Medicina Intensiva 49 (2025) 154–164

medicina intensiva

http://www.medintensiva.org/en/

UPDATE IN INTENSIVE CARE MEDICINE: CRITICAL PATIENT SAFETY

Current perspectives on the use of artificial intelligence in critical patient safety

Jesús Abelardo Barea Mendoza[a,*], **Marcos Valiente Fernandez**[a], **Alex Pardo Fernandez**[b], **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, Spain*
[b] *Universitat Rovira i Virgili, Tarragona, Spain*
[c] *Hospital Universitari de Tarragona Joan XXIII. Universitat Rovira i Virgili. Institut d'Investigació Sanitària Pere i Virgili, Tarragona, Spain*

Received 19 December 2023; accepted 11 March 2024
Available online 26 April 2024

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

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.

DOI of original article: https://doi.org/10.1016/j.medin.2024.03.007
* Corresponding author.
E-mail address: jesusabelardo.barea@salud.madrid.org (J.A. Barea Mendoza).

2173-5727/

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

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

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 presenta 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 es crucial.

Artificial intelligence and safety in the health care setting

Artificial Intelligence (AI) is neither a new nor recent concept. The first time this term was used was back in the 1950s.[1] However, it has been in recent years that its large-scale implementation has become possible due to exponential technological growth.[2] AI is perceived as a powerful tool with enormous potential to change the way we live; however, AI is no stranger to controversy.[3] Among all types of AI, those capable of reaching higher levels of predictive or even creative skills are based on deep learning models.[4] These models are nonlinear and rely on a large number of mathematical operations, which complicate their interpretation.[5,6] It is precisely this type of AI that generates more concern and distrust in society and is being analyzed from ethical and legal perspectives.[7]

While the introduction of AI requires a risk-benefit analysis in any scenario, in the health care field, this analysis must be deep and comprehensive. It has been proposed as a solution to many of the problems of current health care systems, contributing to decreasing mortality, shortening the length of stay, or preventing the occurrence of adverse events.[8–16]

Despite recent efforts to increase the culture of safety in health care systems, adverse events still significantly contribute to morbidity, mortality, and health care spending. The addition of new technologies emerges as a promising strategy in this regard. The critical care setting is an ideal scenario for the implementation of such technologies. Extremely ill patients undergoing numerous procedures and highly complex treatments on a daily basis, which happen to be prone to errors coexist in this context. Additionally, there is pressure for quick decision-making as it is a typical area for time-dependent conditions. The volume of information generated in an ICU can be overwhelming. In fact, it has been suggested that it exceeds the capacity of an expert clinician.[17] Thus, the increase in available information in different formats (images, lab test results, genetics, invasive physiological monitoring, etc.) is not always associated with better decisions. Therefore, the addition of AI-derived technologies in critical care areas could help clinicians increase diagnostic or therapeutic capabilities and contribute to improving outcomes by facilitating better integration of information.[18] There is growing controversy due to the proliferation of AI publications with poor methodological quality and questionable validity, which limits its implementation. A recent systematic review of more than 400 studies reporting models developed in the critical care setting found that more than 20% were associated with preventing the occurrence of adverse events. However, the authors highlight that the methodological quality of the articles was very poor, most of them being retrospective (96.4%) and highly biased.[19]

For all these reasons, the aim of this manuscript is to review the potential contribution of AI to patient safety in the critical care setting by summarizing technical aspects and providing examples. Secondly, aspects of safety related to the implementation processes of AI-derived technologies, which will undoubtedly be part of the future of our units, have also been reviewed.

Main applications of AI in patient safety

Decision support - clinical decision support systems (CDSS)

Technological innovations allow for the implementation of CDSS aimed at assisting physicians in decision-making by quickly identifying patterns of potential problems (not easily perceptible by humans) and suggesting optimal treatment plans. These tools can analyze and synthesize large sets of clinical data quickly (health records, vital signs, or imaging).[20] CDSS are not new. Their development dates back to the first advances made in computing in the 1970s.[21] As noted, ICUs are a niche for CDSS due to their inherent

characteristics (high availability of data, monitoring, clinical complexity), as well as new AI and Machine Learning (ML)-based technological implementations.

