TYPE Review
PUBLISHED 05 August 2025
DOI 10.3389/fdgth.2025.1633458

OPEN ACCESS

EDITED BY
Kirti Sundar Sahu,
Canadian Red Cross, Canada

REVIEWED BY
Wei Jun Dan Ong,
National University Health System, Singapore
Muhammad Muneeb Khawar,
King Edward Medical University, Pakistan

*CORRESPONDENCE
Tanja Stamm
✉ tstamm34@gmail.com

RECEIVED 22 May 2025
ACCEPTED 15 July 2025
PUBLISHED 05 August 2025

CITATION
Stamm T, Bader-El-Den M, McNicholas J, Briggs J and Zhao P (2025) Applications of generative artificial intelligence in outcome prediction in intensive care medicine—a scoping review.
Front. Digit. Health 7:1633458.
doi: 10.3389/fdgth.2025.1633458

# Applications of generative artificial intelligence in outcome prediction in intensive care medicine—a scoping review

Tanja Stamm[1,2]*, Mohamed Bader-El-Den[2], James McNicholas[3], Jim Briggs[2] and Peng Zhao[2]

[1]Institute of Outcomes Research, Centre for Medical Data Science, Medical University of Vienna, Vienna, Austria, [2]Portsmouth AI and Data Science Centre, Faculty of Technology, University of Portsmouth, Portsmouth, United Kingdom, [3]Department of Critical Care, Queen Alexandra Hospital, Portsmouth, United Kingdom

When a patient survives the first 24 h in intensive care, outcome prediction is crucial for further treatment decisions. As recent advances have shown that Artificial Intelligence (AI) outperforms clinicians in prognostication, and especially generative AI has developed rapidly in the past ten years, this scoping review aimed to explore the use of generative AI models for outcome prediction in intensive care medicine. Of the 481 records found in the search, 119 studies were subjected to abstract screening and, when necessary, full-text review for eligibility assessment. Twenty-two studies and two review articles were finally included. The studies were categorized into three prototypical use cases for generative AI in outcome prediction in intensive care: (i) data augmentation, (ii) feature generation from unstructured data, and (iii) prediction by the generative model. In the first two use cases, the generative models worked together with downstream predictive models. In the third use case, the generative models made the predictions themselves. The studies within data augmentation either fell into the area of compensation for class imbalances by producing additional synthetic cases or imputation of missing values. Overall, Generative Adversarial Network (GAN) was the most frequently used technology (8/22 studies; 36%), followed by Generative Pretrained Transformer (GPT) (7/22 studies; 32%). All publications except one were from the last four years. This review shows that generative AI has immense potential in the future, and continuous monitoring of new technologies is necessary to ensure that patients receive the best possible care.

KEYWORDS

large language model, generative adversarial network, critical care, survival, mortality, comorbidity

## 1 Introduction

The first medical decision in critical care is whether or not to admit a patient to the Intensive Care Unit (ICU). If this decision is positive and the patient survives the first 24 h, outcome prediction is essential for future treatment over the next few days (1). At this stage, the outcomes of interest include not only short- and long-term survival in and out of the hospital, but also organ or body functioning, the occurrence of new comorbidities and symptoms, quality of life, and the ability to master everyday activities after hospital discharge (2). The latter is vital for patients and their caregivers.

Until now, mainly clinical scores were used to estimate the probabilities of certain outcomes in intensive care. However, in recent years, several studies have shown that prognostications based on Artificial Intelligence (AI) deliver better results than clinical scores (3–5).

Previous AI studies used primarily predictive models to forecast outcomes. However, even state-of-the-art predictive AI methods, such as gradient-boosted trees, are still far from perfect outcome prediction in intensive care (6). With the increasing development of generative models and the more extensive availability of the necessary computing power and large enough datasets, the question arises as to whether generative models or a combination of predictive and generative methods could significantly improve prognostication in intensive care medicine (7). Generative AI comprises data synthesis models, such as Generative Adversarial Networks (GANs) (8), Autoencoders, Variational Autoencoders (VAEs) (9), diffusion models (10), and transformer-based Large Language Models (LLMs) (11).

Two recent reviews on generative AI and outcome prediction in intensive care were published. However, one limited its scope to LLMs (12) and the other focused specifically on critical care nursing (13). No review described the general tasks and applications of generative AI in predicting outcomes in intensive care medicine. Therefore, this was defined as the objective of the present study.

# 2 Methods

## 2.1 Study design and search strategy

A scoping review (14) was conducted due to the exploratory nature of this study and our aim to provide an overview of the diverse use of generative AI in predicting outcomes in ICU, rather than synthesizing all relevant empirical evidence to answer a specific research question, which would be better suited to a systematic review.

Our goal was to provide information on the use of generative models to predict outcomes in intensive care. The search strategy was based on a two-level keyword tree (15) covering the areas of intensive care, outcome prediction, generative AI in general, and specific generative AI applications (Figure 1). The search strings were adapted to the respective databases and ran between February 28 and March 17, 2025, in IEEE Xplore, PubMed, CINAHL, Google Scholar, and arXiv (Supplementary Table S1). No limit for publication date was set, and only articles in English were included. We did not limit our scoping review to any level of technical progression. We also included studies dealing with unimodal and multimodal data, time series analyses, as well as studies that used images or medical text. Genetic studies were excluded due to the rare availability of these data in electronic health records. Although outcome predictions in intensive care came most often from tabular data, we did not exclude studies that used images or text. Furthermore, all study designs were included, except for commentaries and opinion letters.

## 2.2 Study selection

Duplicates were removed. The first author (TS) screened the titles and abstracts for eligibility. Full-text records were consulted when necessary. A second automated abstract screening was conducted for quality control using ChatGPT's freely accessible GPT-3.5 model. The prompt used for this computerized eligibility assessment was as follows: "*You are a critical researcher. Below is an abstract of a study. Please tell me if you would include this in a literature review. This review aims to*

