Medicina Intensiva 49 (2025) 154–164

medicina intensiva

http://www.medintensiva.org/

PUESTA AL DÍA EN MEDICINA INTENSIVA: SEGURIDAD DEL PACIENTE CRÍTICO

Perspectivas actuales sobre el uso de la inteligencia artificial en la seguridad del paciente crítico

Jesús Abelardo Barea Mendoza[a,*], **Marcos Valiente Fernandez**[a], **Alex Pardo Fernandez**[b] y **Josep Gómez Álvarez**[c]

[a] *UCI de Trauma y Emergencias, Servicio de Medicina Intensiva, Hospital Universitario 12 de Octubre, Instituto de Investigación Hospital 12 de Octubre, Madrid, España*
[b] *Universidad Rovira i Virgili, Tarragona, España*
[c] *Hospital Universitario de Tarragona Joan XXIII, Universidad Rovira i Virgili, Instituto de Investigación Sanitaria Pere i Virgili, Tarragona, España*

Recibido el 19 de diciembre de 2023; aceptado el 11 de marzo de 2024
Disponible en Internet el 20 de abril de 2024

PALABRAS CLAVE
Cuidados críticos;
Seguridad del paciente;
Predicción;
Evaluación de riesgos;
Algoritmos;
Inteligencia artificial;
Aprendizaje automático;
Eventos adversos

Resumen Las unidades de cuidados intensivos (UCI) han sido objeto de mejoras en la seguridad del paciente y la inteligencia artificial (IA) se presenta como una tecnología disruptiva que ofrece nuevas oportunidades. Aunque la evidencia publicada es limitada y tiene problemas metodológicos, algunas áreas resultan prometedoras como los sistemas de ayuda a la decisión, detección de eventos adversos o errores de prescripción. El uso de la IA en seguridad puede tener un objetivo diagnóstico o predictivo. La implementación de sistemas basados en IA requiere procedimientos para garantizar la asistencia segura, enfrentando desafíos como la confianza en dichos sistemas, sesgos, calidad de los mismos, escalabilidad y consideraciones éticas y de confidencialidad. El desarrollo y la aplicación de la IA demandan pruebas exhaustivas, incluyendo testeo sobre datos retrospectivos, validación con cohortes prospectivas en tiempo real y demostración de eficacia en ensayos clínicos. La transparencia y explicabilidad algorítmica resultan esenciales siendo la participación activa de profesionales clínicos en la implementación crucial.

* Autor para correspondencia.
Correo electrónico: jesusabelardo.barea@salud.madrid.org (J.A. Barea Mendoza).

https://doi.org/10.1016/j.medin.2024.03.007
0210-5691/

KEYWORDS
Critical care;
Patients safety;
Prediction;
Risk assessment;
Algorithms;
Artificial intelligence;
Machine learning;
Adverse events

Current Perspectives on the Use of Artificial Intelligence in Critical Patient Safety

Abstract Intensive Care Units (ICUs) have undergone enhancements in patient safety, and artificial intelligence (AI) emerges as a disruptive technology offering novel opportunities. While the published evidence is limited and presents methodological issues, certain areas show promise, such as decision support systems, detection of adverse events, and prescription error identification. The application of AI in safety may pursue predictive or diagnostic objectives. Implementing AI-based systems necessitates procedures to ensure secure assistance, addressing challenges including trust in such systems, biases, data quality, scalability, and ethical and confidentiality considerations. The development and application of AI demand thorough testing, encompassing retrospective data assessments, real-time validation with prospective cohorts, and efficacy demonstration in clinical trials. Algorithmic transparency and explainability are essential, with active involvement of clinical professionals being crucial in the implementation process.

Inteligencia artificial y seguridad en el ámbito sanitario

La inteligencia artificial (IA) no es un concepto nuevo ni reciente. La primera vez que se empleó este término fue en la década de los 50[1]. Sin embargo, es en los últimos años cuando su implementación a gran escala ha sido posible gracias al crecimiento exponencial de la tecnología[2]. La IA es percibida como una herramienta muy poderosa con un potencial enorme para cambiar la forma en la que vivimos, sin embargo, no está exenta de controversia[3]. De todos los tipos de IA que existen, la que es capaz de llegar a niveles más altos de capacidad predictiva o incluso creativa está basada en modelos de aprendizaje profundo[4]. Estos modelos son modelos no lineales basados en un gran número de operaciones matemáticas que complican su interpretación[5,6]. Es precisamente este último tipo de IA el que genera más inquietud y desconfianza para la sociedad, y se está tratando de analizar desde la ética y el derecho[7].

Si bien la introducción de la IA requiere de un análisis de riesgo-beneficio en cualquier escenario, en el ámbito de la salud este debe ser profundo y exhaustivo. Se ha propuesto como la solución a muchos de los problemas de los actuales sistemas de salud contribuyendo a evitar muertes, reducir días de hospitalización o prevenir eventos adversos[8–16].

A pesar de los esfuerzos recientes por incrementar la cultura de seguridad en los sistemas sanitarios, los eventos adversos siguen contribuyendo de forma muy relevante a la morbimortalidad y el gasto sanitario. La incorporación de nuevas tecnologías emerge como una estrategia prometedora en ese sentido. El entorno del paciente crítico resulta un escenario ideal para la implementación de este tipo de tecnologías. En ellas, coexisten pacientes extremadamente graves sobre los que se realizan diariamente numerosas intervenciones y tratamientos de alta complejidad susceptibles al error. Se une además la presión por la toma de decisiones rápidas al ser un área típica de atención a patologías tiempo dependientes. El volumen de información que se genera en una Unidad de Cuidados Intensivos (UCI) puede ser abrumador, habiéndose argumentado que excede a la capacidad de un clínico experto[17]. Así el incremento de información disponible en diferentes formatos (imágenes, pruebas de laboratorio, genética, monitorización invasiva de la fisiología, etc.) no siempre se asocia a mejores decisiones. En las áreas de cuidados críticos la incorporación de tecnologías derivadas de la IA podría ayudar a los clínicos a incrementar las capacidades diagnósticas o terapéuticas y contribuir a la mejora de los desenlaces facilitando una mejor integración de la información[18]. Existe una creciente controversia debido a la proliferación de publicaciones de IA con escasa calidad metodológica y validez más que dudosa, lo cual limita su implementación. Una reciente revisión sistemática que incorporó más de 400 estudios que reportaron modelos desarrollados en el entorno del paciente crítico determinó que más del 20% se encontraban relacionados con la prevención de eventos adversos. Sin embargo, los autores destacan que la calidad metodológica de los artículos fue muy escasa siendo la mayoría retrospectivos (96,4%) con un alto riesgo de sesgo[19].

A la luz de estas razones, el objetivo del presente manuscrito es revisar la potencial contribución de la IA a la seguridad del paciente en el entorno del enfermo crítico resumiendo los aspectos técnicos y ofreciendo ejemplos. En un segundo lugar se han revisado aspectos de la seguridad propios de los procesos de implementación de tecnologías derivadas de la IA que sin duda formarán parte del futuro de nuestras unidades.

