



INTELIGENCIA ARTIFICIAL EXPLICABLE EN EL LABORATORIO CLÍNICO: OPORTUNIDADES Y BARRERAS PARA MEJORAR LA PRECISIÓN DIAGNÓSTICA Y LA SEGURIDAD DEL PACIENTE

EXPLAINABLE ARTIFICIAL INTELLIGENCE IN THE CLINICAL LABORATORY: OPPORTUNITIES AND BARRIERS TO IMPROVING DIAGNOSTIC ACCURACY AND PATIENT SAFETY

Marco Vinicio Flores Blacio¹, Noemí Elizabeth Aroca Rivera², Jeniffer Paola Tonato Alava³, Gabriela Ginely Lojano Ocampo⁴

{mflores1@utmachala.edu.ec¹, naroca1@utmachala.edu.ec², paola97tonato@gmail.com³, gabyginely@gmail.com⁴}

Fecha de recepción: 21/07/2026 / Fecha de aceptación: 10/09/2026 / Fecha de publicación: 16/09/2026

RESUMEN: La inteligencia artificial (IA) se ha incorporado progresivamente al laboratorio clínico como herramienta de apoyo para el procesamiento de datos, interpretación de resultados y toma de decisiones diagnósticas. Sin embargo, la utilización de modelos complejos ha generado preocupaciones relacionadas con la transparencia, interpretabilidad y posibilidad de supervisar las decisiones algorítmicas, particularmente cuando estas pueden influir sobre la atención y seguridad del paciente. En este contexto, la inteligencia artificial explicable (XAI) ha adquirido relevancia al proporcionar mecanismos destinados a facilitar la comprensión de las predicciones realizadas por los modelos de IA. El objetivo de esta revisión fue analizar las oportunidades y barreras asociadas con la aplicación de XAI en el laboratorio clínico, considerando su posible contribución a la precisión diagnóstica y a la seguridad del paciente. Se realizó una revisión bibliográfica descriptiva y analítica de literatura científica publicada principalmente entre enero de 2021 y septiembre de 2026, mediante búsquedas en PubMed/MEDLINE, Scopus, Web of Science, ScienceDirect, SpringerLink y Google Scholar. Se consideraron estudios relacionados con inteligencia artificial explicable, medicina de laboratorio, diagnóstico, precisión, validación, seguridad del paciente e interpretabilidad. La evidencia revisada mostró aplicaciones potenciales de XAI en las fases preanalítica, analítica y postanalítica, así como en hematología, bioquímica, microbiología, inmunología y patología digital. Entre las principales oportunidades se identificaron una mayor transparencia de los modelos, trazabilidad de las predicciones, apoyo a la identificación de errores y fortalecimiento de la supervisión profesional. Las principales barreras correspondieron a la calidad y heterogeneidad de los datos, interoperabilidad, validación externa, sesgos, protección de la

¹Universidad Técnica de Machala, Ecuador, ORCID: <https://orcid.org/0009-0006-4381-7776>

²Universidad Técnica de Machala, Ecuador, ORCID: <https://orcid.org/0009-0004-0943-0573>

³Investigador independiente, Salcedo, Ecuador, ORCID: <https://orcid.org/0009-0001-5827-9257>

⁴Investigador independiente, Ecuador, ORCID: <https://orcid.org/0009-0001-0061-2694>



información, ausencia de criterios estandarizados de explicabilidad y necesidad de capacitación profesional. Se concluyó que XAI constituyó una herramienta complementaria con potencial para favorecer la precisión diagnóstica y la seguridad del paciente, siempre que su implementación se acompañara de validación rigurosa, monitorización continua, supervisión humana y criterios éticos y regulatorios adecuados.

Palabras clave: *inteligencia artificial explicable, laboratorio clínico, inteligencia artificial, precisión diagnóstica, seguridad del paciente, aprendizaje automático, interpretabilidad*

ABSTRACT: Artificial intelligence (AI) has been progressively incorporated into clinical laboratories as a tool to support data processing, result interpretation, and diagnostic decision-making. However, the use of complex models has raised concerns regarding transparency, interpretability, and the ability to supervise algorithmic decisions, particularly when these decisions may influence patient care and safety. In this context, explainable artificial intelligence (XAI) has gained relevance by providing mechanisms intended to facilitate the understanding of predictions generated by AI models. This review aimed to analyze the opportunities and barriers associated with the application of XAI in the clinical laboratory, considering its potential contribution to diagnostic accuracy and patient safety. A descriptive and analytical bibliographic review was conducted of scientific literature published mainly between January 2021 and September 2026 through searches in PubMed/MEDLINE, Scopus, Web of Science, ScienceDirect, SpringerLink, and Google Scholar. Studies addressing explainable artificial intelligence, laboratory medicine, diagnosis, accuracy, validation, patient safety, and interpretability were considered. The reviewed evidence showed potential applications of XAI across the preanalytical, analytical, and postanalytical phases, as well as in hematology, biochemistry, microbiology, immunology, and digital pathology. The main opportunities included increased model transparency, traceability of predictions, support for error identification, and strengthened professional oversight. The main barriers involved data quality and heterogeneity, interoperability, external validation, bias, information protection, the lack of standardized explainability criteria, and the need for professional training. It was concluded that XAI represented a complementary tool with potential to support diagnostic accuracy and patient safety, provided that its implementation was accompanied by rigorous validation, continuous monitoring, human oversight, and appropriate ethical and regulatory criteria.

Keywords: *explainable artificial intelligence, clinical laboratory, artificial intelligence, diagnostic accuracy, patient safety, machine learning, interpretability.*



INTRODUCCIÓN

La inteligencia artificial (IA) ha ido adquiriendo una relevancia creciente en el ámbito de las ciencias de la salud, gracias a su capacidad para procesar enormes volúmenes de información en las diferentes etapas del proceso diagnóstico, identificar patrones complejos y atender a elaborar predicciones apropiadas. En el laboratorio clínico, estas capacidades son especialmente relevantes porque el laboratorio clínico genera un elevado número de datos durante la etapa analítica, preanalítica y posanalítica. Los sistemas de aprendizaje automático pueden llegar a ser de utilidad en el procesamiento e interpretación de los resultados, en clasificación de las muestras, identificación de los patrones de asociación entre enfermedades y en apoyo a la toma de decisiones (1). En este escenario, la utilización de herramientas de IA supondrá no solo una transición tecnológica sino también un proceso gradual en la forma cómo se obtienen, interpretan y utilizan los resultados del laboratorio clínico en beneficio del paciente.

La creación de modelos de aprendizaje automático o aprendizaje profundo permite obtener resultados muy altos en determinadas tareas diagnósticas, pero gran parte de estos modelos tiene una estructura muy compleja para poder saber cómo se ha llegado a una determinada predicción, razón esta por la cual reciben a menudo la denominación de "cajas negras", lo que constituye uno de los principales impedimentos para su uso en el mundo clínico, donde las decisiones tienen que ser comprensibles, verificables y justificables.

Una revisión sistemática publicada en 2022 destacó que la falta de transparencia sigue siendo uno de los principales inhibidores para la adopción de sistemas de IA en salud, sobre todo en el caso de que los profesionales tengan que entender cómo los factores de entrada llevan a una determinada predicción (2). Por ello, la inteligencia artificial explicable (Explainable Artificial Intelligence, XAI) ha suscitado un especial interés como método para ofrecer información interpretable acerca del funcionamiento y de las predicciones de los modelos. La XAI se puede definir como un conjunto de procedimientos y técnicas diseñadas para ayudar a entender los/as resultados obtenidos por sistemas de IA.

Las aproximaciones que se emplean son las de métodos de importancia de variables, explicaciones locales y globales, modelos reemplazo, reglas interpretables y métodos que hacen hincapié en particularidades de los datos. La XAI no pretende necesariamente sustituir modelos complejos por algoritmos simples, sino que su objetivo es proporcionar explicaciones que permiten a los/as usuarios/as saber por qué se obtuvo un determinado resultado (3). De una revisión de 2024 se desprende que la transparencia, confianza, integración en flujos de trabajo y la satisfacción de requisitos reglamentarios son aspectos clave para la integración responsable de la XAI en la atención sanitaria (4).

La medicina de laboratorio tiene particular interés por esta problemática en particular. Los resultados obtenidos a partir de un análisis constituyen una fuente principal e importante de información sobre el diagnóstico, el seguimiento y el pronóstico de un elevado número de



enfermedades; por lo que los errores en su interpretación pueden incidir directamente en las decisiones clínicas que se puedan llegar a tomar. En 2024 se publicaba una revisión donde se describían aplicaciones de IA en las distintas etapas del proceso de un análisis clínico, indicando el alto potencial de estas tecnologías para cambiar la Historias Clínica de un laboratorio (5). Igualmente, pero en otro contexto, los modelos de machine learning sobre datos de laboratorio pueden usar muchas variables a la vez para articular sistemas de apoyo a la decisión clínica, aunque su efectividad está ligada a la calidad de los datos, a la formación, a la validación y a la correcta integración de los modelos dentro de los procesos asistenciales (6).