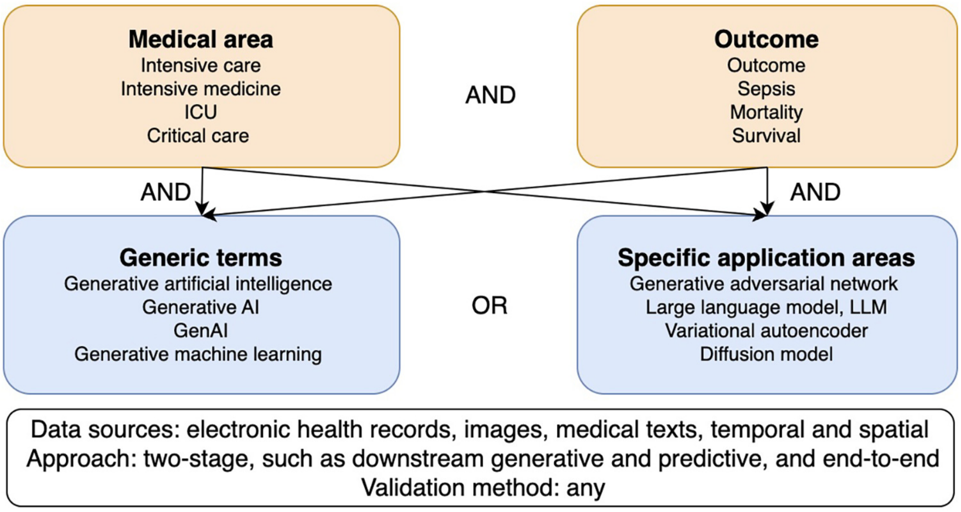
Medical area
Intensive care
Intensive medicine
ICU
Critical care
AND
Outcome
Outcome
Sepsis
Mortality
Survival
AND
AND
Generic terms
Generative artificial intelligence
Generative AI
GenAI
Generative machine learning
OR
Specific application areas
Generative adversarial network
Large language model, LLM
Variational autoencoder
Diffusion model
Data sources: electronic health records, images, medical texts, temporal and spatial
Approach: two-stage, such as downstream generative and predictive, and end-to-end
Validation method: any

FIGURE 1
Two-level keyword tree and research field classification diagram.

*describe the applications of generative artificial intelligence in outcome prediction in critical care. Only studies that used generative AI models and focused on outcome prediction in critical care must be included.*" In case of disagreement between the human and machine assessments, ChatGPT's arguments were discussed in a study team meeting, and a joint decision was made. Human judgment was prioritized.

## 2.3 Data extraction and reporting

Full texts of suitable publications were obtained. The publication year, the name of the journal or conference, the aim and outcome domains of the research, the dataset, and the generative models were extracted from the full text records using a custom data extraction sheet. There was no overlap across all three use cases in any one study. We used the Prediction model Risk Of Bias ASsessment Tool (PROBAST) +AI (16) to assess the quality of the studies in terms of the selection of participants and data sources, predictors, outcomes and analysis. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (17) were followed when presenting the results.

# 3 Results

Of the 481 records found in the search, 119 studies were subjected to abstract screening and, where necessary, also full-text review. Twenty-two studies and two review articles were finally included in the present scoping review (Figure 2; Tables 1, 2). The

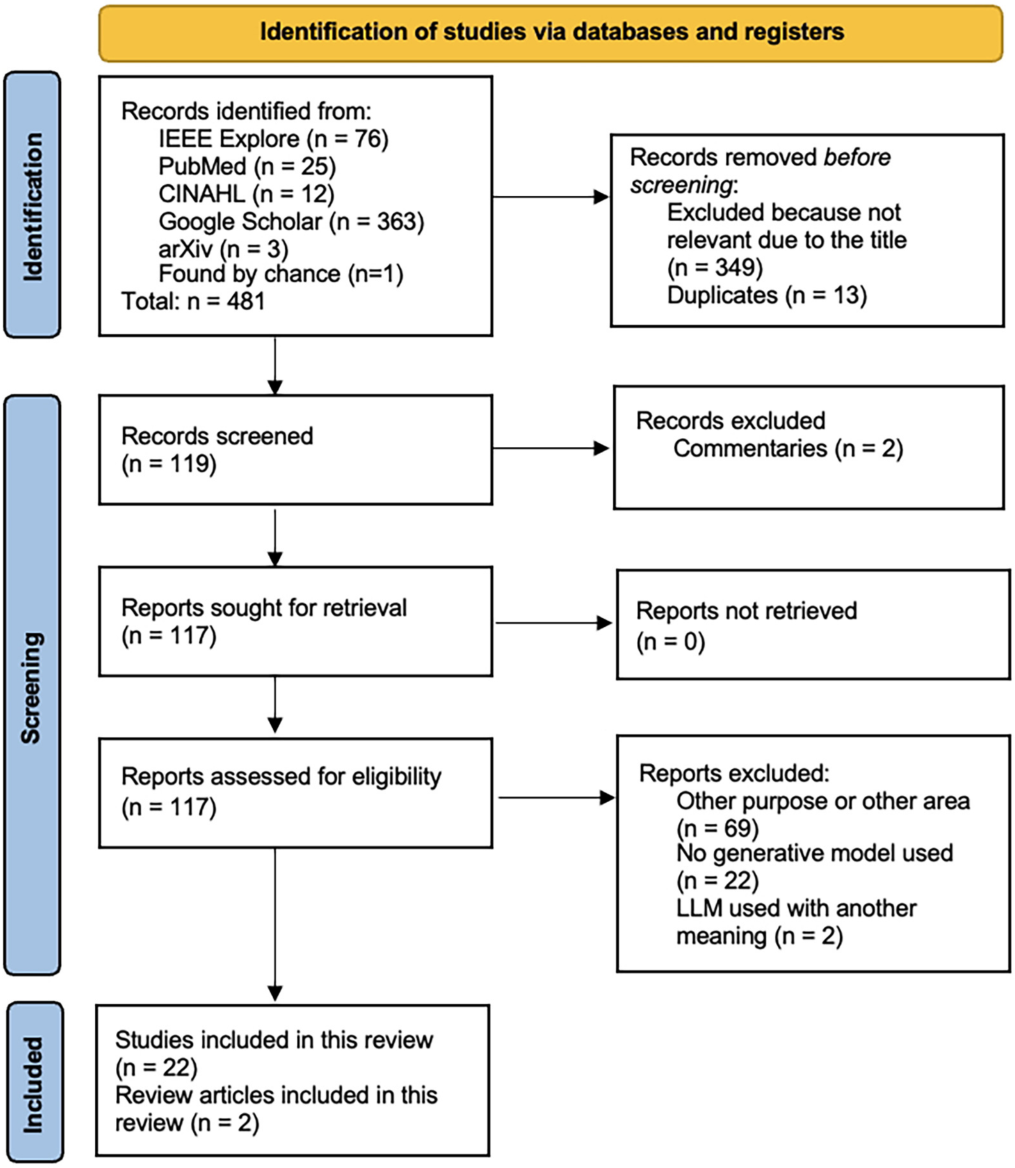
Identification of studies via databases and registers
Identification
Records identified from:
IEEE Explore (n = 76)
PubMed (n = 25)
CINAHL (n = 12)
Google Scholar (n = 363)
arXiv (n = 3)
Found by chance (n=1)
Total: n = 481
Records removed *before screening*:
Excluded because not relevant due to the title (n = 349)
Duplicates (n = 13)
Screening
Records screened (n = 119)
Records excluded
Commentaries (n = 2)
Reports sought for retrieval (n = 117)
Reports not retrieved (n = 0)
Reports assessed for eligibility (n = 117)
Reports excluded:
Other purpose or other area (n = 69)
No generative model used (n = 22)
LLM used with another meaning (n = 2)
Included
Studies included in this review (n = 22)
Review articles included in this review (n = 2)

FIGURE 2
PRISMA Flow Chart for study selection and assessment.