Principales aplicaciones de la inteligencia artificial en la seguridad del paciente

Ayuda a la toma de decisión. Sistemas de soporte a la decisión

Las innovaciones tecnológicas permiten la implementación de Sistemas de Soporte a la Decisión Clínica (CDSS) que pretenden asistir a los médicos en la toma de decisiones

ayudando a identificar rápidamente patrones de problemas potenciales (que no son fácilmente perceptibles por los humanos) y a sugerir planes de tratamiento óptimos. Estas herramientas pueden analizar y sintetizar rápidamente grandes conjuntos de datos clínicos (historia clínica, constantes vitales o imagen)[20]. Los CDSS no son nuevos remontándose su desarrollo a los avances informáticos desde 1970[21]. Como se ha señalado, las UCI constituyen un nicho para los CDSS debido a sus características inherentes (alta disponibilidad de datos, monitorización, complejidad clínica), así como las nuevas implementaciones tecnológicas basadas en IA y *Machine Learning* (ML).

¿Por qué pueden ser útiles y qué papel pueden desempeñar los CDSS?

La implementación de los CDSS busca mejorar la calidad en todas sus dimensiones, especialmente la seguridad. Aunque se ha demostrado su utilidad en ciertas disciplinas médicas, su papel óptimo sigue siendo objeto de debate[22]. Resultan útiles por diferentes motivos: 1) incorporación de la medicina personalizada gracias al uso de modelos basados en ML. Estos modelos han demostrado ser iguales o superiores a profesionales experimentados en varios escenarios: predicción de mortalidad, reingreso, fracaso renal, sepsis y distrés respiratorio entre otros. 2) Generación de propuestas de plan terapéutico a demanda. 3) Reducción de la sobrecarga de información permitiendo a los equipos tomar mejores decisiones basadas en el gran volumen de datos disponibles (tabla 1)[21].

¿Qué tipos de modelos utilizan los CDSS?

Los CDSS hasta hace unos años se han basado en el conocimiento previo empleando reglas del tipo «si A – entonces B», lo que suponía una simplificación de la praxis médica. Los CDSS basados en IA permiten almacenar y procesar fuentes de datos muy variables, específicas del paciente, y finalmente proponer recomendaciones con las que podemos dar *feedback* al sistema. Así se mitiga la simplificación de la praxis médica de los CDSS previos[23]. Es clave que la UCI facilite el acceso a un flujo constante de datos permitiendo la realización de análisis específicos (series temporales) que pueden evaluar las tendencias y anticiparnos potencialmente a sus problemas[24].

¿Es fácil su implantación en el ámbito clínico?

Su implantación clínica no está exenta de limitaciones, siendo algunos de los elementos clave[25]:

- *Confianza:* tanto pacientes como clínicos deben confiar en los modelos de IA utilizados. También preocupa un exceso de confianza por parte del clínico ya que estos sistemas requieren de la supervisión humana en todo el proceso.
- *Sesgo:* los conjuntos de datos utilizados para entrenar modelos de IA pueden contener sesgos en función del origen de los mismos, contexto epidemiológico o tratamiento de los datos.
- *Escalabilidad:* la implementación de CDSS debe ser fácilmente escalable y adaptable a una variedad de entornos clínicos. La implementación requiere de un proceso gradual junto a retroalimentación entre desarrolladores y profesionales clínicos. También son necesarias la mejora de la calidad de los datos, la realización de numerosas iteraciones y ajustes, así como la optimización del flujo de trabajo.
- *Despliegue:* enfrenta desafíos regulatorios debido al acceso a datos de carácter personal especialmente sensibles, así como por la escasa reproducibilidad de los resultados. Sin embargo, en el caso de los CDSS el interés puede recaer en explotar las características locales, con reevaluaciones de los datos y refinamientos periódicos[26,27].
- *Éticos:* la implementación de estos sistemas puede suponer un desafío cultural y ético, afectando a la concepción de la autonomía de médicos y pacientes en función de las sugerencias de los CDSS[28].
- *Perspectiva de los clínicos:* la evidencia muestra que los clínicos se muestran favorables a la integración de CDSS, especialmente para ciertas preguntas clínicas como la probabilidad de reingreso (fig. 1). También buscan entender los factores que contribuyen a las predicciones. Resulta así de especial interés el desarrollo de la explicabilidad algorítmica (XAI, *eXplanaible Artificial Intelligence*)[29].

Predicción de eventos adversos en la Unidad de Cuidados Intensivos

Recientemente se ha propuesto el empleo de la IA en los principales dominios de los eventos adversos destacando la predicción, prevención y detección temprana de pacientes en riesgo de deterioro. De esta manera la IA, basada en la automatización de registros y el uso de ML, ofrece nuevas estrategias para mitigar la aparición de evento adverso (EA)[30]. Actualmente existen algoritmos que utilizan los datos de pacientes en tiempo real acumulando información para personalizar el tratamiento durante su ingreso. Estas herramientas, por ejemplo, permitirían iniciar o discontinuar la terapia antitrombótica según el riesgo de sangrado en un momento concreto del ingreso[31]. Otros algoritmos permiten resumir toda la información de un paciente respecto a un evento de interés. Al condensarla, facilitan la comparación con otros pacientes y la creación de cohortes pacientes con perfiles de riesgo equiparables para un determinado evento de interés (mortalidad)[32]. Son múltiples los entornos clínicos en los que se han publicado resultados en este sentido destacando: 1) predicción y estratificación del riesgo de reingreso, ayudando en la gestión del flujo de pacientes y evitando EA asociados al reingreso no programado[33,34]. 2) Predicción de EA relacionados con el fallo renal[35,36]. 3) Predicción de extubación no planeada permitiendo implementar medidas preventivas y reduciendo la carga de trabajo para el personal[37]. Aunque su implementación puede centrarse en mitigar los eventos adversos de pacientes de UCI, no debe reducirse al ámbito de UCI. Esta tecnología puede utilizarse en cualquier punto del sistema sanitario para predecir el deterioro clínico temprano o las transferencias inmediatas a UCI que permitirían instaurar

Tabla 1 Posibles funcionalidades de los CDSS, sus potenciales riesgos y estrategias de mitigación de riesgos