La explicabilidad, por lo tanto, adquiere una dimensión de seguridad del paciente. Una predicción que presenta un alto nivel de exactitud estadística no implica necesariamente que pueda utilizarse de una forma segura desde el punto de vista clínico si el profesional no puede determinar las variables que han influido en ella, saber cuáles son sus restricciones o identificar aquellas situaciones en las que el modelo puede errar. Recientes revisiones han puesto de manifiesto que la XAI todavía conlleva importantes inconvenientes para ser trasladada a la asistencia sanitaria en tiempo real y que la evaluación de una explicación no debería limitarse a saber si el modelo puede llegar a producir información explicativa, sino que también debería alcanzar si dicha información es comprensible, de utilidad y clínica (7).

Además, la escasa evidencia disponible permite saber que algunos modelos explicativos pueden derivar en simplificaciones o en contradicciones lo que requiere que se interpreten con mucho cuidado los resultados que se obtienen. La necesidad de la transparencia también se relaciona con la responsabilidad profesional y con la gobernanza respecto a las tecnologías sanitarias. Los temas relacionados con la supervisión de las personas, el resguardo de la privacidad de los datos, los sesgos, la responsabilidad ante los errores y el seguimiento del proceso decisional son cuestiones a tener en cuenta cuando se usa una herramienta de IA para la toma de decisiones diagnósticas.

En este sentido, los estudios que han ido surgiendo sobre la IA que se aplica a los algoritmos diagnósticos han ido evidenciando que el uso de estas tecnologías genera interrogantes éticos y jurídicos en lo que se refiere a la atribución de responsabilidad en el caso en que un sistema automatizado esté en la base de una decisión errónea (8). En el ámbito de la patología y el laboratorio, se ha puesto también de manifiesto que el uso de la IA debería ir acompañado de criterios de gestión de datos, de validación y de uso responsable de los sistemas (9).

La XAI puede tener además un papel que favorezca una relación más cooperativa entre las/los profesional/es y los sistemas automatizados. Las investigaciones actuales sugieren que las explicaciones pueden proveer al usuario de información extra para valorar una predicción, detectar contradicciones y usar la IA como herramienta de soporte, en lugar de reemplazar el criterio profesional (10). Esto resulta importante en el laboratorio, donde la interpretación de los resultados depende no sólo de los valores en sí mismos, sino que también del contexto clínico,



de las características del paciente, de la calidad de la muestra, de los intervalos de referencia y de otros elementos que pueden variar el significado clínico de un resultado.

A pesar del interés creciente en los estudios relacionados con IA y XAI en medicina, sigue existiendo una fisura particular respecto a su aplicación en el laboratorio. La mayor parte de la literatura existente ha estado centrada en imágenes médicas, oncología, la predicción, o sistemas generales de soporte a la decisión. Por el contrario, hay menos integración de la evidencia en los procesos de medicina de laboratorio y sus implicaciones en la precisión diagnóstica y la seguridad del paciente (11). Por tanto, se hace necesario que se lleve a cabo un análisis conjunto de las oportunidades que presenta la XAI y de las barreras que pueden limitar su aplicación en los entornos reales.

Por consiguiente, este documento tenía como objetivo analizar la evidencia científica reciente relacionada con la inteligencia artificial explicable en el laboratorio clínico, identificando sus mayores oportunidades para mejorar la precisión diagnóstica y la seguridad del paciente; y las barreras técnicas, clínicas, éticas y reglamentarias que pueden condicionar su aplicación. La revisión que se realizó ofreció la posibilidad de integrar los principales avances que se describen en la literatura y poder establecer los aspectos que se deberían tener en cuenta para permitir una XAI válida, transparente, validada y con supervisión profesional.

MATERIALES Y MÉTODOS

Tipo de método de investigación

Se llevó a cabo una revisión bibliográfica del tipo descriptivo-analítico orientada al análisis de la evidencia científica existente sobre el uso de la inteligencia artificial explicable (Explainable Artificial Intelligence, XAI) en el laboratorio clínico, dedicando especial interés a sus oportunidades para mejorar la precisión diagnóstica y a su contribución a la seguridad del paciente. La revisión se dirigió a la literatura científica reciente en el ámbito del desarrollo, la validación, la interpretabilidad, la transparencia y el uso clínico de modelos de inteligencia artificial en el ámbito del diagnóstico y el laboratorio.

Población o muestra

La búsqueda bibliográfica se realizó en bases de datos y plataformas de literatura científica de uso habitual en salud y ciencias biomédicas, lo que incluye PubMed/MEDLINE, Scopus, Web of Science, ScienceDirect, SpringerLink y Google Scholar. Los artículos preferidos son los que han sido publicados entre enero de 2021 y septiembre de 2026, con el fin de identificar evidencias recientes acerca de la evolución de la IA explicativa y sus potenciales vías de incorporación en el entorno clínico. Solo de forma complementaria se han considerado las publicaciones más



antiguas en los momentos en que han sido necesarias para poder fundamentar los conceptos más importantes o los antecedentes metodológicos que son relevantes para la comprensión de la XAI.

La estrategia de búsqueda se definió mediante la combinación de los términos de búsqueda de la inteligencia artificial, de aprendizaje automático, de inteligencia artificial explicativa, de laboratorio clínico, de diagnóstico, de precisión diagnóstica, de interpretabilidad, de transparencia y de seguridad del paciente. Se utilizaron entre los términos empleados "explainable artificial intelligence", "XAI", "artificial intelligence", "machine learning", "clinical laboratory", "laboratory medicine", "diagnostic accuracy", "clinical diagnosis", "patient safety", "interpretability" y "transparency". Los descriptores se combinaron mediante operadores booleanos AND y OR, adaptando las expresiones de búsqueda a las características concretas de cada base de datos.

Los artículos aceptados como criterios de inclusión fueron artículos científicos originales, revisiones sistemáticas, revisiones narrativas y estudios de evaluación de los modelos desarrollados en relación a la inteligencia artificial publicados en revistas científicas que cuenten con revisión por pares; trabajos que se relacionen específicamente con aplicaciones diagnósticas de la inteligencia artificial; trabajos que aborden de manera clara la explicabilidad o interpretabilidad de los modelos; obras que reflexionen sobre precisión, fiabilidad, validación, utilidad clínica o seguridad de los sistemas que se basan en inteligencia artificial y trabajos que estuviesen disponibles en español o inglés. Se les concedió especial importancia a las investigaciones publicadas en relación a los últimos cinco años, así como a aquellos artículos ya que provenían de revistas indexadas a bases de datos internacionales.

Se excluyeron documentos duplicados, publicaciones que no presentaban relación directa con el objetivo de revisión, estudios que se ocupaban exclusivamente de aplicaciones de inteligencia artificial desde perspectivas administrativas, trabajos que no presentaban un componente diagnóstico y/o clínico; documentos de opinión que no presentaban suficiente soporte científico; literatura no arbitrada y publicaciones cuyo contenido completo impidiera identificar la aplicación, la metodología o los principales resultados en relación con la XAI y otras publicaciones en donde la inteligencia artificial sólo fuera una ayuda, sin relación con procedimientos de interpretación, predicción o apoyo diagnóstico.

El proceso de selección se dividió en las siguientes cuatro etapas: identificación, cribado, evaluación de elegibilidad e inclusión.

Entorno

La investigación se desarrolló en un **entorno documental y científico**, centrado en la literatura internacional relacionada con la aplicación de la inteligencia artificial explicable en el laboratorio clínico, el diagnóstico y la seguridad del paciente. La búsqueda se efectuó mediante bases de datos científicas especializadas en salud, ciencias biomédicas y literatura académica.



Mediciones

La extracción de la información se realizó tomando las siguientes categorías como referencia: año de publicación, autores, país o región, tipología de estudio, área clínica o de laboratorio, tipología de modelo de inteligencia artificial, técnica de explicabilidad utilizada, objetivo diagnóstico, variables o datos analizados, indicadores de rendimiento, mecanismos de validación, transferibilidad clínica, beneficios esperados y principales limitaciones o barreras para su uso.

La información extraída se estructuró mediante matrices de análisis para realizar comparaciones entre estudios y determinar tendencias compartidas.

El análisis fue de naturaleza cualitativa, descriptivo y comparativo. Los resultados se agrupaban en categorías temáticas de interés relacionadas con la utilidad de la XAI en el diagnóstico, su posible aporte a la precisión diagnóstica, la transparencia de los modelos, la seguridad del paciente, las principales técnicas para lograr la explicabilidad y las barreras en términos técnicos, clínicos, éticos y regulatorios.

Análisis estadísticos

Debido a la heterogeneidad previsible de los diseños, poblaciones, conjuntos de datos y métricas utilizadas en los estudios, no se planteó un metaanálisis estadístico, sino una síntesis narrativa de la evidencia disponible.

La calidad y relevancia de los estudios se halló considerando: la claridad del objetivo, la idoneidad metodológica, la descripción de los datos utilizados, las características del modelo de inteligencia artificial, el procedimiento de validación, la presentación de las métricas de rendimiento de los modelos, la explicación del mecanismo de interpretabilidad y la posibilidad de trasladar los resultados al contexto clínico. Considerar este aspecto permitió distinguir los resultados obtenidos en condiciones experimentales de la mayoría de la evidencia más cercana a la práctica clínica.

