTIFICIAL INTELLIGENCE

Autores: [1]Dayana Nicole Ramón Sánchez, [2]Luis Fabián Salazar Garcés.
[1]ORCID ID: https://orcid.org/0009-0002-3441-5880
[2]ORCID ID: https://orcid.org/0000-0002-5128-7211
[1]E-mail de contacto: dramon0276@uta.edu.ec
[2]E-mail de contacto: lf.salazar@uta.edu.ec
Afiliación: [1*]Facultad de Ciencias de la Salud, Universidad Técnica de Ambato, (Ecuador). [2*]Allergy and Acarology Laboratory, Institute of Health Sciences, Federal University of Bahia
Artículo recibido: 1 de Agosto del 2026.
Artículo revisado: 3 de Agosto del 2026.
Artículo aprobado: 3 de Agosto del 2026.
[1]Estudiante de la Universidad Técnica de Ambato, (Ecuador).
[2]Licenciado en Laboratorio Clínico, graduado de la Universidad Técnica de Ambato, (Ecuador). Magíster en Inmunología, graduado de la Universidad de Federal Da Bahia, (Brasil). Doctor en Inmunología Universidad de Federal Da Bahia, (Brasil).

Resumen

Actualmente no existen criterios estandarizados que permitan seleccionar de manera objetiva a las pacientes con cáncer de mama elegibles para su inclusión en sistemas de inteligencia artificial. En este contexto, el estudio tuvo como objetivo desarrollar un sistema de puntuación basado en criterios de elegibilidad clínica, biológica, terapéutica, tecnológica y de disponibilidad de datos, obtener evidencia inicial de su validez de contenido y examinar la factibilidad de su aplicación en expedientes clínicos. Se desarrolló un estudio metodológico e instrumental. El instrumento estuvo conformado por 22 ítems distribuidos en cinco dimensiones, con una puntuación total de 100 puntos. Su contenido fue evaluado por siete expertos, quienes valoraron cada ítem según los criterios de pertinencia, claridad, coherencia y adecuación mediante una escala de cuatro puntos. Para el análisis se calcularon la V de Aiken con su intervalo de confianza del 95 %, los índices de validez de contenido por ítem y por escala, I-CVI y S-CVI, el porcentaje de acuerdo y el coeficiente AC1 de Gwet. Además, las observaciones abiertas formuladas por los expertos fueron examinadas mediante un análisis temático. La factibilidad del instrumento se evaluó mediante su aplicación a 50 expedientes clínicos. Los resultados mostraron una V de Aiken global de 0,983, con un intervalo de confianza del 95 % comprendido entre 0,976 y 0,988, sin que ningún ítem presentara un valor inferior a 0,90. El índice S-CVI/Ave alcanzó un valor de 0,994, mientras que el acuerdo relacionado con la retención de los ítems fue del 98,7 %, con un coeficiente AC1 de Gwet de 0,973. A partir de estos resultados se conservaron los 22 ítems originales. Las observaciones cualitativas de los expertos permitieron elaborar una versión 2.0 del instrumento, en la cual se incorporaron definiciones operativas destinadas a mejorar su precisión y facilitar su aplicación. Durante el estudio piloto se comprobó que las variables clínicas y documentales necesarias para completar el sistema de puntuación se encontraban disponibles y podían extraerse en la mayoría de los expedientes evaluados. En conclusión, el instrumento presentó una adecuada validez de contenido y mostró condiciones favorables de factibilidad para su aplicación inicial. Estos resultados permiten considerar que el sistema de puntuación se encuentra en condiciones de avanzar hacia posteriores procesos de validación psicométrica y hacia la realización de un piloto clínico que permita evaluar su desempeño en contextos reales de atención.

Palabras clave: Neoplasias de la mama, Inteligencia artificial, Validez de contenido, Estudios de validación, Sistemas de puntuación.

Abstract

Currently, there are no standardized criteria for objectively selecting breast cancer patients eligible for inclusion in artificial intelligence systems. In this context, the study aimed to develop a scoring system based on clinical, biological, therapeutic, technological, and data availability eligibility criteria, obtain initial evidence of its content validity, and examine the feasibility of its application in medical records. A methodological and instrumental study was conducted. The instrument consisted of 22 items distributed across five dimensions, with a total score of 100 points. Its content was evaluated by seven experts, who assessed each item according to the criteria of relevance, clarity, coherence, and appropriateness using a four-point scale. For the analysis, Aiken's V with its 95% confidence interval, content validity indices per item and per scale (I-CVI and S-CVI), percentage of agreement, and Gwet's AC1 coefficient were calculated. In addition, the open-ended observations made by the experts were examined through thematic analysis. The instrument's feasibility was assessed by applying it to 50 medical records. The results showed an overall Aiken's V of 0.983, with a 95% confidence interval between 0.976 and 0.988, and no item had a value lower than 0.90. The S-CVI/Ave index reached a value of 0.994, while the agreement regarding item retention was 98.7%, with a Gwet AC1 coefficient of 0.973. Based on these results, the original 22 items were retained. Qualitative feedback from experts led to the development of version 2.0 of the instrument, which incorporated operational definitions designed to improve its accuracy and facilitate its application. During the pilot study, it was verified that the clinical and documentary variables necessary to complete the scoring system were available and could be extracted from most of the records evaluated. In conclusion, the instrument demonstrated adequate content validity and showed favorable conditions for initial application. These results suggest that the scoring system is ready to proceed to further psychometric validation and a clinical pilot study to evaluate its performance in real-world healthcare settings.

Keywords: Breast cancer, Artificial intelligence, Content validity, Validation studies, Scoring systems.

Sumário

Atualmente, não existem critérios padronizados para a seleção objetiva de pacientes com câncer de mama elegíveis para inclusão em sistemas de inteligência artificial. Nesse contexto, o estudo teve como objetivo desenvolver um sistema de pontuação baseado em critérios de elegibilidade clínicos, biológicos, terapêuticos, tecnológicos e de disponibilidade de dados, obter evidências iniciais de sua validade de conteúdo e examinar a viabilidade de sua aplicação em prontuários médicos. Foi realizado um estudo metodológico e instrumental. O instrumento consistiu em 22 itens distribuídos em cinco dimensões, com uma pontuação total de 100 pontos. Seu conteúdo foi avaliado por sete especialistas, que avaliaram cada item de acordo com os critérios de relevância, clareza, coerência e adequação, utilizando uma escala de quatro pontos. Para a análise, foram calculados o V de Aiken com seu intervalo de confiança de 95%, os índices de validade de conteúdo por item e por escala (I-CVI e S-CVI), a porcentagem de concordância e o coeficiente AC1 de Gwet. Além disso, as observações abertas feitas pelos especialistas foram examinadas por meio de análise temática. A viabilidade do instrumento foi avaliada por meio de sua aplicação a 50 prontuários médicos. Os resultados mostraram um V de Aiken geral de 0,983, com um intervalo de confiança de 95% entre 0,976 e 0,988, e nenhum item apresentou valor inferior a 0,90. O índice S-CVI/Ave atingiu o valor de 0,994, enquanto a concordância em relação à retenção de itens foi de 98,7%, com um coeficiente Gwet AC1 de 0,973. Com base nesses resultados, os 22 itens originais foram mantidos. O feedback qualitativo de especialistas levou ao desenvolvimento da versão 2.0 do instrumento, que incorporou definições operacionais elaboradas para aprimorar sua precisão e facilitar sua aplicação. Durante o estudo piloto,

verificou-se que as variáveis clínicas e documentais necessárias para completar o sistema de pontuação estavam disponíveis e podiam ser extraídas da maioria dos registros avaliados. Em conclusão, o instrumento demonstrou validade de conteúdo adequada e apresentou condições favoráveis para a aplicação inicial. Esses resultados sugerem que o sistema de pontuação está pronto para prosseguir com a validação psicométrica e um estudo piloto clínico para avaliar seu desempenho em contextos reais de assistência à saúde.

Palavras-chave: Câncer de mama, Inteligência artificial, Validade de conteúdo, Estudos de validação, Sistemas de pontuação.

