*Intensive Care Med (2021) 47:750–760*
https://doi.org/10.1007/s00134-021-06446-7

SYSTEMATIC REVIEW

# Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit

Davy van de Sande[1], Michel E. van Genderen[1*], Joost Huiskens[2], Diederik Gommers[1] and Jasper van Bommel[1]

## Abstract

**Purpose:** Due to the increasing demand for intensive care unit (ICU) treatment, and to improve quality and efficiency of care, there is a need for adequate and efficient clinical decision-making. The advancement of artificial intelligence (AI) technologies has resulted in the development of prediction models, which might aid clinical decision-making. This systematic review seeks to give a contemporary overview of the current maturity of AI in the ICU, the research methods behind these studies, and the risk of bias in these studies.

**Methods:** A systematic search was conducted in Embase, Medline, Web of Science Core Collection and Cochrane Central Register of Controlled Trials databases to identify eligible studies. Studies using AI to analyze ICU data were considered eligible. Specifically, the study design, study aim, dataset size, level of validation, level of readiness, and the outcomes of clinical trials were extracted. Risk of bias in individual studies was evaluated by the Prediction model Risk Of Bias ASsessment Tool (PROBAST).

**Results:** Out of 6455 studies identified through literature search, 494 were included. The most common study design was retrospective [476 studies (96.4% of all studies)] followed by prospective observational [8 (1.6%)] and clinical [10 (2%)] trials. 378 (80.9%) retrospective studies were classified as high risk of bias. No studies were identified that reported on the outcome evaluation of an AI model integrated in routine clinical practice.

**Conclusion:** The vast majority of developed ICU-AI models remain within the testing and prototyping environment; only a handful were actually evaluated in clinical practice. A uniform and structured approach can support the development, safe delivery, and implementation of AI to determine clinical benefit in the ICU.

**Keywords:** Artificial intelligence, Machine learning, Intensive care unit, Clinical trials

## Background

Intensive care unit (ICU) physicians treat patients with complex and severe conditions who often require life-saving interventions. There is a need for adequate and efficient decision-making in the ICU due to the increasing demand for ICU treatment [1, 2]. Clinical decision-making is impeded by factors such as the increasing availability of large amounts of data, the increasing diagnostic and therapeutic opportunities and the increasing

*Correspondence: m.vangenderen@erasmusmc.nl
[1] Department of Adult Intensive Care, Erasmus MC University Medical Center, Room Ne-413, Doctor Molewaterplein 40, 3015 GD Rotterdam, The Netherlands
Full author information is available at the end of the article

complexity of care [3–6]. Treatment protocols are developed to support clinical decision-making but are often based on a simplified representation of reality that in individual cases may not reflect the complexity of illness.

More data are easily supposed to provide more insight but today's ICU physicians already have difficulties to interpret the enormous quantities of conventional clinical data. Often this data do not contain apparently useful information for clinical decision-making [7]. For instance, when caring for patients on ventilatory support, the number of variables ICU physicians have to consider may exceed two hundred [8]. Even though ICU physicians excel at analyzing snapshots of clinical information to determine the best treatment options, the ability to process all these data on a continuous basis lies well beyond the capabilities of the most experienced and knowledgeable ICU physicians [7].

In recent years, medicine witnessed the emergence of artificial intelligence (AI) and machine learning (ML). ML is a domain of AI and engages on the way computers ('machines') learn from data [9]. These technologies do not act upon preprogrammed rules but instead, they learn and improve from exposure to examples. AI models can catalog, classify, and correlate large amounts of data on a continuous basis in order to generate patient-specific predictions [10]. Studies across multiple specialties already demonstrated potential benefits of employing AI in the detection and classification of diseases [11–14]. In the end, the aim is to use AI models to aid clinical decision-making and to improve quality and efficiency of care [15].

In the ICU, AI might aid clinicians on diagnostic, prognostic, and therapeutic levels to improve patient outcomes. The number of publications on ICU-AI models has increased rapidly in the recent years, most commonly aimed at predicting complications, predicting mortality, and improving prognostic models [16]. A recent systematic review demonstrated that ML models can accurately predict onset of sepsis in ICU patients [17]. Although this analysis mainly comprised retrospective cohorts, it is a good example how algorithm performance is able to outperform traditional scoring tools.