Funciones y utilidades de los CDSS	Daños potenciales de los CDSS	Estrategias mitigación	Explicación
Seguridad del Paciente Minimizar la incidencia de errores y eventos adversos.	Fatiga de alertas Acontece cuando se presentan demasiadas alertas insignificantes. Correría el riesgo de descartar las alarmas independientemente de su importancia.	Priorizar alertas críticas, minimizar el uso de alertas disruptivas para indicaciones no críticas.	La fatiga de alertas podría ser minimizada priorizando y seleccionando alertas que sean críticamente importantes, que tengan el mayor impacto, y personalizando las alertas según escenarios clínicos.
Manejo clínico Favorecer adherencia a guías clínicas, recordatorios de seguimiento y tratamiento, etc.	Impacto negativo en las habilidades del usuario Un ejemplo es la dependencia o confianza excesiva en la precisión de un sistema. Conflicto con la autonomía del médico.	Evitar la prescriptividad sistemática en el diseño del sistema. Evaluar el impacto del sistema de manera continua.	Los sistemas deberían ser implementados para ser útiles a los clínicos, sin comprometer la autonomía ni ser demasiado 'prescriptivos' y definitivos.
Función administrativa Selección de códigos de diagnóstico, documentación automatizada y autocompletado de notas.	Desafíos de mantenimiento del sistema y del contenido A medida que cambian las prácticas, puede haber dificultades para mantener actualizados el contenido y las reglas de conocimiento que impulsan el CDSS.	Se podrían implementar dos estrategias: 1) gestión del conocimiento establecido, con un enfoque en la traducción a sistemas CDSS. 2) Sistema para la medición y análisis del rendimiento evolutivos del CDSS.	1) Facilitar la revisión programada, métodos para adquirir e implementar nuevos conocimientos. Implementar medidas de retroalimentación de los médicos sobre el sistema, así como capacitar a los usuarios sobre la introducción adecuada de datos. 2) Es importante identificar cambios en el rendimiento y uso a lo largo del tiempo.
Soporte diagnóstico Sugerir diagnósticos basados en datos del paciente y automatizar la salida de resultados de pruebas.	Desconfianza del usuario hacia el CDSS Desacuerdo con la guía proporcionada por el CDSS.	Incluir referencias científicas en los mensajes cuando sea apropiado.	Facilitar una fuente de información verificable al usuario sobre por qué existe la recomendación. Además de aumentar la confianza, esto puede dar orientación a los usuarios para actualizar su conocimiento.
Soporte de decisión para pacientes Ayudar en la decisión a los pacientes a través de registros de salud personales y otros sistemas.	Dependencia de la alfabetización informática de los usuarios Los CDSS pueden requerir un alto nivel de competencia tecnológica para su uso.	1) Adecuarse a la funcionalidad existente. 2) Proporcionar capacitación adecuada disponible en el lanzamiento.	1) Mantener la consistencia con la interfaz de usuario del sistema preexistente (si lo hay) es crucial para asegurar que los usuarios no tengan una curva de aprendizaje larga para usar el sistema. 2) Debe estar disponible una capacitación adecuada y de fácil acceso para los usuarios.

Adaptado de Sutton R.[21]
CDSS: Sistemas de Soporte a la Decisión Clínica.

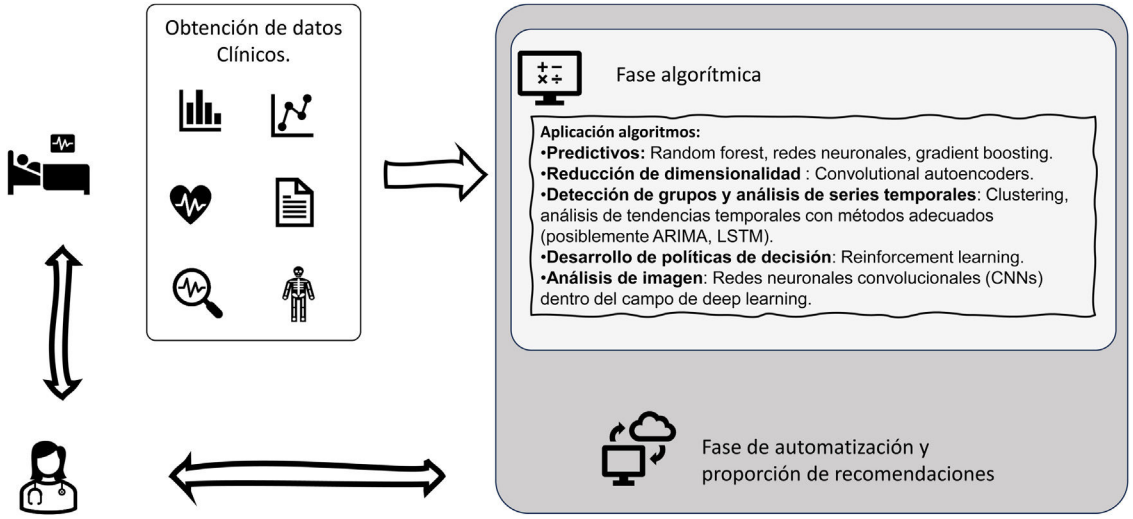

Figura 1 Representa la transmisión de la información y la generación de recomendaciones. El paciente genera datos que son aprovechados clínicamente por el médico, pero que también pueden ser aprovechados para ser valorados digitalmente. El cada vez mayor número de datos, en pacientes complejos, en ambientes con presión temporal puede hacer que se tomen decisiones sesgadas. Es por ello que el procesamiento de las múltiples fuentes de información biomédica del paciente (historia clínica, constantes pruebas radiológicas, laboratorio o medicación) puede ser transferido a sistemas que almacenen y procesen automáticamente esta información, sobre la que aplicar algoritmos. Tras la aplicación de algoritmos obtendremos una recomendación sobre la cual nosotros podríamos retroalimentar al sistema.

tratamiento precoz y organizar los recursos[38,39]. La efectividad de un algoritmo puede estar limitada fuera de su entorno original, pues los datos que utiliza reflejan la cultura y prácticas específicas de cada UCI. Así, lo que funciona en un entorno puede no ser aplicable en otro, especialmente si las condiciones de trabajo y ratios de personal varían, afectando a los resultados. Los ámbitos de aplicación de la IA en seguridad son múltiples, dentro o fuera de la UCI (estrategia UCI sin paredes) y cada vez disponemos de algoritmos más novedosos que permiten captar y adaptar mejor la información obtenida a partir de los datos de los pacientes para hacer predicciones más precisas.

Prescripción y eventos adversos asociados a medicación

Los incidentes relacionados con los medicamentos se mantienen entre los EA más frecuentes[40]. Hasta un 25% de los eventos adversos son considerados prevenibles[41]. Además, las UCI por su complejidad técnica y vinculación a la patología tiempo dependiente son más susceptibles a la aparición de errores de prescripción[42]. Las aplicaciones de la IA en esta área son diversas pudiendo destacar los modelos de predicción de riesgo para el desarrollo de reacciones adversas, la detección de eventos vinculados a la polifarmacia, el desarrollo de modelos de interacción y alergia *in silico*, la aplicación de CDSS así como la explotación de la historia clínica electrónica (EHR) para la detección de reacciones adversas inadvertidas[43]. En función del momento en que se apliquen los modelos, estos pueden ayudar a predecir el riesgo reduciendo la incidencia de los mismos (estrategias de prevención) o posteriormente contribuir a la detección precoz disminuyendo su gravedad y duración (estrategias mitigación del daño)[44].

