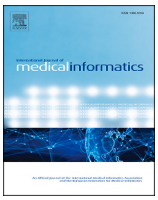
International Journal of Medical Informatics 219 (2026) 106545

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



From development to clinical practice: deployment of an interoperable and secure ML-based CDSS to aid in the early detection of sepsis

Ana Serrano García [a], David López [a], Eric Macias-Fassio [a,b,*], Santiago Salas-Sosa [a], Iván Pascual [a], Cristina Pruenza [a], Marcio Borges-Sa [d,e,f,g,h], Andrés Giglio [d,f,g,h,i,j], Jaime Cruz-Rojo [c], Rodrigo Pacheco-Puig [c]

[a] Instituto de Ingeniería del Conocimiento, Madrid, Spain
[b] BiometricsAI, Universidad Autónoma de Madrid, Madrid, Spain
[c] Hospital Universitario 12 de Octubre, Madrid, Spain
[d] Multidisciplinary Sepsis Group, Son Llatzer Hospital Palma, Palma, Spain
[e] Intensive Care Unit, Son Llatzer Hospital, Palma, Spain
[f] Health Research Institute of the Balearic Islands (IdISBa), Palma, Spain
[g] Balearic Islands University, Palma, Spain
[h] Fundación Código Sepsis, Valencia, Spain
[i] Faculty of Medicine, Finis Terrae University, Santiago, Chile
[j] Critical Care Center, Clínica Las Condes Hospital, Santiago, Chile

HIGHLIGHTS

- Novel framework with standards-based, EHR-agnostic design for high interoperability.
- Flexible CDSS providing early ML-alerts integrated into clinical workflows.
- Deployed on hospital side servers to fully preserve data privacy.
- Operating in real time across two hospital settings.

ARTICLE INFO

Keywords:
CDSS
Machine learning
Sepsis

ABSTRACT

Recent research has increasingly focused on machine learning (ML) models for early disease prediction, yet practical frameworks for integrating these models into clinical workflows remain limited. BIAlert is a microservices-based framework designed to operate as a real-time early-warning system for ML-driven disease prediction in hospitalised patients. It can be deployed remotely on physical or virtual servers and is composed of coupled microservices that communicate through Apache Kafka queues, using HL7 FHIR resources as the message format. The system comprises four core components: (1) the Connector, which ingests raw hospital data and converts it into standardised healthcare formats; (2) the Writer, which stores FHIR-formatted data in an internal database and triggers the prediction pipeline; (3) the Predictor, which hosts ML models and generates patient-specific alerts; and (4) the Model Evaluator, which supports prospective monitoring of model performance. Alerts are displayed through the BIAlert user interface and can also be integrated directly into the electronic health record (EHR). BIAlert is currently deployed and operating in real-time clinical settings in two hospitals, demonstrating its feasibility as a scalable and interoperable solution for ML-based clinical decision support.

1. Introduction

Intervening in health issues at an initial stage often leads to better prognoses and reduced complications. Sepsis, a dangerous and complex condition resulting from an extreme host response to infection, exemplifies the need for early detection and intervention to enhance patient outcomes [1,2]. Yet, sepsis is a heterogeneous entity that develops

* Corresponding author at: Instituto de Ingeniería del Conocimiento, Madrid, Spain.
Email addresses: ana.serrano@iic.uam.es (A. Serrano García), eric.macias@iic.uam.es; eric.macias@estudiante.uam.es (E. Macias-Fassio).

https://doi.org/10.1016/j.ijmedinf.2026.106545
Received 15 December 2025; Received in revised form 12 March 2026; Accepted 9 June 2026
Available online 17 June 2026

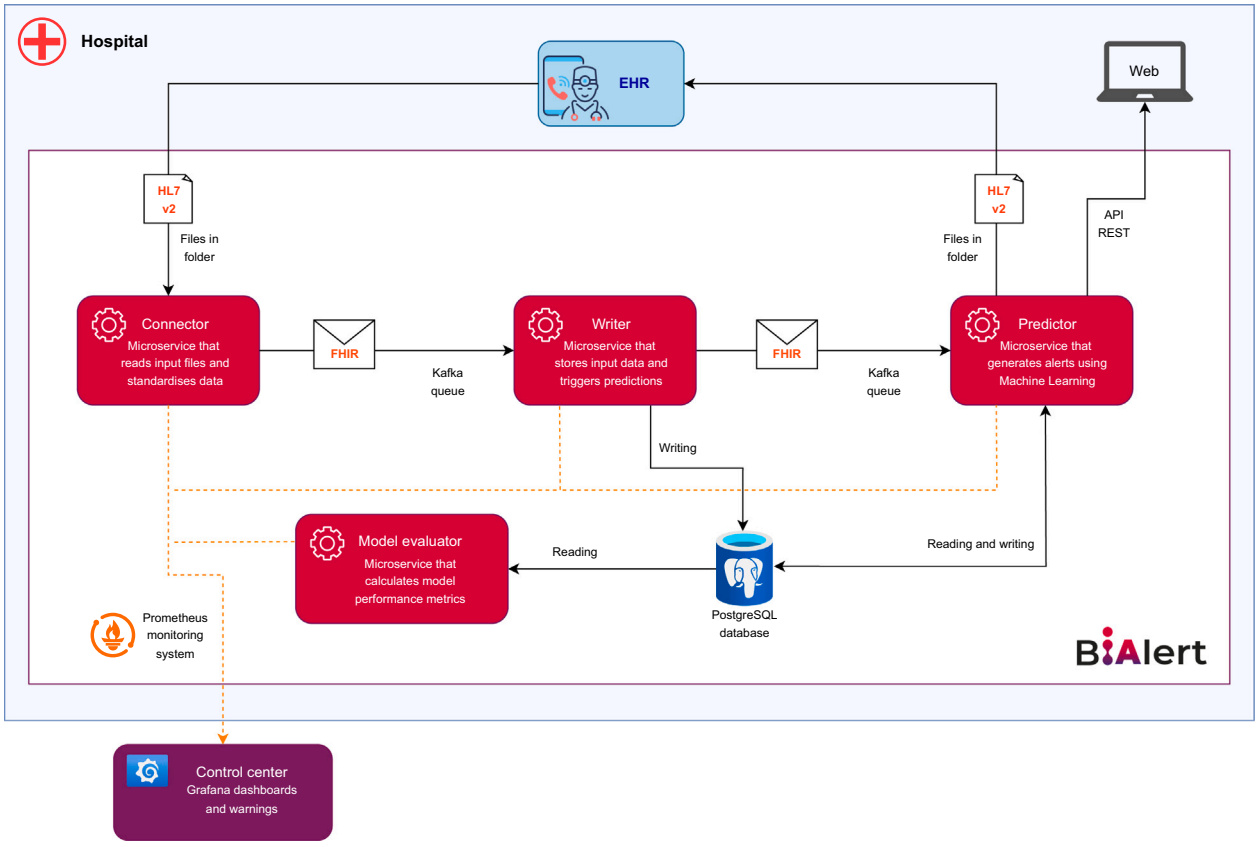


Fig. 1. High-level architecture of BIAlert. Red boxes represent the tool's microservices, while arrows indicate information flows. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

gradually and may be difficult to recognise during its preliminary phases [3]. To address this, some computerised strategies have been reported as valuable by clinical staff [4] and demonstrated their potential to decrease patient lengths of stay and direct medical costs [5].