Why can CDSS be useful, and what role can they play?
The implementation of CDSS seeks to improve quality in all dimensions, especially safety. Although their utility has been demonstrated in certain medical disciplines, their optimal role is still under discussion.[22] They are useful for various reasons: 1- Addition of individualized medicine through the use of ML-based models, which have proven equal or even better than experienced professionals in various scenarios: mortality prediction, readmissions, renal failure, sepsis, and respiratory distress, among others. 2- Generation of therapeutic plans on demand. 3- Reduction of information overload, allowing teams to make better decisions based on the vast amount of available data (Table 1).[21]

What types of models do CDSS use?
Until a few years ago, CDSS were based on prior knowledge using rules of the type ''if A - then B,'' which simplified medical practice. AI-based CDSS allow for the storage and processing of very variable patient-specific data sources, finally proposing recommendations with which we can provide feedback to the system. This mitigates the simplification of medical practice by previous CDSS.[23] It is crucial for the ICU to facilitate access to a constant flow of data, allowing for specific analyses (time series) that can evaluate trends and potentially anticipate problems.[24]

Is it easy to implement CDSS in the routine clinical practice?
The clinical implementation of CDSS is still limited, with some key elements we should take into consideration[25]:

- Trust: both patients and clinicians must trust the AI models used. There is also concern about clinicians' overconfidence as these systems require human supervision throughout the entire process.
- Bias: datasets used to train AI models may contain biases depending on their origin, epidemiological context, or data treatment.
- Scalability: the implementation of CDSS must be easily scalable and adaptable to a variety of clinical environments. Implementation requires a gradual process along with feedback between developers and health professionals. Improvement in data quality, numerous iterations and adjustments, and workflow optimization are also necessary.
- Deployment: CDSS face regulatory challenges because of the access to highly sensitive personal data involved, and the low reproducibility of results. However, in the case of CDSS, interest may lie in exploiting local characteristics, with data reevaluations and periodic refinements.[26,27]
- Ethical considerations: the implementation of these systems may pose a cultural and ethical challenge, which may impact how we see the autonomy of physicians and patients based on the suggestions made by these CDSS.[28]
- Clinicians' perspective: evidence shows that clinicians are favorable to the integration of CDSS, especially for certain clinical questions such as the probability of readmissions (Fig. 1). They also seek to understand those factors that contribute to predictions. Therefore, the development of algorithmic explainability (XAI, eXplainable Artificial Intelligence) is of special interest.[29]

Prediction of adverse events at the ICU setting

Recently, the use of AI has been proposed in the main domains of adverse events (AE), highlighting prediction, prevention, and early detection of patients at risk of deterioration. In this way, AI - based on the automation of records and the use of ML - offers new strategies to mitigate the occurrence of AE.[30] Currently, there are algorithms that use real-time patient data accumulating information to personalize treatment during their admission. These tools, for example, would allow initiating or discontinuing antithrombotic therapy based on the risk of bleeding at a specific time during admission.[31] Other algorithms allow summarizing all the information about a patient regarding an event of interest. By condensing it, they facilitate comparison with other patients, as well as the creation of patient cohorts with comparable risk profiles for a specific event of interest (mortality).[32] There are multiple clinical settings in which results have been published in this regard, highlighting: 1- prediction and stratification of the risk of readmission, helping in patient flow management and avoiding AE associated with unplanned readmissions.[33,34] 2- prediction of kidney failure-related AE.[35,36] 3- prediction of unplanned extubation allowing the implementation of preventive measures and reducing the workload for the health care personnel.[37] Although its implementation may focus on mitigating adverse events in ICU patients, it should not be limited to the ICU setting. This technology can be used anywhere in the health care system to predict early clinical deterioration or immediate transfers to the ICU that would allow early treatment initiation and resource organization.[38,39] The effectiveness of an algorithm may be limited outside its original environment, as the data it uses reflect the culture and specific practices of each ICU. Consequently, what works in one ICU setting may not be applicable to another, especially if working conditions and staff ratios vary, thus affecting outcomes. The application areas of AI regarding safety are varied, whether inside or outside the ICU sttting (wall-less ICU strategy), and more and more innovative algorithms are becoming available that allow capturing and better adapting the information obtained from patient data to make more precise predictions.

