https://politecnicadelaciencia.com/

ISSN: 3091-2008

Edición Especial:
Vol. 4 Num. 10, PP 176-193
DOI:
https://doi.org/10.56519/r6vc1j80

INTELIGENCIA ARTIFICIAL EN FARMACOVIGILANCIA: PREDICCIÓN Y DETECCIÓN DE REACCIONES ADVERSAS A MEDICAMENTOS: REVISIÓN SISTEMÁTICA

ARTIFICIAL INTELLIGENCE IN PHARMACOVIGILANCE: PREDICTION AND DETECTION OF ADVERSE DRUG REACTIONS: A SYSTEMATIC REVIEW

Mariela Alexandra Idrovo Vallejo[1], María del Cisne Díaz Guerrero[2], Jairo Wladimir Almeida Castillo[3], Adriana Guadalupe Villarreal Sánchez[4]

{ mariela.idrovo@unl.edu.ec[1], mdiaz825@puce.edu.ec[2], jairoalmeidacastillo@hotmail.com[3], adrivii996@gmail.com[4]}

Fecha de recepción: 03/10/2026 / Fecha de aceptación: 17/11/2026 / Fecha de publicación: 24/11/2026

RESUMEN: La farmacovigilancia afronta importantes dificultades relacionadas con la infranotificación de reacciones adversas a medicamentos, la fragmentación de los registros sanitarios y los retrasos en la identificación de riesgos asociados al uso de tratamientos farmacológicos. En este contexto, la heterogeneidad de las fuentes de información, los desenlaces evaluados y los métodos de validación dificulta establecer la utilidad clínica de los modelos de inteligencia artificial. La presente investigación tuvo como objetivo analizar las aplicaciones de la inteligencia artificial en la detección y predicción de reacciones adversas a medicamentos, así como examinar las principales limitaciones para su implementación en la práctica asistencial. Se realizó una revisión crítica de ocho investigaciones de síntesis, complementada con estudios originales y documentos de organismos sanitarios. La información se obtuvo mediante la consulta de registros bibliográficos y fuentes editoriales verificables, tomando como referencia las recomendaciones PRISMA 2020 para favorecer la transparencia en la presentación metodológica. Los resultados evidenciaron que una revisión de 67 estudios identificó la detección de señales de seguridad como la principal aplicación de la inteligencia artificial, con una frecuencia del 50,5 %. Asimismo, otra revisión de 59 estudios informó que únicamente el 18,6 % había realizado una validación externa de los modelos, mientras que un metaanálisis de nueve estudios registró un área bajo la curva de 0,841, con un intervalo de confianza del 95 % entre 0,789 y 0,883. Se concluyó que la inteligencia artificial representa una herramienta prometedora para optimizar la identificación de señales de

[1]Docente de la Carrera de Enfermería, Universidad Nacional de Loja, Ecuador, https://orcid.org/0000-0002-5555-0099
[2]Facultad de Salud y Bienestar, Pontificia Universidad Católica del Ecuador, Ecuador, https://orcid.org/0000-0002-2400-5821
[3]Investigador Independiente, https://orcid.org/0009-0001-5437-6690
[4]Investigador Independiente, https://orcid.org/0000-0001-9985-2446

seguridad, el procesamiento de información clínica y la predicción de riesgos farmacológicos. Sin embargo, su incorporación a los sistemas de farmacovigilancia requiere fortalecer la validación independiente, desarrollar evaluaciones prospectivas y garantizar la supervisión de profesionales sanitarios para favorecer su aplicación segura y confiable.

Palabras clave: farmacovigilancia, inteligencia artificial, reacciones adversas a medicamentos, aprendizaje automático, seguridad del paciente

ABSTRACT: Pharmacovigilance faces significant challenges related to the underreporting of adverse drug reactions, the fragmentation of healthcare records, and delays in identifying risks associated with pharmacological treatments. In this context, the heterogeneity of information sources, assessed outcomes, and validation methods makes it difficult to establish the clinical utility of artificial intelligence models. This study aimed to analyze the applications of artificial intelligence in the detection and prediction of adverse drug reactions and to examine the main limitations affecting their implementation in clinical practice. A critical literature review was conducted, including eight evidence synthesis studies, supplemented by original research articles and documents published by health authorities. Information was obtained through bibliographic databases and verifiable editorial sources, using the PRISMA 2020 recommendations as a reference to promote transparency in methodological reporting. The findings revealed that a review of 67 studies identified safety signal detection as the most common application of artificial intelligence, accounting for 50.5% of the studies. Furthermore, another review involving 59 studies reported that only 18.6% had performed external model validation, while a meta-analysis of nine studies reported an area under the receiver operating characteristic curve of 0.841, with a 95% confidence interval ranging from 0.789 to 0.883. It was concluded that artificial intelligence represents a promising tool for improving safety signal identification, clinical information processing, and the prediction of drug-related risks. However, its integration into pharmacovigilance systems requires stronger independent validation, prospective evaluations, and appropriate supervision by healthcare professionals to ensure its safe and reliable application.

Keywords: pharmacovigilance, artificial intelligence, adverse drug reactions, machine learning, patient safety

INTRODUCCIÓN

La seguridad de los fármacos es un aspecto fundamental dentro de la asistencia sanitaria. Si bien los ensayos clínicos permiten caracterizar muchos efectos adversos antes de la autorización de un producto, las condiciones de inclusión de los participantes, el tiempo y el número de pacientes no reproducen toda la variedad de situaciones que se observan durante la práctica clínica. Las personas de edad avanzada, las personas con varias enfermedades, las personas que se encuentran en tratamientos concomitantes pueden desembocar en problemas de seguridad no

suficientemente reflejados en la investigación anterior a comercializar. Por esta razón, la Organización Mundial de la Salud establece la farmacovigilancia como el conjunto de actividades destinadas a detectar, evaluar, entender y prevenir efectos adversos u otros problemas relacionados con el uso de medicamentos (1).