TABLE 1 Details of the 21 studies included in this review.

| # | Authors | Year | Journal or conference | Aim | Dataset | Generative model |
|---|---|---|---|---|---|---|
| 1 | Wei et al. (19) | 2021 | IEEE Access | In-ICU mortality prediction | MIMIC-IV | GAN |
| 2 | Yang et al. (20) | 2023 | Heliyon | In-ICU mortality prediction | MIMIC-III | Variant of GAN (c-med GAN) |
| 3 | Shariat et al. (21) | 2024 | International Conference on Web Research (ICWR) | Prediction of neonatal infections | 113,378 neonates admitted in the year 2022 | GAN |
| 4 | Wang et al. (22) | 2024 | Inter. Conf. on Biomed. Engineering and Applications (ICBEA) | Acute pain prediction | UNBC-McMaster shoulder pain dataset | GAN |
| 5 | Ravikumar et al. (23) | 2024 | IEEE Access | Skin infection prediction | HAM10000, ISIC 2018 challenget | GAN |
| 6 | Ryan et al. (24) | 2013 | Biomedical Sciences and Engineering Conference (BSEC) | In-ICU mortality prediction | Physionet/CinC 2012 Challenge data | Deep Boltzmann machine |
| 7 | Apalak and Kiasaleh (25) | 2022 | IEEE Access | Sepsis prediction | 2019 PhysioNet Computing in Cardiology Challenge dataset | Recurrent conditional GAN |
| 8 | Kim et al. (26) | 2020 | Intern. Conf. on Pattern Recognition (ICPR) | In-ICU mortality prediction | Physionet Challenge 2012 – 4,000 ICU stays with 80.5% missings) | GAN |
| 9 | Zhang et al. (18) | 2023 | International Conference on Tools with Artificial Intelligence (ICTAI) | In-ICU mortality prediction | MIMIC-III | MedCT-BERT |
| 10 | Mesinovic et al. (27) | 2024 | Journal of the American Medical Informatics Association | Survival analysis | MIMIC-IV | Conditional VAE |
| 11 | Ramos et al. (28) | 2021 | Ann. Inter. Conf. of the IEEE Engin. in Med. and Bio. Soc. (EMBC) | Sepsis prediction | MIMIC-III | VAE |
| 12 | Rao et al. (29) | 2024 | IEEE Int. Conf. on Industry 4.0, AI and Comm. Tech. (IAICT) | Anomaly detection | MIMIC-III/IV | GAN |
| 13 | Vurgun et al. (30) | 2024 | Advan. in Med. Found. Mod.: Explainab., Robustn., Secur., a. B. | Cardiac arrest identification | Data from the Hospital of the University of Pennsylvania | GPT-4 and 51 open-source LLMs |
| 14 | Pathak et al. (31) | 2024 | IEEE Journal of Biomedical and Health Informatics | Acute respiratory distress syndrome identification | Data from two hospital in Altanta | NLP Pipeline with BERT model |
| 15 | Madden et al. (32) | 2023 | Intensive care medicine | Creation of patient summaries | Two sets of medical notes | GPT-4 |
| 16 | Lin et al. (33) | 2025 | Journal of the American Medical Informatics Association | In-ICU mortality prediction | MIMIC-IV | BERT |
| 17 | Pabon et al. (34) | 2024 | European Journal of Heart Failure | Feature extraction | 6,263 patients enrolled in the DELIVER trial | GPT-3.5 |
| 18 | Parizad et al. (35) | 2024 | Intern. Conf. on Soft Comput. and Mach. Intell. (ISCMI) | Prediction of hospital readmission | MIMIC-III | ChatGPT |
| 19 | Chung et al. (36) | 2024 | JAMA surgery | Prediction of perioperative risks and prognosis | Retrosp. collected data from electronic health records | GPT-4 Turbo |
| 20 | Amacher et al. (37) | 2024 | Resuscitation Plus | Prediction of outcomes after cardiac arrest | Data of Swiss cardiac arrest patients admitted to ICU | ChatGPT-4 |
| 21 | Yoon et al. (38) | 2025 | Journal of the American Medical Informatics Association | Prediction of 30-day out-of-hospital mortality | MIMIC-IV | GPT-4 |
| 22 | Contreras et al. (39) | 2024 | arXiv | Prediction of delirium in the ICU | Three ICU datasets: eICU, MIMIC, and UFH | New LLM-based model (DeLLiriuM) |

TABLE 2 Details of the 2 review papers included in this review.

| # | Authors | Year | Journal or conference | Scope | Number of papers included | Design |
|---|---|---|---|---|---|---|
| 1 | Shi et al. (12) | 2024 | arXiv | LLMs in critical care | 24 | Scoping review |
| 2 | Porcellato et al. (13) | 2025 | Nursing reports | Generative AI in critical care nursing | 24 | Systematic review |

papers excluded during abstract screening and full text review are shown in Supplementary Table S2.

The studies were categorized into three prototypical use cases for the use of generative AI models in the prediction of outcomes in intensive care medicine: (i) data augmentation, (ii) feature generation from unstructured data, and (iii) prediction by the generative model. In the first two use cases, the generative models worked together with downstream predictive models that performed the prediction. In the third use case, the generative models made the predictions themselves (Figure 3). While studies predominantly fitted one use case, some integrated multiple generative approaches [e.g., (18)].