RESULTADOS

Aplicaciones de la inteligencia artificial explicable en el laboratorio clínico

La revisión realizada por los autores permitió constatar que la IA en medicina de laboratorio se ha desarrollado progresivamente desde aplicaciones de una naturaleza más experimental a una serie de diferentes aplicaciones dirigidas a procesos relacionados con la gestión, interpretación y el uso de resultados diagnósticos. Una revisión publicada por Hou y colaboradores presentó aplicaciones de IA en las distintas etapas de proceso de laboratorio clínico, y mostraron que las aplicaciones de IA pueden contribuir a optimizar el flujo de trabajo del laboratorio como



un todo aunque mostraba limitaciones en relación a costes, a precisión y a necesidad de revisión manual de los resultados generados por los modelos (12).

La evidencia mostrada nos hizo observar que las aplicaciones de IA pueden ser clasificadas en función del proceso del laboratorio que considera la fase(s) preanalítica, analítica y posanalítica. Los algoritmos pueden utilizarse durante la fase preanalítica para hallar factores asociados con errores en la identificación y clasificación de muestras en condiciones de transporte inadecuadas, así como otros factores que pueden alterar la calidad del resultado. Lippi et al. afirman que existen oportunidades de automatizar y mejorar diferentes procesos preanalíticos utilizando inteligencia artificial, particularmente aquellos que se caracterizan por la producción de grandes volúmenes de datos y la necesidad de lograr patrones que pueden perderse en procedimientos ordinarios (13).

Los modelos de aprendizaje automático pueden ser empleados para identificar patrones en los resultados de laboratorio, y clasificar la información dando soporte a procedimientos automatizados en la fase analítica. La revisión de la literatura realizada por Kim et al. identificó 144 estudios de aplicación del aprendizaje automático en medicina de laboratorio a partir de 779 registros recuperados, evidenciándose una heterogeneidad de muestras, tipos de datos, modelos y objetivos. Los modelos más utilizados están relacionados con las redes neuronales convolucionales, los perceptrones multicapa y los algoritmos basados en árboles, los modelos del caso que más se han utilizado y los modelos que principalmente estaban relacionados con las características de los datos analizados (14).

La Inteligencia Artificial (IA) cobra especial interés en la fase posanalítica para ayudar a la interpretación integrada de los resultados y la generación de predicciones para ayudar a la decisión clínica. En la actualidad no se analiza cada uno de los parámetros de forma independiente: los algoritmos analizan las múltiples variables a la vez y establecen vínculos que ayudan en la clasificación o predicción de determinadas condiciones clínicas: este interés se hace especialmente interesante cuando los resultados de laboratorio se convierten en el fetiche de interés junto la información demográfica, los antecedentes clínicos, etc., o cualquier otra variable disponible en los sistemas electrónicos de salud. La introducción de XAI añade un elemento más a estas posibilidades. Mientras que los modelos tradicionales del aprendizaje automático son capaces de realizar una predicción sin ofrecer, por otro lado, una explicación suficientemente comprensible o suficiente en cuanto a su predicción, las técnicas de la XAI buscan por el contrario saber cuáles son los factores que han contribuido a dicha predicción. En este sentido, la revisión sistemática llevada a cabo por Loh et al. ha analizado un total de 99 artículos publicados en revistas Q1 y demuestra que existe un uso amplio de métodos como SHAP, LIME, Grad-CAM o Layer-wise Relevance Propagation (15).

En el laboratorio clínico, tener esta capacidad puede ser especialmente útil en el caso de que el modelo se haya utilizado para realizar una predicción que debe ser revisada por el profesional. Por ejemplo, un resultado calificado como potencialmente anómalo podría contar a su vez con



una explicación que viene dada por la contribución relativa que aportan algunos de los parámetros a dicha clasificación; argumentar o explicar a posteriori una predicción basada en la contribución de estos parámetros a la clasificación puede ayudar a conocer qué parámetros son los que han tenido mayor peso en la misma. Si bien es cierto que la explicación algorítmica no debe ser entendida como un sustituto del juicio del profesional, sino como una información que puede ayudar a que el resultado final pueda ser revisado y evaluado críticamente. Asimismo, la evidencia observada ha puesto de manifiesto una importante interrelación entre la explicabilidad y la confianza profesional. Por su parte, Ali et al. afirman que la interpretabilidad tiene un significado especial para el área médica, ya que una predicción errónea puede tener un impacto relevante para la clínica. Por consiguiente, algunas técnicas como SHAP y LIME pueden aportar información en relación con la contribución de determinadas características para obtener una predicción, si bien persisten dificultades en relación con la interpretación de las explicaciones (16).

La patología puede ser otro de los campos muy cercanos a la medicina de laboratorio en el que la XAI ha cobrado especial protagonismo. Klauschen et al. Argumentaron que los modelos de IA pueden procesar datos histológicos y moleculares complejos para la clasificación de diversas enfermedades, la cuantificación de biomarcadores y la predicción de resultados clínicos, pero también han indicado que el carácter de "caja negra" de determinados modelos son un obstáculo para su uso. Las técnicas XAI se proponen a la manera de una opción para hacer más transparentes los procesos de decisión de este tipo de sistemas (17).

Una característica relevante de la reciente evidencia es que la incorporación de XAI no se debe concentrar únicamente en el desarrollo de algoritmos. Para que una explicación posea utilidad clínica es necesario que ésta sea presentada de forma entendible y contextualizada para el usuario. Estudiosos como Cálem et al. en su análisis de 42 interfaces explicables en el ámbito de la salud, remarcaron la relevancia dada por el modo en que la información generada por XAI es expuesta al usuario, así como una serie de propiedades de diseño de interfaces explicables (18).

Por otra parte, la real puesta en marcha de sistemas de IA en laboratorios clínicos está por debajo de las expectativas marcadas por la investigación experimental. Una encuesta europea hecha por Cadamuro et al. tuvo 195 respuestas válidas y mostró que solo un 25.6 % de los laboratorios estudiados presentan proyectos de IA. Las principales barreras han sido consecuencias de la falta de infraestructura digital, acceso a datos completos y la falta de competencias en IA (19).

Estos resultados muestran que el avance tecnológico no es una garantía suficiente de implementación. La efectividad de una herramienta XAI dependerá de la calidad de los datos, el rendimiento del modelo, la validación local, interoperabilidad, capacitación del personal o la compatibilidad con los procesos del laboratorio. En una revisión sobre validación, implementación y monitorización de soluciones de aprendizaje automático en los laboratorios clínicos, precisamente se ha remarcado la necesidad de evaluar el diseño de estudio, la ingeniería



de datos, la interoperabilidad, las métricas de rendimiento, la generalización, la equidad algorítmica y la explicabilidad antes de la puesta en marcha de una solución (20).

Tabla 1. Principales aplicaciones de ia y xai relacionadas con el laboratorio clínico.

Área de aplicación	Función de la IA	Aporte potencial de XAI	Utilidad clínica esperada	Principal desafío
Fase preanalítica	Identificación y clasificación de factores asociados con errores	Identificación de variables que influyen en la predicción de riesgo	Reducción de errores preanalíticos	Calidad y disponibilidad de datos
Hematología	Clasificación y reconocimiento de patrones celulares	Visualización de características responsables de la clasificación	Apoyo a la identificación de alteraciones	Validación con poblaciones diversas
Bioquímica clínica	Análisis multivariable de resultados	Determinación de parámetros con mayor contribución a una predicción	Apoyo a la interpretación integrada	Complejidad de las relaciones entre variables
Microbiología	Clasificación y reconocimiento de patrones	Identificación de variables o características relevantes	Apoyo a clasificación diagnóstica	Variabilidad de muestras y microorganismos
Inmunología	Reconocimiento de patrones en perfiles analíticos	Explicación de variables asociadas con una clasificación	Apoyo a interpretación de perfiles complejos	Heterogeneidad clínica
Patología digital	Análisis de imágenes y clasificación	Localización de regiones o características relevantes	Apoyo diagnóstico y cuantificación	Validación, estandarización e interoperabilidad
Fase posanalítica	Predicción y apoyo a la decisión	Explicación de factores que determinan la predicción	Mejor comprensión del resultado generado por IA	Integración con criterio profesional
Control del proceso	Detección de patrones anómalos	Identificación de variables relacionadas con la anomalía	Detección temprana de problemas	Falsos positivos y necesidad de revisión

Técnicas de explicabilidad más relevantes

La revisión realizada permitió observar el uso de diversas técnicas de XAI que pueden agruparse, a grandes rasgos, según el ámbito de la explicación y la relación con el modelo. Las explicaciones pueden ser globales, es decir, buscan describir el comportamiento global de un modelo, o locales, ya que explican un resultado concreto. La distinción es importante en medicina de laboratorio, ya que una explicación para intentar describir el comportamiento de un algoritmo no necesariamente responde a la necesidad del experto revisando un resultado en concreto.



Entre las técnicas más demandadas se encuentra SHAP (SHapley Additive exPlanations), basada en valores ingenieros a partir de la teoría de juegos cooperativos, para poder estimar la contribución de las características a la predicción. En el ámbito de la inteligencia artificial explicativa (XAI), halla mención también LIME (Local Interpretable Model-Agnostic Explanations), que ofrece una aproximación interpretable en torno a una predicción en particular; Loh et al. exaltaron ambas metodologías como las más utilizadas entre las estrategias de la XAI dentro de la literatura sanitaria junto con el Grad-CAM, LRP y otros métodos (15). En el caso de las aplicaciones relacionadas con imágenes, el Grad-CAM produce mapas que indican las zonas de una imagen que han contribuido para una predicción del modelo, lo cual resulta de especial valor en la patología digital, donde la explicación puede relacionarse fácilmente con zonas celulares o tisulares, lo que también fue señalado por Klauschen et al. al considerar que las técnicas explicativas pueden ayudar a salvar algunas de las limitaciones asociadas con la opacidad que presentan los modelos de deep learning aplicados a la práctica diagnóstica en patología (17).