La notificación espontánea ha sido una de las principales fuentes para detectar posibles reacciones adversas a partir del momento que los fármacos empiezan a ser utilizados de una forma más habitual. Este proceso no deja de depender de que los propios profesionales o pacientes reconozcan un evento, consideren su posible relación con el fármaco y opten por comunicarlo. La infra notificación en los artículos de revisión seleccionados por Hazell y Shakir es del 94 % como mediana, aunque presenta cierta variación según el contexto y el tipo de reacción (2). Más allá de este tema de la infra notificación, los sistemas de notificación pueden incluir duplicados, información incompleta y diferencias en la codificación, y la autoridad reguladora estadounidense señala, explícitamente, que las notificaciones no sirven para calcular la incidencia de una reacción adversa ni para demostrar que exista causalidad respecto al producto sospechoso (3).
El aumento de los depósitos de información de farmacovigilancia y de las historias clínicas electrónicas ha creado nuevas oportunidades para tratar estos problemas. En el contexto europeo, EudraVigilance recoge informes individuales con sospechas de reacciones adversas, así como tiene procedimientos de procesamiento, detección de duplicados e identificación de desproporcionalidad de las notificaciones (4). Estas herramientas permiten clasificar y encontrar en gran información, si bien la interpretación depende de la calidad del dato y del criterio farmacológico. La señal detectada representaría la sospecha a investigar, pero no significaría la prueba irrefutable de daño por abuso de un medicamento.

La inteligencia artificial en este ámbito ofrece procedimientos capaces de reconocer relaciones complejas en datos estructurados y de obtener información clínica a partir de información escrita en lenguaje natural. De forma más concreta, el aprendizaje automático para clasificaciones de eventos, el aprendizaje profundo para el reconocimiento de patrones de mayor complejidad y el procesamiento del lenguaje natural para detectar menciones de medicamentos, manifestaciones clínicas y relaciones temporales. Ya desde fechas tempranas, Harpaz y colaboradores dieron cuenta de cómo las técnicas de la minería de datos (por las cuales en un conjunto de datos no estructurados es posible descubrir y evaluar acontecimientos adversos) podrían ampliar las posibilidades de descubrimiento y de evaluación de eventos adversos (5); poco después, Tatonetti y colaboradores mostraron que el análisis de grandes repositorios de notificación (esto es, conjuntos de datos sobre eventos adversos) podría tener utilidad para estudiar interacciones entre medicamentos, aunque detectaron la necesidad de controlar variables de confusión (6).

En otra vía complementaria, los registros asistenciales. LePendu y colaboradores utilizaron las notas clínicas para estudiar eventos de seguridad que no necesariamente se encuentran estructurados en bases de datos (7). De igual modo, Sarker y colaboradores sistematizaban los

primeros avances en el aprovechamiento de mensajes que usuarios de Twitter dejan publicados (8), y Nikfarjam y colaboradores probaban la viabilidad del reconocimiento de expresiones de reacciones adversas en esta forma de lenguaje informal (9).Dichas aproximaciones ampliaron la variedad de fuentes potenciales útiles, aunque también suscitaban problemas de ambigüedad lingüística, de privacidad, de sesgo de selección y de verificación de las afirmaciones que hacían las personas.

En este sentido, la publicación de investigaciones con modelos predictivos cada vez más complejos ha incrementado el interés por incorporar inteligencia artificial a la vigilancia de la salud cotidiana. Sin embargo, la capacidad de discriminar casos en una base de datos de entrenamiento no asegura que el mismo rendimiento se mantenga en otro hospital, en otro tiempo o en pacientes con características diferentes; así, ser más sensible en identificar menciones de posibles efectos adversos no implica que el daño clínico sea menor.

La normativa europea de los medicamentos señala que el uso de IA a lo largo del ciclo de los medicamentos debe ser entendido a la luz de los requisitos de calidad, de seguridad y de protección de datos existentes (24). Por lo tanto, persiste una cuestión de investigación importante: los trabajos publicados dan cuenta de beneficios en la identificación y la predicción de un evento adverso, pero unas y otros no hacen uso de fuentes, poblaciones, definiciones de resultados y técnicas de evaluación que sean, en todos los casos, comparables. Tal heterogeneidad dificulta llegar a la conclusión de en qué tareas la IA es capaz de aportar valor objetivo, así como las condiciones que deben ser satisfechas para la introducción de la IA en la atención clínica y en los procesos de toma de decisión regulatorios. Por lo tanto, el objetivo que persigue esta revisión es el de realizar un análisis acerca de las principales aplicaciones de la IA en la predicción y en la detección de las reacciones adversas a los medicamentos, el contrastar los resultados comunicados por diferentes autores y el de señalar las limitaciones metodológicas que los mismos presentan y las condiciones mínimas para una utilización responsable.

La evidencia cuantitativa más reciente pone de manifiesto que el problema va más allá de la exactitud de la clasificación. En un metaanálisis multinivel que incluía 30 estimaciones provenientes de nueve trabajos la superficie bajo la curva obtenida era de 0,841, pero se observaba una alta heterogeneidad lo que llevó a clasificar la certeza de dicha estimación como baja (25). Por lo tanto, la interpretación de la eficacia ha de ser complementada con una valoración de qué clase de datos se han utilizado, cuál ha sido el procedimiento de validación o el contexto asistencial bajo el que fue construido cada uno de los algoritmos. Esta necesidad de interpretación crítica justifica el que se concluyan trabajos de diferentes contextos antes que emplear una única estimación como índice universal de rendimiento.

ISSN: 3091-2008
Edición Especial:
Vol. 4 Num. 10, PP 176-193
DOI: https://doi.org/

MATERIALES Y MÉTODOS

Tipo de método de investigación

Se llevó a cabo una revisión crítica de la literatura, dispuesta en un enfoque cualitativo y en un alcance descriptivo-comparativo, en torno a los resultados publicados sobre detección y predicción de reacciones adversas mediante inteligencia artificial. La estructura de información se llevó a cabo tomando como referencia los apartados adecuados de la declaración PRISMA 2020 (23), que orientan la descripción de los criterios, fuentes, mecanismos de realización y limitaciones.

Pregunta y criterios de inclusión y exclusión

La pregunta guiadora que se llevó a cabo fue la siguiente: "¿Qué aplicaciones y/o resultados han sido estos que han estado comunicados por parte de la inteligencia artificial en la detección y predicción de reacciones adversas a los fármacos y qué limitaciones son las que restringe su uso clínico?