Prescription and drug-related adverse events

Drug-related incidents remain among the most frequent adverse events.[40] Up to 25% of adverse events are considered preventable.[41] Furthermore, intensive care units (ICUs), due to their technical complexity and association with time-dependent diseases, are more susceptible to prescription errors.[42] The applications of AI in this area are diverse, including risk prediction models for the development of adverse reactions, detection of polypharmacy-related events, development of in silico interaction and allergy models, application of Clinical Decision Support Systems (CDSS), and exploitation of electronic health records for the

Table 1 Possible functionalities of CDSS, potential risks, and risk mitigation strategies.

Functions and utilities of CDSS	Potential harms of CDSS	Risk mitigation strategies	Explanation
Patient safety	Alert fatigue	Prioritize critical alerts, minimize use of disruptive alerts for non-critical indications.	Alert fatigue could be minimized by prioritizing and selecting critically important alerts, having the greatest impact, and customizing alerts according to clinical scenarios.
Minimize the incidence of errors and adverse events.	Occurs when too many insignificant alerts are presented. There's a risk of disregarding alarms regardless of their importance.		
Clinical management	Negative impact on user skills	Avoid systematic prescriptiveness in system design. Continuously evaluate system impact.	Systems should be implemented to be useful to clinicians without compromising autonomy or being overly 'prescriptive' and definitive.
Favor compliance with clinical practice guidelines, reminders for follow-up and treatment, etc.	An example is the dependency or excessive trust in the accuracy of a system.		
	Conflict with physician autonomy		
Administrative function	Challenges in maintaining the system and its content	2 strategies can be implemented here:	(1) Facilitate scheduled review, methods for acquiring and implementing new knowledge. Implement physician feedback measures on the system, and train users on proper data input.
Selection of diagnostic codes, automated documentation, and note auto-completion.	As practices change, there may be difficulties in keeping the content and knowledge rules that drive the CDSS up to date.	(1) Manage established knowledge, with a focus on translation to CDSS.	(2) It is important to identify changes in performance and usage over time.
		(2) System for evolutionary performance measurement and analysis of CDSS.	
Diagnostic support	User distrust towards CDSS	Include scientific references in messages when appropriate.	Provide a verifiable source of information to the user about why the recommendation exists.
Suggest diagnoses based on patient data and automate test result output.	Disagreement with the guidance provided by the CDSS.		In addition to increasing confidence, this can provide guidance to users to update their knowledge.
Decision support for patients	Dependence on user computer literacy	(1) Adapt to existing functionality.	(1) Maintain consistency with the existing system user interface (if any) is crucial to ensure users do not have a long learning curve to use the system.

Table 1 (*Continued*)

Functions and utilities of CDSS	Potential harms of CDSS	Risk mitigation strategies	Explanation
Assist patients in decision-making through personal health records and other systems.	CDSS may require a high level of technological competence for their use.	(2) Provide adequate training available at launch.	(2) Adequate and easily accessible training should be available for users.

Adapted from Sutton R.[21] CDSS: Clinical Decision Support System.