## 3.1 Use case 1 – data augmentation

As expected, generative AI was used to compensate for class imbalances by producing additional synthetic data to improve the performance of downstream classical predictive models. Two studies (19, 20) focused on in-ICU mortality prediction and used

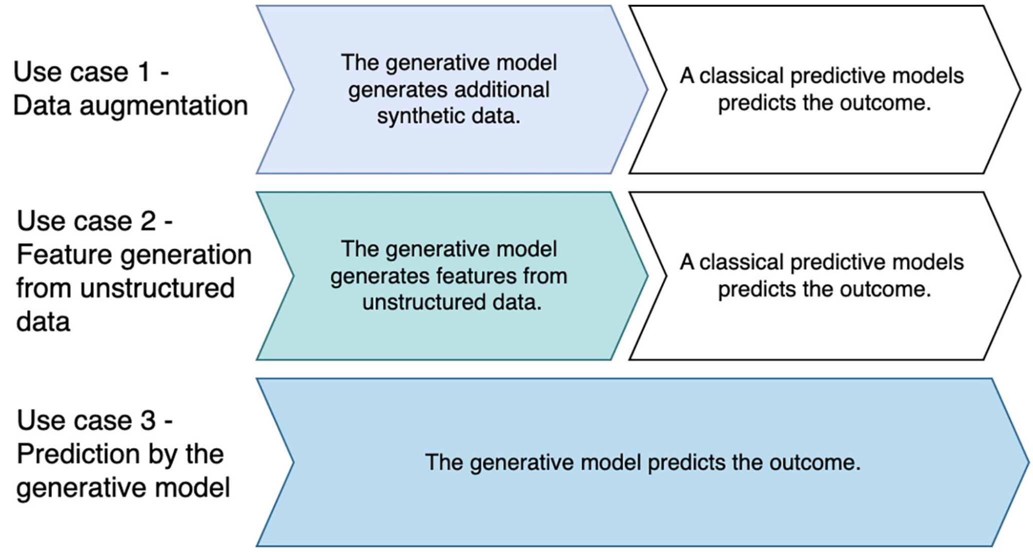
Use case 1 - Data augmentation
The generative model generates additional synthetic data.
A classical predictive models predicts the outcome.
Use case 2 - Feature generation from unstructured data
The generative model generates features from unstructured data.
A classical predictive models predicts the outcome.
Use case 3 - Prediction by the generative model
The generative model predicts the outcome.

FIGURE 3
Prototypical use cases for the use of generative AI in outcome prediction in intensive care medicine.

GAN to supplement data in the minority class. In the first study, Wei et al. (19) applied GAN for tabular data augmentation to improve the performance of a downstream extreme gradient boost tree (XGBoost), leading to an Area Under the Receiver Operating Curve (AUROC) of greater than 0.90. Additional synthetic cases were needed in this study because the number of deceased patients in the original dataset was considerably lower than that of survivors. Moreover, a classifier layer was added to the GAN that reduced falsely non-binary GAN-generated values to binary ones. The missing values were imputed by filling in mean values or the majority subtype for the categorical features. Finally, age and blood urea nitrogen (BUN) contributed the most to the XGBoost model, shown in a Shapley additive explanation (SHAP) analysis. From a medical perspective, the results are meaningful, as age and kidney function play an important role in survival. In the second study on mortality prediction in critical care, Yang et al. (20) developed and evaluated the conditional medical GAN (c-med GAN) to improve the synthesis of discrete parameters consisting of an additional autoencoder that generated the new data separately from their labels and a functionality that added the class labels as constraints. The autoencoder network was used because GAN generally had difficulties in synthesizing discrete data that are often included in electronic health records to a satisfactory quality. Similar to the first study, the AUROC of the model that included the c-med GAN was >0.90. Shariat et al. (21) used imbalanced tabular data to predict neonatal infections and applied a simple GAN to synthesize additional data in the minority class to improve the performance of downstream predictive models, including Random Forest (RF), Support Vector Machine (SVM), Bagging, and Boosting. RF and Bagging outperformed the other models achieving an F1-score of >0.95.

Two further studies utilized GAN to augment image datasets. In the first study, Wang et al. (22) used synthetic, GAN-created pictures of facial expressions of ventilated ICU patients whose faces were often obscured by intubation or masks to predict acute pain. However, while a tool to evaluate pain would be useful both for patient experience and potential avoidance of some physical complications of critical illness, sedation regimes may include deep muscle relaxation, and thus, facial movement could be attenuated or abolished altogether. In the second study, Ravikumar et al. (23) reviewed approaches to create synthetic images of skin lesions and tested a conditional GAN that, together with a downstream DenseNet-201, outperformed Visual Geometry Group 16 and Support Vector Machine in predicting skin infections, achieving an accuracy of approximately 82%. While this approach could indicate specific symptoms, it would probably contribute little to the prediction of the general state of health of critically ill patients.

Another task of generative models in the area of data augmentation was imputation. A major problem with ICU data is the frequently large proportion of missing measured values. The first of the studies in this area and, at the same time, the oldest of the ones included in this entire review, used a generative model, a deep Boltzmann machine, to learn the distribution of time series data and impute missing values to improve the performance of a downstream neural network to predict in-hospital mortality in ICU patients (24). Apalak and Kiasaleh (25) applied a conditional GAN where the generator and discriminator were Long Short-Term Memory networks (LSTMs) to impute missing tabular time series data. They showed that this approach improved the performance of another downstream predictive LSTM in forecasting sepsis, leading to AUROC values between 0.93 and 0.94. LSTM could be a promising approach, as it can process the sequential medical measurement data. Kim et al. (26) used GAN to impute missing values by applying a slightly different approach. By randomly dropping a certain percentage $\alpha$ of values from an original multivariate time series input dataset $\mathcal{X}$, they constructed a corrupted dataset $\bar{\mathcal{X}}$ while retaining the original labels [Equation 1; (26)]:

$$\bar{\mathcal{X}} \sim \ Drop_{\alpha}(\bar{\mathcal{X}}/\mathcal{X}) \tag{1}$$

$\bar{\mathcal{X}}$ was then used as input data instead of $\mathcal{X}$, and the generator replaced the dropped values with the average values in each iteration, thus reducing noise during imputation. Kim et al. (26) showed that the synthesized dataset worked similarly in a downstream predictive model to forecast in-ICU mortality compared to the original data. They measured imputation performance using mean squared error and mean absolute error and achieved the highest values with their new approach of 0.48 and 0.37, respectively. Zhang et al. (18) proposed a multimodal learning model for in-ICU mortality prediction in critical care. They first used a Bidirectional Encoder Representations from Transformers (BERT) model to generate text embeddings of the medical records of each patient. This created missing values due to irregularities in intervals in time series data. Therefore, the authors adapted a GAN to interpolate the time series data and used feature correlations in the original dataset to assess the appropriateness of the interpolated data. This new approach, called MedCT-BERT outperformed several other generative and predictive approaches, achieving an AUROC of 0.89, including BioBERT. The textual embeddings of the data could capture nuances that plain structured data might miss.