Se consideraron pertinentes los estudios publicados en relación con la detección de señales de seguridad, predicción de acontecimientos adversos a fármacos, extracción de datos clínicos, procesamiento de reportes y análisis de interacciones farmacológicas. Se otorgaron prioridad a revisiones con métodos explícitos, resultados cuantitativos atribuidos y artículos originales representativos. No se consideraron para la interpretación cuantitativa cifras no rastreables, artículos con escasa relación con la seguridad de los medicamentos y afirmaciones de causalidad sólo derivadas de informes espontáneos.

Población o muestra

En cuanto a la población o muestra, al ser una revisión sistemática, la muestra estuvo constituida por artículos científicos originales publicados en revistas indexadas. Se realizó una búsqueda sistemática en bases de datos científicas de reconocido prestigio internacional como son MEDLINE, EMBASE y Google Scholar. Se emplearon combinaciones de términos sobre farmacovigilancia, reacciones adversas, inteligencia artificial, machine learning, deep learning, procesado del lenguaje natural, predicción y detección. La consideración se realizó para revisiones con el método descrito y resultados interpretables, así como los estudios originales representativos para dar contexto a aplicaciones concretas. No fue posible realizar exportaciones integrales ni recuentos de búsqueda desde Scopus o Web of Science, con lo cual el tipo de búsqueda que se realizó no se considera terminal en dichas bases.

Unidad de análisis, muestra y entorno

La unidad de análisis fueron los informes científicos y documentos normativos. En ningún momento se llevaron a cabo reclutamientos de pacientes o se generaron muestras clínicas primarias. Se llevó a cabo la comparación de ocho revisiones científicas sobre predicción de eventos, detección de señales, automatización de notificaciones posiblemente emergentes y análisis de textos, sumándose además los estudios originales y documentos de referencia para ofrecer la contextualización de Extracción, variables y valoración crítica

Para todas las ocho revisiones primarias nucleares se recogió el autor, el año, el objeto de estudio, el número de estudios incluidos, las fuentes de datos, los indicadores de rendimiento, la validación comunicada y las limitaciones reconocidas. También se recogieron artículos originales sobre relatos clínicos, el informe de reacciones adversas espontáneas, el uso de interacciones y los duplicados. A su vez, se separaron los resultados de la validación desarrollada, de la validación interna y de la validación externa. La valoración crítica empleó las recomendaciones metodológicas consignadas en las publicaciones originales, pero no se imitó una nueva valoración independiente del riesgo de sesgo a partir de PROBAST u otras herramientas.

Síntesis y análisis estadístico

Se realizó una síntesis narrativa, así como una comparación descriptiva de los indicadores que cada fuente refería. Las tablas transcribían estimaciones, tamaños de muestra y diseños sin agregar artificialmente resultados que pudieran corresponder a estudios primarios. Las figuras fueron elaboradas a partir de porcentajes publicados y se registró la fuente. No se hicieron nuevas pruebas de hipótesis ni metaanálisis. Los intervalos de confianza, los valores de heterogeneidad y los desempeños fueron atribuidos a sus investigaciones origen- primera. Para la elaboración de gráficos se usaron Python y Matplotlib.

Aplicación de la guía PRISMA 2020 y límites de trazabilidad

Se aplicó PRISMA 2020 como guía de transparencia del relato, no como la catadura de una exhaustividad automática. El propio material de partida no contenía protocolo, ni historias completas de recuperación-exclusión, ni selección dual documentada; tampoco se obtuvo un listado completo que replicara esos pasos. Para tanto, no se elaboraron totales de registros, de duplicados o de textos excluidos, ni se incluyó un diagrama de selección con cantidades supuestas. La síntesis se refiere explícitamente a una revisión crítica de la literatura reconocida y verificada, no a una revisión sistemática de novo completada (23).

https://politecnicadelaciencia.com/

ISSN: 3091-2008

Edición Especial:
Vol. 4 Num. 10, PP 176-193
DOI: https://doi.org/

RESULTADOS

El panorama general de la literatura analizada

Se estableció una matriz de ocho revisiones centradas en el uso de métodos computacionales para identificar o predecir el posible daño asociado a medicamentos. Sus diseños, períodos de búsqueda y unidades de análisis no fueron equivalentes: algunas investigaciones integraron modelos aplicados específicamente a historias clínicas, mientras que otras trabajaron con la farmacovigilancia poscomercialización y otra realizó un trabajo centrado en grandes modelos de lenguaje. Por esta razón, el número de estudios incluidos en cada revisión se informa de forma separada y no se sumó, debido a la posibilidad de solapamiento de estudios primarios (tabla 1).

Tabla 1. Características y principales resultados de las ocho revisiones comparadas.

Revisión	Ámbito	Estudios (n)	Resultado principal
Hu y colaboradores, 2024 (15)	Predicción de eventos específicos con historias clínicas	59	AUC promedio de 0,7668; AUC resumida de 0,8069.
Hu y colaboradores, 2024 (16)	Predicción de múltiples eventos con historias clínicas	10	AUC combinada de 0,72 (IC 95 %: 0,68–0,75).
Dsouza y colaboradores, 2025 (17)	Predicción de reacciones adversas en hospitalizados	13	Validación externa en el 23 %; sensibilidad y especificidad agrupadas según diseño.
Chalabianloo y colaboradores, 2026 (18)	Modelos predictivos de eventos adversos	59	191 implementaciones; 18,6 % con validación externa.
Karampatea y colaboradores, 2026 (19)	Modelos en datos clínicos reales	15	Validación temporal/externa poco frecuente; sin patrón uniforme de comparación.
El Sayed y colaboradores, 2026 (20)	Aplicaciones amplias de inteligencia artificial	67	50,5 % de aplicaciones destinadas a detección de señales.
Kim y colaboradores, 2026 (21)	Modelos extensos de lenguaje en seguridad	83	Aplicaciones de extracción y clasificación; rendimiento heterogéneo.
Cappello y colaboradores, 2026 (25)	Predicción multidroga: metaanálisis multinivel	11 (9 cuant.)	AUC principal de 0,841 (IC 95 %: 0,789–0,883).