Recently, research has focused on the use of Machine Learning (ML) to predict diseases before they become clinically apparent. Its ability to find complex, multivariate, non-linear patterns in real examples of large historical datasets has shown great accuracy in the prediction of cardiac arrests [6], COVID-19-induced respiratory distress [7], or sepsis and septic shock [8].

However, integrating ML into healthcare systems requires overcoming several challenges within the software architecture. These include ensuring patient privacy, achieving standardisation and interoperability, and designing systems that are flexible and maintainable. For these reasons, despite the abundance of evidence that ML models may help clinical practice, only a small proportion of them has been successfully embedded in real-world Clinical Decision Support Systems (CDSS) [9,10].

This paper introduces BIAlert (Big-data Intelligent Alerting), a framework built on a microservices architecture [11] that operates as an early warning system for the ML-based prediction of diseases in hospitalised patients. BIAlert has been carefully designed to overcome some common limitations in CDSS reaching production level:

- It is highly parametrisable to adapt to the hospital's clinical practice and operational preferences.
- It relies on medical standards to maximise interoperability and become agnostic to the Electronic Health Records (EHR) system of the hospital.
- It is remotely deployed on the hospital's premises, so that patient data never need to leave the hospital and privacy is preserved.

While this work focuses on the architectural scalability of BIAlert, the system's clinical relevance including its impact on ICU admissions and length of stay has been preliminarily validated in a previous single-centre study [12]. In the following sections, the CDSS's architecture will be described together with the experience of its implementation and real-world deployment for sepsis prediction in two distinct settings: all services in Son Llàtzer Hospital (HSLL) and the Emergency Room (ER) service in 12 de Octubre Hospital (H12O).

2. Materials and methods

2.1. High-level architecture

2.1.1. Overview

A diagram of the CDSS's architecture can be found in Fig. 1. It is composed of several microservices that communicate with each other through Apache Kafka [13] queues where messages constitute HL7 FHIR objects [14].

Lately, microservices-based architectures are replacing traditional monolithic approaches [15], as they facilitate scalability and maintainability. In BIAlert, each microservice has its own life cycle, and can be developed, debugged, tested and updated in production on its own. Other CDSS with microservices-based architectures have been implemented for infectious disease prediction [16,17] or EHR exploitation [18].

2.1.2. Remote deployment and monitoring

The tool is designed to be installed on a physical or virtual server within the hospital, although cloud installation is also possible. In this way, patient data remains inside the hospital at all times. The installation, however, is done remotely through secure means, like Virtual Private Networks, so that access to the server remains restricted. Then,

Fig. 2. Dummy example of BIAlert's interface for HSLL and sepsis prediction.

the tool is updated automatically using ArgoCD [19] whenever a new version is released, pulling the release through Hypertext Transfer Protocol Secure (HTTPS) from a unique, central repository, hosted in GitLab [20], without the need to open ports for incoming connections.

BIAlert is monitored in real time using Prometheus and Grafana [21]. Operational and performance metrics are continuously collected from the diverse microservices and exported through an outgoing connection. Such metrics are visualised on dashboards, and alerts are configured to indicate issues such as missing input data or poor model performance.

2.1.3. Connector component

The connector is the component that reads original data from the hospital, converts them to match BIAlert's standard and sends them to the writer. BIAlert has an internal catalog that is based on SNOMED CT [22], LOINC [23] and ATC [24] standards. As part of the installation process, the hospital must fill out the data manifest: a structured map where, for each variable in BIAlert, the hospital's identifier and unit are linked. This manifest is used by the connector to standardise and scale hospital's data in an online manner.

Input data are structured, and they consist of demographic variables for the patient, together with clinical, analytical, pharmacological and microbiological variables. The hospital constantly exports new data from its EHR to BIAlert's server. The information is agreed to follow the HL7 v2 standard [25], as it is widely adopted and syntactically simple.

2.1.4. Writer component

The writer receives new standardised data from the connector, stores them in an internal relational [26] database and triggers the predictor. Such trigger is executed when a patient's predictor variable has been updated. Obviously, this includes the patient's first measurements upon their arrival.

The writer microservice works with FHIR data structures (objects) that facilitate transmission efficiency. A hospital could do without the connector component if they directly sent standardised FHIR objects to the writer's input queues.

2.1.5. Predictor component

The predictor is the microservice that embeds the ML model and uses it to generate alerts. It gets triggered by the writer when a new prediction needs to be made. The predictor fetches the patient's standardised data and predicts their risk of developing the disease in the following hours. If this risk is high, an alert is created (or updated if the patient's last prediction was positive as well).

The predictor reads from and writes to the internal database. Thus, it exploits both new and historical patient data to make its predictions. The details regarding the predictions are also stored in the database.

2.1.6. Outputs

When an alert is generated, the predictor sends it to the hospital through two outputs:

- A credentialed web interface that can be accessed via a browser inside the hospital's network. The interface presents detailed patient information and highlights the key features contributing to the alert, based on SHAP value explanations [27]. Figs. 2 and 3 show example screenshots of the interface, illustrating how clinicians can review active alerts and interpret the key variables driving each prediction.
- HL7 v2 files on the server where BIAlert is installed. This mechanism allows the hospital to read them online and integrate the information with their EHR system in a format of their choice. An example of integration in the form of EHR pop-ups is displayed in Fig. 4.

2.1.7. Model evaluator component

In ML-based tools where output interpretation is not completely straightforward, it is necessary to prospectively monitor the performance of the model. Apart from predictive information, the hospital also sends validation information for the predictions. These can be diagnoses or standard diagnostic criteria. Examples of such standard criteria for sepsis are sepsis-2 [28] and sepsis-3 [29] definitions.