Two further studies used VAE to augment tabular data. Mesinovic et al. (27) imputed serial lab value data from the ICU using a conditional VAE and combined this model with a right-censored time-to-event survival analysis instead of just a discrete outcome label. The conditional VAE was used in this study to learn the latent representation of the input dataset. This approach led to a mean AUROC > 0.67. In another study, Ramos et al. (28) first imputed missing data with a forward-filling strategy, then applied VAE to learn the latent representation of the input data and, lastly, used unsupervised clustering to detect rare abnormal events and, thus, predicted septic shocks; this model showed a comparable performance as a supervised LSTM network, achieving an AUROC of 0.82. This study differs from many others because no labeled dataset was used. Overall, it proposed a potentially useful tool to prompt the initiation of preventive measures, such as the prescription of antibiotics.

Rao et al. (29) attempted to filter out abnormal values in physiological parameters that could lead to false alarms during data synthesis. The authors used distance-related anomaly scores calculated at the generator and the discriminator of the GAN, leading to an accuracy of 0.97. The new model outperformed a Convolutional Neural Network-autoencoder-based anomaly detection model. From a medical perspective, this new method could have the potential to detect risks of events other than sepsis and shock onset, for example, sudden cardiac arrest.

## 3.2 Use case 2 – feature generation using unstructured data

In this use case, generative AI was used to create new features based on unstructured textual data for a downstream predictive model. Vurgun et al. (30) tested 51 open-source LLMs against GPT-4.0 to extract in-hospital cardiac arrest events using discharge summaries, progress notes, and tabular data, with several other open-source models demonstrating competitive results, such as Mistral-Nemo-Instruct-2407. The highest AUROCs achieved were between 0.91 and 0.90. However, this was not a prediction of a future event, as the cardiac arrest had already been determined by the care team. Pathak et al. (31) used BERT to classify radiology reports to predict acute respiratory distress syndrome (ARDS). However, while technically novel, applications like ARDS identification from radiology reports may offer limited added value if clinicians already documented these diagnoses. Maden et al. (32) used GPT-4 to create patient summaries from daily free-text medical notes, thereby making this information accessible for critical care decisions and outcome prediction. The study concluded that the clarity of the prompts determined the quality of the summaries. The highest AUROCs were between 0.75 and 0.88. Moreover, writing patient summaries would tie up considerable resources and could therefore be usefully replaced by AI. Lin et al. (33) predicted in-hospital mortality based on radiology reports, chest x-rays, and clinical ICU data. Convolutional neural network processed the image data. The token embeddings from the last layer of a BERT model were used as latent representations of the radiology reports. Feature fusion from all three data sources, clinical data, chest x-rays and radiology reports, led to a slightly better AUROC than only one or two feature sources alone (+0.01). Pabon et al. (34) applied GPT-3.5 to extract information on the left ventricular ejection fraction from medical records. However, tabular data were used in this study for data extraction and not the unstructured text.

Parizad et al. (35) used ChatGPT in a different way than the studies described above. The authors of this study first asked ChatGPT for advice on which features to include in a frailty index, but then extracted the actual data from the clinical notes using non-generative natural language processing techniques.

## 3.3 Use case 3 – prediction by the generative model

Generative models were also used to forecast outcomes in intensive care medicine. GPT-4 Turbo showed an acceptable performance in predicting mortality (F1 = 0.86) and clinical scores of the American Society of Anesthesiologists (F1 = 0.50) using medical notes on the instructions. However, temporal predictions (e.g., ICU stay duration) performed poorly (MAE = 1.1 days) (36). While the two F1 scores were 76% and 194% better than a random classifier at baseline, the mean absolute error of 1.1 days was the same as the dummy regressor baseline.

Moreover, a comparison with the actual decision of an intensive care physician would have been desirable here, too. Amacher et al. (37) asked ChatGPT-4 to predict the occurrence of poor neurological outcome and the likelihood of survival of cardiac arrest patients from a Swiss ICU dataset, but used, in contrast to the previous study, the tabular data in the prompts. In Amacher's study, ChatGPT-4 showed only a similar performance (AUROC = 0.85) as clinical scores derived from health professionals (AUROC = 0.83). Furthermore, Yoon et al. (38) tested LLMs tuned with instructions to predict mortality from discharge notes at 30 days after the patients had left the hospital; in this study, GPT-4 showed the best result (32.2% in F1 metrics compared to 28.9% for best-performing supervised model). Contreras et al. (39) trained a novel LLM-based delirium prediction model using electronic health record data from the first 24 h of ICU admission from three openly accessible databases. The new model used a clinical 345 million-parameter LLM (GatorTronS) as the backbone and performed better (AUROC ranging being 0.77 and 0.82 in two external validation datasets) than three other deep learning models, namely a Neural Network, a Transformer model, and Mamba. The features identified in the SHAP analysis were consistent with the usual accepted risk factors for delirium, and only urine specific gravity was unexpected and new.

Furthermore, the two narrative reviews (12, 13) outlined several clinical areas where LLMs and other AI methods could show their advantages, such as the integration of multimodal and unstructured data, the creation of patient summaries and the prediction and prognostication, the second review specifically highlighting how these applications could support critical care nursing.

## 3.4 Technology used and bibliometric findings

GAN was the most frequently used technology (8/22 studies; 36%), followed by GPT (7/21 studies; 32%). All publications except one were from the last four years; the oldest was published in 2013. Medical Information Mart for Intensive Care (MIMIC) in different versions was the most frequently used dataset (10/22 studies; 45%). No study predicted long-term outcomes in intensive care medicine using generative AI.

## 3.5 Risk of bias assessment

While the predictors, outcome and analysis were described transparently in most studies (Table 3), information regarding the selection of the study participants and datasets was often lacking (22, 27, 31). Moreover, pragmatically defined sample sizes (26) and the use of commercial models on a smaller scale, due to cost constraints, compared to the unrestricted use of freely accessible models (30), could have introduced additional bias.

TABLE 3 Risk of bias assessment according to Prediction model Risk Of Bias ASsessment Tool (PROBAST) +AI (16).