Nota: *AUC: área bajo la curva; IC: intervalo de confianza.*

Panorama general de las aplicaciones. La evidencia investigada indica que la inteligencia artificial se ha usado en diversos trabajos dentro de la farmacovigilancia. Entre ellos consta la extracción automática de información contenida en reportes clínicos, la clasificación y priorización de notificaciones, la detección de combinaciones de medicamentos potencialmente problemáticas,

la predicción de los eventos adversos como un subproducto de características del paciente y la ayuda al seguimiento documental de casos. Estas funciones no tienden necesariamente a la misma meta: algunas desean detectar una mención textual; otras desean estimar la posibilidad de un evento futuro. Como consecuencia de ello, sus resultados no han de ser comparables como si midieran una capacidad tecnológica única (5,14,20).El Sayed y colaboradores publicaron en 2026 una revisión sistemática en la que se incluyen 67 estudios en los que se documenta que la detección de señales de seguridad es responsable del 50,5 % de las aplicaciones, a continuación la predicción de efectos adversos (21 %), la eficiencia del procesamiento de casos (15 %), la identificación de las interacciones de fármacos (7,5 %) y la detección de pacientes con mayor riesgo (6 %) (20). La distribución que se muestra en este estudio corresponde a la muestra y a los criterios de esa revisión, no es representativa de una frecuencia general de los estudios existentes. El hecho de que la detección de señales sea la predominante forma parte de la necesidad de procesar un importante número de informes y de centrar la revisión humana en los casos de mayor interés.

Detección de señales en sistemas de notificación. Los métodos clásicos de desproporcionalidad establecen una comparación entre la frecuencia relativa de una combinación medicamento-evento y la de las otras combinaciones notificadas. En EudraVigilance se aplica, entre otros, el método de la razón de probabilidades de notificación para apoyar la vigilancia de señales (4). Al ser planteado de esta manera, los algoritmos de machine learning pueden añadir relaciones adicionales entre variables, diferenciando patrones y ayudando a priorizar los expedientes. Sin embargo, un resultado estadísticamente relevante no es más que una pista para posteriores investigaciones; la inexistencia de un denominador fiable de las personas expuestas impide la interpretación del número de notificaciones en términos de riesgo absoluto (3,5).

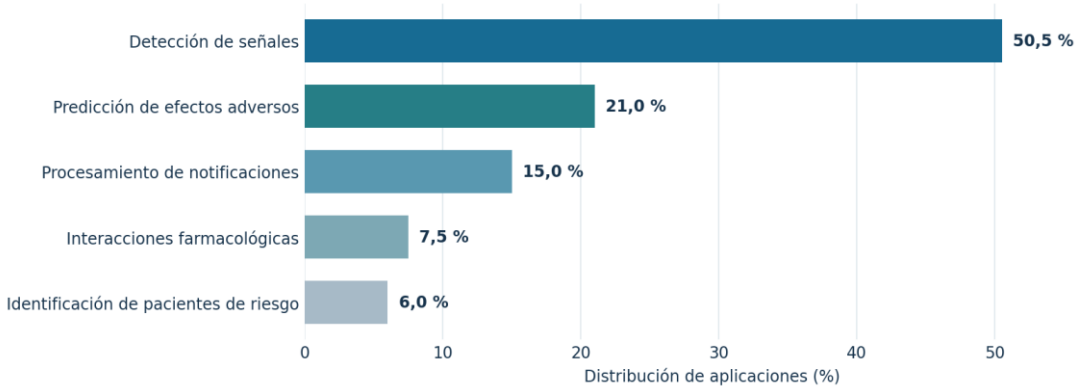


Figura 1. Distribución de las aplicaciones de inteligencia artificial en farmacovigilancia.

El planteamiento de un método de corrección de factores de confusión a partir de datos de notificaciones que exploran efectos asociados a las interacciones medicamentosas de clases farmacológicas fue realizado por Tatonetti, entre otros, el cual, además de evidenciar 47 asociaciones entre clases de fármacos en función de una revisión independiente de historias

clínicas electrónicas, identificó una asociación potencial entre el uso combinado de ciertos antidepresivos y determinados diuréticos en los que pudiese producirse una prolongación del intervalo QT (6). Esta publicación pone de manifiesto las posibilidades de integrar diversas fuentes de datos, pero también revela que una señal computacional debe, en su caso, exigirse una validación adicional y que las asociaciones observacionales no son iguales a las causas importantes del clínico.

Otro uso operativo reside la existencia de las notificaciones duplicadas, capaces de modificar el aparente interés de una señal determinada, como demuestran Kreimeyer, et al., con una propuesta de agrupamiento mediante redes para encontrar patrones en las notificaciones y los posibles duplicados en el sistema estadounidense de notificación de eventos adversos investigados (13).La importancia de esta aplicación se da para obtener una mejor cualidad de los registros que se evaluarán antes de un análisis, y no para mejorar la identificación exclusivamente de los acontecimientos que son efectivamente provocados por un medicamento. Y además, la estandarización terminológica y la revisión lógica de constancias irregulares son igualmente necesarias para prevenir conclusiones erróneas.

Reconocimiento de acontecimientos en textos clínicos. Las notas de evolución, los informes de alta o de progresos, así como p.ej. las historias clínicas están repletas de información respecto de síntomas, de cambios de dosis y de secuenciación temporal, información que muchas veces no se traduce en campos estructurados. LePendu y colaboradores han mostrado la conveniencia de manejar notas clínicas para extraer información útil para la vigilancia farmacológica (7). Bagattini y colaboradores propusieron un marco de referencia de clasificación que utiliza características temporales que están fuera de los registros de los campos estructurados, y que muestra una vía para reconocer patrones clínicos a partir de informaciones longitudinales (11). Dichas investigaciones enfatizan que la estructura temporal del dato puede llegar a ser tan importante como la mera identificación aislada de una palabra o un diagnóstico.

Las técnicas de procesamiento del lenguaje natural se han aplicado a la detección de entidades (por ejemplo, los nombres de medicación, los síntomas, los posibles efectos adversos) así como a la identificación de relaciones entre entidades. En un artículo revisado de 12 artículos, Li et al. informan ventajas de los enfoques de aprendizaje profundo en estas tareas de identificación de entidades, pero también los resultados elogiando el rendimiento de los modelos de lenguaje contextualizados en el procesamiento completo de texto clínico (14). La comparación entre el estado de la cuestión estaba marcada por conjuntos de referencia, definiciones de etiqueta y estrategias de evaluación.

El reconocimiento de una expresión clínica no se debe considerar suficiente para asegurar que el acontecimiento esté causado por el tratamiento de salud referido.