Introducción

El cáncer de mama constituye la neoplasia maligna más frecuente y una de las principales causas de mortalidad oncológica en mujeres a nivel mundial. De acuerdo con las estimaciones de GLOBOCAN 2022, se registraron cerca de 2,3 millones de casos nuevos y aproximadamente 670 000 muertes atribuibles a esta enfermedad, con una carga particularmente elevada y en aumento en los países de ingresos bajos y medianos (Bray et al., 2024), donde además persisten importantes desigualdades en la incidencia, el acceso al diagnóstico y los resultados clínicos (Giaquinto et al., 2024; World Health Organization, 2024). A pesar de los avances en el diagnóstico temprano y en las estrategias terapéuticas, la marcada heterogeneidad biológica y molecular del cáncer de mama expresada en subtipos con comportamientos clínicos y pronósticos diferentes y la variabilidad en la respuesta al tratamiento continúan representando un desafío para la personalización de la atención (Bray et al., 2024; Giaquinto et al., 2024). En este escenario, la inteligencia artificial ha emergido como una herramienta prometedora para el análisis de datos médicos complejos y el apoyo a la toma de decisiones clínicas (Esteva et al., 2019; Topol, 2019). El aprendizaje automático y el aprendizaje profundo permiten integrar información clínica, imagenológica y molecular para generar predicciones relacionadas con el diagnóstico, el pronóstico y la respuesta terapéutica (Esteva et al., 2019; Rajpurkar et al., 2022).

En oncología y, específicamente, en cáncer de mama, se han desarrollado modelos aplicados a la detección temprana mediante mamografía, la interpretación de información tumoral y la estratificación del riesgo y del pronóstico, con resultados prometedores en tareas clínicas definidas (Kourou et al., 2015; McKinney et al., 2020). No obstante, la traslación de estas herramientas a la práctica clínica rutinaria enfrenta limitaciones relevantes. El desempeño de los modelos depende de la disponibilidad, calidad, representatividad y estandarización de los datos, así como de las condiciones tecnológicas del entorno asistencial. Además, su rendimiento puede disminuir cuando son implementados en poblaciones o escenarios diferentes de aquellos en los que fueron desarrollados y validados, lo que constituye uno de los principales desafíos para su generalización y adopción clínica (Rajpurkar et al., 2022).

En consecuencia, no todas las pacientes reúnen necesariamente las condiciones clínicas, documentales y tecnológicas requeridas para la utilización apropiada de sistemas basados en inteligencia artificial. Sin embargo, persiste la necesidad de establecer procedimientos explícitos y reproducibles que permitan valorar previamente la idoneidad de los casos y de los datos disponibles para este tipo de aplicaciones. La ausencia de criterios estructurados puede favorecer procesos de selección heterogéneos y limitar la reproducibilidad, transparencia y

posterior generalización de los modelos. Esta necesidad adquiere especial relevancia en contextos con recursos limitados, particularmente en sistemas de salud de países de ingresos bajos y medianos, donde las diferencias en infraestructura tecnológica, digitalización y calidad de los registros clínicos pueden dificultar la implementación de herramientas de inteligencia artificial y profundizar desigualdades preexistentes.

Por ello, disponer de un instrumento que permita estratificar de manera reproducible la elegibilidad de las pacientes podría contribuir a una utilización más racional, segura y transparente de estas tecnologías en oncología. Por lo anterior, el presente estudio tuvo como objetivo desarrollar y establecer la validez de contenido de un sistema de puntuación destinado a evaluar la elegibilidad clínica, biológica, terapéutica, tecnológica y de datos de pacientes con cáncer de mama para su inclusión en sistemas de inteligencia artificial predictiva, así como examinar la factibilidad de su aplicación en el contexto hospitalario ecuatoriano.

Materiales y Métodos

Se realizó un estudio metodológico, de tipo instrumental, orientado al desarrollo y la validación de contenido de un instrumento de puntuación destinado a determinar la elegibilidad de pacientes con cáncer de mama para su inclusión en sistemas de inteligencia artificial (IA) predictiva. El proceso se estructuró en tres fases: construcción del instrumento, validación de contenido por juicio de expertos y piloto de factibilidad de aplicación. El reporte de la validez de contenido se orientó según principios metodológicos consolidados para este propósito (Terwee et al., 2018). El estudio se desarrolló en el marco de la línea de investigación en oncología e IA de la Facultad de Ciencias de la Salud de la Universidad Técnica de Ambato (UTA), en colaboración con el Hospital de Especialidades Carlos Andrade Marín (HECAM), Ecuador. La población objetivo estuvo constituida por mujeres diagnosticadas con cáncer de mama atendidas en el HECAM. Para la fase de validación de contenido, la unidad de análisis fueron los ítems del instrumento, valorados por un panel de expertos.

La construcción del instrumento se desarrolló a partir de una revisión de la literatura orientada a identificar las variables clínicas, biológicas, terapéuticas, tecnológicas y de gobernanza de datos consideradas relevantes para la utilización de modelos de inteligencia artificial (IA) en oncología. Como resultado de este proceso se establecieron cinco dimensiones y un total de 22 ítems. El instrumento fue estructurado sobre una puntuación máxima de 100 puntos, distribuidos entre la sección clínico-biológica, con 40 puntos; la sección terapéutica, con 20 puntos; la dimensión tecnológico-documental, con 15 puntos; la sección de elegibilidad, con 15 puntos; y una escala global de valoración tipo Likert, con 10 puntos. A partir de la puntuación total obtenida se establecieron tres categorías de elegibilidad: alta, para puntuaciones comprendidas entre 80 y 100 puntos; media, entre 60 y 79 puntos; y baja, para puntuaciones inferiores a 60 puntos.

Posteriormente, se realizó la validación de contenido mediante juicio de expertos. Para ello, se conformó un panel integrado por siete especialistas (n = 7), seleccionados de acuerdo con su formación académica y experiencia profesional en las áreas clínica, oncológica o metodológica relacionadas con el propósito del instrumento. El panel estuvo constituido predominantemente por profesionales de medicina familiar y comunitaria, además de una

experta en metodología de la investigación. La experiencia profesional de los participantes osciló entre 7 y 30 años, con una mediana de 11 años. Cada experto evaluó los 22 ítems del instrumento considerando cuatro criterios: pertinencia clínica, claridad, coherencia del sistema y adecuación tecnológica. Para la valoración se utilizó una escala ordinal de cuatro puntos, donde 1 correspondió a "no cumple" y 4 a "cumple totalmente". Asimismo, los jueces emitieron para cada ítem una decisión orientada a mantenerlo, modificarlo o eliminarlo. De manera complementaria, evaluaron el sistema general de puntuación, la suficiencia de cada una de las dimensiones y registraron observaciones cualitativas abiertas destinadas a mejorar la estructura y contenido del instrumento.

Para determinar cuantitativamente la validez de contenido, se calculó para cada ítem el coeficiente V de Aiken (Aiken, 1985), junto con su respectivo intervalo de confianza al 95 %, estimado mediante el método de puntuación propuesto por Penfield y Giacobbi (2004). Adicionalmente, la relevancia de los ítems se evaluó mediante el índice de validez de contenido por ítem (I-CVI) y el índice de validez de contenido de la escala, tanto en su modalidad de promedio (S-CVI/Ave) como de acuerdo universal (S-CVI/UA). Para estos análisis se consideraron favorables las puntuaciones de 3 o 4 otorgadas por los expertos, de acuerdo con los criterios planteados por Polit y Beck (2006). Se establecieron como puntos de referencia un I-CVI ≥ 0,78 y un S-CVI/Ave ≥ 0,90, valores considerados adecuados para paneles conformados por seis o más jueces. La concordancia entre los expertos se estimó mediante el porcentaje de acuerdo y el coeficiente AC1 de Gwet, seleccionado por su comportamiento estadístico favorable en situaciones caracterizadas por niveles elevados de concordancia entre evaluadores (Gwet, 2008). No se utilizó el coeficiente W de Kendall debido a que la varianza entre las puntuaciones emitidas por los jueces fue prácticamente nula, condición que limitaba su pertinencia estadística. Paralelamente, las observaciones abiertas proporcionadas por los expertos fueron analizadas mediante análisis temático, identificándose patrones, sugerencias y aspectos susceptibles de modificación. Los temas emergentes fueron posteriormente traducidos en cambios concretos en la redacción, organización y estructura del sistema de puntuación, dando lugar a la versión 2.0 del instrumento. Todos los análisis estadísticos fueron realizados mediante Python, versión 3.12.