Por tanto, la mera existencia de una explicación no quiere decir que esta sea correcta, suficiente o clínicamente válida. Una revisión sistemática de las propiedades y eficacia de las explicaciones de la XAI reportó un total de seis estudios elegibles de un total de 882 artículos recuperados, lo que pone de manifiesto una significativa escasez de estudios empíricos específicamente diseñados para evaluar la efectividad de las explicaciones. Además, se observó cómo recaen diferencias en la manera de valorar el apartado de fidelidad, la validez explicativa, la interpretabilidad, plausibilidad, satisfacción del usuario, confianza y la posibilidad de que lo explicado sea corregido (21).

Este hallazgo es una pieza clave para el laboratorio clínico. No bastaría demostrar que un modelo tiene una elevada exactitud, sensibilidad, especificidad o área bajo la curva. También se debería determinar si las explicaciones que entrega son útiles para que el profesional entienda el resultado, visualice posibles errores y se sienta más seguro al tomar decisiones. Por ello, la evaluación de los sistemas XAI a futuro debería analizar tanto métricas de rendimiento del algoritmo como métricas relacionadas con la interacción humano-inteligencia artificial.

Tabla 2. Técnicas XAI y posibles aplicaciones en laboratorio clínico.

Técnica	Tipo de explicación	Aplicación potencial	Ventaja	Limitación
SHAP	Principalmente local y global	Interpretación de variables de resultados de laboratorio	Permite estimar contribución de variables	Puede ser complejo de interpretar
LIME	Local	Explicación de predicciones individuales	Modelo independiente	La explicación puede depender de la configuración



Grad-CAM	Visual	Patología digital e imágenes microscópicas	Señala regiones relevantes	Principalmente aplicable a modelos de imágenes
LRP	Local/visual	Redes neuronales y clasificación	Permite analizar relevancia de características	Requiere conocimientos técnicos
Modelos basados en reglas	Global	Apoyo a decisiones y clasificación	Alta interpretabilidad	Pueden limitar la complejidad representable
Modelos intrínsecamente interpretables	Global	Predicción con variables de laboratorio	Interpretabilidad integrada	En algunos problemas pueden presentar menor flexibilidad
Explicaciones contrafactuales	Local	Análisis de cambios necesarios para modificar una predicción	Facilita razonamiento orientado a escenarios alternativos	Requiere generar escenarios clínicamente plausibles

Los resultados de esta primera categoría evidencian, en conjunto, un importante potencial de la XAI para hacer más próximos los sistemas de inteligencia artificial con las necesidades de la medicina de laboratorio, pero la evidencia disponible también muestra que existe una disjunción que sucede entre, por un lado, el desarrollo de modelos técnicamente avanzados y, por otro lado, su implementación clínica adecuadamente ajustada a la práctica. Esta disjunción se tendrá que considerar en el análisis posterior de las oportunidades y de las barreras.

Aportación de la inteligencia artificial explicativa a la precisión en el diagnóstico

La evidencia científica revisada sugiere que los sistemas de inteligencia artificial pueden lograr un elevado rendimiento diagnóstico en determinadas tareas, fundamentalmente cuando se emplean grandes conjuntos de datos y modelos de aprendizaje automático o de aprendizaje profundo. En la medicina de laboratorio, estos sistemas permiten la integración simultánea de muchas variables clínicas y analíticas, incluyendo relaciones no lineales que probablemente sean difíciles de identificar con técnicas estadísticas convencionales (12,14). Además, la precisión en el diagnóstico no debería ser solamente evaluada y medida en función de la proporción total de clasificaciones adecuadas; sólo se debe considerar indudablemente mediante indicadores que la miden como sensibilidad, especificidad, valores predictivos, área bajo la curva ROC (AUC), calibración y su capacidad de generalización.

La revisión de estudios sobre aprendizaje automático en medicina de laboratorio puso de evidencia una gran heterogeneidad en el repertorio de aplicaciones y de métricas



consideradas. Los algoritmos fueron utilizados para clasificar -por ejemplo, clasificar el útil diagnóstico de una condición patológica-, predecir -predicciones e inferencias acerca de resultados del proceso diagnosticador- y para el reconocimiento de patrones -el reconocimiento de patrones que favorezcan el cuidado de poblaciones- de los diferentes tipos de datos. Se ha constatado que las redes neuronales y otros modelos complejos permiten obtener un alto rendimiento en tareas concretas (14). Sin embargo, el rendimiento observado en el momento en el que se elabora un modelo no necesariamente se sostiene en el momento en que se aplica a una población diferente. Las diferencias en los instrumentos analíticos, en las características demográficas, en la prevalencia de las enfermedades, en los procedimientos de laboratorio o en la calidad de los datos pueden alterar de forma drástica el rendimiento del algoritmo.

En este contexto, la XAI puede aportar información extra para evaluar la plausibilidad de las predicciones. Cuando un modelo predice una condición clínica, una técnica de explicación puede permitir reconocer cuáles son las variables que más aportan a la predicción. En un modelo que se basa exclusivamente en los resultados de laboratorio, por ejemplo, puede determinarse si parámetros hematológicos, bioquímicos o inmunológicos intervienen significativamente en toda probabilidad en la clasificación. Esta información puede llegar a ayudar al profesional a reconocer predicciones coherentes con el contexto clínico y detectar de manera apropiada unas ciertas situaciones problemáticas.

SHAP es una de las metodologías más populares para este objetivo. Su implementación permite la estimación de la contribución de cada variable a una predicción dada y, por medio de análisis agregados, estudiar el comportamiento general del modelo (15). Con ello, la XAI podría complementar las métricas estándar de rendimiento al exponer información sobre el pensamiento del algoritmo. Ahora bien, los valores de importancia de las variables no están lingüísticamente obligados a interpretarse como relaciones causales: una característica podría tener una elevada contribución estadística sin dejar de ser el mecanismo biológico responsable de una condición patológica.

LIME, en cambio, es una técnica de explicación local para predicciones individuales (16). Para el laboratorio clínico, esta característica podría ser especialmente relevante para repasar unos determinados resultados, catalogados como de riesgo alto o discordantes. La explicación local podría permitir analizar qué variables contribuyeron a clasificar correctamente a un paciente determinado o permitirle al profesional seguir una revisión. Sin embargo, dada su dependencia de la aproximación adoptada en torno a la observación, su robustez o estabilidad, así como su grado de fidelidad ha de ser verificable antes de ser fundamentar decisiones clínicas en ellas.

Por tanto, la relación entre aplicabilidad y diagnóstico ha de concebirse como complementaria. La XAI no necesariamente incrementa simultáneamente en sensibilidad o especificidad un determinado algoritmo. Su principal aporte será dar cuenta del funcionamiento del modelo a partir de las entradas y permitir llevar a cabo una evaluación crítica de las predicciones propuestas. Un algoritmo puede tener un alto AUC y a mismo tiempo proporcionar explicaciones



de poco interés; o al contrario, un modelo muy explicable puede operar de forma subóptima para determinadas aplicaciones clínicas. Un uso seguro requiere la que ambas dimensiones han de ser consideradas al mismo tiempo (20,21). Una cuestión altamente relevante consiste en aprovechar las explicaciones para identificar patrones potencialmente espurios. Los modelos de aprendizaje profundo son capaces de aprender asociaciones que están en los datos de entrenamiento sin tener que ver con relaciones clínicamente significativas. En el caso de las imágenes médicas, por ejemplo, un algoritmo podría apoyarse en características relacionadas con el proceso de adquisición de la imagen en lugar de las características patológicas de interés. Las técnicas de XAI pueden ayudar a identificar estos patrones, mostrando cuál fue el peso de las regiones o las variables en las predicciones (17). Este principio puede extrapolarse a los datos de laboratorio cuando se poseen variables correlacionadas con la enfermedad que no necesariamente representan mecanismos diagnósticos directos.

La generalización es otro elemento fundamental. Una revisión reciente sobre la validación e implementación de alternativas de aprendizaje automático en el laboratorio clínico enfatiza la necesidad de tener en cuenta aspectos como la calidad del dato, interoperabilidad, rendimiento, generalización, equidad y explicabilidad antes de poner modelos a funcionar en situaciones reales (20). De esta forma, una herramienta de XAI orientada al laboratorio debería ser validada externamente y luego evaluada localmente antes de introducirse en los procesos rutinarios.

La calibración también es importante. La discriminación permite saber cuán bien un modelo discrimina entre categorías diagnósticas, mientras que la calibración mide el grado de concordancia entre las probabilidades asignadas y la frecuencia real del evento. En medicina clínica, un modelo puede discriminar bien entre pacientes con o sin una determinada condición, pero asignar probabilidades mal calibradas. La explicación, por sí misma, no soluciona este problema de modo que la XAI debe formar parte de una evaluación más global que incluya medidas de discriminación, calibración y utilidad clínica.