Such positive studies tend to feed the hype regarding AI, although they have heterogeneous designs and methodologies potentially leading to low quality and risk of bias in some studies. There is thus a risk that the interest in AI may outpace the development of a uniform and structured approach to safely develop and deliver AI to patients [18]. Moreover, it remains undetermined whether patients already clinically benefit from AI.

The current systematic review seeks to give a contemporary overview of the current maturity of AI in the ICU, the research methods behind these studies, and the risk of bias in these studies. Specifically, we sought to describe the study design and aim, the size of used datasets, and the level of clinical readiness as part of the maturity definition.

**Take-home message**

At this moment, the majority of the published artificial intelligence (AI) models designed for use in the intensive care unit (ICU) do not reach beyond the prototyping and development environment. There are a number of barriers to overcome before AI can aid clinical decision-making in the ICU.

## Methods

This manuscript has been prepared according to the Preferred Reporting Items for Systematic reviews and Meta-analysis (PRISMA) guideline and was registered in the online PROSPERO database (Record ID: 199,683; reference number: CRD42020199863) before initiation of the literature search [19, 20].

### Inclusion criteria and study identification

Publications were eligible for review inclusion if they only included adult patients (≥ 17 years of age), assessed AI or ML algorithms, defined as computational models that are able to learn from exposure to large amounts of data, for clinical impact, used data that was gathered during ICU stay, were published as original research, and were available in English in full text. Publications were excluded when they solely used data gathered from a general ward, emergency room, operating theatre or post-anesthetic care unit. Candidate publications were identified through a comprehensive search in Embase, MEDLINE ALL, Web of Science Core Collection, Cochrane Central Register of Controlled Trials and Google Scholar from March to July 28, 2020. Our local librarian helped to further polish and update the electronic search strategies. The following terms were used as index terms or free-text words: 'artificial intelligence', 'machine learning', 'intensive care unit' and 'decision support' to identify eligible studies.

The full search terms used for literature search in Embase, MEDLINE ALL, Web of Science Core Collection, Cochrane Central Register of Controlled Trials and Google Scholar are noted in Online Resource 1.

### Study selection

After study selection, duplicates were identified and removed using EndNote X(9) (Clarivate Analytics, Philadelphia, PA, USA). An individual author (DvdS) screened all title and abstracts and judged whether a paper met inclusion criteria. Two authors (MvG and JvB) independently reviewed the included publications by abstract screening. Disagreement was resolved by consensus of a third reviewer (JH). Full-text publications were then screened and the final decision on eligibility was made by author (DvdS), reasons for exclusion were recorded per article. Excluded studies were reviewed using the same criteria for consensus.

### Data collection and review process

We extracted information on the following study characteristics: (1) design (categorized as retrospective-, prospective observational- and clinical [categorized as the following study designs: non-randomized clinical trials and randomized clinical trials/randomized controlled trials (RCT)] designs); (2) aim (categorized as alarm reduction, assessing clinical notes, classifying sub-populations, detecting spurious recorded values, determining physiological thresholds, improving prognostic models/risk scoring system, improving upon previous methods, predicting complications, predicting health improvement, predicting length of stay, predicting medication administration, predicting mortality and predicting readmissions) (in case studies had more than one aim, all aims were recorded); (3) size of the dataset (the total number of patients used for data analysis); (4) level of validation (categorized as internal validation [models are validated on patients who are included in studies' own dataset], external validation [models are validated on data of patients from other geographical locations or times], prospective observational validation, clinical validation and no reported validation); (5) AI level of readiness, which was assessed over time by applying the general concept of technology readiness levels introduced by National Aeronautics and Space Administration (NASA), which previously has been translated to the ICU environment [the consecutive levels increase from development to the clinical implementation of AI: problem identification (level 1), proposal of solution (level 2), model prototyping and development (level 3 and 4), model validation (level 5), real-time testing (level 6), workflow integration (level 7), clinical testing (level 8), and integration in clinical practice (level 9)] [21, 22]; (6) clinical study design and effects on patient outcome measures were extracted [categorized as reduced length of ICU stay, reduced overall mortality, reduced time on mechanical ventilation, reduced rate of complications and other (with details)]. Clinical study designs were considered to be either pre–post-implementation trials, non-randomized clinical trials or randomized clinical trials. Retrospective and prospective observational designs were considered to be non-clinical study designs, since treatment of patients or clinical decision-making was not being influenced by the use of AI.