Una vez obtenida la versión 2.0, se procedió a evaluar su factibilidad de aplicación mediante su utilización en una muestra de 50 historias clínicas de pacientes con cáncer de mama atendidas en el HECAM. Los expedientes fueron seleccionados mediante un muestreo estratificado de acuerdo con el estadio clínico, incluyendo pacientes elegibles correspondientes a los estadios I, II y III. La recolección de la información se desarrolló mediante un procedimiento mixto. Las variables clínicas, biológicas, terapéuticas y documentales de naturaleza objetiva fueron extraídas de los expedientes utilizando un procedimiento de extracción estructurada asistido por un modelo de lenguaje de gran escala, siguiendo aproximaciones descritas recientemente para el procesamiento automatizado de información clínica (Chen et al., 2025). La información obtenida mediante este procedimiento fue sometida a verificación manual en una submuestra con la finalidad de comprobar su correspondencia con los datos registrados originalmente en los expedientes.

En contraste, las variables que dependían directamente del juicio del evaluador, entre ellas la calidad del registro clínico, la consideración de un perfil compatible con IA, las condiciones de conectividad, la presencia del consentimiento informado y la escala global de valoración, fueron establecidas para ser aplicadas de manera independiente por dos evaluadores.

Durante esta fase se analizaron principalmente la disponibilidad de las variables requeridas dentro de las historias clínicas y la distribución de las puntuaciones que podían ser extraídas a partir de la información documental disponible. Debido a que la mayor parte del instrumento presenta una estructura de naturaleza formativa, el análisis de consistencia interna se consideró pertinente únicamente para la escala global correspondiente a la Sección G. En relación con los aspectos éticos, el estudio contó con la aprobación del Comité de Ética de la Investigación en Seres Humanos de la Universidad Técnica de Ambato (CEISH-UTA), bajo el código de aprobación 113-CEISH-UTA-2025. La información procedente de las historias clínicas fue manejada de manera anonimizada y codificada, garantizando la protección de la identidad y confidencialidad de los pacientes. Asimismo, el instrumento incorporó específicamente un ítem relacionado con el consentimiento informado para la utilización de datos con fines vinculados con sistemas de inteligencia artificial, en correspondencia con los principios de autonomía, transparencia, responsabilidad y protección de datos sensibles que deben orientar el desarrollo y aplicación de estas tecnologías en el ámbito sanitario.

Resultados y Discusión

El panel estuvo conformado por siete expertos, predominantemente del área de medicina familiar y comunitaria, con una experta en metodología de la investigación y una experiencia profesional entre 7 y 30 años (mediana = 11 años) (Tabla 1).

Tabla 1. *Caracterización del panel de expertos.*

Código	Formación / cargo	Área de especialización	Institución	Años
E1	Médico	Medicina Familiar y Comunitaria	UTA	12
E2	Médico	Medicina Familiar	U. Indoamérica	7
E3	Docente universitario	Medicina	UTA	8
E4	Doctora en Medicina / Docente	Medicina Familiar y Comunitaria	UTA	20
E5	Docente-investigadora (PhD)	Ciencias Pedagógicas (metodología)	UTA	30
E6	Profesor docente	Medicina Familiar y Comunitaria	UTA	8
E7	Médico	Emergencia y Desastres	UTA	11

Fuente: Elaboración propia

La V de Aiken global fue 0,983 (IC 95 %: 0,976–0,988) y ningún ítem descendió por debajo de 0,90; dieciséis de los 22 ítems alcanzaron el valor máximo (V = 1,000). Los seis ítems con valores inferiores —edad al diagnóstico (C1), receptores hormonales (C4), tratamiento recibido (D1), completitud del expediente (E2), calidad del registro (F1) y perfil compatible con IA (F3)— reflejaron exclusivamente las puntuaciones de un evaluador. El índice de validez de contenido a nivel de escala fue elevado (S-CVI/Ave = 0,994; S-CVI/UA = 0,955), con un único ítem por debajo del acuerdo unánime en relevancia (C4; I-CVI = 0,86) (Tabla 2). La concordancia entre jueces sobre la decisión de retención fue casi perfecta (acuerdo = 98,7 %; AC1 de Gwet = 0,973); la W de Kendall no se calculó por resultar inapropiada ante la varianza prácticamente nula entre jueces.

Tabla 2. *Validez de contenido por ítem: coeficiente V de Aiken (IC 95 %), I-CVI y decisión.*

Cód.	Sección	Variable / ítem	V de Aiken (IC 95 %)	I-CVI	Decisión
C1	Clínica	Edad al diagnóstico	0,940 (0,868–0,974)	1,00	Modificar
C2	Clínica	Tipo histopatológico	1,000 (0,956–1,000)	1,00	Mantener
C3	Clínica	Estadio clínico	1,000 (0,956–1,000)	1,00	Mantener
C4	Clínica	Receptores (ER, PR, HER2)	0,952 (0,884–0,981)	0,86	Modificar
C5	Clínica	Ki-67	1,000 (0,956–1,000)	1,00	Mantener
C6	Clínica	Subtipo biológico	1,000 (0,956–1,000)	1,00	Mantener
C7	Clínica	Tamaño tumoral (T)	1,000 (0,956–1,000)	1,00	Mantener
C8	Clínica	Ganglios (N)	1,000 (0,956–1,000)	1,00	Mantener
C9	Clínica	Comorbilidades	1,000 (0,956–1,000)	1,00	Mantener
C10	Clínica	Antecedentes familiares	1,000 (0,956–1,000)	1,00	Mantener
D1	Terapéutica	Tratamiento recibido	0,929 (0,853–0,967)	1,00	Mantener
D2	Terapéutica	Respuesta terapéutica	1,000 (0,956–1,000)	1,00	Mantener
D3	Terapéutica	Adherencia	1,000 (0,956–1,000)	1,00	Mantener
E1	Tecnológica	Historia clínica electrónica	1,000 (0,956–1,000)	1,00	Mantener
E2	Tecnológica	Completitud del expediente	0,929 (0,853–0,967)	1,00	Mantener
E3	Tecnológica	Reportes DICOM	1,000 (0,956–1,000)	1,00	Mantener
E4	Tecnológica	Estudios digitalizados	1,000 (0,956–1,000)	1,00	Mantener
E5	Tecnológica	Conectividad institucional	1,000 (0,956–1,000)	1,00	Mantener
F1	Elegibilidad	Calidad del registro	0,952 (0,884–0,981)	1,00	Mantener
F2	Elegibilidad	Datos longitudinales	1,000 (0,956–1,000)	1,00	Mantener
F3	Elegibilidad	Perfil compatible con IA	0,929 (0,853–0,967)	1,00	Mantener
F4	Elegibilidad	Consentimiento informado IA	1,000 (0,956–1,000)	1,00	Mantener
Global (22 ítems)			0,983 (0,976–0,988)	—	—

Fuente: Elaboración propia.

El sistema de puntuación y clasificación obtuvo validez de contenido satisfactoria (V global = 0,981), con valores ligeramente inferiores en los puntos de corte y la utilidad clínica por sugerencias de fundamentación (Tabla 3). Las cinco dimensiones fueron consideradas suficientes (media ≥ 3,86); dos evaluadores señalaron ajustes en las variables clínicas y uno en las terapéuticas (Tabla 4).

Tabla 3. *Validez de contenido del sistema de puntuación y clasificación.*

Aspecto evaluado	V de Aiken	Valoración
Distribución de pesos por sección (C–G)	0,984	Adecuada
Adecuación de puntos de corte	0,952	Adecuada*
Equilibrio clínico–tecnológico	1,000	Óptima
Utilidad clínica de la clasificación	0,968	Adecuada*
Riesgo de sesgo o exclusión	1,000	Óptima
Global	0,981	—

Fuente: Elaboración propia.

Tabla 4. *Suficiencia por dimensión.*

Dimensión	Suf. media	V de Aiken	¿Ajustes?
Variables clínicas	3,86	0,952	Sí (2/7)
Variables terapéuticas	4,00	1,000	Sí (1/7)
Variables tecnológicas/documentales	4,00	1,000	No
Variables de elegibilidad	4,00	1,000	No
Escala global y clasificación	4,00	1,000	No

Fuente: Elaboración propia.

Las observaciones cualitativas no cuestionaron la pertinencia de las variables, sino que orientaron mejoras concretas que dieron origen a la versión 2.0 (Tabla 5).