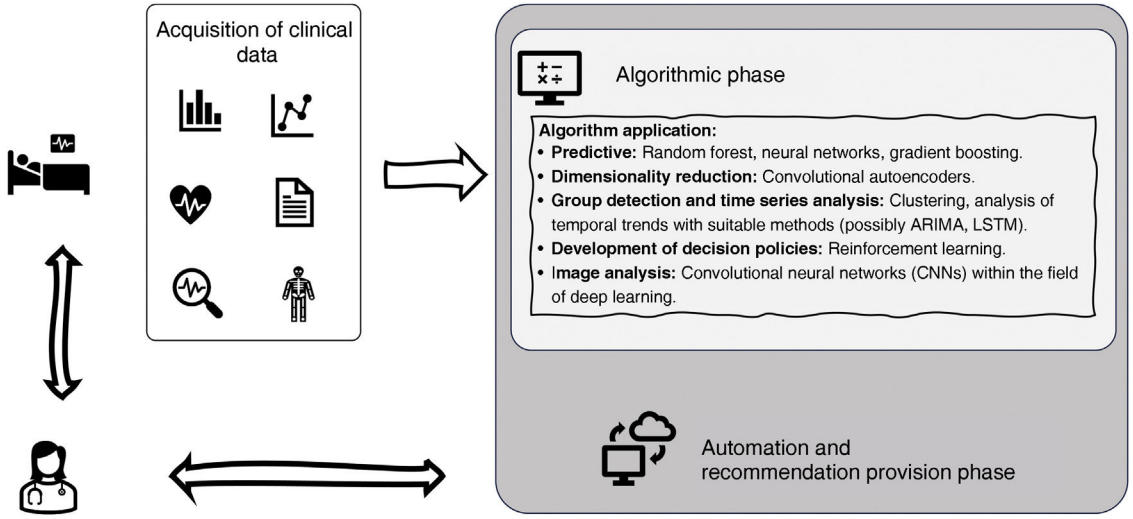

Figure 1 Represents the transmission of information and generation of recommendations. The patient generates data that are clinically utilized by the physician, but which can also be leveraged for digital assessment. The increasing amount of data, especially in complex patient cases and environments with temporal pressure, can lead to biased decision-making. Therefore, processing the multiple sources of biomedical information from the patient (medical history, vital signs, radiological tests, laboratory results, or drus) can be transferred to systems that automatically store and process this information, upon which algorithms are applied. After applying algorithms, we obtain a recommendation, which can then be fed back into the system.

detection of unintended adverse reactions.[43] Depending on when these models are applied, they can help predict risk, reducing their incidence (prevention strategies), or subsequently contribute to early detection, reducing severity and duration (damage mitigation strategies).[44]

Regarding predictive use cases, various strategies and examples are found in the literature. The most common is the prediction of patients at high risk of adverse drug reactions. The most studied event groups in this regard are renal, cardiovascular, and opioid-related overdoses.[44] A second group consists of predicting therapeutic response, which could prevent the use of drugs in non-responsive patients, which is particularly relevant in pharmaceutical groups with a high incidence of adverse events and poor tolerance, such as antineoplastics or some antivirals.[45] Predicting the optimal dose is also an area of interest, with suitable drugs for these purposes being anticoagulants or antineoplastics.[31,46,47] We should mention that these models integrate information from the patient's entire medical history, including past health records, current free-text history, or lab test results. Although promising, we should mention that the addition of genetic results into the models has only marginally improved predictive capability.[45] From a technical standpoint, the most widely used models are decision trees, natural language processing techniques, and neural networks, although both the variety and variability of these techniques used are extensive.

In the early incident detection block, we distinguish between detecting reactions and detecting medication errors that have already occurred (inappropriate prescription, interactions, and duplications). Undoubtedly, one of the most interesting areas is related to prescription errors. From a safety standpoint, although these occur after an individual action (prescriber, administrator, or consumer), they are considered system failures. Some authors even argue that they should be considered failures of clinical information systems.[48] In recent years, some CDSS with ML algorithms have been tested to aid in real-time prescription error detection.[49] These models add electronic health record information and subsequently detect prescriptions considered atypical by the model. These atypicalities can be due to infrequent prescriptions (contraceptive for an infant), discordance with the medical history (antidiabetic for a patient without a history of diabetes), or uncommon dosages. These systems can intervene in 2 different way, by generating synchronous (at the time of prescription) or asyn-

Table 2 Types of most common CDSS-generated alerts based on models.