Desde el punto de vista de los casos de uso vinculados a la predicción encontramos diversas estrategias y ejemplos en la literatura. La más frecuente es la predicción de aquellos pacientes con riesgo elevado de reacciones adversas a medicamentos. Los grupos de eventos más estudiados en este sentido son los renales, cardiovasculares, así como la sobredosis vinculada al empleo de opioides[44]. Un segundo grupo consiste en la predicción de la respuesta terapéutica. Así podría evitarse el empleo en pacientes no respondedores. Este es especialmente interesante en grupos farmacéuticos con alta incidencia de eventos adversos y mala tolerancia como los antineoplásicos o algunos antivirales[45]. También la predicción de la dosis óptima es un área de interés siendo fármacos idóneos para estos usos los anticoagulantes o los antineoplásicos[31,46,47]. Es interesante reseñar que estos modelos integran información de toda la historia clínica incorporando antecedentes, historia actual en texto libre o resultados de laboratorio. Aunque resulta prometedor es importante resaltar que la incorporación de resultados genéticos a los modelos solo ha conseguido mejorar la capacidad predictiva de forma discreta[45]. Desde un punto de vista técnico los modelos más frecuentemente empleados son los árboles de decisión, técnicas de procesamiento de lenguaje natural y de redes neuronales, aunque la variedad y variabilidad de técnicas empleadas es alta.

En el bloque de detección temprana de incidentes distinguimos la detección de las reacciones, así como la detección de errores de medicación que ya se han producido (prescripción inadecuada, interacciones y duplicidades). Sin duda, una de las áreas de mayor interés es la relacionada con los errores de prescripción. Desde un punto de vista de la seguridad, aunque estos se producen tras una acción individual (prescriptor, administrador o consumidor) son considerados como un fallo del sistema. Algunos autores incluso afirman que deben considerarse fallos de los sistemas de información

Tabla 2 Tipos de alertas más frecuentes generadas por CDSS basados en modelos

Alerta	Definición	Ejemplos
Tiempo-dependiente (sincrónica)	Datos existentes en el perfil del paciente hacen que el medicamento prescrito sea inapropiado o peligroso.	Antihipertensivo en un paciente en *shock* séptico.
Atipicidad clínica	Prescripción no se ajusta al perfil clínico del paciente	Fármaco hipoglucémico a un paciente sin diagnóstico de diabetes mellitus ni resultados indicativos de tal enfermedad
Atipicidad de dosis	La dosis de cierto medicamento se considera como un valor atípico con respecto a la distribución de dosis aprendida por el modelo de ese medicamento en la población y/o el historial propio del paciente.	Dosis raras, unidades de dosificación inusuales, frecuencia poco común, vía poco común.
Prescripción solapada	Una alerta señalada cuando hay tratamiento simultáneo con 2 medicamentos del mismo grupo.	Prescripción duplicada de perfusiones de noradrenalina con diferente formulación.
Tiempo-dependiente (asincrónica)	Una alerta señalada cuando se producen cambios en el perfil del paciente después de la prescripción, haciendo que la prescripción sea inapropiada o peligrosa para continuar.	Cuando la presión arterial disminuye y la continuación de los medicamentos antihipertensivos es inapropiada.

CDSS: Sistemas de Soporte a la Decisión Clínica.

clínica[48]. En los últimos años se han testado algunos CDSS que incorporan algoritmos de ML ayudando a la detección de errores de prescripción en tiempo real[49]. Estos modelos incorporan la información de la EHR y posteriormente detectan aquellas prescripciones que son consideradas atípicas por el modelo. Estas atipicidades pueden deberse a prescripciones infrecuentes (anticonceptivo a un lactante), discordantes con la historia clínica (antidiabético en un paciente sin antecedentes de diabetes) o con posologías infrecuentes. Estos sistemas son capaces de intervenir en dos puntos generando alertas sincrónicas (en el momento de hacer la prescripción) o asincrónicas (durante el seguimiento cuando cambia la situación clínica del paciente). Los principales tipos de alertas y ejemplos se describen en la tabla 2.

Uno de los sistemas con más desarrollo en este sentido es el *software MedAware* (https://www.medaware.com/about/) que ha sido validado de forma prospectiva. En una validación sobre más 78.000 prescripciones la tasa de alertas fue baja (0,40%). De estas el 40% fueron de tipo sincrónico siendo las más frecuentes las tiempo-dependientes (64,80%). De las alertas generadas el 89% se consideraron adecuadas y un 43% condicionaron un cambio en la prescripción. Estos datos fueron superiores al CDSS basado en reglas que presentó una carga de alertas elevada (37,10%) y de escasa trascendencia clínica (5,30% cambios de prescripción)[50].

Seguridad en los procesos de implementación de herramientas basadas en inteligencia artificial

Tal y como se ha expuesto anteriormente, la IA puede ser de utilidad en varios aspectos de la medicina intensiva. Sin embargo, también plantea retos tanto médicos como éticos y tecnológicos. Resulta relevante no solo analizar la utilidad de la IA en la seguridad del paciente sino establecer los marcos teóricos para que los procesos de uso e implementación de IA también sean seguros. Tras revisar las posibles contribuciones de esta tecnología a la seguridad, a continuación se profundizará en los riesgos que presenta así como en las soluciones en relación con los efectos adversos que pueden generar y a la implementación segura de estos algoritmos.

Cómo generar una inteligencia artificial segura basada en *machine learning*

Los algoritmos de predicción que emplean aprendizaje automático supervisado basan su funcionamiento en el aprendizaje con ejemplos. Gracias a ellos, se modela un sistema que es capaz de asociar nuevos eventos a los datos aprendidos y generar una predicción. Resulta evidente que los datos usados en el aprendizaje serán un punto clave para el éxito del modelo. Es por esto que la recolección de datos es crucial y se debe asegurar que los datos representan correctamente a la población objetivo. Entre los problemas más citados podemos destacar[51]:

- *Poblaciones/casos desbalanceados:* sucede cuando no todos los grupos están igualmente representados. Si no se tiene en cuenta en la fase de entrenamiento, se corre el riesgo de que se favorezca una predicción de un grupo por el simple hecho de tener más casos.
- *No generalización:* tiene lugar cuando la selección de población para el entrenamiento no incluye casos de toda la población objetivo. En este caso, si el sistema entra en producción, fallará al generar predicciones para estos grupos.
- *Infrarrepresentación de un grupo:* como en el caso anterior, se excluye un grupo del conjunto de entrenamiento. No se considera un problema en la selección de la población, sino que es debido a una infrarrepresentación de este grupo por causas sociales o económicas que no se pueden solventar mediante una selección más amplia del grupo de entrenamiento.

En la fase de selección de variables se deberán incluir los factores de confusión y asegurarnos que no se incluyen variables que favorezcan la discriminación de un grupo. Existen varios ejemplos que ilustran la importancia de esta fase. En el transcurso de la pandemia de COVID-19 se observó un menor riesgo de hospitalización por neumonías relacionadas con el virus en individuos asmáticos. Este fenómeno puede deberse a la subestimación del riesgo en una subpoblación específica, a saber, asmáticos que desarrollan neumonía. Esta subestimación puede ser atribuible a la falta de consideración de factores relevantes, como el uso previo de esteroides[52]. Además, se han reportado tasas menores de ingreso por insuficiencia cardiaca en poblaciones de riesgo de exclusión, como las comunidades afrodescendientes y latinas, lo que resalta la necesidad de abordar las disparidades en la evaluación del riesgo[53,54]. Además de definir las variables, tendremos que seleccionar la función de error a optimizar, es decir, la métrica que evaluará nuestro modelo. Esta selección no es trivial y puede inducir sesgos en el momento de la aplicación[55]. Queda claro, por tanto, que es esencial entender el efecto de cada métrica de evaluación en el problema que se trata de resolver[56].