Said validation information is standardised through the connector and stored in the database. The model evaluator microservice fetches predictions generated some days before and uses the posterior diagnostic information to validate them and calculate model performance metrics. These are then exported through the continuous monitoring of the tool.

2.2. Machine learning models

The BIAlert models are extensions of the work established in [30], using custom ensembles of Gradient Boosting-based estimators [31] to

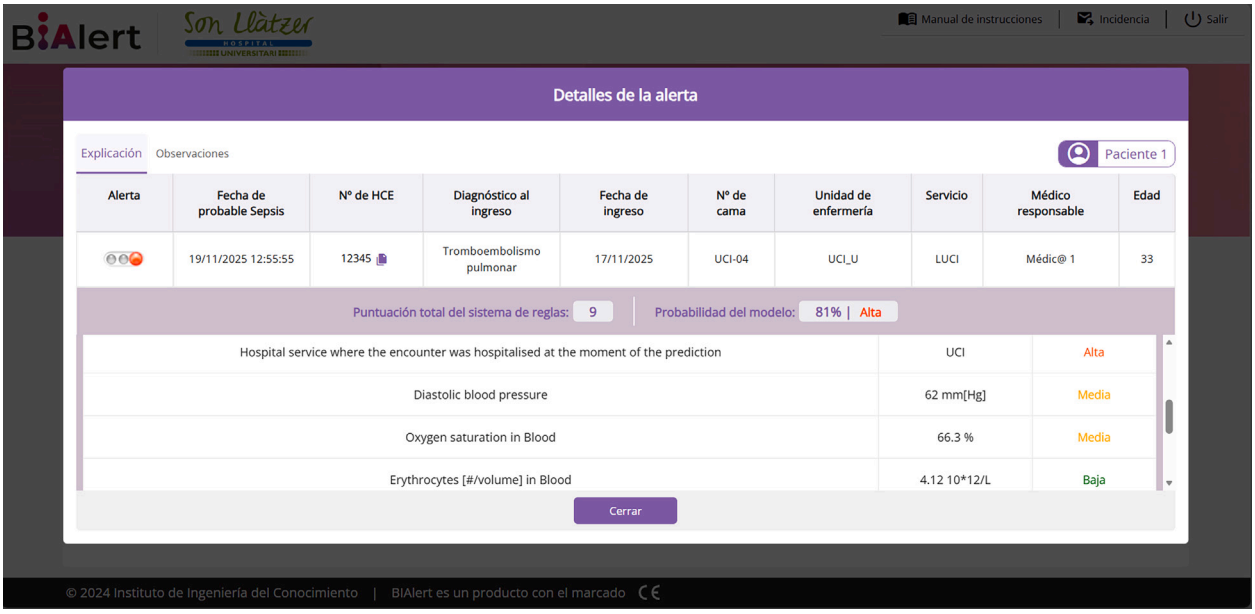


Fig. 3. Dummy example of BIAlert's interface displaying an alert's most influential predictors.

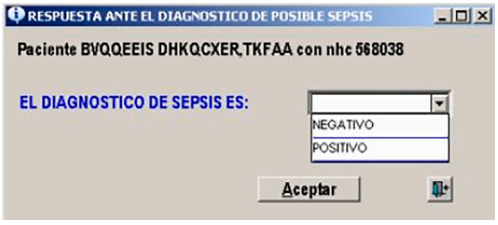


Fig. 4. Dummy example of the pop-up that appears on HSLL's EHR system when BIAlert generates a sepsis alert. The hospital has enabled a feedback system that allows recording answers to alerts in the EHR, which are later also sent to BIAlert as diagnostic information.

predict the risk of disease within the next 24 hours in hospitalised patients. Such architectures natively handle missing values and do not require imputation.

Features were derived from 68 structured EHR variables. Data windows were generated dynamically: a new 72-hour window was created each time a new batch of clinical observations was recorded for a patient. For every such update, the most recent value for each of the 68 variables within the preceding 72 hours was selected. Positive labels were assigned to windows followed by a disease diagnosis within the subsequent 24 hours. Negative samples corresponded to windows not followed by a disease event and where enough data were recorded to ensure that disease criteria were not met (see Section 2.4 for an explanation of the labelling process).

Model training and evaluation were performed ensuring that data points from one same patient could only be in either the training or testing subset to prevent data leakage. Optimal architectures and hyperparameters were selected based on global performance and the hospital's specific sensitivity–precision preferences. To address class imbalance, ROC-AUC and F-score metrics were favored, and the final deployment threshold was determined by maximising the Youden Index on the training set [32].

The development of the ML models in BIAlert follows the TRIPOD-AI reporting guidelines [33] and the methodology is equivalent for all models.

2.3. Real-world deployment

BIAlert was successfully deployed in two settings for patients above 14 years of age:

- ER and ward services (except for Psychiatry) of Son Llàtzer Hospital (HSLL) in Mallorca, Spain.
- ER service of 12 de Octubre Hospital (H12O) in Madrid, Spain.

Each hospital had its particular ML model, trained with its own labelled data and its preferred criteria. Table 1 summarises the performance metrics obtained during model training. HSLL chose a sensitive model based on sepsis-2, while H12O prioritised positive predictive value (PPV) and the sepsis-3 definition of sepsis.

In the Results section, we report the ML performance results for HSLL and H12O over a standard four-month period (January 1st to April 30th,

Table 1
Training metrics of the ML models deployed in HSLL and H12O. Metrics were obtained by averaging the results of a 4-fold cross-validation [34]. *N* represents the total number of labelled data points that were used to train the models. The following row represents the unique patients who produced such data points at different moments in time. The other metrics are typically used for evaluating binary classification models and represent the model's accuracy, specificity, sensitivity, negative predictive value (NPV), positive predictive value (PPV) and area under the Receiver Operating Characteristic (ROC) curve (AUC) [32].

	HSLL	H12O
N	19,045	20,459
Patients, *N*	10,928	7159
Accuracy, %	89.62	88.99
Specificity, %	84.01	95.04
Sensitivity, %	95.21	69.39
NPV, %	98.48	90.97
PPV, %	61.80	81.15
AUC, %	95.70	93.76

2025), during which the latest stable version of BIAlert was deployed in both hospitals.

2.4. Sepsis labelling strategies

The labelling strategies described below are applied both for the construction of the training dataset and for real-time performance monitoring within the model-evaluator microservice. These methods are based on existing literature [35–37] and tailored to each hospital's clinical practice.