Otro aspecto señalado en la literatura es la necesidad de no realizar una interpretación excesivamente optimista de las explicaciones. Ali et al. indican que las técnicas de XAI pueden promover la transparencia a la vez que tienen limitaciones metodológicas relacionadas con la fidelidad y estabilidad de las explicaciones (16). Así que la explicación generada por un sistema no debe ser considerada automáticamente la representación fiel del proceso interno mediante el cual el modelo predijo.

Tabla 3. Indicadores para evaluar modelos de IA aplicados al diagnóstico clínico

Indicador	Finalidad	Importancia para el laboratorio clínico	Relación con XAI
Sensibilidad	Identificar correctamente casos positivos	Fundamental cuando la omisión de una enfermedad puede generar consecuencias clínicas	Permite contextualizar las explicaciones de los casos positivos y falsos negativos



Especificidad	Identificar correctamente casos negativos	Importante para reducir clasificaciones incorrectas de pacientes sanos	Facilita investigar qué variables contribuyeron a falsos positivos
AUC-ROC	Evaluar capacidad discriminativa global	Permite comparar el rendimiento diagnóstico de modelos	No explica por sí sola cómo funciona el modelo
Valor predictivo positivo	Estimar probabilidad de enfermedad ante resultado positivo	Depende de la prevalencia y del contexto clínico	Puede complementarse con explicaciones de predicciones individuales
Valor predictivo negativo	Estimar probabilidad de ausencia de enfermedad ante resultado negativo	Relevante para descartar determinadas condiciones	Puede ayudar a revisar casos con predicciones negativas
Calibración	Comparar probabilidades predichas con resultados observados	Importante cuando el modelo genera probabilidades clínicas	Permite contextualizar la confiabilidad de las predicciones
Importancia de variables	Determinar características relevantes para el modelo	Facilita comprender qué parámetros analíticos influyen en las predicciones	Constituye uno de los principales objetivos de SHAP y otras técnicas
Explicación local	Analizar una predicción individual	Útil para revisar resultados concretos	LIME y SHAP pueden proporcionar explicaciones locales
Explicación global	Comprender el comportamiento general del modelo	Permite identificar patrones generales utilizados por el algoritmo	Facilita auditoría y evaluación del modelo
Validación externa	Evaluar generalización	Esencial antes de utilizar un modelo en una población diferente	Permite comprobar si las explicaciones se mantienen en otros contextos

IA entendida y disminución del riesgo de errores de diagnóstico

La seguridad del paciente constituye uno de los principales motivos para defender modelos transparentes en la medicina. En los sistemas de los laboratorios, un error de categorización puede poder ser el origen de investigación adicional innecesaria, demoras en el diagnóstico, tratamientos inapropiados o, en su caso, no llevar a cabo intervenciones que sí están indicadas. Por ello, se debería buscar que la IA no busque únicamente aumentar la seguridad diagnóstica, sino que, al mismo tiempo, persiga la reducción de los riesgos que derivan de formatters interpretaciones que son erróneas.

La IA entendida puede ser de ayuda en este sentido, dado que, entre otras cosas, es capaz de poder identificar cuáles son los factores que sostienen una predicción. En el caso en que una explicación muestre que una categorización está sostenida por variables que son clínicamente plausibles, el profesional puede contar con información adicional que le permita entender la conclusión. En todo caso, cuando el modelo realiza una predicción basada en variables no previstas, podría ser un indicativo de que la predicción debe ser revisada a la luz de una nueva búsqueda de evidencia.



Esto se hace necesario en el caso de los ya referidos errores de automatización, pues tener confianza en la automatización puede llevar a aceptar una predicción sin una actitud crítica. La transparencia no debe ser utilizada para provocar la aceptación sin crítica de las recomendaciones automáticas, sino para permitir una interacción crítica entre el profesional y el sistema.

Por otra parte, la literatura sobre interacción humano-IA expone que el diseño de explicaciones debe tomar en consideración las necesidades de la persona que ha de utilizarlas y el contexto de uso de la información [10,18]. En el laboratorio clínico, una respuesta excesivamente técnica podría carecer de utilidad para alguien que necesita una respuesta rápida o una respuesta demasiado simplificada podría esconder las incertidumbres que son relevantes por esta razón, la utilidad de XAI también dependerá de cómo se presente la información.

En términos generales, los resultados que se han revisado permiten concluir que la XAI sí puede contribuir a que los sistemas de IA puedan ser evaluables, pero no es suficiente por sí sola para garantizar la seguridad del mismo. La coincidencia de un rendimiento diagnóstico adecuado, la validación externa, la calidad de los datos, la monitorización continua de la IA, la supervisión humana y los mecanismos de detección de errores o del cambio en el comportamiento del modelo (20) son requisitos necesarios para que un sistema pueda ser clínicamente seguro.

Esta consideración es especialmente relevante en laboratorios clínicos porque los procesos analíticos están en constante cambio y pueden cambiar los reactivos, los equipos, los métodos, los intervalos de referencia, los procedimientos operativos y las características de la población atendida. Cambios que pueden suponer una modificación gradual nuevamente del rendimiento del modelo lo que requiere una monitorización posterior.

Por todo ello, la IA explicativa debería ser concebida como un componente de un sistema más amplio de gobernanza clínica de la inteligencia artificial. La IA explicativa debería ser acompañada de protocolos de validación, de criterios de aceptación, de documentación de versiones, de evaluación de sesgos, del seguimiento del rendimiento y de mecanismos de intervención de los profesionales cuando el comportamiento del modelo no sea seguro.

Inteligencia artificial explicable y seguridad del paciente

Un argumento relevante que apoya el desarrollo de mecanismos de explicabilidad en los sistemas de inteligencia artificial (IA) diseñados para auxiliar en procesos diagnósticos es el que hace referencia a la seguridad del paciente, especialmente en el laboratorio clínico, donde el resultado analítico ligada a una determinada decisión médica o de hospitalización, o de seguimiento o incluso de descarte de una enfermedad, puede llegar a suponer la diferencia entre dicha decisión y otras alternativas; ya no se puede considerar como suficiente ni única una predicción obtenida a partir de un solo algoritmo. Una implementación responsable de dicho algoritmo exige conocer su rendimiento, sus condiciones de funcionamiento, sus limitaciones y, en los casos en los que sea importante, algunos de los factores que han contribuido a una determinada predicción [20].



Recientemente se revisó la validación, implementación y monitorización de soluciones de aprendizaje automático en el laboratorio clínico, pero se observó que solamente un número muy reducido de los modelos que se desarrollan en base a la investigación alcanzan a ser implementados en los laboratorios clínicos. Dentro de los elementos que deben estar presentes antes de la implementación hay que encontrar la calidad de los datos, interoperabilidad, correcta definición de las variables objetivo, propuesta de métricas, verificación de la generalización, equidad algorítmica y explicabilidad (20).

Desde este punto de vista, la XAI puede ser también una productividad sobre la seguridad desde el momento en que permite reconocer cuales son las variables o las características que apoyan la predicción. Si un sistema clasifica una muestra como perteneciente a una determinada categoría de riesgo, la explicación es la que permite al experto saber qué parámetros fueron predictores más importantes. Hacer esto puede permitir detectar predicciones que potencialmente son inconsistentes y facilitar la revisión del profesional antes de que el resultado se traduzca en decisiones clínicas.

No obstante, la explicabilidad no debe interpretarse como una garantía de la seguridad automática. Una explicación puede ser clínica y no llegar a ser suficiente. Asimismo, una aclaración que fuera muy sencilla podría inducir a una falsa sensación de conocer la información y hacer que el usuario goce de una confianza excesiva respecto a la que se podría inferir a partir de la capacidad del sistema. La utilidad de una explicación, en consecuencia, dependería del contexto clínico, del usuario y del tipo de decisión que se aspirara a apoyar.

La literatura ética reciente también ha puesto de manifiesto que existe un debate sobre si todos los sistemas de IA sanitaria deberían estar sometidos a un mismo nivel de explicabilidad, pues una revisión sistemática publicada en 2024 (22) había identificado posiciones divergentes en torno a la pretensión de preconizar la explicabilidad y había encontrado distintos tipos de argumentos tanto a favor como en contra de extenderla a todos los contextos. De hecho, de acuerdo con los autores, el nivel de explicabilidad podría depender del contexto de uso y de las particularidades de la decisión clínica que se intentara resolver (22).

Esta consideración es muy significativa en el laboratorio clínico. No todos los programas informáticos conllevan el mismo nivel de riesgo. Por ejemplo, un sistema que sirva para la mejora de un determinado proceso administrativo podría requerir un nivel de explicabilidad distinto del que podría requerir un algoritmo con un objetivo tan distinto como clasificar una muestra o realizar una predicción que alterase una determinada conducta clínica. Por ese motivo, la explicabilidad debería predicarse como un conjunto de opciones determinadas por el riesgo potencial, el nivel de autonomía del sistema y las potenciales consecuencias ante una predicción errónea.