En la tabla 3 se detallan diversas métricas de evaluación, así como sus inconvenientes, los riesgos a nivel clínico, su transparencia y casos de uso. Finalmente, es conveniente definir qué significa que el modelo desarrollado sea justo, entendiendo que individuos con características parecidas sean tratados de forma similar[57,58]. En diversas publicaciones se proponen implementaciones que tienen en cuenta la paridad demográfica y la igualdad de oportunidades, sin embargo, su uso no está normalizado en el área clínica[59–61].

Cómo implementar una inteligencia artificial segura en tiempo real

Disponer de una IA generada de forma segura con datos de un entorno de desarrollo (ED) no implica necesariamente que vaya a funcionar de forma segura cuando se realice como herramienta de soporte a la decisión en la práctica clínica en tiempo real con datos de un entorno de implementación (EI). Es importante aclarar que estos dos entornos pueden ser diferentes debido a la dimensión espacial (dos UCI distintas), pero también pueden ser diferentes debido a la dimensión temporal (misma UCI distintos periodos de tiempo). A día de hoy no existe un protocolo estandarizado de los pasos a seguir para garantizar su éxito, pero sí existen guías consensuadas entre expertos que nos pueden ayudar durante el proceso[62]. Nosotros proponemos un mínimo de cuatro fases necesarias para trasladar una IA del ED al EI de forma segura como herramienta de soporte a la decisión (fig. 2).

Fase 1: testear la inteligencia artificial con datos retrospectivos del entorno de implementación

Las IA necesitan un conjunto de variables predictoras (VP) para devolver la variable respuesta (VR). Un primer paso evidente es asegurar que se pueden obtener de forma automática las VP a partir de la EHR del EI. Cuantas menos VP requiera la IA y menos específicas sean estas, más fácil será su implementación en distintos EI. Actualmente, modelos de IA que han demostrado un gran rendimiento en la literatura no son, en su mayoría, aplicables en la práctica clínica[63]. En caso de poder asegurar su obtención automática, se procederá a evaluar el rendimiento de la IA dentro del EI con datos retrospectivos. Si los resultados no son satisfactorios se deberá tomar la decisión de, o bien finalizar el proceso, o bien abrir un nuevo escenario de reentrenamiento de la IA para mejorar los resultados en el EI, es decir, pasar por un proceso de generalización[64]. Esto último dependerá mucho del marco en el que se esté llevando el proceso de integración, y deberán seguirse siempre todas las directrices éticas y legales que se requieran. Solo en caso de llegar a obtener un buen rendimiento de la IA con los datos retrospectivos del EI tiene sentido pasar a la Fase 2.

Fase 2: testear la inteligencia artificial con datos en tiempo real del entorno de implementación de forma ciega para el clínico

Convertir un proceso de extracción, transformación y carga (ETL) de las VP para ejecutar una IA de forma «*ad-hoc*» en un proceso o *pipeline* estable y escalable a prueba de fallos es una tarea costosa a nivel tecnológico. Sin entrar en detalles técnicos, una vez asegurado que todo funciona en tiempo real y habiendo diseñado un protocolo de actuación en caso de fallos en el sistema, se puede proceder a evaluar la IA de forma prospectiva. Este tipo de evaluación prospectiva es necesaria también en los casos en los que el ED y el EI sean la misma UCI, donde lo que habrá cambiado es la dimensión temporal. Una IA que haya demostrado buen rendimiento en su ED o en la Fase 1 del EI puede verse mermada por los cambios inherentes al tiempo (nuevos profesionales, nuevos hábitos, nuevos fármacos, pandemias, etc.). Es por tanto muy necesario asegurar un buen rendimiento sostenido en esta Fase 2, ya que será tanto indicador de que el EI dispone de la infraestructura tecnológica necesaria para mantener la herramienta de soporte a la decisión como de que la IA es lo suficientemente robusta para funcionar de forma estable en el tiempo.

Fase 3: ensayo clínico considerando el uso de la inteligencia artificial como intervención

En caso de haber superado la Fase 2 con éxito, sabemos que disponemos de una IA robusta y estable capaz de generar una predicción acertada en la mayoría de los casos. Sin embargo, no sabemos qué impacto podría haber tenido en el paciente su utilización por parte del equipo clínico. En esta fase se requiere diseñar un ensayo clínico capaz de evaluar si existen diferencias significativas entre un grupo control sin IA y un grupo de intervención con IA[62]. Es en este punto donde puede ser clave que la IA sea interpretable y no una caja negra[65]. Una IA interpretable puede dar información al clínico sobre las VP que están teniendo peso en la VR, ayudando al clínico a entender que debería modificar para evitar esa VR no deseada en el caso que decida hacerlo. En el caso contrario la tarea de búsqueda del motivo por el cual la IA ofrece una VR no deseada recaerá totalmente en el clínico.

Fase 4: monitorización continua y evolución de la inteligencia artificial

Las IA basadas en ML aprenden de un conjunto de casos en base a unas VP y definiendo las VR. La dimensión tem-