| # | Authors | Year | Risk of bias introduced by the | | | |
|---|---|---|---|---|---|---|
| | | | Selection of participants and data sources | Predictors or their assessment | Outcome or its determination | Analysis |
| 1 | Wei et al. (19) | 2021 | ? | + | + | + |
| 2 | Yang et al. (20) | 2023 | ? | + | + | + |
| 3 | Shariat et al. (21) | 2024 | ? | + | + | + |
| 4 | Wang et al. (22) | 2024 | – | + | + | + |
| 5 | Ravikumar et al. (23) | 2024 | + | + | + | + |
| 6 | Ryan et al. (24) | 2013 | + | + | + | + |
| 7 | Apalak and Kiasaleh (25) | 2022 | + | + | + | + |
| 8 | Kim et al. (26) | 2020 | ? | + | + | + |
| 9 | Zhang et al. (18) | 2023 | ? | + | + | + |
| 10 | Mesinovic et al. (27) | 2024 | – | + | + | + |
| 11 | Ramos et al. (28) | 2021 | + | + | + | + |
| 12 | Rao et al. (29) | 2024 | – | + | + | + |
| 13 | Vurgun et al. (30) | 2024 | ? | ? | + | + |
| 14 | Pathak et al. (31) | 2024 | ? | + | + | + |
| 15 | Madden et al. (32) | 2023 | – | + | ? | + |
| 16 | Lin et al. (33) | 2025 | + | + | + | + |
| 17 | Pabon et al. (34) | 2024 | + | ? | + | ? |
| 18 | Parizad et al. (35) | 2024 | ? | + | + | + |
| 19 | Chung et al. (36) | 2024 | ? | + | + | + |
| 20 | Amacher et al. (37) | 2024 | + | + | + | + |
| 21 | Yoon et al. (38) | 2025 | + | + | + | + |
| 22 | Contreras et al. (39) | 2024 | + | + | + | + |
| Review 1 | Shi et al. (12) | 2024 | NA | NA | NA | NA |
| Review 2 | Porcellato et al. (13) | 2025 | NA | NA | NA | NA |

"+" refers to a low; "–" to high risk of bias; "?" indicates unclear information.

## 4 Discussion

Although generative models have been used in many areas, primarily for images and text, our results clearly show their value in tabular data from electronic health records in outcome prognostication in intensive care medicine. The recent publication dates of the studies highlight the topicality of this research field and the immense innovation potential in using new technologies for unusual tasks, such as making a generative model take over prediction tasks on its own. However, generative AI research still remains narrowly focused on short-term mortality. Future work should target patient-centered long-term outcomes.

Although some technologies have rarely been used in outcome prediction in intensive care medicine, they could be adapted and applied to this field even more in the future. Examples are Retrieval Augmented Generation (RAG) or diffusion models. A study, for example, using RAG to retrieve information from medical reports, was excluded from this review because it did not focus on predicting outcomes in intensive care medicine (40). Likewise, another study applied a novel combination of a diffusion model with an upstream autoencoder block to forecast time series data on heart rate and blood pressure (41), but did not forecast outcomes in critical care either.

Generative AI further showed that it had the potential to enrich tabular data with additional information from different sources. In the reviewed studies, the source of information was mainly medical notes. Images were used less frequently and only in special contexts, e.g., to predict facial expressions of pain or skin lesions. A special category of studies did not use the generative model to extract features from medical notes, but rather asked the generative model, such as ChatGPT for features to extract from medical notes (35). Similarly, other studies also used ChatGPT to obtain medical advice (42). However, there are controversial views on the quality and evidence-based nature of the recommendations derived from ChatGPT (43, 44). Unrelated work (45) suggested potential options for psychosocial feature extraction; however, the actual feature extraction was performed by experts in this study (45).

An interesting distinction is whether the approaches in the studies are two-stage approaches or end-to-end approaches. All studies that we classified as use cases 1 and 2 contain two-stage approaches per se, as the generative model was used in the first case to supplement incomplete or infrequently available data or to generate new parameters from unstructured data. For this purpose, a generative model was always used first, followed by a predictive model. Only in the third use case, the predictive model made the prediction itself (end-to-end approach).

Denoising referred to irregularly sampled time series values, abnormally imputed values, and simple errors (26). Denoising autoencoders were originally designed to prevent the output sequence of the encoder from being equal to the input data as this would make the autoencoder obsolete. The denoising autoencoder would, therefore, use noisy or corrupted input for the decoding, but calculate the encoder loss based on the original input data.

In some studies, although the data science method was new and innovative, the medical aim was questionable. Especially when diagnoses could be made easily or, for example, a chest X-ray must have been viewed by a specialist anyway, as not only was a diagnosis made by that, but other parameters were also assessed. However, it is still possible to use the experiments to further develop the technology or contribute to the wider availability of medical knowledge beyond that of experts. This could be relevant for the automated creation of summaries, but also for doctors in training or when certain experts are unavailable.

Generative models are computationally resource-intensive. This might create additional data protection issues, and secure processing environments would be needed to analyse patient data. To avoid such problems, the publicly accessible MIMIC dataset was probably the most often used database in the studies included. However, perhaps the results should also be validated in different datasets in the future. In addition, generative models can help solve data protection problems. Completing missing data can be extended to a data synthesis problem, and data synthesised in sufficient quality will be accepted more and more as a replacement of the original data. Since data synthesis is a generic task, methods from other fields can also be applied to predict outcomes based on ICU data. For example, Neves et al. developed a GAN with only two layers and a hyperbolic tangent activation function in the output layer for the imputation of medical data, which could synthesize data faster (46).

The clinical implications of using generative AI in intensive care units extend beyond outcome prediction, which was the scope of this review. At present, we are still a long way from automated treatment. Beyond liability issues, ethical considerations, bias, hallucinations, the lack of necessary qualifications in healthcare professionals, and potential changes to work processes, medicine also involves an element of humanity that AI cannot yet easily replace. However, it has already been shown that machines can not only provide effective decision support in treatment, but also engage in more empathetic dialogues than clinicians when interacting with patients and their relatives (47). Nevertheless, until now, the final decision on treatment has always been made by a human, albeit supported by AI.

Ethical implications of using generative AI in real-time critical care settings, including hallucinations, bias, accountability, and data privacy, as well as regulatory issues should be addressed through future research, which should also take into account multi-stakeholder perspectives. Real-time deployment of generative models [e.g., ChatGPT-4 (37)] requires rigorous hallucination safeguards, especially when tabular data inputs may propagate biases.

Practical implementation of generative AI in intensive care medicine will depend heavily on its effectiveness, regulatory approval, technical interoperability with electronic health records and other systems, the skills of health professionals and hospital policies related to patient pathways and digital medicine. The detailed discussion of these aspects is beyond the scope of this review. Further research should focus on addressing them.

Interpretability and explainability of generative AI are crucial for its acceptance, adoption and deployment in clinical practice (48). Further research should address this area, incorporating the perspectives of clinicians, patients and their caregivers.

A limitation of this review might be its static nature, while the technology is advancing at a high speed. Novel technologies and new application areas might be published during or after this review.