Tabla 5. *Temas emergentes del análisis cualitativo y cambios incorporados en la versión 2.0.*

Tema	Observación	Cambio (v2.0)
Coherencia propósito–dimensiones	No se explicitaba lo terapéutico ni lo biológico.	Reformulación del propósito.
Operacionalización de variables	C1, C4, E2, F1 y F3 con alto componente interpretativo.	Definiciones operativas.
Alcance de la elegibilidad	Ambigüedad casos incidentes vs. recidivas.	Especificación del escenario clínico.
Pesos, puntos de corte y sesgo	Fundamentar la ponderación y el sesgo de selección.	Fundamentación y nota de sesgo.
Uso apropiado y gobernanza	La clasificación no es valor clínico ni pronóstico.	Nota de alcance y recalibración.

Fuente: Elaboración propia.

Factibilidad de aplicación. De los 347 registros disponibles del HECAM, 323 correspondieron a cáncer de mama confirmado y 271 resultaron elegibles por estadio clínico I–III. Sobre esta cohorte se aplicó la versión 2.0 a una muestra de 50 pacientes (edad media 54,1 ± 11,8 años; estadio II 54 %, I 26 %, III 20 %; predominio de subtipos luminales). La disponibilidad de las variables extraíbles en los expedientes fue elevada para la mayoría de los componentes, con las mayores brechas en antecedentes familiares (66 %), respuesta terapéutica (62 %) y reportes DICOM (84 %). Los ítems dependientes del juicio del evaluador — conectividad, calidad del registro, perfil compatible con IA, consentimiento y escala global— no constan en el registro y requieren aplicación directa (Tabla 6). El componente extraíble del puntaje fue computable en la totalidad de los casos y, reescalado a 100 puntos, siguió una distribución normal (media 70,1 ± 5,7; Shapiro-Wilk W = 0,994; p = 0,997). La clasificación definitiva de elegibilidad y las medidas de concordancia inter-evaluador se establecerán al completar la aplicación de los ítems de juicio.

Tabla 6. *Disponibilidad de las variables del instrumento en los expedientes (n = 50).*

Variable extraíble de la historia clínica	Disponibilidad
Edad, histopatología, estadio, receptores, subtipo	100 %
Estudios digitalizados, HCE, completitud, datos longitudinales	100 %
Tamaño tumoral (T), ganglios (N)	98 %
Ki-67	96 %
Adherencia	92 %
Reportes DICOM	84 %
Antecedentes familiares	66 %
Respuesta terapéutica	62 %

Fuente: Elaboración propia.

El instrumento alcanzó una validez de contenido alta, con una V de Aiken global de 0,983 y un S-CVI/Ave de 0,994 que superan los umbrales habituales para paneles de seis o más jueces (Aiken, 1985). Los 22 ítems se conservaron. La escasa dispersión entre evaluadores indica que la pertinencia de las variables seleccionadas generó un acuerdo amplio; las puntuaciones menores se concentraron en un solo juez y en ítems con un componente interpretativo apreciable edad al diagnóstico, disponibilidad de receptores, calidad del registro, lo que motivó precisar su definición antes que retirarlos. La elevada

validez de contenido observada en el presente estudio representa un hallazgo metodológicamente relevante, ya que proporciona evidencia inicial de que el sistema de puntuación propuesto logra representar de forma adecuada el constructo de elegibilidad para la implementación de inteligencia artificial en pacientes con cáncer de mama.

De acuerdo con el marco metodológico COSMIN, la validez de contenido es considerada la propiedad de medición más importante durante el desarrollo de un instrumento, puesto que determina si los ítems seleccionados son pertinentes, comprensibles y suficientemente exhaustivos para reflejar el fenómeno que se pretende medir antes de evaluar propiedades como la confiabilidad, la validez de constructo o la capacidad predictiva (Terwee et al., 2018; Mokkink et al., 2025). En consecuencia, los elevados valores obtenidos para la V de Aiken (0,983), el I-CVI y el S-CVI/Ave no deben interpretarse únicamente como un alto nivel de concordancia entre expertos, sino como evidencia de que las dimensiones clínica, biológica, terapéutica, tecnológica y de elegibilidad incorporadas en el instrumento representan de manera integral los elementos actualmente reconocidos como determinantes para la utilización responsable de herramientas de inteligencia artificial en oncología.

Esta interpretación adquiere especial relevancia debido a que el constructo evaluado difiere de los instrumentos clínicos tradicionales; mientras la mayoría de las escalas desarrolladas en oncología buscan cuantificar síntomas, calidad de vida o desenlaces funcionales, el presente sistema pretende estructurar un proceso complejo de toma de decisiones donde confluyen variables clínicas, disponibilidad tecnológica, calidad de los registros y gobernanza de los datos, dominios cuya representación conceptual resulta indispensable para garantizar la utilidad posterior del instrumento. Desde una perspectiva metodológica, la obtención de índices elevados de validez de contenido constituye un requisito indispensable para disminuir el riesgo de error sistemático durante las etapas posteriores de validación. La literatura reciente enfatiza que una adecuada representación conceptual del constructo permite reducir la probabilidad de incorporar variables irrelevantes o de excluir dimensiones esenciales, situaciones que comprometen la interpretación de los resultados aun cuando el instrumento presente posteriormente buenos indicadores de confiabilidad (Mokkink et al., 2025).

En este sentido, diversos autores han señalado que las propiedades psicométricas deben interpretarse como un proceso jerárquico, donde la validez de contenido representa el fundamento sobre el cual se construyen las restantes evidencias de validez. Por ello, iniciar el desarrollo del instrumento mediante una revisión estructurada de la literatura, seguida de un proceso formal de evaluación por expertos, constituye una estrategia metodológica consistente con las recomendaciones internacionales para la construcción de instrumentos aplicables en ciencias de la salud (Polit y Beck, 2006; Terwee et al., 2018; Mokkink et al., 2025; Boateng et al., 2018). Otro hallazgo relevante es que ningún ítem requirió ser eliminado tras la evaluación del panel de expertos. Aunque en ocasiones la eliminación de variables se interpreta como un indicador de rigurosidad metodológica, diversos estudios han demostrado que el verdadero propósito del juicio de expertos consiste en optimizar la calidad conceptual de cada componente más que reducir el número de ítems (Polit y Beck, 2006; Boateng et al., 2018).

Bajo esta perspectiva, la conservación de los 22 ítems observada en el presente estudio no refleja ausencia de crítica por parte del panel, sino la existencia de una adecuada selección inicial sustentada en la evidencia disponible sobre inteligencia artificial aplicada al cáncer de mama. De hecho, las observaciones emitidas por los evaluadores estuvieron dirigidas principalmente hacia la necesidad de precisar definiciones operativas, clarificar criterios de interpretación y fortalecer la fundamentación de algunos puntos de corte, modificaciones que dieron origen a la versión 2.0 del instrumento sin alterar su estructura conceptual. Este tipo de refinamiento incremental ha sido descrito como una de las principales fortalezas del proceso de validación de contenido, ya que permite incrementar la claridad y reproducibilidad del sistema de puntuación antes de iniciar evaluaciones psicométricas de mayor complejidad (Mokkink et al., 2025; Boateng et al., 2018).

Estos resultados también poseen implicaciones relevantes para el desarrollo de herramientas de inteligencia artificial en salud. En los últimos años, múltiples investigaciones han demostrado que el rendimiento de los algoritmos depende no solo de la arquitectura computacional utilizada, sino también de la calidad, completitud, representatividad y estandarización de los datos empleados durante su entrenamiento y validación (Rajpurkar et al., 2022; Collins et al., 2024; Moons et al., 2025). Sin embargo, la mayoría de los estudios publicados se han concentrado en optimizar modelos predictivos sin establecer previamente criterios objetivos que permitan identificar qué pacientes reúnen las condiciones necesarias para ser incluidos en dichos sistemas. Como consecuencia, la selección de la población suele depender de decisiones metodológicas heterogéneas entre investigaciones, dificultando la comparación de resultados, la validación externa y la implementación clínica de los modelos desarrollados. En este contexto, el instrumento propuesto aporta un enfoque novedoso al trasladar el proceso de selección de pacientes desde una decisión implícita del investigador hacia un procedimiento explícito, reproducible y sustentado en criterios previamente validados. Esta aproximación podría contribuir a disminuir la heterogeneidad observada entre estudios, mejorar la calidad de las bases de datos destinadas al entrenamiento de algoritmos y favorecer procesos de implementación más transparentes, reproducibles y acordes con los principios actuales de inteligencia artificial responsable (Collins et al., 2024; Moons et al., 2025; Fletcher et al., 2021).