Información expresada por pacientes en redes sociales.Los mensajes efectuados por los usuarios son una fuente alternativa para investigar a los pacientes, con aquellas experiencias que no

aparecen entre las notificaciones formales. En este sentido, Sarker et al. señalan tanto las oportunidades como los inconvenientes de esta modalidad, haciendo hincapié en las diferentes formas lingüísticas que podrían ser utilizadas para describir el medicamento o el síntoma en cuestión (8); otras investigaciones, como la de Nikfarjam et al., empleó la etiquetación secuencial y las representaciones de palabras con el objetivo de recuperar las menciones a las reacciones adversas en mensajes de redes sociales (9). O'Connor et al. también investigaron el reconocimiento de las expresiones de reacciones adversas en publicaciones breves (10). El valor añadido de estas técnicas es ayudar a la categorización previa de grandes colecciones de texto no formal, pero no ayudar a proporcionar diagnósticos clínicos confirmados.

Predicción de eventos adversos procedentes de historias clínicas. Los modelos predictivos suelen incluir en su estructura datos demográficos, antecedentes, comorbilidades, tratamientos concomitantes y diagnóstico de complementarias.Su objetivo es el de estimar la probabilidades de que ocurra un evento determinado en un seguimiento. Datta et al. recurrieron a registros clínicos retrospectivos para identificar interacciones farmacológicas asociadas al daño hepático en pacientes tratados con antiinflamatorios no esteroides, mostrando la posibilidad de investigar riesgos asociados al uso combinado de medicamentos (12). La aplicabilidad clínica depende, sin embargo, de la exactitud del registro y de la capacidad de separar una relación particular a partir de otros factores de riesgo.

Hu, Chen et al. evaluaron 59 estudios de predicción de eventos adversos específicos a partir de historias clínicas electrónicas. Además se identificaron 38 algoritmos y reportaron un área bajo la curva media de 0,7668, con una desviación estándar de 0,1073; en la síntesis cuantitativa, la sensibilidad media agrupada fue de 0,65 y la especificidad, de 0,89; y el área bajo la curva media comunicada fue de 0,8069 (15). La discrepancia entre sensibilidad y especificidad indica que un modelo puede clasificar correctamente muchos casos, pero no ser capaz de identificar todos los eventos adversos que deberían ser vigilados. Los resultados deben entenderse con relación a los desenlaces, poblaciones y procedimientos de validación incluidos en dicha revisión.

En una segunda revisión, Hu, Li et al. revisaron 10 estudios desarrollados para predecir múltiples eventos adversos de un conjunto de historias médicas. El área bajo la curva media comunicada fue de 0,76, con un amplio intervalo de confianza y se reportaron resultados heterogéneos entre los algoritmos y las estrategias de remuestreo (16). Ambas revisiones identificaron rendimientos con potencial valor clínico, pero la diferencia entre el número de estudios y en la definición de los desenlaces imposibilita leer los promedios como si de una comparación directa entre técnicas se tratase; tampoco sería correcto sumar sus muestras, pues ambas pueden incluir estudios primarios coincidentes.

D'Souza y colaboradores revisaron 13 estudios sobre la predicción de reacciones adversas en pacientes ingresados y refirieron que el 77 % estaban orientados a la construcción de modelos y el 23 %, a la validación externa. Para los trabajos orientados a la construcción estrictamente,

informaron de sensibilidades y especificidades agrupadas de 78,1 % y 70,6 %, respectivamente. Para los trabajos con validación externa fueron de 81,5 % y 79,5 % (17). Estos porcentajes son solamente representativos de los subconjuntos que iban siendo analizados por los autores, sin poder concluir que la validación externa aumente por sí misma el rendimiento de los algoritmos, pues los grupos pueden diferir también en otras características metodológicas y su composición.

Comparación de las métricas reportadas por las revisiones

El área bajo la curva en las tareas predictivas proporciona un resumen de discriminación entre los pacientes con o sin un evento adverso, pero no nos dice nada sobre calibración, utilidad clínica o prevalencia. Hu y colaboradores reportaron una AUC media de 0,7668 en el caso de estudios de eventos concretos, y una AUC resumida de 0,8069 mediante curva característica operativa del receptor; en otra revisión de múltiples desenlaces el efecto conjunto fue de 0,72; el metaanálisis multinivel de Cappello y colaboradores obtuvo una estimación de 0,841. Esto no implica una progresión temporal ni permite afirmar que un algoritmo sea mejor que otro, ya que sus criterios de selección, sus fuentes de datos y sus desenlaces clínicos son totalmente diferentes (15,16,25).

Tabla 2. Resultados cuantitativos reportados por investigaciones de síntesis.

Estudio	Indicador	Valor publicado	Interpretación clínica
Hu y col., 2024 (15)	AUC media	0,7668 ± 0,1073	Discriminación media de modelos para eventos específicos.
Hu y col., 2024 (15)	AUC resumida	0,8069	Estimación de un modelo de síntesis, no de la totalidad de herramientas.
Hu y col., 2024 (16)	AUC combinada	0,72 (IC 95 %: 0,68–0,75)	Predicción de múltiples eventos; población y algoritmos distintos.
Dsouza y col., 2025 (17)	Sensibilidad / especificidad	78,1 % / 70,6 %	Modelos de desarrollo sin validación externa.
Dsouza y col., 2025 (17)	Sensibilidad / especificidad	81,5 % / 79,5 %	Subconjunto de modelos con validación externa; grupos no comparables causalmente.
Cappello y col., 2026 (25)	AUC metaanálisis multinivel	0,841 (IC 95 %: 0,789–0,883)	9 estudios; 30 estimaciones de modelos; certeza baja.

Calidad metodológica y capacidad de generalización. Chalabianloo y colaboradores llevaron a cabo una revisión de 59 estudios con 191 implementaciones de aprendizaje automático. Aunque la mayoría de los modelos con área bajo la curva declarada demostró internamente validez moderada o alta, únicamente un 33,9 % de los estudios tomaron en consideración el desequilibrio entre clases y un 18,6 % realizaron validación externa (18). Esta última limitación es especialmente importante cuando las reacciones adversas de interés son poco frecuentes, porque un sistema puede alcanzar un nivel de exactitud general elevado, si clasifica

correctamente la mayoría de los pacientes y no identifica de manera suficiente los pacientes relevantes.