The primary ground truth is a structured sepsis confirmation recorded in the EHR. If a prediction has been followed by an explicit sepsis diagnosis within 24 hours, the data point is labelled as positive. In the absence of an explicit confirmation from the EHR, standard criteria are used for labelling:

- HSLL: A data point is labelled positive if in its following 24 hours the patient has had a suspected or confirmed infection and an acute score ≥ 6 on the hospital's PIMIS system for sepsis-2 [4]. Conversely, the data point is labelled as negative only if sufficient data have been recorded to confirm the score does not exceed this threshold.
- H12O: A data point is labelled positive if in its following 24 hours the patient has had a suspected or confirmed infection and an acute SOFA score of 2 or more [38]. Again, the data point is labelled as negative only if sufficient data have been recorded to confirm the score does not exceed this threshold.
- For both environments, a suspected or confirmed infection is defined as the recording of both a blood culture sample and antibiotic administration within a $\pm$ 2 day window relative to the data point's timestamp [39].

3. Results

3.1. Clinical population

Some clinical characteristics of the population processed by the tool during the study period can be found in Table 2. On the one hand, the ER service of H12O showed a greater patient rotation than the combined ER and wards services of HSLL. With respect to the total number of patients, approximately 7% of those in H12O's ER had an alert. In comparison, this proportion was higher in HSLL ($\approx$ 19%). Of these patients with alerts in HSLL, the vast majority had their first alert generated in the ER service ($\approx$ 80%). On the other hand, while both hospitals admitted more women than men overall, sepsis alerts were more commonly generated for male patients. Finally, patients with sepsis alerts tended to be older than the average hospitalised patient in both hospitals.

Table 2
Demographics of the patients from HSLL and H12O processed by BIAlert during the four-month period of the study. In H12O, the tool was only deployed for ER, so all patients with an alert had their first alert in ER.

	HSLL	H12O
All patients		
N	17,214	54,159
Women, *N* (%)	9452 (55)	31,148 (58)
Age, mean (SD)	53 (21)	48 (21)
Patients with alert		
N	3328	3834
Women, *N* (%)	1639 (49)	1781 (46)
Age, mean (SD)	63 (21)	68 (21)
Patients with 1st alert in ER		
N	2666	3834
Women, *N* (%)	1277 (48)	1781 (46)
Age, mean (SD)	63 (21)	68 (21)

Table 3
Operational metrics during the study period, consisting of counts of observations received and processed from the hospitals and predictions generated by the tool.

	HSLL	H12O
N° of observations	1,739,347	1,696,002
N° of predictions	343,293	253,867

Table 4
Alert dynamics during the study period, consisting of total numbers of positive predictions, numbers of new alerts among the positive predictions, positive predictions per patient, new alerts per day and hours between patient admission and a patient's first alert.

	HSLL			H12O
	All	ER	Wards	ER
N° of positive predictions	67,941	21,879	46,062	29,844
N° of new alerts	3874	2338	1536	4312
Positive predictions P.P., median (IQR)	9 (21)	5 (9)	8 (17)	6 (8)
Alerts per day, median (IQR)	27 (8)	20.5 (7)	6 (4)	31 (10.5)
Hours to 1st alert, median (IQR)	3 (9)	1 (3)	32 (86)	3 (6)

3.2. Operational performance

Table 3 displays some counts related to BIAlert's operation in both HSLL and H12O settings during the study. We define an observation as any time-stamped structured clinical entry in the EHR – such as vital signs, laboratory results, pharmacological prescriptions, logistical updates, or diagnoses – that is transmitted to BIAlert's connector.

Predictions are triggered whenever any patient observation corresponding to a predictor variable is recorded. They may result in a new alert, an alert update if the patient had a previous alert, or in no alert if no risk of the disease is estimated by the ML model. In HSLL, $\approx 1,7$ million observations from $17,214$ patients produced $343,293$ predictions, while in H12O $\approx 1,7$ million observations from $54,159$ patients produced $253,867$ predictions.

3.3. Alert dynamics

Table 4 shows that positive predictions accounted for approximately 19.8% of all predictions at HSLL and 11.8% at H12O. Regarding alert dynamics, 6% of daily positive predictions at HSLL corresponded to first alerts in previously non-alerted patients, whereas approximately 14% represented new alerts at H12O. The remaining positive predictions consisted of alert updates. The median number of positive predictions (initial alerts or updates) per patient was 9 at HSLL and 6 at H12O; HSLL exhibited a higher frequency of patients with more positive predictions. Detailed histograms of positive predictions per patient are provided in Fig. A.5 (Appendix A).

The alert burden was comparable across both sites, with a median of 27 alerts per day at HSLL and 31 at H12O. While medians were similar, H12O experienced a higher frequency of high-volume alert days. Detailed alert burden histograms are available in Fig. A.6 (Appendix A). The median time from admission to the first sepsis alert was approximately 3 hours at both institutions.

Analysis of HSLL alert dynamics by service type revealed that, while the ER generated fewer total positive predictions than the wards, the majority of initial alerts occurred in the ER, representing 11% of the positive predictions. In contrast, only 3% of positive predictions in the wards were new alerts, with the remainder being updates. This is also reflected in the daily alert frequency: the median number of new alerts per day was approximately 3.5 times higher in the ER than in the wards. Lastly, the median time from admission to the first alert was 1h for ER alerts, compared to 32 hours for those initiated in the wards, the latter also showing a significantly higher IQR.

Table 5
Prospective performance metrics of the deployed ML models. Metrics from HSLL are disaggregated by service type.

	HSLL			H12O
	All services	ER	Wards	ER
N	87,905	28,735	59,170	20,289
Patients, *N*	7304	5530	4898	8361
Accuracy, %	89.26	89.78	89.01	89.50
Specificity, %	90.30	89.79	90.55	96.26
Sensitivity, %	84.65	89.97	82.22	70.08
NPV, %	96.32	97.52	95.75	90.23
PPV, %	66.25	66.20	66.28	86.70
AUC, %	93.65	96.19	92.50	95.51

3.4. Sepsis model performance

Following the logic in Section 2.1.7, the evaluation microservice labelled 87,905 predictions from HSLL (≈ 26% of the total generated) and 20,289 predictions from H12O (≈ 8% of the total generated). A fairness analysis concerning the labelled predictions can be found in Tables B.6 and B.7 (see Appendix B). These labels served for the prospective model performance evaluation in Table 5.

A greater number of distinct data points have been labelled in HSLL compared to H12O, despite H12O encompassing more unique individuals generating those data points. Overall, prospective metrics align with the training results from Table 1. In HSLL, PPV is slightly higher in production than during training, at the expense of a reduction in sensitivity. Furthermore, there is no significant difference in performance across the two subsettings of HSLL. As during the training phase, the model in HSLL shows higher sensitivity than PPV, whereas the model in H12O is more precise than sensitive.