Alerttype	Definition	Examples
Time-dependent (synchronous)	Existing data in the patient's profile makes the prescribed medication inappropriate or dangerous.	Antihypertensive in a patient in septic shock.
Clinical atypicality	Prescription does not fit the patient's clinical profile.	Hypoglycemic drug to a patient without a diagnosis of diabetes mellitus or indicative results of such disease.
Dose atypicality	The dose of a certain drug is considered an outlier value with respect to the dose distribution learned by the model for that drug in the population and/or the patient's past medical history.	Rare doses, unusual dosing units, uncommon frequency, uncommon route.
Overlapping prescription	An alert signaled when there is simultaneous treatment with 2 drugs from the same group.	Duplicate prescription of noradrenaline infusions with different formulations.
Time-dependent (asynchronous)	An alert signaled when changes occur in the patient's profile after the prescription, making the prescription inappropriate or dangerous to continue.	When blood pressure decreases, and continuation of antihypertensive drugs is inappropriate.

CDSS, Clinical Decision Support System.

chronous alerts (during follow-up when the patient's clinical situation changes). The main types of alerts and examples are described in Table 2.

One of the most developed systems in this regard is the MedAware software, which has been prospectively validated. In a validation on over 78 000 prescriptions, the alert rate was low (0.40%). Of these, 40% were synchronous alerts, with time-dependent alerts being the most frequent (64.80%). Of the generated alerts, 89% were considered appropriate, and 43% resulted in prescription changes. These data were superior to rule-based CDSS, which presented a high alert burden (37.10%) and low clinical significance (5.30% prescription changes).[50]

Safety in the implementation processes of AI-based tools

As previously mentioned, AI can be useful in various aspects at the ICU setting. However, it also poses challenges, both medical and ethical, as well as technological. Therefore, it is relevant not only to analyze the utility of AI regarding patient safety, but also establish theoretical frameworks for the safe use and implementation processes of AI. After reviewing the possible contributions of this technology to safety, we will now delve into the risks it presents, the possible solutions regarding adverse events it may generate, and the safe implementation of these algorithms.

Generating safe ML-based AI

Prediction algorithms employing supervised machine learning rely on learning from examples. Through them, a system is modeled capable of associating new events with learned data and generate a prediction. It is evident that the data used in learning will be a key point for the model's success. Therefore, data collection is essential, making sure that data accurately represent the target population. The following are among the most cited issues on this regard[51]:

- Imbalanced populations/cases: this occurs when not all groups are equally represented. If not considered during training, there's a risk of favoring predictions towards one group simply because it includes more cases.
- Non-generalization: this occurs when the population selection for training does not include cases from the entire target population. In this case, if the system goes into production, it will fail to generate predictions for these groups.
- Underrepresentation of a group: as in the previous case, a group is excluded from the training set. This is not a problem in population selection as much as an underrepresentation of this group for social or economic reasons that cannot be solved by broader selection of the training group.

During variable selection, confounding factors should be included, while making sure that variables favoring discrimination of a group have not been included. There are several examples illustrating the importance of this phase. During the COVID-19 pandemic, a lower risk of admissions due to virus-related pneumonias was observed in asthmatic individuals. This phenomenon may be due to risk underestimation in a specific subpopulation, namely, asthmatics developing pneumonia. This underestimation can be attributed to the lack of consideration of relevant factors, such as prior steroid use.[52] Additionally, lower rates of heart failure-related admissions have been reported in exclusion risk populations, such as African American and Latino communities, which stresses the need for addressing disparities in risk assessment.[53,54] In addition to defining the variables, we will have to select the error function that we want to optimize, i.e., the metric that will eventually evaluate our model. This selection is not trivial and can induce applica-

tion biases.[55] Therefore, it seems evident that it is essential to understand the effect of each evaluation metric on the problem being addressed.[56]