### Data analysis

We used the PROBAST method, a tool to assess the risk of bias for prediction model studies, to assess the risk of bias in retrospective development studies [23]. The risk of bias judgement (categorized as high, unclear or low) was based on the four PROBAST domains (categorized as participants, predictors, outcomes, and analysis), was plotted for each individual domain of bias assessment, and is reported as percentages. No quantitative synthesis was conducted. We did not assess applicability, since no specific therapeutic question was defined for this systematic review.

Study aims were tabulated according to the corresponding study design. Dataset sizes were plotted against the proportion (%) of studies with the corresponding study design, and the level of readiness was plotted against the number of studies with the corresponding level and year of publication.

## Results

### Identification of studies

A total of 8645 studies were identified through our electronic search of which 4978 were identified via Embase, 1542 via Medline, 1743 via Web of Science Core Collection, 178 via Cochrane Central Register of Controlled Trials, and 204 studies through an additional search in Google Scholar, from July 1991 to July 2020. We reviewed 860 full-text studies of which 494 studies were finally included (Fig. 1). Main reason for ineligibility was that studies did not use AI or data was not collected in the ICU. A reference list of all included studies and the list with collected study items can be found in Online Resource 2 and Online Resource 3, respectively.

### Study design and purpose of AI

Most studies had a retrospective study design [476 studies (96.4%)], 8 studies (1.6%) had a prospective observational design and 10 studies (2%) had a clinical design, from which 5 (1%) were non-randomized trials and 5 (1%) were randomized clinical trials. The most common study aims were predicting complications [110 studies (22.2%)] and predicting mortality [102 studies (20.6%)] followed by improving prognostic models/risk scoring systems [86 studies (17.4%)] and classifying sub-populations [57 studies (11.7%)] (Table 1).

The median sample size across all retrospective studies was 1010 patients (IQR 149–7817) [for studies reporting on internal validation the median sample size was 968 (IQR 144–7794) and for external validation 1528 (IQR 235–7894)]. In addition, the median sample size was 179 (IQR 94–1411) and 142 (IQR 40–380) across all prospective observational and clinical studies, respectively.

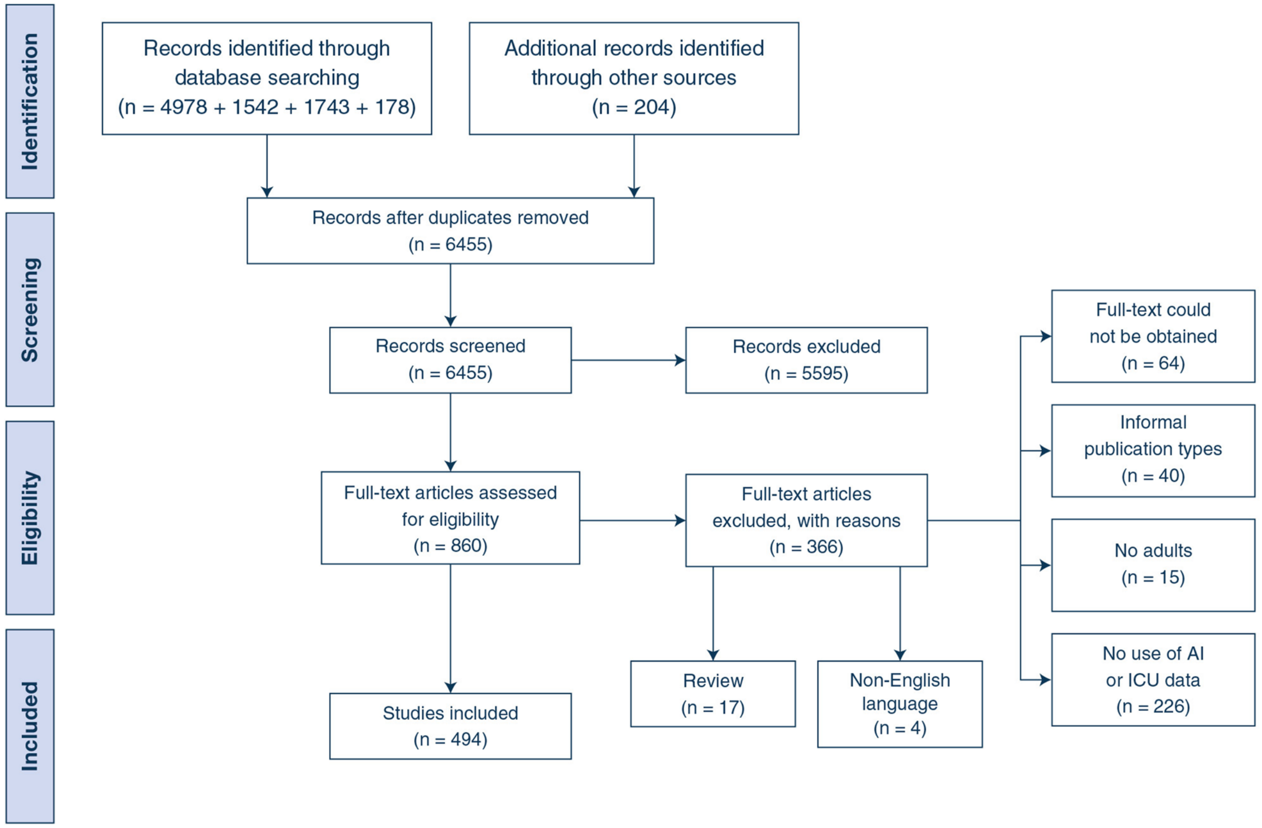
Identification
Screening
Eligibility
Included
Records identified through database searching (n = 4978 + 1542 + 1743 + 178)
Additional records identified through other sources (n = 204)
Records after duplicates removed (n = 6455)
Records screened (n = 6455)
Records excluded (n = 5595)
Full-text articles assessed for eligibility (n = 860)
Full-text articles excluded, with reasons (n = 366)
Full-text could not be obtained (n = 64)
Informal publication types (n = 40)
No adults (n = 15)
No use of AI or ICU data (n = 226)
Review (n = 17)
Non-English language (n = 4)
Studies included (n = 494)