La seguridad pasaría por mantener la supervisión humana. La XAI debería permitir que el profesional entienda, contradiga y, llegado el caso, rechace la recomendación que realiza un



determinado algoritmo. En este sentido, el modelo debería actuar como un complemento y no como un relevo del criterio del profesional. Así, la interacción entre el humano y la IA adopta un papel vital para evitar la automatización acrítica de las decisiones. La incorporación de las herramientas informáticas para la identificación y comunicación de resultados críticos es un ejemplo paradigmático de la interrelación entre tecnología de laboratorio y seguridad del paciente. Recientes estudios han observado que el hecho de identificar y de comunicar resultados críticos en la atención clínica constituye una parte integrante del proceso de validación clínica de los resultados analíticos, además de que se relaciona estrechamente con la seguridad asistencial (23).

En estos contextos, un sistema de IA podría identificar patrones de resultados potencialmente críticos y establecer la priorización de alertas de forma prioritaria. Antes de su posible utilización, evidentemente, debería demostrarse que las alertas son suficientemente sensibles para evitar alarmas emisoras y que son suficientemente específicas para evitar falsas alarmas que sobrecarguen con alertas al personal de salud. La adición de explicaciones podría ser la vía para ayudar al profesional a saber por qué una determinada combinación de resultados fue considerada de riesgo.

Oportunidades de XAI en el laboratorio clínico

Los estudios que se han analizado permitieron determinar diversas oportunidades considerando el uso de XAI en medicina de laboratorio, las cuales fueron agrupadas en cinco dimensiones: soporte a la precisión de diagnóstico, transparencia, seguridad, eficiencia y refuerzo de la interacción profesional-tecnología.

Una primera oportunidad es la integración multivariable de información. Los modelos de aprendizaje automático pueden de forma simultánea considerar múltiples parámetros e identificar asociaciones complejas. Cuando estos modelos se combinan con técnicas de aplicabilidad, el profesional de la salud puede obtener información adicional acerca de la importancia de cada variable en la predicción. Esto podría ser particularmente sustentador en escenarios donde la interpretación aislada de cada parámetro no permite identificar un patrón diagnóstico con claridad. La segunda oportunidad se refiere a la auditoría de los modelos. Las explicaciones pueden ser aprovechadas para evaluar cómo se comporta un sistema tras ser desarrollado y a partir de ese momento en su utilización. Si se puede verificar que el modelo basa sistemáticamente sus predicciones en variables que no son justificables clínicamente, se puede aprovechar esta información para revisar el modelo, la base de datos o el proceso de entrenamiento.

La tercera oportunidad se asocia con la formación y confianza de los profesionales. Un sistema que brinda explicaciones de las predicciones puede facilitar la relación de especialistas de laboratorio con las tecnologías de IA, pero incrementarla no debe equivaler a calibrarla. El



objetivo debería ser que sepa en qué ocasiones puede depositar confianza en el sistema y en qué otras necesita realizar una revisión adicional.

La reciente literatura de IA de carácter sanitario 675675 sugiere que la explicabilidad es una característica importante para desarrollar sistemas dignos de confianza, pero que también depende de otros aspectos, como la equidad, la privacidad o la fiabilidad y transparencia.

Una cuarta oportunidad se refiere a la detección de errores y sesgos. Los modelos pueden llegar a comportamientos opuestos cuando se aplican a poblaciones no suficientemente representadas durante el entrenamiento. Las explicaciones pueden ayudar a detectar diferencias en las variables que se están utilizando por el algoritmo y facilitar procesos de auditoría. De todas maneras, la detección de sesgos requiere análisis específicos y no puede ser delegada únicamente a métodos XAI.

Finalmente, la explicabilidad de la IA puede facilitar la trazabilidad de las decisiones algorítmicas. En un entorno clínico, la capacidad de documentar las características que han afectado a una predicción puede servir como información útil en auditorías, evaluaciones de incidentes y procesos de mejora continua. La capacidad de trazabilidad también puede ser importante en aquellos procesos donde el sistema tiene que participar y que de un modo u otro pueden afectar directamente al paciente.

Tabla 4. Oportunidades de la XAI para el laboratorio clínico

Oportunidad	Aplicación potencial	Beneficio esperado
Interpretación multivariable	Análisis simultáneo de resultados	Identificación de patrones complejos
Transparencia	Explicación de predicciones	Mayor comprensión del funcionamiento del modelo
Apoyo diagnóstico	Predicción de condiciones clínicas	Complemento al criterio profesional
Detección de errores	Análisis de predicciones discordantes	Identificación de posibles fallos
Auditoría	Revisión de variables utilizadas	Evaluación del comportamiento algorítmico
Seguridad del paciente	Revisión de predicciones de riesgo	Reducción potencial de errores
Trazabilidad	Registro de factores asociados con decisiones	Facilita auditoría y seguimiento
Formación profesional	Visualización del razonamiento algorítmico	Favorece alfabetización en IA
Monitorización	Evaluación continua del rendimiento	Detección de pérdida de desempeño
Personalización	Adaptación de modelos a contextos específicos	Mayor pertinencia clínica



Obstáculos técnicos para el despliegue

A pesar de las oportunidades detectadas, se evidencia cómo la transición de los modelos experimentales a los que se utilizan rutinariamente en los laboratorios en la práctica sigue siendo escasa; una encuesta internacional reciente entre expertos en IA aplicada al laboratorio indicó que la implementación sigue escasa y ser heterogénea. Los encuestados indicaron que era necesario realizar una verificación local incluso en lo que los propios encuestados declaran como sistemas introducidos externamente para aplicaciones clínicas; además, se indicaron como prioridades centradas en la evaluación del desempeño local, la concordancia del uso esperado y la verificación de las medidas de privacidad y seguridad [25].

Una de las barreras es la infraestructura tecnológica: los sistemas XAI requieren capacidad de cálculo, almacenamiento, interoperabilidad y acceso a los datos. Cuando los sistemas de información del laboratorio no están suficientemente integrados, la implementación de modelos puede ofrecer obstáculos para obtener información de forma estructural y en tiempo suficiente. El nivel de calidad de los datos representa otra restricción importante. Los algoritmos aprenden a partir de la información que tienen disponible y pueden aprender a reproducir errores, sesgos o patrones del conjunto de datos de entrenamiento. Información incompleta, desbalanceada, contradictoria o solamente aportada por una población pueden llegar a limitar la capacidad de generalización del modelo. Miller y Valdes observaron que la validación de modelos de aprendizaje automático en medicina de laboratorio debe incluir una serie de aspectos metodológicos, de calidad de los datos, de validez, de generalización, y de aplicabilidad, entre otros aspectos [26].

Una tercera barrera es la interoperabilidad. Los laboratorios se apoyan en múltiples analizadores, sistemas de información, plataformas de Hospitales, y fuentes de datos. La integración de una herramienta XAI requiere información que sea transferida y procesada de manera homogénea. La falta de estándares o estructuras puede impedir añadir sistemas de IA en el flujo de trabajo habitual. También se constató una barrera asociada con la validación local. Un modelo que se valida en una institución puede tener un comportamiento distinto al puesta en práctica en otra debido a existencia de diferencias en la población, en instrumentos, en reactivos, en procedimientos y en prevalencias de enfermedades. De este modo para esta situación la aprobación o validación que ha tenido lugar en la fase de su definición no necesariamente hace innecesaria la verificación del sistema en un contexto específico en el que se aplica.

Barreras éticas, regulatorias y profesionales

Más allá de las barreras tecnológicas, la literatura última nos hace ver cómo son igualmente difíciles las barreras éticas y reguladoras; aparecen problemas como la privacidad, la seguridad de los datos, el sesgo algorítmico, la responsabilidad profesional, la transparencia y la equidad.



La posibilidad de proteger la privacidad de los datos es fundamental en un ámbito como el de la medicina de laboratorio ya que los sistemas de IA necesitan grandes volúmenes de información clínica y analítica. La cantidad de datos requeridos para entrenar modelos y la obligación de proteger la privacidad de los pacientes deben encontrar un delicado equilibrio. El sesgo algorítmico es, a su vez, un problema. Si los conjuntos de entrenamiento no reflejan adecuadamente la diversidad de los pacientes, el sistema puede tener un rendimiento desigual entre grupos. Una revisión reciente sobre los retos de la implementación de la IA en la práctica médica citaba precisamente la explicabilidad, el sesgo y la equidad, infraestructura, privacidad, seguridad, regulación y formación como los principales obstáculos para incorporar efectivamente esta tecnología [27]. La responsabilidad profesional es también un reto. A medida que una decisión clínica es apoyada mediante IA, habrá que precisar las responsabilidades que pueden corresponder al desarrollador, al proveedor, a la institución y a los profesionales de la salud. Contar con una explicación algorítmica no elimina la necesidad de contar con mecanismos de supervisión y de ejecución claramente establecidos. La regulación también avanza más lentamente que el ritmo del desarrollo tecnológico. La literatura reciente señala que existe una carencia de armonización entre la innovación tecnológica y los marcos regulatorios, en particular en lo referente a la transparencia, la responsabilidad y la supervisión humana [28]. En este marco, iniciar la implementación de XAI requiere un enfoque de la gobernanza que incluya profesionales de laboratorio, especialistas clínicos, informáticos, responsable de calidad, expertos en protección de datos y reguladores. La incorporación de la tecnología no se debe hacer como un proceso sólo informático, sino como un proceso organizacional y clínico.