Karampatea y colaboradores publicaron en 2026 otra revisión en los que seleccionaron 15 estudios de 281 registros revisados. Su metaanálisis fue principalmente sobre historias clínicas estructuradas y bases administrativas y también, concluyó que existía poca validación externa o temporal, mala aplicación de los métodos de explicación y ausencia de criterios de comparación homogéneos (19). Asimismo, El Sayed y sus coautores determinaron que la validación externa resultó ser el componente con menor puntuación registrado en su evaluación metaanalítica, con un promedio cercano al 52 % en el dominio correspondiente (20). Estos hallazgos ilustran una barrera común: la evidencia de desempeño que se obtiene de forma retrospectiva no aporta de una vez por todas sustento en la atención habitual.

Modelos de lenguaje recientes y automatización documental. Las herramientas que permiten analizar textos muy extensos, por otro lado, se encuentran en estudio para la extracción de información clínica, la clasificación de casos, la búsqueda de literatura pertinente y facilitar la administración de tareas. Kim et al. identificaron un total de 83 estudios empíricos asociados a modelos de lenguaje muy extensos en farmacovigilancia, destacando un rendimiento variable por función y tipo de evaluación. La revisión hizo hincapié en ejemplos de modelos que brindan buenos resultados en la extracción de datos, pero también trajo a la atención problemas de generación de contenido no fundamentado, escasa especificidad e intolerancia a la formulación de las instrucciones (21). Por lo tanto, su aplicación más plausible está en situaciones bien definidas, auditables y pasibles de verificación profesional.

Síntesis de los hallazgos. En síntesis, la evidencia recuperada revela tres hallazgos principales: ampliación de las fuentes de información analizadas, una automatización parcial de las tareas repetitivas y el desarrollo de modelos de predicción individual. A la vez se repiten cuatro condiciones limitantes: calidad desigual de los datos, riesgo de sesgos, escasa validación independiente y dificultad para trasladar la medida del desempeño del modelo informático en beneficio clínico demostrable (18-22). Por lo que la conclusión central no es la superioridad universal de un algoritmo, sino la necesidad de elegir cada herramienta según la tarea de farmacovigilancia que se deba cumplir, los datos de que se disponga, y la posibilidad de validar sus decisiones.

Limitaciones metodológicas encontradas en la evidencia

Chalabianloo y colaboradores observaron validación externa en el 18,6 % de las estudios incluidos y tratamiento explícito del desbalance entre categorías en el 33,9 %. Ambas cifras reportan problemas metodológicos diferentes: la primera hace referencia a la capacidad del modelo de generalización; la segunda, a la forma de manejar una distribución desigual entre eventos

adversos y no adversos. No deben interpretarse como proporciones complementarias de la otra ni más que como probabilidades de error de los algoritmos (18).

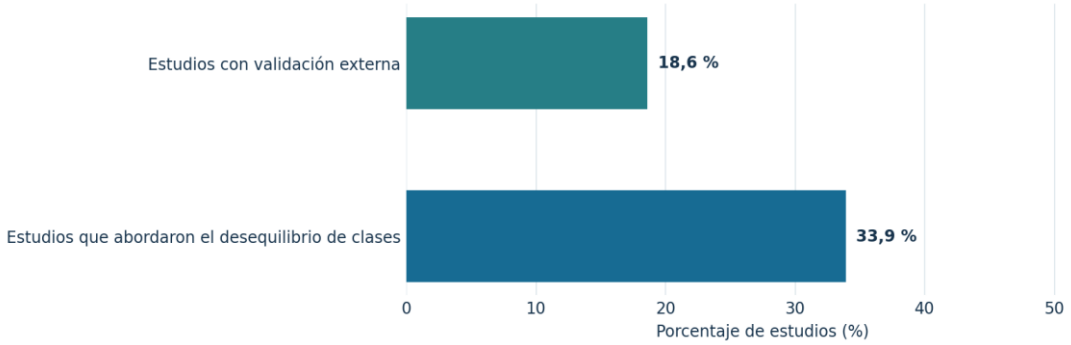


Figura 2. Proporción de estudios que documentaron dos prácticas metodológicas importantes.

Tabla 3. Aplicaciones, ventajas y riesgos de interpretación de la inteligencia artificial.

Fuente de información	Aplicación principal	Contribución potencial	Limitación fundamental
Notificaciones espontáneas	Detección de asociaciones medicamento-evento	Priorización de señales y revisión de gran volumen	Notificación selectiva; ausencia de denominadores; no demuestra causalidad (3–6).
Historias clínicas electrónicas	Predicción e identificación de eventos	Integra antecedentes, tratamientos y resultados de laboratorio	Sesgo de registro, cambios asistenciales y validación externa limitada (15–19).
Narrativas clínicas	Extracción de entidades y acontecimientos	Recupera información fuera de campos estructurados	Negaciones, contexto temporal y errores de interpretación (7,14).
Redes sociales	Detección exploratoria de experiencias	Añade expresiones espontáneas de pacientes	Autoselección, falta de verificación clínica y duplicaciones (8–10).
Modelos extensos de lenguaje	Clasificación, extracción y apoyo documental	Ayuda en tareas acotadas con supervisión	Respuestas inventadas y reproducibilidad dependiente de instrucciones (21).

DISCUSIÓN

Los hallazgos analizados corroboran la posible función de la inteligencia artificial como apoyo de la farmacovigilancia convencional, pero también manifiestan un distanciamiento entre obtener una buena actuación sobre una base retrospectiva y mejorar la seguridad del paciente. La revisión

realizada por El Sayed y coautores indicó que la señalización reunía casi la mitad de las aplicaciones localizadas (20), lo cual es coherente con el interés histórico por extraer partido a los registros de notificación para detectar patrones que no llegan a ser evidentes mediante la revisión manual; pero la advertencia de la autoridad reguladora de EE.UU. sobre la imposibilidad de inferir sobre la incidencia y la causalidad de esos informes se mantiene vigente para los análisis automatizados (3), -la técnica no elimina las limitaciones de la fuente que utiliza.

La confrontación entre los estudios predictivos pone de manifiesto una heterogeneidad considerable. Hu, Chen y sus colaboradores indicaron haber encontrado un área bajo la curva promedio alrededor de 0,77 en modelos orientados a eventos específicos (15), mientras que Hu, Li y sus colaboradores indicaron un área media de 0,76 al examinar la predicción de múltiples acontecimientos (16). La similitud entre ambos estimados no implica una equivalencia metodológica, dado que las revisiones tuvieron diferentes criterios de selección y de desenlace. Además, la sensibilidad en conjunto hallada de 0,65 entre Hu, Chen y sus colaboradores sugeriría que una capacidad discriminativa aceptable puede coexistir con una tasa de acontecimientos no detectados (15). Tal disconformidad tiene relevancia en la farmacovigilancia, puesto que la urgencia clínica, normalmente, depende de detectar precozmente acontecimientos adversos graves o poco frecuentes.