Work on the first update of this review is therefore scheduled to begin one year after the publication date. Following these arguments, more studies could also have been expected in this review. However, in a review of scoping reviews, Tricco et al. (48) showed that the average number of studies included in 494 scoping reviews was 118, ranging from 1 to 2800. This is comparable to the 119 records that were assessed in detail in the present study. Moreover, the scoping review on LLMs in intensive care (but not limited to outcome prediction) also included in this review (13) ended with a similar number of 24 articles as our final selection. In addition, as also done in the aforementioned paper, we limited the keyword search to titles and abstracts in some databases, as we were specifically looking for studies that used generative methods in outcome prediction and did not just mention generative AI, for example, in the discussion. A further limitation of our study could be that preprint studies without peer review were also considered. However, we were transparent about this, and the publication origin of the studies is indicated accordingly in Tables 1, 2. Another limitation of our study is that the second abstract screening was conducted using ChatGPT's 3.5 model. Inclusion of a second human review could have enhanced the methodological rigor of this process. This review shows that generative AI has immense potential in the future, and continuous monitoring of new technologies is necessary to ensure that patients receive the best possible care.

## Author contributions

TS: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Visualization, Writing – original draft, Writing – review & editing. MB-E-D: Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing. JM: Conceptualization, Formal analysis, Supervision, Writing – original draft, Writing – review & editing. JB: Conceptualization, Supervision, Writing – review & editing. PZ: Writing – original draft, Writing – review & editing.

## Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

## Generative AI statement

The author(s) declare that Generative AI was used in the creation of this manuscript. Artificial Intelligence (AI) was not used to write text in this manuscript; only a few phrases were translated with the online version of DeepL (https://www.deepl.com/de/translator) from German to English to check their exact meaning. In addition to the abstract screening by the first author (TS), a second automated abstract screening was conducted using ChatGPT's freely accessible GPT-3.5 model. This is described in the manuscript, and the prompt used for this computerized eligibility assessment is displayed in Supplementary Table S2.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdgth.2025.1633458/full#supplementary-material

## References

1. Masud FN, Sasangohar F, Ratnani I, Fatima S, Hernandez MA, Riley T, et al. Past, present, and future of sustainable intensive care: narrative review and a large hospital system experience. *Crit Care.* (2024) 28:154. doi: 10.1186/s13054-024-04937-9

2. Haines KJ, Hibbert E, McPeake J, Anderson BJ, Bienvenu OJ, Andrews A, et al. Prediction models for physical, cognitive, and mental health impairments after critical illness: a systematic review and critical appraisal. *Crit Care Med.* (2020) 48:1871–80. doi: 10.1097/CCM.0000000000004659

3. Awad A, Bader-El-Den M, McNicholas J, Briggs J. Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach. *Int J Med Inform.* (2017) 108:185–95. doi: 10.1016/j.ijmedinf.2017.10.002

4. 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. doi: 10.1177/1460458219850323

5. Iwase S, Nakada T, Shimada T, Oami T, Shimazui T, Takahashi N, et al. Prediction algorithm for ICU mortality and length of stay using machine learning. *Sci Rep.* (2022) 12:12912. doi: 10.1038/s41598-022-17091-5

6. Ibrahim M, Al Khalil Y, Amirrajab S, Sun C, Breeuwer M, Pluim J, et al. Generative AI for synthetic data across multiple medical modalities: a systematic review of recent developments and challenges. *Comput Biol Med.* (2025) 189:109834. doi: 10.1016/j.compbiomed.2025.109834

7. Fang X, Xu W, Tan FA, Zhang J, Hu Z, Qi Y, et al. Large language models (LLMs) on tabular data: prediction, generation, and understanding–a survey. *arXiv* [Preprint]. *arXiv:2402.17944* (2024).

8. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. *arXiv* [Preprint]. *arXiv:1406.2661 [stat.ML]* (2014).

9. Kingma DP, Welling M. Auto-encoding variational bayes. *arXiv* [Preprint]. *arXiv:1312.6114v11[stat.ML]* (2013). doi: 10.48550/arXiv.1312.6114

10. Song Y, Sohl-Dickstein J, Kingma DP, Kumar A, Ermon S, Poole B. Score-based generative modeling through stochastic differential equations. *arXiv* [Preprint]. *arXiv:2011.13456* (2020).

11. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. *arXiv* [Preprint]. *arXiv:1706.03762 [cs.CL]* (2017) 30. doi: 10.48550/arXiv.1706.03762

12. Shi T, Ma J, Yu Z, Xu H, Xiong M, Xiao M, et al. Stochastic parrots or ICU experts? Large language models in critical care medicine: a scoping review. *arXiv* [Preprint]. *arXiv:2407.19256* (2024).

13. Porcellato E, Lanera C, Ocagli H, Danielis M. Exploring applications of artificial intelligence in critical care nursing: a systematic review. *Nurs Rep*. (2025) 15:55. doi: 10.3390/nursrep15020055

14. Arksey H, O'malley L. Scoping studies: towards a methodological framework. *Int J Soc Res Methodol*. (2005) 8:19–32. doi: 10.1080/1364557032000119616

15. D'Amico S, Sauta E, Bersanelli M, Dall'Olio D, Sala C, Dall'Olio L, et al. Synthetic data generation by artificial intelligence to accelerate translational research and precision medicine in hematological malignancies. *Blood*. (2022) 140:9744–6. doi: 10.1182/blood-2022-168646

16. Moons KG, Damen JA, Kaul T, Hooft L, Navarro CA, Dhiman P, et al. Probast+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ*. (2025) 388:e082505. doi: 10.1136/bmj-2024-082505

17. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. Prisma 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ*. (2021) 372:n160. doi: 10.1136/bmj.n160

18. Zhang K, Niu K, Zhou Y, Tai W, Lu G. MedCT-BERT: multimodal mortality prediction using medical ConvTransformer-BERT model. In: *2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE (2023). p. 700–7.

19. Wei M, Huang Z, Yuan D, Yang L. Predicting ICU mortality based on generative adversarial nets and ensemble methods. *IEEE Access*. (2023) 11:76403–14. doi: 10.1109/ACCESS.2023.3296147

20. Yang W, Zou H, Wang M, Zhang Q, Li S, Liang H. Mortality prediction among ICU inpatients based on MIMIC-III database results from the conditional medical generative adversarial network. *Heliyon*. (2023) 9:e13200. doi: 10.1016/j.heliyon.2023.e13200

21. Shariat S, Kargari M, Shariat N, Valiollahi A, Alavi M. Prediction of neonatal infections using machine learning techniques. In: *2024 10th International Conference on Web Research (ICWR)*. (2024). p. 244–9.

22. Wang L, Wang Z, Xu A, Liu S. A generative adversarial network-based approach for facial pain assessment. In: *2024 8th International Conference on Biomedical Engineering and Applications (ICBEA)*. (2024). p. 44–9.