Tabla 5. Principales barreras para la implementación de XAI

Categoría	Barrera	Posible consecuencia
Técnica	Datos insuficientes o de baja calidad	Predicciones poco fiables
Técnica	Falta de interoperabilidad	Dificultad de integración
Técnica	Infraestructura limitada	Implementación incompleta
Clínica	Escasa validación externa	Baja generalización
Clínica	Explicaciones poco comprensibles	Uso inadecuado del sistema
Ética	Sesgos algorítmicos	Desigualdad en el desempeño
Ética	Privacidad	Riesgo de exposición de información
Regulatoria	Marcos normativos en evolución	Incertidumbre sobre requisitos
Profesional	Falta de capacitación	Dependencia o rechazo del sistema



Organizacional	Falta de protocolos	Implementación heterogénea
Económica	Costos de infraestructura y mantenimiento	Dificultad de sostenibilidad
Seguridad	Falta de monitorización	Detección tardía del deterioro del modelo

DISCUSIÓN

La información analizada dio pie a concluir que la inteligencia artificial explicable se convierte en una opción de creciente interés para aumentar la transparencia de los sistemas de IA aplicados al laboratorio clínico. Los trabajos que han sido revisados coinciden en que los modelos de aprendizaje automático pueden conseguir un buen diagnóstico; no obstante, su puesta en marcha en la práctica clínica debe abordar y superar la condición de "caja negra" que caracteriza a algunos modelos complejos (15,16). La contribución fundamental que puede hacerse de la XAI no es incrementar por sí misma la precisión y la exactitud de un algoritmo, sino permitir el entendimiento, la evaluación y la interrogación en torno a las predicciones que haga este.

La propuesta es relevante especialmente en medicina de laboratorio, ya que los resultados analíticos son el pilar de numerosas decisiones clínicas. El uso de modelos que permitan incluir múltiples variables puede suponer una mejora en el análisis cuando se compara la evaluación aislada de parámetros; aun así, un alto nivel de predicción no hace que el modelo que lo ofrece utilice asociaciones clínicamente correctas. Los resultados revisados indican que técnicas como SHAP, LIME, y otras formas de interpretación visual pueden ayudar a determinar las variables o las regiones que influyen más en el proceso de predicción (15,17). Por tanto, la explicabilidad puede representar un sistema de comprobación del comportamiento algorítmico.

Por otro lado, los resultados también mostraron que se debe evitar una lectura casi optimista de la explicabilidad. El hecho de que exista una explicación no significa que el modelo haya reproducido con precisión o causalmente el razonamiento interno. Las técnicas post hoc son capaces de proporcionar aproximaciones útiles, aunque la fidelidad, la estabilidad, y la utilidad de las mismas han de ser evaluadas para el caso del modelo y del contexto (16,21). Esta consideración tiene relevancia en el laboratorio clínico, donde una interpretación de una explicación errónea podría dar lugar a tener una confianza sin fundamento en una predicción.

La literatura revisada también sirvió para identificar una diferencia importante entre el desarrollo de modelos de IA y su implementación efectiva: en el caso de la multitud de estudios experimentales que informan de resultados favorables, las encuestas en medicina de laboratorio muestran que la adopción rutinaria sigue siendo escasa (19,25). Esta diferencia podría corresponder a la existencia de unas barreras que no sólo tienen que ver con la capacidad predictiva de los propios algoritmos. Infraestructura tecnológica, interoperabilidad, disponibilidad de datos, formación profesional, validación local, privacidad y requerimientos



regulatorios también son condicionantes en el uso de éstas en escenarios reales [20,26]. Así, la validación constituye, por tanto uno de los elementos más importantes para garantizar una implementación segura. Un modelo que ha sido desarrollado a partir de una población en concreto puede mostrar peores resultados en otro laboratorio distinto como consecuencia de diferencias en las características de los pacientes y en los instrumentos, reactivos, métodos analíticos y procedimientos de trabajo. La literatura más reciente recomienda la evaluación no solamente en términos de discriminación, sino que también requiere considerar la calibración, la generalización, la equidad y la aplicabilidad clínica [20,26]. Como resultado, los resultados obtenidos durante el desarrollo de un modelo no pueden considerarse suficientes como para dar por hecho que el mismo puede ser utilizado de manera generalizada.

La relación entre XAI y seguridad del paciente debe ser interpretada desde esta misma perspectiva; la aplicabilidad puede ayudar a identificar predicciones propensas a ser inesperadas, facultar la aplicabilidad de las decisiones tomadas para permitir la revisión profesional o contribuir a la trazabilidad; no puede sustituir a los mecanismos tradicionales previos de control de calidad, ni puede sustituir el papel de la supervisión humana. La seguridad de los pacientes depende de la interacción de la tecnología, de los profesionales, de los procesos organizacionales y de los sistemas de gestión del riesgo. Por eso, la XAI se tiene que añadir como parte de un sistema global de gobernanza de la IA y no como una solución por sí sola [22,24].

Otro aspecto que es igualmente relevante en el ámbito de la IA es la confianza del profesional. La transparencia puede permitir la aceptación de los sistemas de IA cuando las explicaciones son comprensibles y clínicamente relevantes. No obstante, aumentar la confianza desmedidamente puede traer perjuicios si esta confianza condujera a una aceptación sin cuestionamientos de las recomendaciones del algoritmo. El objetivo sería desarrollar una confianza calibrada, es decir, una confianza en la cual el profesional conozca las potencialidades y las limitaciones del sistema e identifique contextos en los que sea necesario cuestionar su predicción [10,18].

Esta consideración también es muy relevante para aquellos sistemas de IA que se aplican a la identificación de resultados extremos o en contextos de riesgos. La automatización puede aumentar la velocidad de procesar y hacer las prioridades sobre la información, mas también puede dar como resultado alarmas falsas o dejar de lado determinadas situaciones. Por consiguiente, cualquier herramienta que se ponga en práctica en el laboratorio debería eficiente umbrales de trabajo, procedimientos para revisar los resultados de la herramienta y canalizar su posible escalado a los profesionales responsables.

La evidencia también plantea una cuestión vinculada a la estandarización de la técnica. Hay multitud de técnicas de XAI, de métricas y de formas de presentar las explicaciones [15,21]. Esta diversidad complica establecer comparaciones entre diferentes sistemas y criterios generales para decidir si una explicación es suficientemente buena para su uso clínico. Por tanto, uno de los retos de futuro será establecer marcos de evaluación que permitan valorar la fidelidad, la estabilidad, la comprensibilidad y utilidad clínica.



Desde la perspectiva profesional, la implementación de XAI también necesita de nuevas competencias. El personal del laboratorio no debe ser un crack en programación o en machine learning; sin embargo, sí que debería ser capaz de desarrollar unos conocimientos suficientes para entender cuestiones de validación, sensibilidad, especificidad, calibración, sesgo algorítmico, incertidumbre y limitaciones de las explicaciones. La tener alfabetización en IA puede animar a los grupos de profesionales a involucrarse activamente en la selección, evaluación y seguimiento de tales tecnologías.

Los resultados acerca de las barreras éticas y regulatorias apuntan en esta dirección. El uso de datos clínicos y de laboratorio incluye responsabilidades en torno a la privacidad, la seguridad, la responsabilidad profesional y la trazabilidad en el que el uso de los modelos que llegan a incluirse en decisiones diagnósticas [22,27,28]. Por eso el desarrollo de XAI debería de llevarse a cabo por equipos de trabajo interdisciplinarios que incluyan profesionales del laboratorio, clínicos, especialistas en informática, responsables de calidad, asesores en protección de datos y representantes de la institución.

Este último punto es especialmente interesante ya que la aplicabilidad debería ir asociada al nivel de riesgo del uso esperado. Las aplicaciones de bajo riesgo necesitarían explicaciones más simples, aunque las que lo son sobre decisiones diagnósticas o terapéuticas deberían estar sujetas a requerimientos más exigentes de validación, transparencia y seguimiento [22]. Con esta alternativa se podría evitar tanto una regulación escasa de aplicaciones de alto riesgo como requerimientos desmesurados para sistemas de bajo impacto clínico.

Globalmente, la evidencia revisada permitiría defender que la XAI tiene un gran potencial para favorecer una incorporación más responsable de la inteligencia artificial en el laboratorio clínico, aunque su verdadero valor depende de que las explicaciones sean técnicamente válidas, clínicamente relevantes y fácilmente comprensibles para los usuarios. El reto no consiste sólo en construir modelos que sean capaces de predecir sino en crear sistemas que sean capaces de predecir, explicar, ser evaluados y que sigan bajo control profesional.

Por último, la principal brecha que se ha identificado ha sido la de la escasa evidencia de implementación en condiciones reales de laboratorio. La literatura presenta muchos modelos experimentales, pero son todavía menos los estudios prospectivos, multicéntricos y longitudinales que evalúen el impacto de la XAI en los resultados clínicos, en la seguridad del paciente, en la eficiencia del laboratorio y en la toma de decisiones profesionales. Este hecho pone de manifiesto la necesidad de futuros trabajos que vayan más allá de la evaluación puramente retrospectiva del rendimiento algorítmico y que aprovechen la capacidad de las tecnologías para estudiar el flujo de trabajo real.