Tabla 3 Principales métricas empleadas en la evaluación de modelos en IA

Métrica	Descripción	Impacto en los resultados clínicos	Transparencia	Ejemplos
Exactitud	Porcentaje de datos correctamente clasificados.	Puede conducir a una mayor tasa de falsos negativos, lo que puede retrasar o impedir el tratamiento adecuado.	Fácil de entender e interpretar.	Predicción de la probabilidad de muerte de pacientes con COVID-19
Precisión	Porcentaje de datos positivos correctamente clasificados.	Puede conducir a una mayor tasa de falsos positivos, lo que puede causar ansiedad o estrés en los pacientes.	Puede ser difícil de interpretar si la prevalencia de la clase positiva es baja.	Predicción de la presencia de cáncer de mama en mamografías
F1-*score*	Media ponderada de la precisión y la exhaustividad.	Puede ser una buena opción para tareas de clasificación equilibrada, pero es importante considerar el impacto potencial en los resultados clínicos.	Puede ser difícil de interpretar si la prevalencia de la clase positiva es baja.	Predicción de la probabilidad de ictus en pacientes con hipertensión
Especificidad	Porcentaje de datos negativos correctamente clasificados.	Puede conducir a una mayor tasa de falsos negativos, lo que puede retrasar o impedir el tratamiento adecuado.	Puede ser difícil de interpretar si la prevalencia de la clase positiva es baja.	Biomarcadores (usada con la precisión)
Curva ROC	Representa la relación entre la TPR y FPR.	Puede ser una buena opción para comparar el rendimiento de diferentes modelos, pero otorga la misma importancia a la precisión y a la especificidad.	Puede ser difícil de interpretar si la prevalencia de la clase positiva es baja.	Predicción de la probabilidad de supervivencia de pacientes con cáncer
AUC ROC	Valor de la curva ROC en el punto (0, 0) y (1, 1).	Puede ser una buena opción para comparar el rendimiento de diferentes modelos, pero otorga la misma importancia a la precisión y a la especificidad.	Puede ser difícil de interpretar si la prevalencia de la clase positiva es baja.	Predicción de mortalidad en pacientes de UCI
Precisión/*Recall* AUC	Representa la relación entre la precisión y la exhaustividad (*recall*).	Puede ser difícil de interpretar si la prevalencia de la clase positiva es baja.	Puede ser una buena opción para tareas de clasificación desequilibrada.	Predicción de hospitalizaciones agudas en personas mayores que reciben atención domiciliaria.
Pérdida logarítmica	Suma de los logaritmos de las probabilidades de las predicciones correctas.	Puede ser una buena opción para comparar el rendimiento de diferentes modelos, pero es sensible a datos desbalanceados y carece de explicabilidad. Además, no considera la gravedad del error.	Es una métrica difícil de interpretar, ya que requiere un conocimiento de los logaritmos. Sin embargo, es una métrica objetiva que puede ser utilizada para comparar el rendimiento de diferentes modelos.	Predicción de readmisión un año después del alta
Índice Jaccard	Relación entre el número de elementos correctamente clasificados frente a la suma del número de elementos correctamente clasificados y el número de elementos mal clasificados.	Puede ser una buena opción para comparar el rendimiento de diferentes modelos, pero otorga la misma importancia a la precisión y a la especificidad. Además, a nivel individual (p. ej., pixel en el caso de imágenes), carece de gradación ya que es una métrica binaria.	Es una métrica fácil de entender e interpretar. Sin embargo, es menos sensible a los falsos negativos que otras métricas, como la exactitud o el F1-*score*.	Predicción de la presencia de daño cerebral en imágenes de resonancia magnética

AUC: área bajo la curva; FPR: tasa de falsos positivos; ROC: Receiver Operating Characteristic; TPR: tasa de verdaderos positivos; UCI: Unidad de Cuidados Intensivos.

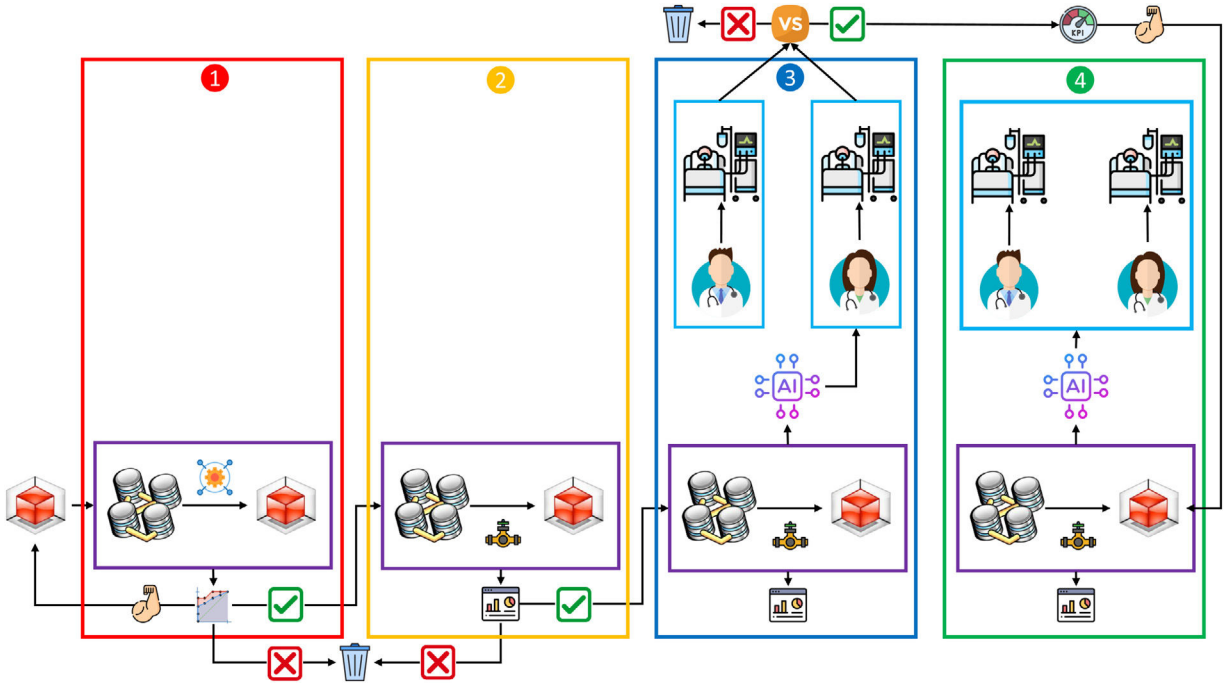

Figura 2 Mapa de las cuatro fases de implementación segura de IA en tiempo real. En la fase 1 se testea que la IA se adapte a la realidad de los datos del centro donde se implanta. En la fase 2 se construye el flujo de datos que permita evaluar su rendimiento en tiempo real. En la fase 3 se comparan los resultados de utilizar o no la IA con relación al beneficio del paciente. Finalmente, en la fase 4 se monitoriza continuamente el rendimiento de la IA y se aplican las mejoras necesarias para que esta evolucione en beneficio de todos.

poral convierte inevitablemente cualquier EI en ED con el paso del tiempo. Nuevos contextos socioeconómicos, nuevos equipos, nuevos fármacos incluso nuevos hábitos adquiridos de las futuras sinergias IA-humano harán que las IA queden obsoletas si no evolucionan de forma dinámica. Por ejemplo, una IA entrenada para predecir un determinado evento adverso en un entorno donde los protocolos no usaban esa misma IA puede dejar de funcionar en el momento en que se aplique la propia IA para evitarlo, ya que se habrán generado nuevos protocolos que incluyan la IA modificando completamente el contexto en el que se entrenó. En esta última fase se deberá definir un conjunto de indicadores de acciones humanas motivadas por la IA cuya monitorización permita asegurar que la convivencia de ambas inteligencias (humana-artificial) es beneficiosa para el paciente. Este conjunto de indicadores dependerá del tipo de IA y su objetivo. Finalmente, de forma periódica y mediante los constantes *inputs* clínicos se deberán reentrenar las IA con nuevas VP a fin de irse adaptando a nuevos EI y mejorar su rendimiento[66].

Conclusiones

La integración de la IA en el ámbito de la seguridad, aunque prometedora, enfrenta desafíos clave. La predicción de eventos adversos y ayuda a la prescripción segura representan oportunidades significativas. Sin embargo, la falta de calidad metodológica en las investigaciones y la necesidad de abordar preocupaciones éticas, como la confianza y el sesgo, son imperativos. La implementación exitosa requiere no solo robustez técnica sino también una transición cuidadosa, asegurando la comprensión y aceptación de los profesionales de la salud. La seguridad continua y la adaptabilidad emergen como cimientos cruciales para una colaboración efectiva entre la IA y la atención médica, asegurando beneficios tangibles para la seguridad del paciente.