23. Ravikumar A, Sriraman H, Chadha C, Kumar Chattu V. Alleviation of health data poverty for skin lesions using ACGAN: systematic review. *IEEE Access*. (2024) 12:122702–23. doi: 10.1109/ACCESS.2024.3417176

24. Ryan DP, Daley BJ, Wong K, Zhao X. Prediction of ICU in-hospital mortality using a deep boltzmann machine and dropout neural net. In: *2013 Biomedical Sciences and Engineering Conference (BSEC)*. IEEE (2013). p. 1–4.

25. Apalak M, Kiasaleh K. Improving sepsis prediction performance using conditional recurrent adversarial networks. *IEEE Access*. (2022) 10:134466–76. doi: 10.1109/ACCESS.2022.3230324

26. Kim J, Kim T, Choi JH, Choo J. End-to-end multi-task learning of missing value imputation and forecasting in time-series data. In: *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE (2021). p. 8849–56.

27. Mesinovic M, Watkinson P, Zhu T. DySurv: dynamic deep learning model for survival analysis with conditional variational inference. *J Am Med Inform Assoc*. (2024):ocae271. doi: 10.1093/jamia/ocae271

28. Ramos G, Gjini E, Coelho L, Silveira M. Unsupervised learning approach for predicting sepsis onset in ICU patients. In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE (2021). p. 1916–9.

29. Rao VA, Rao R, Hota C. Anomaly detection in wireless body area networks using generative adversarial networks. In: *2024 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*. IEEE (2024). p. 60–5.

30. Vurgun U, Hwang S, Mowery DL. Reliability in AI-assisted critical care: assessing large language model robustness and instruction following for cardiac arrest identification. In: Advancements in medical foundation models: explainability, robustness, security, and beyond. (2024). Available online at: https://openreview.net/pdf?id=psOWQZbI6Z (Accessed January 07, 2025).

31. Pathak A, Marshall C, Davis C, Yang P, Kamaleswaran R. RespBERT: a multi-site validation of a natural language processing algorithm, of radiology notes to identify Acute Respiratory Distress Syndrome (ARDS). *IEEE J Biomed Health Inf* (2025) 29(2):1455–63. doi: 10.1109/JBHI.2024.3502575

32. Madden MG, McNicholas BA, Laffey JG. Assessing the usefulness of a large language model to query and summarize unstructured medical notes in intensive care. *Intensive Care Med*. (2023) 49:1018–20. doi: 10.1007/s00134-023-07128-2

33. Lin M, Wang S, Ding Y, Zhao L, Wang F, Peng Y. An empirical study of using radiology reports and images to improve intensive care unit mortality prediction. *JAMIA open*. (2024) 8:ooae137. doi: 10.1093/jamiaopen/ooae137

34. Pabon MA, Vaduganathan M, Claggett BL, Chatur S, Siqueira S, Marti-Castellote P, et al. In-hospital course of patients with heart failure with improved ejection fraction in the DELIVER trial. *Eur J Heart Fail*. (2024) 26:2532–40. doi: 10.1002/ejhf.3410

35. Parizad SH, Seifollahi S, Taheri S, Abdollahian M. A generative frailty index based explainable approach for hospital readmission prediction. In: *2024 11th International Conference on Soft Computing & Machine Intelligence (ISCMI)*. IEEE (2024). p. 163–7.

36. Chung P, Fong CT, Walters AM, Aghaeepour N, Yetisgen M, O'Reilly-Shah VN. Large language model capabilities in perioperative risk prediction and prognostication. *JAMA Surg*. (2024) 159:928–37. doi: 10.1001/jamasurg.2024.1621

37. Amacher SA, Arpagaus A, Sahmer C, Becker C, Gross S, Urben T, et al. Prediction of outcomes after cardiac arrest by a generative artificial intelligence model. *Resusc Plus*. (2024) 18:100587. doi: 10.1016/j.resplu.2024.100587

38. Yoon W, Chen S, Gao Y, Zhao Z, Dligach D, Bitterman DS, et al. LCD benchmark: long clinical document benchmark on mortality prediction for language models. *J Am Med Inform Assoc*. (2025) 32:285–95. doi: 10.1093/jamia/ocae287

39. Contreras M, Kapoor S, Zhang J, Davidson A, Ren Y, Guan Z, et al. DeLLiriuM: a large language model for delirium prediction in the ICU using structured EHR. *arXiv* [Preprint]. *arXiv:2410.17363* (2024).

40. Mahalakshmi M, Bharadwaj S, Bhuyan AN. A real-time medical report analysis and AI-powered diagnosis: a cloud-based solution for improved patient care. In: *2024 Second International Conference on Advances in Information Technology (ICAIT)*. IEEE (2024). Vol. 1. p. 1–6.

41. Chang P, Li H, Quan SF, Lu S, Wung SF, Roveda J, et al. A transformer-based diffusion probabilistic model for heart rate and blood pressure forecasting in intensive care unit. *Comput Methods Programs Biomed*. (2024) 246:108060. doi: 10.1016/j.cmpb.2024.108060

42. Kucukkaya A, Arikan E, Goktas P. Unlocking ChatGPT's potential and challenges in intensive care nursing education and practice: a systematic review with narrative synthesis. *Nurs Outlook*. (2024) 72:102287. doi: 10.1016/j.outlook.2024.102287

43. Haverkamp W, Tennenbaum J, Strodthoff N. Chatgpt fails the test of evidence-based medicine. *Eur Heart J Digit Health*. (2023) 4:366–7. doi: 10.1093/ehjdh/ztad043

44. Huang J, Yang DM, Rong R, Nezafati K, Treager C, Chi Z, et al. A critical assessment of using chatgpt for extracting structured data from clinical notes. *npj Digit Med*. (2024) 7:106. doi: 10.1038/s41746-024-01079-8

45. Noaeen M, Amini S, Bhasker S, Ghezelsefli Z, Ahmed A, Jafarinezhad O, et al. Unlocking the power of EHRs: harnessing unstructured data for machine learning-based outcome predictions. In: *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE (2023). p. 1–4.

46. Neves DT, Alves J, Naik MG, Proença AJ, Prasser F. From missing data imputation to data generation. *J Comput Sci*. (2022) 61:101640. doi: 10.1016/j.jocs.2022.101640

47. Chen D, Chauhan K, Parsa R, Liu ZA, Liu FF, Mak E, et al. Patient perceptions of empathy in physician and artificial intelligence chatbot responses to patient questions about cancer. *npj Digit Med*. (2025) 8:1–5. doi: 10.1038/s41746-025-01671-6

48. Tricco AC, Lillie E, Zarin W, O'brien K, Colquhoun H, Kastner M, et al. A scoping review on the conduct and reporting of scoping reviews. *BMC Med Res Methodol*. (2016) 16:1–10. doi: 10.1186/s12874-016-0116-4