Various evaluation metrics, their drawbacks, clinical risks, transparency, and use cases are detailed in Table 3. Finally, it is advisable to define what it means for the developed model to be fair, understanding that individuals with similar characteristics are treated similarly.[57,58] Several publications propose implementations considering the demographic parity and equality of opportunities; however, their use has not been standardized in the clinical field.[59–61]

How to implement secure AI in real-time

Having an AI generated securely with data from a development environment (DE) does not necessarily imply that it will function securely when implemented as a decision support tool in the routine clinca practice in real-time with data from an implementation environment (IE). It is important to clarify that, although these 2 environments can be different due to spatial dimension (2 different ICUs), they can also differ due to temporal dimension (same ICU, different time periods). Currently, there is no standardized protocol of the steps that should be followed to ensure its success. However, there are consensus guidelines among experts to help us during the process.[62] Therefore, we propose a minimum of 4 necessary phases to safely transition an AI from a DE to a IE as a decision support tool (Fig. 2).

Phase 1: testing the AI with retrospective data from the IE

AI requires a set of predictor variables (PV) to return the response variable (RV). An obvious first step is to ensure that PVs can be automatically obtained from the electronic health record (EHR) of the IE. The fewer PVs the AI requires and the less specific they are, the easier its implementation in different IE will be. Currently, AI models that have demonstrated high performance in literature are mostly not applicable to the routine clinical practice.[63] If they can be automatically obtained, the AI's performance within the IE with retrospective data will be evaluated. If results are unsatisfactory, the decision should be made either to terminate the process or to open a new scenario for AI retraining to improve results in the IE, i.e., undergo a process of generalization.[64] This latter step will depend greatly on the framework in which the integration process is being carried out, and all required ethical and legal guidelines must always be followed. It only makes sense to move on to phase 2 if good AI performance is achieved with retrospective data from the IE.

Phase 2: testing the AI with real-time data from the IE blindly for the clinician

Turning the extraction, transformation, and loading (ETL) process of PVs to run an AI in an 'ad-hoc' manner into a stable and scalable process or pipeline resistant to failures is a costly technological task. Without going into detail, once there is confirmation that everything works in real-time and that a protocol for handling system failures has been designed, the AI can be evaluated prospectively. This type of prospective evaluation is also necessary in cases where the DE and the IE are the same ICU, where what has changed is the temporal dimension. An AI that has shown good performance in its DE or in Phase 1 of the IE may be impaired by inherent time changes (new professionals, new habits, new drugs, pandemics, etc.). Therefore, ensuring sustained good performance in this Phase 2 is crucial, as it will indicate both that the IE has the necessary technological infrastructure to maintain the decision support tool and that the AI is robust enough to function stably over time.

Phase 3: clinical trial considering the use of AI as intervention

If Phase 2 is successfully passed, we know that we have a robust and stable AI capable of making accurate predictions in most cases. However, we do not know what impact its use by the clinical team could have had on the patient. This 3rd phase requires the design of a clinical trial capable of evaluating if there are significant differences between a control group without AI and an intervention group with AI.[62] At this point, it can be crucial for the AI to be interpretable rather than a black box.[65] An interpretable AI can give information to the clinician on the PVs that are impacting the RV, helping the clinician understand what should be changed to avoid that unwanted RV if they choose to do so. Conversely, if the AI is not interpretable, the task of finding out why the AI provides an unwanted RV will entirely fall on the clinician.