Contribución de los autores

Jesús Abelardo Barea Mendoza: conceptualización, redacción, edición y revisión de manuscrito final. Josep Gómez Álvarez: conceptualización, redacción, edición y revisión de manuscrito final. Alex Pardo Fernandez: redacción y revisión de manuscrito final. Marcos Valiente Fernandez: redacción y revisión de manuscrito final.

Durante la preparación de este trabajo, los autores utilizaron Chat-GPT 3.5 para solicitar sinónimos y mejorar la traducción de expresiones técnicas del inglés al español. Después de usar este servicio, los autores revisaron y editaron el contenido según fuera necesario, asumiendo la plena responsabilidad del contenido de la publicación.

Financiación

No se ha recibido financiación para la realización del presente manuscrito.

Conflicto de intereses

Jesús Abelardo Barea Mendoza ha trabajado para la empresa de inteligencia artificial Savana Médica. El resto de los auto-

res no presenta conflicto de intereses relacionado con el presente estudio.

Bibliografía

1. Mintz Y, Brodie R. Introduction to artificial intelligence in medicine. Minim Invasive Ther Allied Technol. 2019;28:73–81.
2. Kaul V, Enslin S, Gross SA. History of artificial intelligence in medicine. Gastrointest Endosc. 2020;92:807–12.
3. Keskinbora KH. Medical ethics considerations on artificial intelligence. J Clin Neurosci Off J Neurosurg Soc Australas. 2019;64:277–82.
4. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface. 2018;15:20170387.
5. Rueda J, Rodríguez JD, Jounou IP, Hortal-Carmona J, Ausín T, Rodríguez-Arias D. «Just» accuracy? Procedural fairness demands explainability in AI-based medical resource allocations. AI Soc. 2022:1–12. Online ahead of print.
6. London AJ. Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. Hastings Cent Rep. 2019;49:15–21.
7. Finocchiaro G. The regulation of artificial intelligence. AI Soc. 2023.
8. Li F, Xin H, Zhang J, Fu M, Zhou J, Lian Z. Prediction model of in-hospital mortality in intensive care unit patients with heart failure: machine learning-based, retrospective analysis of the MIMIC-III database. BMJ Open. 2021;11:e044779.
9. Johnson AEW, Mark RG. Real-time mortality prediction in the Intensive Care Unit. AMIA Annu Symp Proc. 2018;2017:994–1003.
10. Awad A, Bader-El-Den M, McNicholas J, Briggs J, El-Sonbaty Y. Predicting hospital mortality for intensive care unit patients: Time-series analysis. Health Informatics J. 2020;26:1043–59.
11. Verburg IWM, Atashi A, Eslami S, Holman R, Abu-Hanna A, de Jonge E, et al. Which Models Can I Use to Predict Adult ICU Length of Stay? A Systematic Review. Crit Care Med. 2017;45:e222–31.
12. Peres IT, Hamacher S, Cyrino Oliveira FL, Bozza FA, Salluh JIF. Data-driven methodology to predict the ICU length of stay: A multicentre study of 99,492 admissions in 109 Brazilian units. Anaesth Crit Care Pain Med. 2022;41:101142.
13. Fabregat A, Magret M, Ferré JA, Vernet A, Guasch N, Rodríguez A, et al. A Machine Learning decision-making tool for extubation in Intensive Care Unit patients. Comput Methods Programs Biomed. 2021;200:105869.
14. Kim J, Chae M, Chang H-J, Kim Y-A, Park E. Predicting Cardiac Arrest and Respiratory Failure Using Feasible Artificial Intelligence with Simple Trajectories of Patient Data. J Clin Med. 2019;8:1336.
15. Ma X, Si Y, Wang Z, Wang Y. Length of stay prediction for ICU patients using individualized single classification algorithm. Comput Methods Programs Biomed. 2020;186:105224.
16. Alfieri F, Ancona A, Tripepi G, Rubeis A, Arjoldi N, Finazzi S, et al. PloS One. 2023;18:e0287398.
17. Morris AH. Human Cognitive Limitations. Broad, Consistent, Clinical Application of Physiological Principles Will Require Decision Support. Ann Am Thorac Soc. 2018;15:S53–6.
18. Ocampo-Quintero N, Vidal-Cortés P, Del Río Carbajo L, Fdez-Riverola F, Reboiro-Jato M, Glez-Peña D. Enhancing sepsis management through machine learning techniques: A review. Med Intensiva. 2022;46:140–56.
19. van de Sande D, van Genderen ME, Huiskens J, Gommers D, van Bommel J. Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit. Intensive Care Med. 2021;47:750–60.
20. Moazemi S, Vahdati S, Li J, Kalkhoff S, Castano LJV, Dewitz B, et al. Artificial intelligence for clinical decision support for monitoring patients in cardiovascular ICUs: A systematic review. Front Med. 2023;10:1109411.
21. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med. 2020;3:17.
22. El-Kareh R, Sittig DF. Enhancing Diagnosis Through Technology: Decision Support. Artificial Intelligence, and Beyond. Crit Care Clin. 2022;38:129–39.
23. Hak F, Guimarães T, Santos M. Towards effective clinical decision support systems: A systematic review. PloS One. 2022;17:e0272846.
24. Hong N, Liu C, Gao J, Han L, Chang F, Gong M, et al. State of the Art of Machine Learning-Enabled Clinical Decision Support in Intensive Care Units: Literature Review. JMIR Med Inform. 2022;10:e28781.
25. Mittermaier M, Raza M, Kvedar JC. Collaborative strategies for deploying AI-based physician decision support systems: challenges and deployment approaches. NPJ Digit Med. 2023;6:137.
26. Kindle RD, Badawi O, Celi LA, Sturland S. Intensive Care Unit Telemedicine in the Era of Big Data Artificial Intelligence, and Computer Clinical Decision Support Systems. Crit Care Clin. 2019;35:483–95.
27. Pinsky MR, Dubrawski A, Clermont G. Intelligent Clinical Decision Support. Sensors. 2022;22:1408.
28. Hendriks M, Willemsen MC, Sartor F, Hoonhout J. Respecting Human Autonomy in Critical Care Clinical Decision Support. Front Comput Sci. 2021;3:1–10.
29. van der Meijden SL, de Hond AAH, Thoral PJ, Steyerberg EW, Kant IMJ, Cinà G, et al. Intensive Care Unit Physicians' Perspectives on Artificial Intelligence-Based Clinical Decision Support Tools: Preimplementation Survey Study. JMIR Hum Factors. 2023;10:e39114.
30. Bates DW, Levine D, Syrowatka A, Kuznetsova M, Craig KJT, Rui A, et al. The potential of artificial intelligence to improve patient safety: a scoping review. NPJ Digit Med. 2021;4:54.
31. Chen D, Wang R, Jiang Y, Xing Z, Sheng Q, Liu X, et al. Application of artificial neural network in daily prediction of bleeding in ICU patients treated with anti-thrombotic therapy. BMC Med Inform Decis Mak. 2023;23:171.
32. Zhu Y, Venugopalan J, Zhang Z, Chanani NK, Maher KO, Wang MD. Domain Adaptation Using Convolutional Autoencoder and Gradient Boosting for Adverse Events Prediction in the Intensive Care Unit. Front Artif Intell. 2022;5:640926.
33. Hegselmann S, Ertmer C, Volkert T, Gottschalk A, Dugas M, Varghese J. Development and validation of an interpretable 3 day intensive care unit readmission prediction model using explainable boosting machines. Front Med. 2022;9:960296.
34. Hosein FS, Bobrovitz N, Berthelot S, Zygun D, Ghali WA, Stelfox HT. A systematic review of tools for predicting severe adverse events following patient discharge from intensive care units. Crit Care Lond Engl. 2013;17:R102.
35. Wang L, Duan S-B, Yan P, Luo X-Q, Zhang N-Y. Utilization of interpretable machine learning model to forecast the risk of major adverse kidney events in elderly patients in critical care. Ren Fail. 2023;45:2215329.
36. McKown AC, Wang L, Wanderer JP, Ehrenfeld J, Rice TW, Bernard GR, et al. Predicting Major Adverse Kidney Events among Critically Ill Adults Using the Electronic Health Record. J Med Syst. 2017;41:156.
37. Hur S, Min JY, Yoo J, Kim K, Chung CR, Dykes PC, et al. Development and Validation of Unplanned Extubation Prediction Models Using Intensive Care Unit Data: Retrospective, Comparative, Machine Learning Study. J Med Internet Res. 2021;23:e23508.