4. Discussion

The proposed CDSS has been deployed for sepsis prediction in two distinct healthcare settings, both on-site yet managed remotely. One deployment covers all wards and ER services of a medium-sized hospital located on the island of Mallorca, while the other is limited to the ER of a large hospital in Madrid, Spain's capital and largest city. HSLL serves a reference population of approximately 270,000 inhabitants [40], whereas H12O serves a catchment population of roughly 450,000 inhabitants [41]. Each institution operates its own EHR system with different variable codes and units of measure. Moreover, HSLL adheres to Sepsis-2 criteria, whereas the H12O employs Sepsis-3 guidelines. Their alerting preferences also diverge, with the former favoring a more sensitive strategy to capture a broader range of potential cases, and the latter prioritising a more precise approach to reduce false positives.

These deployments have been made possible thanks to technologies that enable remote installation and monitoring of software systems, hence avoiding the need to export any patient data outside hospital premises and preserving privacy and compliance with regulatory requirements. Additionally, the CDSS benefits from a highly parameterisable architecture, which allows it to adapt to diverse clinical environments and preferences. Central to this flexibility is the data manifest built on SNOMED CT, LOINC, ATC and HL7 FHIR interoperability standards, facilitating integration with heterogeneous EHR systems.

The results reflect that each hospital operates with its own dynamics. For instance, although a similar number of observations are received from both hospitals, there is a greater number of predictions triggered in HSLL. This difference can be attributed to the observations from H12O being recorded in batches with equal timestamps. Besides, BIAlert processes a higher number of unique patients in H12O compared to HSLL. This is attributable to H12O's larger size and the tool being deployed for ER only, as ER stays tend to be shorter. Notably, H12O's ER's number of beds is comparable to the total bed capacity of HSLL. The longer stays tracked in HSLL are also visible in alert dynamics, as there exist many patients with a great number of positive predictions and cases with a long span between admission to the hospital and a new sepsis alert. With respect to the process of prospective labelling: longer stays in HSLL allow for more predictions per patient to be later labelled based on diagnostic data.

HSLL prefers a conservative alerting strategy, ensuring that as few septic patients as possible go undetected, even if this results in a higher number of false positives. In contrast, H12O aims for a higher precision to avoid overwhelming clinical staff with too many alerts. Training and deploying models locally allows optimising sensitivity in HSLL and PPV in H12O, which is manifested in prospective performance metrics, as well as in the proportion of patients with an alert, which is much higher in HSLL than in H12O. Along these lines, the alert burden is comparable in both settings, although the rotation of patients in H12O is significantly greater. The daily frequencies of new alerts seem manageable in both settings; together with the PPV metrics, the results suggest a low risk of alert fatigue [42,43]. Additionally, the two hospitals follow different diagnostic criteria for sepsis, which implies that a patient may be classified as septic in one hospital but not in the other. Training separate models on each hospital's own data not only improves predictive capacity but also enables generating predictions that replicate particular diagnostic practices. The ML models demonstrate a satisfactory performance which is consistent across training and production environments. In HSLL, where the tool is deployed across both the ER and wards, similar metrics are observed in both settings, suggesting model robustness.

Patients who trigger alerts tend to be older men compared to those who do not, which aligns with existing literature that identifies advanced age and masculine gender as risk factors for sepsis [44]. In HSLL, the majority of alerts are triggered in the ER. Across the ER services of both HSLL and H12O, alerts are typically generated rapidly after patient admission, which implies that most sepsis cases develop prior to arrival (community-acquired sepsis). This observation is consistent with previous studies reporting that approximately 70–80% of sepsis cases originate outside the hospital [45,46]. In contrast, patients in hospital wards may develop sepsis later during hospitalisation (nosocomial sepsis), leading to a longer time interval between admission and the first alert in HSLL.

While our findings provide valuable insights into the interoperability of BIAlert, certain limitations should be acknowledged. In terms of the labelling process, a shortcoming is that not all predictions are subsequently labelled, partly due to the subjective and gradual nature of sepsis. This can introduce bias when evaluating the models' predictive capacity, as clearer cases are more likely to be formally diagnosed, while ambiguous cases may be underrepresented in the labelled data points, potentially leading to inflated metrics. Although this report suggests that most sepsis cases begin before hospital admission, detailed analyses of the time span between new alerts and the clinical diagnosis of sepsis remain to be investigated, as such work was beyond the scope of this study due to the inherent complexity and subjectivity of determining a true sepsis onset.

At HSLL, we conducted a clinical impact analysis using a retrospective before–after design including all patients admitted with sepsis between April 2011 and June 2024. Outcomes during the baseline period (January 2011–March 2019) were compared with those observed after the implementation of BIAlert (April 2019–May 2024) [12]. Following deployment of the AI system, ICU admissions decreased from 34.4% to 30.4%, accompanied by reductions of 0.35 ICU days and 0.59 ward days per patient. These findings suggest a potential benefit of earlier sepsis identification and management facilitated by the system. However, this analysis was conducted at a single centre, so future work should evaluate the clinical impact of BIAlert using multicentre studies to better assess its generalisability and real-world effectiveness across different hospital settings.

Another limitation of the current system architecture is the need for an initial anonymised data extraction to train each hospital's model. To

address this, a planned evolution of BIAlert involves including a microservice within the tool to enable on-site model training, where training metrics would be supervised remotely via the existing monitoring system.

In conclusion, this paper proposes an architecture for disease prediction with ML that goes beyond retrospective studies and enables real-world deployment in heterogeneous clinical environments. The system has been successfully deployed and prospectively evaluated, demonstrating its adaptability to diverse EHR systems and operational protocols, while protecting sensitive health information. By emphasising a flexible customisation, interoperability and data security, ML can be effectively translated into CDSS that assist daily clinical practice.

CRediT authorship contribution statement

Ana Serrano García: Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **David López:** Writing – review & editing, Methodology, Conceptualization. **Eric Macias-Fassio:** Writing – review & editing, Investigation, Conceptualization. **Santiago Salas-Sosa:** Writing – review & editing, Investigation. **Iván Pascual:** Writing – review & editing, Methodology, Conceptualization. **Cristina Pruenza:** Writing – review & editing, Conceptualization. **Marcio Borges-Sa:** Writing – review & editing, Investigation. **Andrés Giglio:** Writing – review & editing, Investigation. **Jaime Cruz-Rojo:** Writing – review & editing, Investigation. **Rodrigo Pacheco-Puig:** Writing – review & editing, Investigation.