Con todo ello, la evidencia existente da soporte a una incorporación progresiva y controlada de la IA explicable en medicina de laboratorio, la cual debería estar respaldada por validación externa y local, evaluación continua, protección de datos, formación profesional, control humano y



criterios regulatorios ajustados al riesgo. En estas condiciones, la XAI podría contribuir no solamente a aumentar la transparencia de los modelos, sino también a la confianza profesional, a la trazabilidad y a la seguridad del uso de inteligencia artificial en el diagnóstico clínico.

CONCLUSIONES

La revisión mostró que la inteligencia artificial explicable (XAI), tuvo oportunidades importantísimas para ayudar a explicar los resultados que se obtienen a partir de las predicciones que dan modelos de inteligencia artificial en el laboratorio clínico. Su uso y aplicación permitió, entre otras cosas, identificar y visualizar todas aquellas variables que favorecen los resultados obtenidos mediante tal o cual predicción, contribuyendo a una mejor comprensión del proceso diagnóstico en el que se soporta o se utiliza. Las técnicas de explicabilidad, tales como la XAI y SHAP, LIME, Grad-CAM o mediante métodos basados en Relevancia o sobre modelos intrínsecamente explicables potencialmente muestran cierto valor para el laboratorio. Las áreas en las que examinar las aplicaciones de la inteligencia artificial explicable podrían ser la hematología, bioquímica, microbiología, inmunología y patología digital. Sin embargo, la explicabilidad no fue por sí misma suficiente como para permitir una precisión diagnóstica mejorada, por lo que el propio valor tuvo que ser examinado junto con el funcionamiento, la calibración, la validación externa y la utilidad clínica del conjunto de cada modelo.

La integración de XAI fue equiparada a oportunidades de mejora de la seguridad del paciente mediante una mayor trazabilidad de las decisiones algorítmicas, la identificación de potenciales errores, la ayuda de la revisión profesional y el establecimiento de mecanismos de control de decisiones. Sin embargo, la evidencia revisada mostró que se mantuvieron barreras técnicas, organizativas, éticas y regulatorias que limitaron su integración rutinaria. Entre ellas identificamos heterogeneidad y calidad de los datos, interoperabilidad escasa, falta de validación externa, sesgos, interpretaciones de determinadas explicaciones difíciles, protección de datos, escasez de protocolos estandarizados y necesidad de formación especializada para el personal del laboratorio. Por ello, la integración de XAI tuvo que ser llevada a cabo tal que poseyera una participación activa del profesional sanitario, evitando que las predicciones automatizadas pudieran sustituir el juicio clínico.

Se concluyó que la integración segura de inteligencia artificial explicable en el laboratorio clínico tuvo que ser fruto de un movimiento gradual mediante un enfoque local y externo basado en la validación, una evaluación continua del funcionamiento, una monitorización de los cambios de los datos y de los modelos, la protección de la información, la documentación de las decisiones y la supervisión humana. Igualmente, resultó necesario establecer criterios estandarizados no solo para evaluar la precisión del modelo, sino así también la fidelidad, la estabilidad, la utilidad clínica y la comprensión de las explicaciones producidas. En este sentido, XAI constituye una herramienta adicional como apoyo a la precisión diagnóstica y a la seguridad del paciente, cuyo beneficio va a depender de la manera en la que se incluya de manera adecuada en los procesos del laboratorio, la infraestructura, las competencias profesionales o bien el aspecto ético y regulatorio aplicable.



REFERENCIAS BIBLIOGRÁFICAS

1. Baron JM. Artificial intelligence in the clinical laboratory: an overview with frequently asked questions. *Clin Lab Med.* 2023;43(1):1-16. doi:10.1016/j.cll.2022.09.002.
2. Loh HW, Ooi CP, Seoni S, Barua PD, Molinari F, Acharya UR. Application of explainable artificial intelligence for healthcare: a systematic review of the last decade (2011-2022). *Comput Methods Programs Biomed.* 2022;226:107161. doi:10.1016/j.cmpb.2022.107161.
3. Biswas AA. A comprehensive review of explainable AI for disease diagnosis. *Array.* 2024;22:100345. doi:10.1016/j.array.2024.100345.
4. Mienye ID, Obaido G, Jere N, Mienye E, Aruleba K, Emmanuel ID, Ogbuokiri B. A survey of explainable artificial intelligence in healthcare: concepts, applications, and challenges. *Inform Med Unlocked.* 2024;51:101587. doi:10.1016/j.imu.2024.101587.
5. Hou H, Zhang R, Li J. Artificial intelligence in the clinical laboratory. *Clin Chim Acta.* 2024;559:119724. doi:10.1016/j.cca.2024.119724.
6. Çubukçu HC, Topcu Dİ, Yenice S. Machine learning-based clinical decision support using laboratory data. *Clin Chem Lab Med.* 2023. doi:10.1515/cclm-2023-1037.
7. Jung J, Lee H, Jung H, Kim H. Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: a systematic review. *Heliyon.* 2023;9(5):e16110. doi:10.1016/j.heliyon.2023.e16110.
8. Cestonaro C, Delicati A, Marcante B, Caenazzo L, Tozzo P. Defining medical liability when artificial intelligence is applied on diagnostic algorithms: a systematic review. *Front Med (Lausanne).* 2023;10:1305756. doi:10.3389/fmed.2023.1305756.
9. Pennestrì F, Banfi G. Artificial intelligence in laboratory medicine: fundamental ethical issues and normative key-points. *Clin Chem Lab Med.* 2022;60(12):1867-1874. doi:10.1515/cclm-2022-0096.
10. Subramanian HV, Canfield C, Shank DB. Designing explainable AI to improve human-AI team performance: a medical stakeholder-driven scoping review. *Artif Intell Med.* 2024;149:102780. doi:10.1016/j.artmed.2024.102780.
11. Klauschen F, Dippel J, Keyl P, Jurmeister P, Bockmayr M, Mock A, et al. Toward explainable artificial intelligence for precision pathology. *Annu Rev Pathol.* 2024;19:541-570. doi:10.1146/annurev-pathmechdis-051222-113147.
12. Hou H, Zhang R, Li J. Artificial intelligence in the clinical laboratory. *Clin Chim Acta.* 2024;559:119724. doi:10.1016/j.cca.2024.119724.
13. Lippi G, Mattiuzzi C, Favaloro EJ. Artificial intelligence in the pre-analytical phase: state-of-the-art and future perspectives. *J Med Biochem.* 2024;43(1):1-10. doi:10.5937/jomb0-45936.
14. You J, Seok HS, Kim S, Shin H, et al. Advancing laboratory medicine practice with machine learning: swift yet exact. *Ann Lab Med.* 2025;45(1):22-35. doi:10.3343/alm.2024.0354.
15. Loh HW, Ooi CP, Seoni S, Barua PD, Molinari F, Acharya UR. Application of explainable artificial intelligence for healthcare: a systematic review of the last decade (2011-2022). *Comput Methods Programs Biomed.* 2022;226:107161. doi:10.1016/j.cmpb.2022.107161.



16. Ali S, Akhlaq F, Imran AS, Kastrati Z, Daudpota SM, Moosa M. The enlightening role of explainable artificial intelligence in medical & healthcare domains: a systematic literature review. *Comput Biol Med.* 2023;166:107555. doi:10.1016/j.compbimed.2023.107555.
17. Klauschen F, Dippel J, Keyl P, Jurmeister P, Bockmayr M, Mock A, et al. Toward explainable artificial intelligence for precision pathology. *Annu Rev Pathol.* 2024;19:541-570. doi:10.1146/annurev-pathmechdis-051222-113147.
18. Cálem J, Moreira C, Jorge J. Intelligent systems in healthcare: a systematic survey of explainable user interfaces. *Comput Biol Med.* 2024;180:108908. doi:10.1016/j.compbimed.2024.108908.
19. Cadamuro J, Carobene A, Cabitza F, Debeljak Z, De Bruyne S, van Doorn W, et al. A comprehensive survey of artificial intelligence adoption in European laboratory medicine: current utilization and prospects. *Clin Chem Lab Med.* 2025;63(4):692-703. doi:10.1515/cclm-2024-1016.
20. Spies NC, Farnsworth CW, Wheeler S, McCudden CR. Validating, implementing, and monitoring machine learning solutions in the clinical laboratory safely and effectively. *Clin Chem.* 2024;70(11):1334-1343. doi:10.1093/clinchem/hvae126.
21. Jung J, Lee H, Jung H, Kim H. Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: a systematic review. *Heliyon.* 2023;9(5):e16110. doi:10.1016/j.heliyon.2023.e16110.
22. Freyer N, Groß D, Lipprandt M. The ethical requirement of explainability for AI-DSS in healthcare: a systematic review of reasons. *BMC Med Ethics.* 2024;25:104. doi:10.1186/s12910-024-01103-2.
23. García-de la Torre A, Sánchez-Lora FJ, Navarrete-Ortiz MC, Trillo López P, Ortiz-García C, Méndez Alarcón G, et al. Implementation of an automated alert system of critical results in hospitalized and emergency patients. *Clin Chem Lab Med.* 2026;64(4):838-844. doi:10.1515/cclm-2025-1312.
24. Ning Y, Liu M, Liu N. Advancing ethical AI in healthcare through interpretability. *Patterns (N Y).* 2025;6(6):101290. doi:10.1016/j.patter.2025.101290.