La elección de los estadísticos de concordancia merece un comentario. Cuando el acuerdo entre jueces es casi total, la varianza tiende a cero y el coeficiente W de Kendall pierde estabilidad. Por esa razón se prefirieron la V de Aiken con intervalos de confianza, los índices I-CVI y S-CVI, y el coeficiente AC1 de Gwet, que se mantiene estable ante acuerdo elevado y no incurre en la paradoja de los índices corregidos por azar (Penfield & Giacobbi, 2004). Informar la concordancia con estadísticos ajustados a la estructura de los datos evita conclusiones distorsionadas por un artefacto métrico. La utilización de múltiples indicadores para evaluar la validez de contenido también representa una fortaleza metodológica del estudio. Aunque la V de Aiken continúa siendo uno de los coeficientes más empleados para cuantificar el grado de acuerdo entre expertos, la evidencia metodológica reciente recomienda complementarla con índices de validez de contenido y medidas de concordancia que permitan obtener una valoración más integral del instrumento. Esta estrategia reduce la

dependencia de un único estadístico y proporciona evidencia convergente sobre la estabilidad de los resultados, especialmente cuando los niveles de acuerdo son muy elevados. En consecuencia, la combinación de la V de Aiken, los índices I-CVI y S-CVI y el coeficiente AC1 de Gwet ofrece una evaluación más robusta del proceso de validación, al integrar medidas de relevancia de los ítems y de concordancia entre evaluadores desde perspectivas metodológicas complementarias, fortaleciendo la confianza en la calidad del instrumento desarrollado (Penfield & Giacobbi, 2004; Polit & Beck, 2006; Gwet, 2008; Terwee et al., 2018; Mokkink et al., 2025).

La validación no se limitó a cuantificar la relevancia de los ítems. El análisis temático de las observaciones abiertas condujo a una segunda versión del instrumento con definiciones operativas, mayor correspondencia entre el propósito declarado y las dimensiones evaluadas, y una justificación explícita de la ponderación y de la gobernanza de los datos. Concebir la validez de contenido como un procedimiento iterativo, y no como un sello de aprobación, es coherente con las guías metodológicas vigentes (Polit & Beck, 2006) y aporta trazabilidad al vínculo entre cada crítica del panel y su corrección.

La incorporación sistemática de las observaciones formuladas por los expertos permitió, además, fortalecer la utilidad práctica del instrumento. En estudios metodológicos recientes se ha señalado que la participación del panel de expertos no solo contribuye a evaluar la pertinencia de los ítems, sino también a identificar ambigüedades, mejorar la redacción y optimizar la definición operativa de las variables antes de su implementación en escenarios reales. Este proceso resulta especialmente importante en instrumentos destinados a apoyar la toma de decisiones clínicas, donde pequeñas diferencias en la interpretación de un criterio pueden afectar la reproducibilidad de las mediciones. En este estudio, las modificaciones introducidas en la versión 2.0 se orientaron precisamente a disminuir esa variabilidad potencial, fortaleciendo la claridad de las definiciones y facilitando una aplicación más uniforme entre distintos evaluadores. Estas mejoras incrementan la factibilidad de utilización del instrumento en futuros estudios multicéntricos y favorecen su adaptación a diferentes contextos asistenciales, sin modificar el constructo originalmente planteado (Terwee et al., 2018; Mokkink et al., 2025; Boateng et al., 2018).

El principal aporte del presente estudio radica en abordar un aspecto escasamente desarrollado dentro de la investigación en inteligencia artificial aplicada al cáncer de mama: la estandarización de la elegibilidad de las pacientes antes de la implementación de modelos predictivos. Durante la última década, la mayor parte de la producción científica en este campo se ha concentrado en optimizar algoritmos destinados al diagnóstico por imágenes, la clasificación histopatológica, la predicción pronóstica o la estimación de la respuesta terapéutica mediante técnicas de aprendizaje automático y aprendizaje profundo (Esteva et al., 2019; Rajpurkar et al., 2022; Kourou et al., 2015; McKinney et al., 2020). Aunque estos avances han demostrado desempeños comparables e incluso superiores a los obtenidos por especialistas en tareas específicas, la atención se ha centrado predominantemente en mejorar el rendimiento matemático de los modelos, mientras que el proceso mediante el cual se seleccionan los pacientes incluidos para el entrenamiento, validación o implementación clínica continúa

siendo poco uniforme y rara vez se describe mediante criterios objetivos (Rajpurkar et al., 2022; Collins et al., 2024; Moons et al., 2025). Esta situación posee implicaciones metodológicas importantes. El desempeño de un algoritmo depende no solo de su arquitectura computacional, sino también de la calidad del conjunto de datos utilizado durante su desarrollo. Numerosos estudios han demostrado que los modelos entrenados con bases de datos incompletas, sesgadas o poco representativas pueden presentar una importante pérdida de rendimiento cuando son aplicados en poblaciones diferentes de aquellas donde fueron desarrollados, fenómeno conocido como dataset shift o pérdida de validez externa (Rajpurkar et al., 2022; Collins et al., 2024; Moons et al., 2025).

Este problema constituye actualmente una de las principales limitaciones para la implementación clínica de la inteligencia artificial, ya que explica por qué muchos algoritmos que muestran excelentes resultados durante su validación interna experimentan reducciones significativas en sensibilidad, especificidad o capacidad discriminativa al incorporarse a hospitales con características epidemiológicas, tecnológicas o demográficas distintas. En consecuencia, la selección adecuada de los pacientes deja de ser una etapa exclusivamente logística para convertirse en un componente fundamental de la calidad metodológica de cualquier sistema basado en inteligencia artificial.

En este contexto, el sistema de puntuación desarrollado aporta una perspectiva complementaria a la predominante en la literatura. En lugar de asumir que todas las pacientes poseen condiciones equivalentes para beneficiarse de modelos de inteligencia artificial, propone un proceso estructurado que integra variables clínicas, biológicas, terapéuticas, tecnológicas y relacionadas con la disponibilidad de datos para establecer un nivel de elegibilidad previamente definido. Este enfoque resulta consistente con las recomendaciones emitidas recientemente por iniciativas internacionales como TRIPOD+AI (Collins et al., 2024) y PROBAST+AI (Moons et al., 2025), las cuales enfatizan que la transparencia en la selección de la población de estudio constituye un requisito indispensable para interpretar adecuadamente el desempeño de los modelos predictivos y facilitar su reproducibilidad. No obstante, dichas guías se orientan principalmente al reporte y evaluación del riesgo de sesgo de los modelos ya desarrollados, mientras que ofrecen escasa orientación acerca de cómo seleccionar de manera estandarizada la población candidata antes del entrenamiento o de la implementación clínica del algoritmo. Desde esta perspectiva, el presente sistema de puntuación no pretende sustituir estas guías internacionales, sino complementarlas mediante una herramienta que operacionaliza uno de los componentes menos desarrollados dentro del proceso de implementación de la inteligencia artificial en salud.

Otro aspecto que diferencia este instrumento de la mayoría de las publicaciones disponibles es su integración explícita de variables relacionadas con la infraestructura tecnológica y la gobernanza de los datos. Tradicionalmente, la elegibilidad de pacientes para estudios clínicos ha dependido casi exclusivamente de características biológicas o terapéuticas; sin embargo, en aplicaciones de inteligencia artificial la disponibilidad de imágenes digitales, la interoperabilidad de la historia clínica electrónica, la completitud de los registros y la calidad de la información constituyen factores igualmente determinantes

para garantizar un funcionamiento adecuado de los algoritmos (Rajpurkar et al., 2022; Chen et al., 2025; Fletcher et al., 2021). La incorporación simultánea de estos dominios reconoce que la implementación de la inteligencia artificial representa un proceso sociotécnico más que exclusivamente tecnológico, donde el rendimiento final depende de la interacción entre las características del paciente, la calidad del ecosistema digital y la gobernanza institucional de los datos.

Esta aproximación amplía la visión tradicional de elegibilidad utilizada en investigación clínica y la adapta a las necesidades actuales de la medicina digital. Desde una perspectiva práctica, la existencia de criterios explícitos y reproducibles para determinar la elegibilidad podría generar beneficios que trascienden el desarrollo de un algoritmo específico. La utilización de un sistema estandarizado facilita la comparación entre estudios, favorece la reproducibilidad metodológica, mejora la calidad de las cohortes utilizadas para entrenamiento y validación, y proporciona un procedimiento transparente para justificar la inclusión o exclusión de pacientes.