Ethical approval

The Ethics and Health Research Committee of the Balearic Community (CEIC-Ib) approved the study with a patient-consent waiver. After review by the Legal Departments of the Hospital, Idisba, and the Hospital's Research Commission, authorisation was obtained from hospital management to conduct the study in compliance with European and Spanish Data Protection Law. In addition, the Ethics Committee of 12 de Octubre Hospital reviewed and approved the study under the same regulatory and data-protection standards.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this manuscript, the authors used ChatGPT solely to improve grammar, clarity, and sentence structure. After using this tool, all authors carefully reviewed and edited the content and took full responsibility for the final version of the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Alert burden

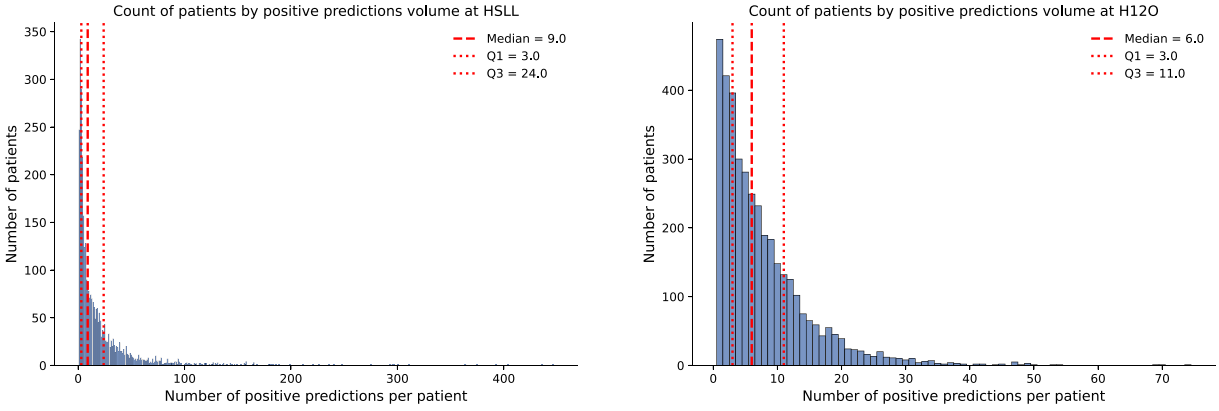


Fig. A.5. Distribution of positive predictions per patient across both hospitals (left panel: HSLL; right panel: H12O).

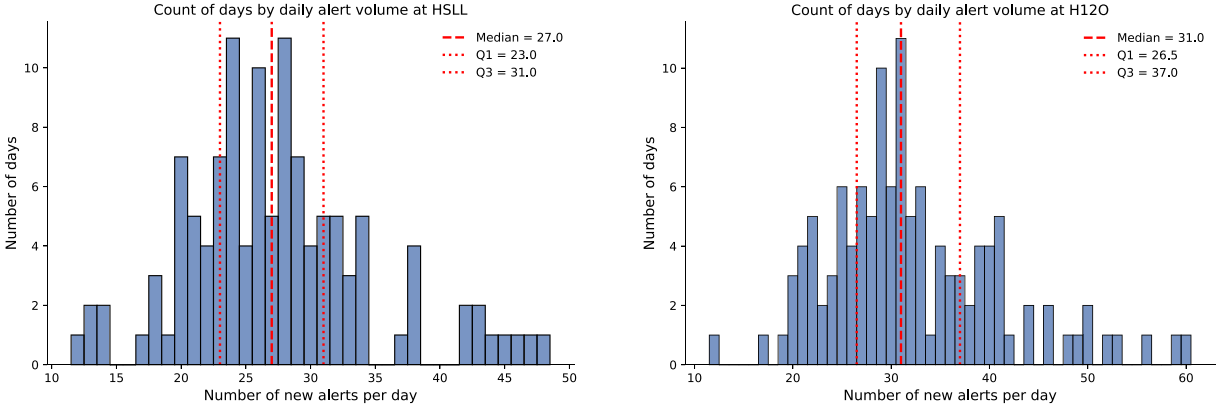


Fig. A.6. Distribution of daily alert frequencies across both hospitals (left panel: HSLL; right panel: H12O).

Appendix B. Prediction labelling

It is important to note that a conservative approach was used for labelling: cases were labelled as non-septic only when sufficient clinical observations were available to confidently rule out sepsis.

Tables B.6 and B.7 report the number of predictions and the proportion of labelled cases across patient subgroups for the HSLL and H120 models, respectively. Overall, labelling rates were substantially higher for HSLL predictions (25.6%) than for H120 predictions (8.0%). These differences may be attributable to differences in the clinical contexts in which the models generate predictions (all hospital services vs ER), as well as differences in the availability of diagnostic information required to confirm or rule out sepsis under the conservative labelling strategy.

In both models, positive predictions were labelled more frequently than negative predictions. For HSLL, 30.4% of positive predictions were labelled compared with 24.4% of negative predictions. A similar pattern was observed for H120, where labelling rates were 14.2% for positive predictions and 7.2% for negative predictions. This pattern is consistent with the conservative labelling strategy, as cases flagged as positive by the model may be more likely to undergo additional diagnostic evaluation, increasing the likelihood that a definitive label is recorded. Furthermore, the volume of total negative predictions is much greater than that of positive predictions.

Moderate variation was observed across demographic subgroups. For the HSLL model, labelling rates were slightly higher among men than women (26.7% vs 24.4%), whereas for the H120 model the overall labelling rate was higher among women (8.8% versus 7.1%). However, for positive predictions the labelling rates were similar across sexes in both models. Differences across age groups were modest in both models, with only small variations in labelling rates and no consistent pattern across age categories.

Table B.6
Counts and labelling rates of HSLL predictions stratified by service type, gender, age group and model prediction type.

Subgroup	Prediction type	Predictions, *N*	Labelled, %
All patients	All	343,293	25.6
All patients	Negative	275,352	24.4
All patients	Positive	67,941	30.4
ER	All	99,575	28.9
Wards	All	243,718	24.3
ER	Negative	77,696	27.8
Wards	Negative	197,656	23.1
ER	Positive	21,879	32.4
Wards	Positive	46,062	29.4
Women	All	165,973	24.4
Men	All	177,320	26.7
Women	Negative	137,084	23.5
Men	Negative	138,268	25.3
Women	Positive	28,889	28.5
Men	Positive	39,052	31.7
Ages in (14, 45]	All	57,166	21.1
Ages in (45, 65]	All	89,641	27.3
Ages > 65	All	196,486	26.1
Ages in (14, 45]	Negative	50,531	20.5
Ages in (45, 65]	Negative	70,441	25.4
Ages > 65	Negative	154,380	25.3
Ages in (14, 45]	Positive	6635	26.1
Ages in (45, 65]	Positive	19,200	34.3
Ages > 65	Positive	42,106	29.2

Table B.7
Counts and labelling rates of H12O predictions stratified by gender, age group, and model prediction type.