La comparación entre las sintéticas de Hu y colaboradores y el metaanálisis de Cappello y colaboradores sugiere que, a nivel conceptual, existe una diferencia metodológica fundamental. Si bien las publicaciones iniciales analizaron modelos asociados a eventos o fuentes clínicas concretos, en la segunda publicación se estudió el problema de las dependencias entre estimaciones obtenidas a partir de un mismo estudio. Si bien el valor conjunto de 0,841 puede parecer notablemente superior a los promedios de las otras revisiones, esta constatación no implica una mejora tecnológica; la heterogeneidad de las poblaciones y de los desenlaces impide una comparación de tipo causal entre cifras agregadas. De hecho, la heterogeneidad del 90,7 %, expuesta en el análisis convencional de sensibilidad de Cappello y colaboradores, así como la certeza de la evidencia calificada como baja, invitan a la cautela (15,16,25).

Hay una buena aproximación entre los estudios de Dsouza, Chalabianloo y Karampatea. Los tres investigadores coinciden en que el porcentaje de evaluaciones realizadas fuera de la muestra de desarrollo o en períodos diferentes es escaso, aunque no tienen los mismos criterios para decir que han realizado una validación y para contarla. La ausencia de esta fase impide que un modelo se pueda trasladar a hospitales con un sistema de prescripción distinto, distintas formas de registrar o diferentes perfiles de paciente. En el fondo, un buen rendimiento local no implica que la herramienta sea útil en general: antes de su implementación hay que comprobar su discriminación, su calibración, los efectos de su prevalencia y el rendimiento por subgrupos clínicos (17-19).

Sayed y colaboradores encontraron que la detección de señales agrupa aproximadamente la mitad de las aplicaciones de inteligencia artificial que se han analizado. Esto coincide con el valor operativo de clasificar un gran número de informes, pero no implica que estas herramientas sean capaces de disminuir per se la morbimortalidad por medicamentos. Un sistema automatizado puede establecer una asociación dentro de un contexto determinista en el que, al mismo tiempo, puede estar afectado por confusión por indicación, duplicación de notificaciones o sesgo de notoriedad. En esta fase, la evaluación farmacológica de temporalidad, plausibilidad, diagnóstico alternativo y exposición sigue siendo fundamental para configurar una señal de una relación causal (3,5,20).

Las conclusiones sobre modelos extensos de lenguaje requieren una precaución adicional. Kim et al. sostuvieron que había aplicaciones prometedoras para extraer entidades, revisar textos y ayudar a la clasificación, pero también apuntaron problemas de especificidad, producción de respuestas incorrectas y escasa evaluación prospectiva. Por tanto, el soporte computacional es más razonable en tareas bien definidas, que tienen verificación del dato fuente, control de versiones y trazabilidad de las decisiones, que en una automatización integral de las conclusiones de seguridad (21,22,24).

Respecto a las limitaciones del presente artículo, el punto fuerte se refiere a la comparación explícita de los resultados y el respeto a los denominadores, las métricas y los diseños originales. Sin embargo, la ausencia de un registro completo de búsquedas y exclusiones, la falta de doble cribado documentado y la ausencia de un nuevo examen independiente del riesgo de sesgo no nos permite afirmar que se ha completado una revisión sistemática de novo. Es más, pueden existir estudios primarios incluidos en más de una de las ocho revisiones comparadas. Esta limitación condiciona las inferencias sobre la totalidad de la literatura y señala el procedimiento indispensable para eventuales presentaciones bajo la categoría formal de revisión sistemática (23).

CONCLUSIONES

El conjunto de indicios analizados nos lleva a lograr el propósito de aportar pruebas que demuestran que la inteligencia artificial tiene aplicaciones concretas en la detección y predicción de reacciones adversas a medicamentos. Su capacidad se centra en la extracción de texto de documentos clínicos, la identificación y priorización de las señales desde grandes repositorios, la ayuda al tratamiento de informes duplicados o la estimación de riesgos a partir de las historias clínicas electrónicas; todo ello permite contestar a la pregunta por sus principales aportes en relación a la farmacovigilancia, sin atribuir a estas herramientas la capacidad independiente para determinar causalidad.

La comparación de la evidencia publicada muestra que los modelos pueden llegar a obtener resultados predictivos alentadores, pero las cifras que se consideran son muy variables en

relación con las distintas poblaciones, las distintas definiciones de eventos adversos y las distintas técnicas de validación. La no-suficiencia de la evidencia externa y prospectiva es una limitación para su aplicación a otros centros y contextos clínicos. Por tanto, los índices de sensibilidad, especificidad u área bajo la curva no deben ser considerados como cifras independientes de la calidad de los datos y de la pertinencia clínica de los eventos que se pretenden detectar.

Finalmente, la incorporación responsable de la inteligencia artificial en los sistemas de farmacovigilancia exige una evaluación independiente, los mecanismos de supervisión profesional, la protección de los datos personales o la ejecución de procedimientos de seguimiento transparentes. Los estudios siguientes deberían comparar los modelos con las prácticas convencionales, medir su rendimiento en distintos entornos y conocer si el empleo de estos algoritmos permite disminuir el tiempo requerido para investigar las señales o, más bien, si la utilización de los mismos ayuda a tomar decisiones sobre la seguridad. Por lo tanto, la inteligencia artificial se convierte en una herramienta de apoyo con un gran potencial, cuyo momento de verdad dependerá de una puesta en marcha metodológicamente correcta y clínicamente verificable.