Asimismo, podría contribuir a disminuir parte de la variabilidad introducida por decisiones subjetivas durante la selección de participantes, fortaleciendo la comparabilidad entre investigaciones multicéntricas y facilitando procesos posteriores de validación externa. Aunque estas ventajas deberán confirmarse mediante estudios prospectivos, representan una contribución potencialmente relevante para la consolidación de estrategias de inteligencia artificial clínicamente robustas y metodológicamente reproducibles, especialmente en contextos donde la heterogeneidad de los registros continúa limitando la transferencia de modelos entre instituciones (Collins et al., 2024; Moons et al., 2025; Fletcher et al., 2021). La posibilidad de alimentarlo con datos reales se apoyó en la extracción asistida por un modelo de lenguaje, procedimiento cuya utilidad en oncología se ha descrito recientemente (Chen et al., 2025), y en la disponibilidad elevada de la mayor parte de las variables clínicas en los expedientes del HECAM.

La evaluación de factibilidad constituye un complemento fundamental de la validación de contenido, ya que permite determinar si un instrumento metodológicamente sólido puede aplicarse de manera efectiva en escenarios clínicos reales. En el presente estudio, la aplicación piloto demostró que la mayoría de las variables clínicas, biológicas y documentales necesarios para calcular la puntuación de elegibilidad estuvieron disponibles en los expedientes electrónicos del HECAM, lo que respalda la viabilidad operativa del instrumento dentro de un hospital de alta complejidad. Este hallazgo es particularmente relevante porque uno de los principales obstáculos para la implementación de herramientas de inteligencia artificial en salud continúa siendo la heterogeneidad de los registros clínicos y la disponibilidad incompleta de información estructurada, factores que afectan tanto el entrenamiento como la validación y la aplicación clínica de los algoritmos (Rajpurkar et al., 2022; Chen et al., 2025; Hernandez-Boussard et al., 2020).

La elevada disponibilidad de variables relacionadas con el estadio clínico, subtipo molecular, historia clínica electrónica y estudios digitalizados sugiere que la infraestructura de información existente permite recuperar una proporción importante de los datos requeridos sin incrementar

significativamente la carga de trabajo del personal sanitario. Desde una perspectiva práctica, este resultado indica que el instrumento podría incorporarse dentro del flujo habitual de atención utilizando información ya registrada durante la práctica clínica cotidiana, favoreciendo procesos de implementación progresiva y reduciendo la necesidad de recopilar datos adicionales exclusivamente para fines de inteligencia artificial. Esta característica constituye una ventaja potencialmente relevante para hospitales con recursos limitados, donde la sostenibilidad de nuevas herramientas depende en gran medida de su integración con los sistemas de información ya existentes (Rajpurkar et al., 2022; Fletcher et al., 2021; Hernández et al., 2020).

No obstante, el piloto también permitió identificar limitaciones que aportan información valiosa para futuras implementaciones. Las menores tasas de disponibilidad observadas en antecedentes familiares, respuesta terapéutica y reportes DICOM evidencian áreas susceptibles de fortalecimiento dentro de los sistemas de registro institucional. Más que representar una debilidad del instrumento, estos hallazgos reflejan deficiencias inherentes a la calidad de los datos disponibles, situación ampliamente descrita como uno de los principales determinantes del rendimiento de la inteligencia artificial en medicina.

La literatura ha señalado de forma consistente que la ausencia de variables relevantes, la documentación incompleta y la falta de estandarización reducen la capacidad predictiva de los algoritmos y limitan su generalización hacia otras poblaciones, independientemente de la sofisticación del modelo computacional utilizado (Rajpurkar et al., 2022; Hernández et al., 2020). En este sentido, el sistema de puntuación propuesto no solo identifica pacientes potencialmente elegibles para la aplicación de inteligencia artificial, sino que también actúa como un mecanismo indirecto para detectar oportunidades de mejora en la calidad de los registros clínicos.

Un hallazgo particularmente interesante fue la diferenciación entre variables objetivas extraíbles automáticamente desde la historia clínica y aquellas que requieren valoración directa por parte del evaluador, como la calidad del registro, el perfil compatible con inteligencia artificial, la conectividad institucional y algunos componentes relacionados con la gobernanza de los datos. Esta distinción reconoce que la implementación de la inteligencia artificial no depende exclusivamente de la disponibilidad de información clínica, sino también de factores organizacionales y tecnológicos que difícilmente pueden capturarse mediante extracción automatizada.

En consecuencia, el instrumento adopta un enfoque híbrido que combina datos objetivos provenientes del expediente electrónico con el juicio estructurado del evaluador, permitiendo una valoración más completa de la elegibilidad. Esta aproximación resulta coherente con los principios actuales de implementación responsable de inteligencia artificial, que promueven la integración equilibrada entre automatización, supervisión humana y evaluación contextual durante la incorporación de estas tecnologías en la práctica clínica (Moons et al., 2025; Fletcher et al., 2021; World Health Organization, 2021). La experiencia obtenida durante el piloto proporciona evidencia preliminar de que el instrumento puede utilizarse como una herramienta de preparación institucional antes de implementar

modelos de inteligencia artificial. En lugar de limitarse a clasificar pacientes, el sistema permite identificar brechas relacionadas con infraestructura digital, completitud de los registros, disponibilidad de imágenes médicas y procesos de documentación clínica, facilitando intervenciones dirigidas a fortalecer la calidad de los datos antes del desarrollo o adopción de algoritmos predictivos.

Desde esta perspectiva, el instrumento trasciende su función inicial como sistema de elegibilidad y adquiere un potencial valor como herramienta de mejora de la preparación digital institucional, aspecto que ha recibido creciente atención dentro de las estrategias internacionales orientadas hacia una implementación segura, transparente y sostenible de la inteligencia artificial en los sistemas de salud (Fletcher et al., 2021; Hernández et al., 2020; World Health Organization, 2021). En entornos de recursos limitados, condicionar el uso de IA a la calidad de los datos y a las condiciones del entorno tiene una lectura de equidad. Trasladar sin filtro algoritmos entrenados en poblaciones distintas puede amplificar sesgos cuando se aplican a sistemas de salud con registros heterogéneos (Collins et al., 2024). Un criterio de elegibilidad explícito, reproducible y auditable ofrece una vía para racionalizar esa incorporación y para hacer transparentes los motivos por los que un caso se considera apto.

Además de sus implicaciones metodológicas, el sistema de puntuación propuesto puede contribuir a promover una implementación más responsable y transparente de la inteligencia artificial en oncología. En los últimos años, diversos organismos internacionales han advertido que la incorporación de modelos de IA en la práctica clínica debe ir acompañada de mecanismos que reduzcan el riesgo de sesgos, favorezcan la equidad y permitan justificar las decisiones derivadas de estas tecnologías. La utilización de criterios explícitos para determinar la elegibilidad de las pacientes aporta un proceso de selección reproducible y auditable, disminuyendo la posibilidad de decisiones arbitrarias y facilitando la identificación de limitaciones relacionadas con la disponibilidad de información clínica o tecnológica. En contextos como el ecuatoriano, donde persisten diferencias en la digitalización de los servicios de salud y en la calidad de los registros médicos, este tipo de herramientas puede favorecer una incorporación progresiva y contextualizada de la inteligencia artificial, priorizando la seguridad del paciente y el uso racional de los recursos disponibles (Fletcher et al., 2021; World Health Organization, 2021; European Commission, 2019).

La relevancia del instrumento adquiere una dimensión adicional cuando se analiza desde la perspectiva de los sistemas de salud de países de ingresos bajos y medios. La mayor parte de los algoritmos de inteligencia artificial utilizados en oncología han sido desarrollados y validados utilizando bases de datos provenientes de instituciones altamente digitalizadas de Norteamérica, Europa y algunos países asiáticos, donde existen historias clínicas electrónicas consolidadas, protocolos homogéneos de adquisición de imágenes y una elevada disponibilidad de datos estructurados (Rajpurkar et al., 2022; Hernández et al., 2020). En contraste, muchos hospitales latinoamericanos continúan enfrentando limitaciones relacionadas con la interoperabilidad de los sistemas de información, la heterogeneidad de los registros clínicos, la disponibilidad variable de estudios complementarios y diferencias importantes en infraestructura tecnológica. Estas condiciones dificultan la transferencia directa de algoritmos

desarrollados en otros contextos y aumentan el riesgo de obtener desempeños inferiores a los esperados cuando dichas herramientas se implementan en poblaciones distintas a aquellas utilizadas durante su entrenamiento (Fletcher et al., 2021; Hernández et al., 2020; World Health Organization, 2021).