Subgroup	Prediction type	Predictions, *N*	Labelled, %
All patients	All	253,867	8
All patients	Negative	224,023	7.2
All patients	Positive	29,844	14.2
Women	All	136,867	8.8
Men	All	117,000	7.1
Women	Negative	123,227	8.2
Men	Negative	100,796	5.9
Women	Positive	13,640	13.7
Men	Positive	16,204	14.6
Ages in (14, 45]	All	77,560	7.9
Ages in (45, 65]	All	66,912	8.2
Ages > 65	All	109,395	7.9
Ages in (14, 45]	Negative	74,310	7.6
Ages in (45, 65]	Negative	59,733	7.3
Ages > 65	Negative	89,980	6.7
Ages in (14, 45]	Positive	3250	13.8
Ages in (45, 65]	Positive	7179	15.8
Ages > 65	Positive	19,415	13.7

Data availability

The data supporting this study's findings are available upon reasonable request to the corresponding author. However, due to privacy restrictions, the data are not publicly available.

References

[1] T. Cardoso, A.H. Carneiro, O. Ribeiro, A. Teixeira-Pinto, A. Costa-Pereira, Reducing mortality in severe sepsis with the implementation of a core 6-hour bundle: results from the Portuguese community-acquired sepsis study (SACiUCI study), Crit. Care 14 (2010) 1–11.
[2] L.Y. Leung, H.-L. Huang, K.K. Hung, C.Y. Leung, C.C. Lam, R.S. Lo, C.Y. Yeung, P.J. Tsoi, M. Lai, M. Brabrand, et al., Door-to-antibiotic time and mortality in patients with sepsis: systematic review and meta-analysis, Eur. J. Intern. Med. 129 (2024).
[3] J. Rello, T. Van Engelen, E. Alp, T. Calandra, V. Cattoir, W. Kern, M. Netea, S. Nseir, S. Opal, F. van de Veerdonk, et al., Towards precision medicine in sepsis: a position paper from the European society of clinical microbiology and infectious diseases, Clin. Microbiol. Infect. 24 (2018) 1264–1272.
[4] B. de Dios, M. Borges, T.D. Smith, A. Del Castillo, A. Socias, L. Gutierrez, J. Nicolas, B. Llado, J.A. Roche, M.P. Diaz, et al., Computerised sepsis protocol management. Description of an early warning system, Enferm. infecc. microbiol. clin. (Engl. ed.) 36 (2018) 84–90.
[5] D.J. Horton, K.K. Graves, P.V. Kukhareva, S.A. Johnson, M. Cedillo, M. Sanford, W.A. Dunson Jr, M. White, D. Roach, J.J. Arego, et al., Modified early warning score-based clinical decision support: cost impact and clinical outcomes in sepsis, JAMIA open 3 (2020) 261–268.
[6] L.S. Crespí, L.G. Madroñal, M.F. Sarubbo, M. Borges-Sa, A.S. García, D.L. Ramos, C.P. Garcia-Hinojosa, E.M. Garijo, Application of a machine learning model for early prediction of in-hospital cardiac arrests: retrospective observational cohort study, Med. intensiva 49 (2025) 88–95.
[7] N. Lazzarini, A. Filippoupolitis, P. Manzione, H. Eleftherohorinou, A machine learning model on real world data for predicting progression to acute respiratory distress syndrome (ARDS) among COVID-19 patients, PLoS One 17 (2022) e0271227.
[8] H.-F. Deng, M.-W. Sun, Y. Wang, J. Zeng, T. Yuan, T. Li, D.-H. Li, W. Chen, P. Zhou, Q. Wang, et al., Evaluating machine learning models for sepsis prediction: a systematic review of methodologies, iScience 25 (2022).
[9] A.P. Susanto, D. Lyell, B. Widyantoro, S. Berkovsky, F. Magrabi, Effects of machine learning-based clinical decision support systems on decision-making, care delivery, and patient outcomes: a scoping review, J. Am. Med. Inform. Assoc. 30 (2023) 2050–2063.
[10] C.K. Corbin, R. Maclay, A. Acharya, S. Mony, S. Punnathanam, R. Thapa, N. Kotecha, N.H. Shah, J.H. Chen, DEPLOYR: a technical framework for deploying custom real-time machine learning models into the electronic medical record, J. Am. Med. Inform. Assoc. 30 (2023) 1532–1542.
[11] J.I. Akerele, A. Uzoka, P.U. Ojukwu, O.J. Olamijuwon, Improving healthcare application scalability through microservices architecture in the cloud, Int. J. Sci. Res. Updates 8 (2024) 100–109.
[12] A. Giglio, E. Macias-Fassio, S. Salas-Sosa, Prospective economic evaluation of a predictive artificial intelligence model for sepsis: effects on hospital costs and return on investment, Available at SSRN 5905239, 2026.
[13] T.P. Raptis, A. Passarella, A survey on networked data streaming with apache kafka, IEEE Access 11 (2023) 85333–85350.
[14] M. Lehne, S. Luijten, P. Vom Felde Genannt Imbusch, S. Thun, The use of FHIR in digital health–a review of the scientific literature, in: German Medical Data Sciences: Shaping Change–Creative Solutions for Innovative Medicine, IOS Press, 2019, pp. 52–58.
[15] A. Saransig, F. Tapia, Performance analysis of monolithic and micro service architectures–containers technology, in: Trends and Applications in Software Engineering: Proceedings of the 7th International Conference on Software Process Improvement (CIMPS 2018) 7, Springer, 2019, pp. 270–279.
[16] H. Calderón-Gómez, L. Mendoza-Pittí, M. Vargas-Lombardo, J.M. Gómez-Pulido, J.L. Castillo-Sequera, J. Sanz-Moreno, G. Sención, Telemonitoring system for infectious disease prediction in elderly people based on a novel microservice architecture, IEEE Access 8 (2020) 118340–118354.