REFERENCIAS BIBLIOGRÁFICAS

1. Organización Mundial de la Salud. Programa de Vigilancia Farmacéutica Internacional [Internet]. Ginebra: OMS; [consultado 31 oct 2026]. Disponible en: https://www.who.int/teams/regulation-prequalification/regulation-and-safety/pharmacovigilance/networks/pidm
2. Hazell L, Shakir SAW. Under-reporting of adverse drug reactions: a systematic review. Drug Saf. 2006;29(5):385-396. doi:10.2165/00002018-200629050-00003.
3. Food and Drug Administration. FDA Adverse Event Reporting System (FAERS): reporting and review [Internet]. Silver Spring: FDA; [consultado 31 oct 2026]. Disponible en: https://www.fda.gov/media/165667/download
4. European Medicines Agency. EudraVigilance system overview [Internet]. Ámsterdam: EMA; [consultado 31 oct 2026]. Disponible en: https://www.ema.europa.eu/en/human-regulatory-overview/research-development/pharmacovigilance-research-development/eudravigilance/eudravigilance-system-overview
5. Harpaz R, DuMouchel W, Shah NH, Madigan D, Ryan P, Friedman C. Novel data-mining methodologies for adverse drug event discovery and analysis. Clin Pharmacol Ther. 2012;91(6):1010-1021. doi:10.1038/clpt.2012.50.
6. Tatonetti NP, Ye PP, Daneshjou R, Altman RB. Data-driven prediction of drug effects and interactions. Sci Transl Med. 2012;4(125):125ra31. doi:10.1126/scitranslmed.3003377.
7. LePendu P, Iyer SV, Bauer-Mehren A, Harpaz R, Mortensen JM, Podchiyska T, et al. Pharmacovigilance using clinical notes. Clin Pharmacol Ther. 2013;93(6):547-555. doi:10.1038/clpt.2013.47.

8. Sarker A, Ginn R, Nikfarjam A, O'Connor K, Smith K, Jayaraman S, et al. Utilizing social media data for pharmacovigilance: a review. J Biomed Inform. 2015;54:202-212. doi:10.1016/j.jbi.2015.02.004.
9. Nikfarjam A, Sarker A, O'Connor K, Ginn R, Gonzalez G. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. J Am Med Inform Assoc. 2015;22(3):671-681. doi:10.1093/jamia/ocu041.
10. O'Connor K, Pimpalkhute P, Nikfarjam A, Ginn R, Smith KL, Gonzalez G. Pharmacovigilance on twitter? Mining tweets for adverse drug reactions. AMIA Annu Symp Proc. 2014;2014:924-933. PMID:25954400.
11. Bagattini F, Karlsson I, Rebane J, Papapetrou P. A classification framework for exploiting sparse multi-variate temporal features with application to adverse drug event detection in medical records. BMC Med Inform Decis Mak. 2019;19:7. doi:10.1186/s12911-018-0717-4.
12. Datta A, Flynn NR, Barnette DA, Woeltje KF, Miller GP, Swamidass SJ. Machine learning liver-injuring drug interactions with non-steroidal anti-inflammatory drugs (NSAIDs) from a retrospective electronic health record (EHR) cohort. PLoS Comput Biol. 2021;17(7):e1009053. doi:10.1371/journal.pcbi.1009053.
13. Kreimeyer K, Spiker J, Dang O, De S, Ball R, Botsis T. Deduplicating the FDA adverse event reporting system with a novel application of network-based grouping. J Biomed Inform. 2025;165:104824. doi:10.1016/j.jbi.2025.104824.
14. Li Y, Tao W, Li Z, Sun Z, Li F, Fenton S, et al. Artificial intelligence-powered pharmacovigilance: a review of machine and deep learning in clinical text-based adverse drug event detection for benchmark datasets. J Biomed Inform. 2024;152:104621. doi:10.1016/j.jbi.2024.104621.
15. Hu Q, Chen Y, Zou D, He Z, Xu T. Predicting adverse drug event using machine learning based on electronic health records: a systematic review and meta-analysis. Front Pharmacol. 2024;15:1497397. doi:10.3389/fphar.2024.1497397.
16. Hu Q, Li J, Li X, Zou D, Xu T, He Z. Machine learning to predict adverse drug events based on electronic health records: a systematic review and meta-analysis. J Int Med Res. 2024;52(12):03000605241302304. doi:10.1177/03000605241302304.
17. Dsouza VS, Leyens L, Kurian JR, Brand A, Brand H. Artificial intelligence (AI) in pharmacovigilance: a systematic review on predicting adverse drug reactions (ADR) in hospitalized patients. Res Social Adm Pharm. 2025;21(6):453-462. doi:10.1016/j.sapharm.2025.02.008.
18. Chalabianloo N, Ahmadi F, Omrani MA, Abdullah SS, Rostamzadeh N, Jafari A, et al. Machine learning methods for predicting adverse drug events: a systematic review. Br J Clin Pharmacol. 2026;92(2):422-444. doi:10.1002/bcp.70377.
19. Karampatea A, Kassandros K, Constantinides T, Kontogiorgis C. Systematic review of AI-based models in pharmacoepidemiology for adverse drug event prediction and detection. Front Drug Saf Regul. 2026;6:1773186. doi:10.3389/fdsfr.2026.1773186.

20. El Sayed I, Bedwany RN, Salama M, Eltrass ASH. Integration of artificial intelligence into pharmacovigilance: a systematic review. Res Social Adm Pharm. 2026;publicación anticipada. doi:10.1016/j.sapharm.2026.09.001.
21. Kim TY, Oh WS, Jeong DH. Large language models in adverse drug reaction detection and pharmacovigilance: a systematic review of current applications, challenges, and future directions. Diagnostics (Basel). 2026;16(15):2435. doi:10.3390/diagnostics16152435.
22. Norén GN, Durand J, Kara V, MacEntee Pileggi E, Carroll H, Hill S, et al. Artificial intelligence in pharmacovigilance: guiding principles from the CIOMS Working Group XIV. Drug Saf. 2026;publicación anticipada. doi:10.1007/s40264-026-01708-z.
23. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. doi:10.1136/bmj.n71.
24. European Medicines Agency. Reflection paper on the use of artificial intelligence in the medicinal product lifecycle [Internet]. Ámsterdam: EMA; 2024 [consultado 31 oct 2026]. Disponible en: https://www.ema.europa.eu/en/scientific-guidelines/use-artificial-intelligence-ai-medicinal-product-lifecycle
25. Cappello G, Oster C, Schmidt T, Grieger J, Jekel L, Kizina K, et al. Machine learning models for adverse drug reaction prediction. A systematic review and three-level multilevel meta-analysis. Front Pharmacol. 2026;17:1906920. doi:10.3389/fphar.2026.1906920.