Phase 4: continuous monitoring and evolution of AI

ML-based AIs learn from a set of cases based on PVs and defining RVs. The temporal dimension inevitably turns any IE into a DE over time. New socio-economic contexts, new teams, new drugs, even new habits acquired from future human-AI synergies will render AIs obsolete if they do not evolve dynamically. For example, an AI trained to predict a certain adverse event in an environment where protocols did not use that same AI may stop working when the very AI is applied to prevent it, as new protocols adding the AI will have been generated, completely changing the context in which it was trained. In this final phase, a set of indicators of human actions motivated by the AI should be defined, whose monitoring will ensure that the coexistence of both intelligences (human-artificial) is beneficial for the patient. This set of indicators will depend on the type of AI and its objective. Finally, periodically and through constant clinical inputs, AIs should be retrained with new PVs to adapt to new IE and improve their performance.[66]

Conclusions

The addition of artificial intelligence (AI) to the realm of security, while promising, faces key challenges. Predicting adverse events and aiding safe prescription represent significant opportunities. However, the lack of methodological quality in research and the need to address ethical concerns, such as trust and bias, are imperatives. Successful

Table 3 Main metrics used in the evaluation of AI models.

Metric	Description	Impact on clinical outcomes	Transparency	Examples
Accuracy	Percentage of data correctly classified.	Can lead to a higher rate of false negatives, which can delay or prevent appropriate treatment.	Easy to understand and interpret.	Prediction of the probability of death in COVID-19 patients
Precision	Percentage of positive data correctly classified.	Can lead to a higher rate of false positives, which can cause anxiety or stress in patients.	Can be difficult to interpret if the prevalence of the positive class is low.	Prediction of breast cancer presence in mammograms
F1-score	Weighted average of precision and recall.	Although it can be a good choice for balanced classification tasks, it is important to consider its potential impact on the clinical outcomes.	Can be difficult to interpret if the prevalence of the positive class is low.	Prediction of stroke probability in hypertensive patients
Specificity	Percentage of negative data correctly classified.	Can lead to a higher rate of false negatives, which can delay or prevent appropriate treatment.	Can be difficult to interpret if the prevalence of the positive class is low.	Biomarkers (used with precision)
ROC curve	Represents the relationship between true positive rate (TPR) and false positive rate (FPR).	Although it can be a good choice for comparing the performance of different models, it gives equal importance to both precision and specificity.	Can be difficult to interpret if the prevalence of the positive class is low.	Prediction of cancer survival probability
Area under ROC curve	Value of the ROC curve at points (0, 0) and (1, 1).	Although it can be a good choice to compare performance of different models, it gives equal importance to both precision and specificity.	Can be difficult to interpret if the prevalence of the positive class is low.	Prediction of mortality in ICU patients
Precision/Recall AUC	Represents the relationship between precision and recall.	Can be difficult to interpret if the prevalence of the positive class is low.	Can be a good choice for imbalanced classification tasks.	Prediction of acute hospitalizations in elderly receiving home care
Logarithmic loss	Sum of the logarithms of the probabilities of correct predictions.	Although it can be a good choice for comparing the performance of different models, it's sensitive to imbalanced data and lacks explainability. It does not take the severity of the error into account either	It's a difficult metric to interpret, as it requires knowledge of logarithms. However, it's an objective metric that can be used to compare performance of different models.	Prediction of readmission 1 year after discharge
Jaccard Index	Relationship between the number of elements correctly classified and the sum of the number of elements correctly classified and the number of elements misclassified.	Although it can be a good choice for comparing the performance of different models, it gives equal importance to both precision and specificity. Additionally, at an individual level (e.g., pixel in the case of images), it lacks gradation as it is a binary metric.	It's an easy metric to understand and interpret. However, it's less sensitive to false negatives than other metrics, such as accuracy or the F1-score.	Prediction of presence of brain damage on MRI images