[17] A. Garcés-Jiménez, M.-L. Polo-Luque, J.A. Gómez-Pulido, D. Rodríguez-Puyol, J.M. Gómez-Pulido, Predictive health monitoring: leveraging artificial intelligence for early detection of infectious diseases in nursing home residents through discontinuous vital signs analysis, Comput. Biol. Med. 174 (2024) 108469.
[18] A. Mavrogiorgou, S. Kleftakis, K. Mavrogiorgos, N. Zafeiropoulos, A. Menychtas, A. Kiourtis, I. Maglogiannis, D. Kyriazis, beHEALTHIER: a microservices platform for analyzing and exploiting healthcare data, in: 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS), IEEE, 2021, pp. 283–288.
[19] E. Utami, H. Al Fatta, et al., Analysis on the use of declarative and pull-based deployment models on gitops using argo CD, in: 2021 4th International Conference on Information and Communications Technology (ICOIACT), IEEE, 2021, pp. 186–191.
[20] M.S. Arefeen, M. Schiller, Continuous integration using GitLab, Undergrad. Res. Nat. Clin. Sci. Technol. J. 3 (2019) 1–6.
[21] Y. Jani, Unified monitoring for microservices: implementing prometheus and grafana for scalable solutions, J. Artif. Intell. Mach. Learn. Data Sci. 2 (2024) 848–852.
[22] E. Chang, J. Mostafa, The use of SNOMED CT, 2013-2020: a literature review, J. Am. Med. Inform. Assoc. 28 (2021) 2017–2026.
[23] A.W. Forrey, C.J. Mcdonald, G. DeMoor, S.M. Huff, D. Leavelle, D. Leland, T. Fiers, L. Charles, B. Griffin, F. Stalling, et al., Logical observation identifier names and codes (LOINC) database: a public use set of codes and names for electronic reporting of clinical laboratory test results, Clin. Chem. 42 (1996) 81–90.
[24] S. Natsch, Y. Hekster, R. De Jong, E. Heerdink, R. Herings, J. Van der Meer, Application of the ATC/DDD methodology to monitor antibiotic drug use, Eur. j. clin. microbiol. infect. dis. 17 (1998) 20–24.
[25] D. Bender, K. Sartipi, HL7 FHIR: an agile and RESTful approach to healthcare information exchange, in: Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems, IEEE, 2013, pp. 326–331.
[26] M. Stonebraker, A. Jhingran, J. Goh, S. Potamianos, On rules, procedure, caching and views in data base systems, ACM SIGMOD Rec. 19 (1990) 281–290.
[27] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. 30 (2017).
[28] M.M. Levy, M.P. Fink, J.C. Marshall, E. Abraham, D. Angus, D. Cook, J. Cohen, S.M. Opal, J.-L. Vincent, G. Ramsay, et al., 2001 sccm/esicm/accp/ATS/SIS international sepsis definitions conference, Crit. Care Med. 31 (2003) 1250–1256.
[29] M. Singer, C.S. Deutschman, C.W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, R. Bellomo, G.R. Bernard, J.-D. Chiche, C.M. Coopersmith, et al., The third international consensus definitions for sepsis and septic shock (Sepsis-3), Jama 315 (2016) 801–810.
[30] M. Borges-Sa, A. Giglio, M. Aranda, A. Socias, A. Del Castillo, C. Pruenza, G. Hernández, S. Cerdá, L. Socias, V. Estrada, et al., Hospital-wide sepsis detection: a machine learning model based on prospectively expert-validated cohort, J. Clin. Med. 15 (2026) 855.
[31] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. stat. (2001) 1189–1232.
[32] A. Tharwat, Classification assessment methods, Appl. Comput. Inform. 17 (2021) 168–192.
[33] G.S. Collins, K.G. Moons, P. Dhiman, R.D. Riley, A.L. Beam, B. Van Calster, M. Ghassemi, X. Liu, J.B. Reitsma, M. Van Smeden, et al., TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods, BMJ 385 (2024).
[34] I.K. Nti, O. Nyarko-Boateng, J. Aning, et al., Performance of machine learning algorithms with different K values in K-fold cross-validation, Int. J. Inf. Technol. Comput. Sci. 13 (2021) 61–71.
[35] F. van Wyk, A. Khojandi, A. Mohammed, E. Begoli, R.L. Davis, R. Kamaleswaran, A minimal set of physiomarkers in continuous high frequency data streams predict adult sepsis onset earlier, Int. J. Med. Inform. 122 (2019) 55–62.
[36] J.-W. Perng, I.-H. Kao, C.-T. Kung, S.-C. Hung, Y.-H. Lai, C.-M. Su, Mortality prediction of septic patients in the emergency department based on machine learning, J. Clin. Med. 8 (2019) 1906.
[37] M.A. Reyna, C.S. Josef, R. Jeter, S.P. Shashikumar, M.B. Westover, S. Nemati, G.D. Clifford, A. Sharma, Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019, Crit. Care Med. 48 (2020) 210–217.
[38] M. Singer, C.S. Deutschman, C.W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, R. Bellomo, G.R. Bernard, J.-D. Chiche, C.M. Coopersmith, et al., The third international consensus definitions for sepsis and septic shock (Sepsis-3), Jama 315 (2016) 801–810.
[39] C. Rhee, R. Dantes, L. Epstein, D.J. Murphy, C.W. Seymour, T.J. Iwashyna, S.S. Kadri, D.C. Angus, R.L. Danner, A.E. Fiore, et al., Incidence and trends of sepsis in US hospitals using clinical vs claims data, 2009–2014, Jama 318 (2017) 1241–1249.
[40] Servei de Salut de les Illes Balears (IB-Salut), Hospital universitario son llàtzer, 2024. https://docusalut.com/communities (Accessed 12 March 2026).
[41] Comunidad de Madrid, Hospital universitario 12 de octubre, 2024. https://www.comunidad.madrid/hospital/12octubre (Accessed 12 March 2026).
[42] D.R. Murphy, B. Reis, D.F. Sittig, H. Singh, Notifications received by primary care practitioners in electronic health records: a taxonomy and time analysis, Am. j. med. 125 (2012) 209–e1.
[43] A.M. Harrison, O. Gajic, B.W. Pickering, V. Herasevich, Development and implementation of sepsis alert systems, Clin. chest med. 37 (2016) 219.
[44] M. Fathi, N. Markazi-Moghaddam, A. Ramezankhani, A systematic review on risk factors associated with sepsis in patients admitted to intensive care units, Aust. Crit. Care 32 (2019) 155–164.
[45] S.A. Novosad, Vital signs: epidemiology of sepsis: prevalence of health care factors and opportunities for prevention, MMWR. Morb. mortal. wkly. rep. 65 (2016).
[46] D.B. Page, J.P. Donnelly, H.E. Wang, Community-, healthcare-, and hospital-acquired severe sepsis hospitalizations in the university HealthSystem consortium, Crit. Care Med. 43 (2015) 1945–1951.