En este escenario, el sistema de puntuación propuesto puede contribuir a reducir parte de esa brecha mediante una evaluación objetiva del grado de preparación de cada paciente y de la información disponible antes de incorporar herramientas de inteligencia artificial. Más que seleccionar pacientes para un algoritmo específico, el instrumento permite identificar si existen las condiciones clínicas, documentales y tecnológicas mínimas para que la aplicación de modelos predictivos resulte metodológicamente apropiada. Esta aproximación puede ser especialmente útil en instituciones que se encuentran en procesos progresivos de digitalización, ya que facilita la identificación de áreas prioritarias para fortalecer la calidad de los registros, optimizar la captura de variables relevantes y mejorar la gobernanza de los datos antes de invertir recursos en el desarrollo o adopción de sistemas de inteligencia artificial.

Desde una perspectiva de salud pública, este enfoque favorece una implementación gradual, segura y adaptada a las capacidades reales de cada institución, promoviendo un uso más eficiente de los recursos disponibles y disminuyendo el riesgo de incorporar tecnologías cuyo desempeño pueda verse comprometido por limitaciones del entorno (Fletcher et al., 2021; World Health Organization, 2021; European Commission, 2019). Asimismo, la experiencia obtenida en el HECAM demuestra que es posible desarrollar herramientas metodológicas de inteligencia artificial desde contextos latinoamericanos utilizando evidencia local y considerando las características propias de los sistemas de salud de la región. Este aspecto resulta particularmente relevante porque la mayoría de los marcos metodológicos disponibles han sido diseñados en países de altos ingresos y su aplicación directa no siempre contempla las diferencias existentes en infraestructura digital, disponibilidad de datos y organización de los servicios sanitarios. La construcción de instrumentos adaptados al contexto regional no solo favorece una implementación más realista de la inteligencia artificial, sino que también contribuye a disminuir la dependencia de modelos desarrollados exclusivamente en poblaciones externas, fortaleciendo la generación de evidencia propia y promoviendo una innovación más equitativa y contextualizada. En este sentido, el presente estudio representa un paso inicial hacia el desarrollo de estrategias metodológicas que permitan incorporar la inteligencia artificial en oncología de manera progresiva, transparente y científicamente fundamentada dentro de los sistemas de salud latinoamericanos.

Aunque los resultados obtenidos proporcionan evidencia sólida sobre la validez de contenido del instrumento, deben interpretarse considerando algunas limitaciones metodológicas. En primer lugar, el proceso de validación se desarrolló mediante un panel de siete expertos con predominio de profesionales de medicina familiar y comunitaria pertenecientes principalmente a una misma institución académica. Si bien este número cumple con las recomendaciones metodológicas para estudios de validez de contenido y permitió alcanzar elevados niveles de concordancia, una mayor diversidad disciplinaria e institucional podría aportar perspectivas complementarias, particularmente desde áreas como oncología médica, cirugía

oncológica, radiología, anatomía patológica, bioinformática e inteligencia artificial clínica. La incorporación de expertos internacionales en futuras etapas permitiría evaluar la pertinencia del instrumento en contextos sanitarios con diferentes niveles de desarrollo tecnológico y fortalecer aún más su validez externa (Polit y Beck, 2006; Terwee et al., 2018; Mokkink et al., 2025; Boateng et al., 2018).

En segundo lugar, la presente investigación constituye únicamente la primera fase del proceso de validación psicométrica. La validez de contenido confirma que los ítems representan adecuadamente el constructo de elegibilidad para inteligencia artificial; sin embargo, no proporciona evidencia acerca de otras propiedades fundamentales como la validez de constructo, la validez de criterio, la reproducibilidad entre evaluadores, la estabilidad temporal o la capacidad discriminativa del sistema de puntuación. En consecuencia, los elevados índices obtenidos en esta etapa no deben interpretarse como evidencia definitiva del desempeño del instrumento durante su aplicación clínica, sino como un requisito metodológico previo para el desarrollo de estudios posteriores orientados a confirmar dichas propiedades de medición (Terwee et al., 2018; Mokkink et al., 2025).

Otra limitación importante radica en que la evaluación de factibilidad se realizó en un único hospital de alta complejidad perteneciente al sistema de salud ecuatoriano. Aunque esta fase permitió demostrar la aplicabilidad operativa del instrumento y documentar una elevada disponibilidad de la mayoría de las variables requeridas, las características particulares del HECAM podrían diferir de las existentes en hospitales con menor grado de digitalización, diferente complejidad asistencial o pertenecientes a otros sistemas sanitarios. Por esta razón, la generalización de los resultados debe realizarse con cautela hasta disponer de estudios multicéntricos que incluyan instituciones con distintos niveles de infraestructura tecnológica y calidad de los registros clínicos. La validación en escenarios heterogéneos permitirá determinar la estabilidad del sistema de puntuación y establecer si los puntos de corte propuestos mantienen su utilidad en diferentes contextos de implementación (Collins et al., 2024; Moons et al., 2025; Fletcher et al., 2021).

Asimismo, algunos componentes del instrumento dependen necesariamente del juicio estructurado del evaluador, como la valoración de la calidad del registro, el perfil compatible con inteligencia artificial y determinados aspectos relacionados con la gobernanza de los datos. Aunque esta característica responde a la naturaleza compleja del constructo evaluado y permite incorporar dimensiones que no pueden extraerse automáticamente desde la historia clínica electrónica, también introduce la posibilidad de variabilidad entre observadores. En consecuencia, será indispensable estimar la concordancia Inter evaluador mediante estudios específicos y desarrollar manuales operativos que estandaricen la aplicación del instrumento antes de su utilización rutinaria en investigación o práctica clínica.

El presente estudio no tuvo como objetivo evaluar el impacto clínico del sistema de puntuación sobre el desempeño de algoritmos de inteligencia artificial ni determinar si la clasificación de elegibilidad mejora efectivamente la precisión, reproducibilidad o capacidad de generalización de modelos predictivos. Estas preguntas requieren investigaciones prospectivas en las cuales el instrumento sea incorporado dentro del proceso

de desarrollo, validación externa e implementación clínica de algoritmos de inteligencia artificial. Confirmar esta relación representará el paso definitivo para establecer la utilidad clínica del sistema de puntuación y determinar su contribución al desarrollo de modelos más robustos, transparentes y aplicables en escenarios reales de atención oncológica (Collins et al., 2024; Moons et al., 2025). Las líneas siguientes comprenden completar el piloto con la aplicación por dos evaluadores independientes y estimar su concordancia, establecer la validez de constructo y de criterio frente a un desenlace clínico, extender el estudio a más centros y recalibrar los puntos de corte a medida que cambien las capacidades tecnológicas y los algoritmos disponibles.

Conclusiones

El presente estudio desarrolló y obtuvo evidencia inicial de validez de contenido de un sistema de puntuación destinado a evaluar la elegibilidad de pacientes con cáncer de mama para su incorporación en sistemas de inteligencia artificial, aportando evidencia metodológica inicial sobre su pertinencia conceptual y factibilidad de aplicación. Los elevados índices de validez de contenido, la conservación de los 22 ítems y la elevada concordancia entre expertos respaldan que las dimensiones clínicas, biológicas, terapéuticas, tecnológicas y relacionadas con la gobernanza de los datos representan de manera adecuada el constructo propuesto.

Más allá de su desempeño psicométrico inicial, el instrumento introduce una aproximación novedosa al proponer un proceso estandarizado para la selección de pacientes antes de la implementación de modelos de inteligencia artificial, un aspecto escasamente abordado en la literatura actual. Su aplicación podría contribuir a mejorar la transparencia, reproducibilidad y calidad metodológica de futuras investigaciones, además de favorecer procesos de implementación más seguros, equitativos y adaptados a las condiciones reales de los sistemas de salud.