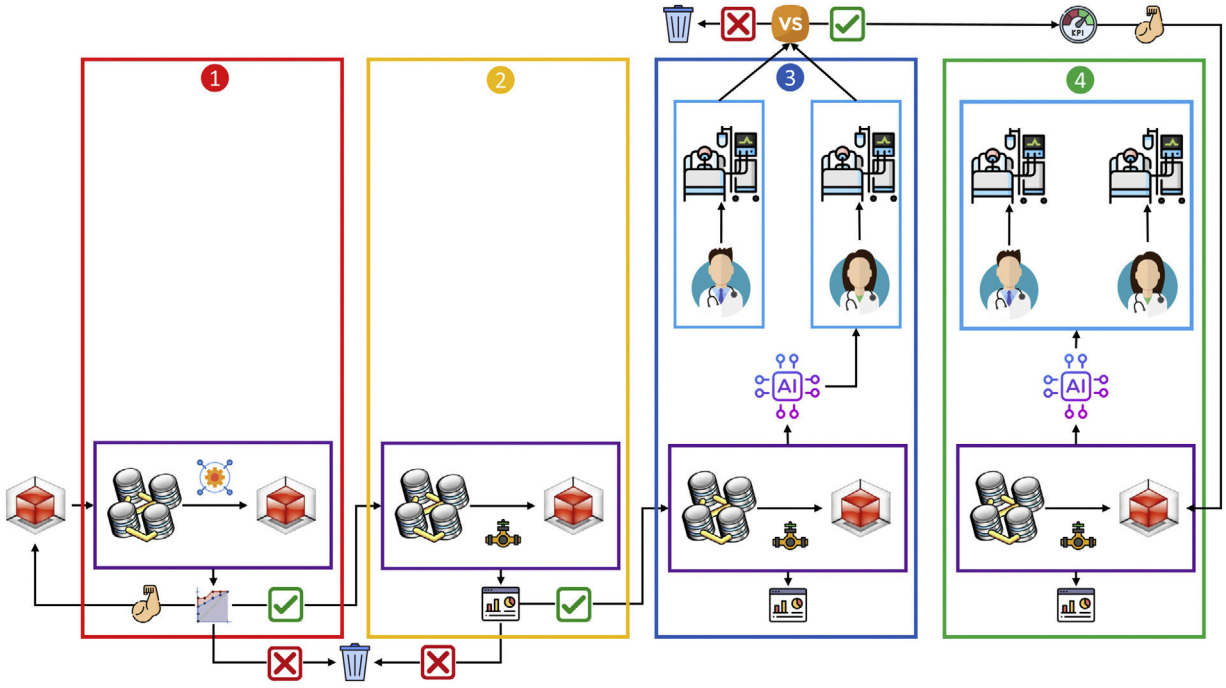

Figure 2 Map of the 4 phases of secure real-time AI implementation. In phase 1, the AI is tested to ensure it adapts to the reality of the data from the center where it is deployed. In phase 2, the data flow is constructed to evaluate its performance in real-time. In phase 3, the results of using AI vs not using it are compared in relation to patient benefit. Finally, in phase 4, the performance of the AI is continuously monitored, and necessary improvements are applied to ensure its evolution for the benefit of all.

implementation requires not only technical robustness but also careful transition, ensuring understanding and acceptance by health care professionals. Continuous security and adaptability emerge as crucial foundations for effective collaboration between AI and health care, ensuring tangible benefits for patient safety.

During the preparation of this work, the authors used Chat-GPT 3.5 to request synonyms and improve the translation of technical expressions from English to Spanish. After using this service, the authors reviewed and edited the content as necessary, assuming full responsibility for the publication's content.

Funding

None declared.

Conflicts of interest

JABM has worked for the artificial intelligence company Savana Médica. The remaining authors declared no conflicts of interest whatsoever.

Authors' contribution

Jesús Abelardo Barea Mendoza: Conceptualization, drafting, editing, and final manuscript review. Josep Gómez Álvarez: Conceptualization, drafting, editing, and final manuscript review.

Alex Pardo Fernandez: Drafting and final manuscript review.

Marcos Valiente Fernandez: Drafting and final manuscript review.

Acknowledgements

None declared.

References

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 AMIA Symp. 2017;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. Continuous and early prediction of future moderate and severe Acute Kidney Injury in critically ill patients: development and multi-centric, multi-national external validation of a machine-learning model. 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 medi-

cations 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 JAMIA. 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 JAMIA. 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. p. 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, Wawira Gichoya J, 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.