38. Veldhuis LI, Woittiez NJC, Nanayakkara PWB, Ludikhuize J. Artificial Intelligence for the Prediction of In-Hospital Clinical Deterioration: A Systematic Review. Crit Care Explor. 2022;4:e0744.
39. Cummings BC, Ansari S, Motyka JR, Wang G, Medlin RP, Kronick SL, et al. Predicting Intensive Care Transfers and Other Unforeseen Events: Analytic Model Validation Study and Comparison to Existing Methods. JMIR Med Inform. 2021;9:e25066.
40. Eldridge N, Wang Y, Metersky M, Eckenrode S, Mathew J, Sonnenfeld N, et al. Trends in Adverse Event Rates in Hospitalized Patients, 2010-2019. JAMA. 2022;328:173–83.
41. Bates DW, Cullen DJ, Laird N, Petersen LA, Small SD, Servi D, et al. Incidence of adverse drug events and potential adverse drug events. Implications for prevention. ADE Prevention Study Group. JAMA. 1995;274:29–34.
42. Leviatan I, Oberman B, Zimlichman E, Stein GY. Associations of physicians' prescribing experience, work hours, and workload with prescription errors. J Am Med Inform Assoc JAMIA. 2021;28:1074–80.
43. Salas M, Petracek J, Yalamanchili P, Aimer O, Kasthuril D, Dhingra S, et al. The Use of Artificial Intelligence in Pharmacovigilance: A Systematic Review of the Literature. Pharm Med. 2022;36:295–306.
44. Syrowatka A, Song W, Amato MG, Foer D, Edrees H, Co Z, et al. Key use cases for artificial intelligence to reduce the frequency of adverse drug events: a scoping review. Lancet Digit Health. 2022;4:e137–48.
45. Sikora A, Rafiei A, Rad MG, Keats K, Smith SE, Devlin JW, et al. Pharmacophenotype identification of intensive care unit medications using unsupervised cluster analysis of the ICURx common data model. Crit Care Lond Engl. 2023;27:167.
46. Poweleit EA, Vinks AA, Mizuno T. Artificial Intelligence and Machine Learning Approaches to Facilitate Therapeutic Drug Management and Model-Informed Precision Dosing. Ther Drug Monit. 2023;45:143–50.
47. Tan BKJ, Teo CB, Tadeo X, Peng S, Soh HPL, Du SDX, et al. Personalised, Rational Efficacy-Driven Cancer Drug Dosing via an Artificial Intelligence SystEm (PRECISE): A Protocol for the PRECISE CURATE.AI Pilot Clinical Trial. Front Digit Health. 2021;3:635524.
48. Velo GP, Minuz P. Medication errors: prescribing faults and prescription errors. Br J Clin Pharmacol. 2009;67:624–8.
49. Schiff GD, Volk LA, Volodarskaya M, Williams DH, Walsh L, Myers SG, et al. Screening for medication errors using an outlier detection system. J Am Med Inform Assoc. 2017;24:281–7.
50. Segal G, Segev A, Brom A, Lifshitz Y, Wasserstrum Y, Zimlichman E. Reducing drug prescription errors and adverse drug events by application of a probabilistic, machine-learning based clinical decision support system in an inpatient setting. J Am Med Inform Assoc. 2019;26:1560–5.
51. Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical Machine Learning in Healthcare. Annu Rev Biomed Data Sci. 2021;4:123–44.
52. Otunla A, Rees K, Dennison P, Hobbs R, Suklan J, Schofield E, et al. Risks of infection, hospital and ICU admission, and death from COVID-19 in people with asthma: systematic review and meta-analyses. BMJ Evid-Based Med. 2022;27:263–73.
53. Vyas DA, Eisenstein LG, Jones DS. Hidden in Plain Sight - Reconsidering the Use of Race Correction in Clinical Algorithms. N Engl J Med. 2020;383:874–82.
54. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible Models for HealthCare. 2015:1721–30.
55. Halligan S, Altman DG, Mallett S. Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: a discussion and proposal for an alternative approach. Eur Radiol. 2015;25:932–9.
56. Erickson BJ, Kitamura F. Magician's Corner: 9 Performance Metrics for Machine Learning Models. Radiol Artif Intell. 2021;3:e200126.
57. Parbhoo S, Gichoya JW, Celi LA, de la Hoz MÁA. Operationalising fairness in medical algorithms. BMJ Health Care Inform. 2022;29:e100617.
58. Fletcher RR, Nakeshimana A, Olubeko O. Addressing Fairness Bias, and Appropriate Use of Artificial Intelligence and Machine Learning in Global Health. Front Artif Intell. 2020;3:561802.
59. Lohaus M, Perrot M, Luxburg UV. Too Relaxed to Be Fair. PMLR. 2020;119:6360–9.
60. Calders T, Karim A, Kamiran F, Ali W, Zhang X. Controlling Attribute Effect in Linear Regression. IEEE. 2013:71–80.
61. Zafar MB, Valera I, Rodriguez MG, Gummadi KP. Fairness Constraints: Mechanisms for Fair Classification. arXiv. 2017.
62. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med. 2022;28:924–33.
63. Panch T, Mattie H, Celi LA. The «inconvenient truth» about AI in healthcare. NPJ Digit Med. 2019;2:77.
64. Sauer CM, Gómez J, Botella MR, Ziehr DR, Oldham WM, Gavidia G, et al. Understanding critically ill sepsis patients with normal serum lactate levels: results from U.S. and European ICU cohorts. Sci Rep. 2021;11:20076.
65. 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.
66. Feng J, Phillips RV, Malenica I, Bishara A, Hubbard AE, Celi LA, et al. Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare. NPJ Digit Med. 2022;5:66.