Los resultados obtenidos respaldan la continuidad del proceso de validación mediante estudios multicéntricos orientados a evaluar la reproducibilidad, la validez de constructo, la validez de criterio y el impacto del sistema sobre el desempeño de algoritmos de inteligencia artificial en escenarios clínicos reales. En este sentido, el instrumento constituye una base metodológica prometedora para avanzar hacia una incorporación más responsable y científicamente fundamentada de la inteligencia artificial en la atención oncológica.

Agradecimientos

Los autores agradecen a la Dirección de Investigación y Desarrollo (DIDE) de la Universidad Técnica de Ambato. El estudio se desarrolló en el marco del Proyecto de Investigación PFCS59 (resolución UTA-CONIN-2025-0060-R).

Referencias Bibliográficas

Aiken, L. (1985). Three coefficients for analyzing the reliability and validity of ratings. Educational and Psychological Measurement, 45(1), 131–142. https://doi.org/10.1177/0013164485451012

Boateng, G., Neilands, T., Frongillo, E., Melgar, H., & Young, S. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. Frontiers in Public Health, 6, 149. https://doi.org/10.3389/fpubh.2018.00149

Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., & Jemal, A. (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians, 74(3), 229–263. https://doi.org/10.3322/caac.21834

Chen, D., Alnassar, S. A., Avison, K. E., Huang, R. S., & Raman, S. (2025). Large language model applications for health information extraction in oncology: Scoping review. JMIR Cancer, 11, e65984. https://doi.org/10.2196/65984

Collins, G., Moons, K.., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., Boulesteix, A.-L., Camaradou, J., Celi, L., Denaxas, S., Denniston, A., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., ... Logullo, P. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ, 385, e078378. https://doi.org/10.1136/bmj-2023-078378

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. Nature Medicine, 25(1), 24–29. https://doi.org/10.1038/s41591-018-0316-z

European Commission. (2019). Ethics guidelines for trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Fletcher, R, Nakeshimana, A., & Olubeko, O. (2021). Addressing fairness, bias, and appropriate use of artificial intelligence and machine learning in global health. Frontiers in Artificial Intelligence, 3, 561802. https://doi.org/10.3389/frai.2020.561802

Giaquinto, A., Sung, H., Newman, L., Freedman, R., Smith, R., Star, J., Jemal, A., & Siegel, R. (2024). Breast cancer statistics, 2024. CA: A Cancer Journal for Clinicians, 74(6), 477–495. https://doi.org/10.3322/caac.21863

Gwet, K. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. British Journal of Mathematical and Statistical Psychology, 61(1), 29–48. https://doi.org/10.1348/000711006X126600

Hernandez, T., Bozkurt, S., Ioannidis, J., & Shah, N. (2020). MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care. Nature Medicine, 26(9), 1369–1375. https://doi.org/10.1038/s41591-020-1041-y

Kourou, K., Exarchos, T., Exarchos, K., Karamouzis, M., & Fotiadis, D. (2015). Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal, 13, 8–17. https://doi.org/10.1016/j.csbj.2014.11.005

McKinney, S., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G., Darzi, A., Etemadi, M., Garcia, F., Gilbert, F., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., ... Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. Nature, 577(7788), 89–94. https://doi.org/10.1038/s41586-019-1799-6

Mokkink, L., Herbelet, S., Tuinman, P.., & Terwee, C. (2025). Content validity: Judging the relevance, comprehensiveness, and comprehensibility of an outcome measurement instrument—A COSMIN perspective. Journal of Clinical Epidemiology, 185, 111879. https://doi.org/10.1016/j.jclinepi.2025.111879

Moons, K., Damen, J.., Kaul, T., Hooft, L., Andaur Navarro, C., Dhiman, P., van Smeden, M., Riley, R. D., Collins, G. S., Reitsma, J. B., Yang, B., Beam, A. L., Van Calster, B., Celi, L., Denaxas, S., Denniston, A., Ghassemi, M., Heinze, G., Kengne, A. P., ... Wynants, L. (2025). PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods.

BMJ, 388, e082505. https://doi.org/10.1136/bmj-2024-082505

Penfield, R., & Giacobbi, P. (2004). Applying a score confidence interval to Aiken's item content-relevance index. Measurement in Physical Education and Exercise Science, 8(4), 213–225. https://doi.org/10.1207/s15327841mpee0804_3

Polit, D., & Beck, C. (2006). The content validity index: Are you sure what's being reported? Critique and recommendations. Research in Nursing & Health, 29(5), 489–497. https://doi.org/10.1002/nur.20147

Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. Nature Medicine, 28(1), 31–38. https://doi.org/10.1038/s41591-021-01614-0

Seneviratne, M. G., Shah, N. H., & Chu, L. (2020). Bridging the implementation gap of machine learning in healthcare. BMJ Innovations, 6(2), 45–47. https://doi.org/10.1136/bmjinnov-2019-000359

Terwee, C., Prinsen, C.., Chiarotto, A., Westerman, M. J., Patrick, D. L., Alonso, J., Bouter, L. M., de Vet, H. C. W., & Mokkink, L. B. (2018). COSMIN methodology for evaluating the content validity of patient-reported outcome measures: A Delphi study. Quality of Life Research, 27(5), 1159–1170. https://doi.org/10.1007/s11136-018-1829-0

Topol, E. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

World Health Organization. (2021). Ethics and governance of artificial intelligence for health. World Health Organization. https://iris.who.int/handle/10665/341996

World Health Organization. (2024). Breast cancer. https://www.who.int/news-room/fact-sheets/detail/breast-cancer

Declaraciones éticas y editoriales del artículo

Contribución de los autores (Taxonomía CRediT)

Dayana Nicole Ramón-Sánchez: redacción del borrador original; recolección de datos e investigación de campo en el HECAM; curación y análisis de los datos; revisión y edición.

Luis Fabián Salazar-Garcés: conceptualización e idea original; diseño metodológico; análisis formal; revisión crítica y corrección; supervisión y dirección del proyecto.

Declaración de conflicto de intereses

Los autores declaran no tener ningún conflicto de interés económico, profesional o personal que pudiera haber influido en el diseño, la ejecución, el análisis o la publicación de este estudio.

Declaración de financiamiento

Los autores agradecen a la Dirección de Investigación y Desarrollo (DIDE) de la Universidad Técnica de Ambato. El estudio se desarrolló en el marco del Proyecto de Investigación PFCS59 (resolución UTA-CONIN-2025-0060-R). No se recibió financiamiento externo adicional.

Declaración del editor

El editor responsable certifica que el proceso editorial del presente artículo se desarrolló conforme a los principios de integridad científica, transparencia y buenas prácticas editoriales. El manuscrito fue sometido a un proceso de evaluación mediante revisión por pares doble ciego, garantizando la confidencialidad de la identidad de los autores y revisores durante todo el proceso de dictamen académico. Asimismo, el editor declara que el artículo cumple con los criterios científicos, metodológicos y éticos establecidos por la revista.

Declaración de los revisores

Los revisores externos que participaron en la evaluación del presente manuscrito declaran haber realizado el proceso de revisión de manera objetiva, independiente y confidencial. Asimismo, manifiestan que no mantienen conflictos de interés con los autores ni con la investigación evaluada, y que sus observaciones y recomendaciones se fundamentan exclusivamente en criterios científicos, metodológicos y académicos.

Declaración ética de la investigación

Los autores declaran que la investigación se desarrolló respetando los principios éticos de la investigación científica, garantizando la confidencialidad de los datos y el respeto a los participantes del estudio. En los casos en que la investigación involucre seres humanos, los procedimientos deben ajustarse a los principios éticos establecidos en la Declaración de Helsinki y a las normativas institucionales correspondientes.

Declaración sobre el uso de inteligencia artificial

Los autores declaran que el uso de herramientas de inteligencia artificial, en caso de haberse utilizado durante el proceso de investigación o redacción del manuscrito, se realizó únicamente como apoyo técnico para mejorar la claridad del lenguaje o el análisis de información, manteniendo siempre la responsabilidad intelectual sobre el contenido del artículo. Las herramientas de inteligencia artificial no fueron utilizadas como autoras del manuscrito ni sustituyen la responsabilidad académica de los investigadores.

Disponibilidad de datos

Los datos que respaldan los hallazgos están disponibles previa solicitud razonada al autor de correspondencia, sujeto a las restricciones éticas aplicables en virtud del consentimiento informado y la aprobación del CEISH-UTA.