**Fig. 1** PRISMA 2009 flow diagram of the study review process and the exclusion of studies; from [19]

Ten studies (2%) analyzed data on more than 100,000 patients, all of which had a retrospective study design. Most studies analyzed data on 100–1000 patients [142 studies (28.7%)]. Of all studies which reported on external validation, most of them analyzed data on 1000–10,000 patients [13 of 35 studies (37.1%)] (Fig. 2).

### Level of readiness

441 studies (89.3%) scored level 4 or below on the 'level of readiness' scale, 35 (7.1%) studies performed external validation (level 5), 8 studies (1.6%) integrated an AI model in the clinical setting without exposing the clinical staff to the results (level 6) and ten studies (2%) clinically evaluated model performance (level 8). Studies reporting on the outcome evaluation of an AI model that has been integrated in routine clinical practice (that is, not in a clinical study setting) were not identified (level 9). In recent years, the total number of studies reporting on model development and prototyping (level 3 and 4), increased rapidly from 30 studies per year in 2017 to 92 studies per year in 2019. Moreover, the number of studies per year reporting on external validation increased from two in 2017 to seven in 2019 (Fig. 3).

### Risk of bias

Risk of bias assessment was restricted to 467 retrospective development studies. Using the PROBAST criteria, the overall risk of bias (ROB) was classified as high in 378 of the 467 (80.9%) studies (Fig. 4). High ROB most often originated in the domains 'participants' (item 1.1 were inappropriate data sources used?) and 'analysis' (items 4.1 were enough patients included? and 4.3 were all enrolled participants included in the analysis?).

### Clinical studies involving AI

A total of ten studies were identified in which the performance of AI was clinically evaluated (Table 2). Five studies were non-randomized clinical trials and the other five were randomized clinical trials [24–33]. Eight out of ten studies provided complete information regarding study characteristics and the effect on patient outcomes. Significant improvement of patient outcomes was observed in seven studies.

## Discussion

The main finding of our systematic review is that the vast majority of current AI models in the ICU still remain

**Table 1 Number and proportion (%) of studies according to the study aim and study design**

| Aim of study | Study design | | | | | | |
|---|---|---|---|---|---|---|---|
| | Number (%) of studies with this aim¥ | Retrospective* | | | Prospective observational | Non-randomized clinical trial | Randomized clinical trial |
| | | Internal | External | Non | | | |
| Predicting complications | 110 (22.2%) | 86 (78.2%) | 12 (10.9%) | 4 (3.6%) | 5 (4.5%) | 2 (1.8%) | 1 (0.9%) |
| Predicting mortality | 102 (20.6%) | 92 (90.2%) | 9 (8.8%) | 1 (1%) | 0 (0%) | 0 (0%) | 0 (0%) |
| Improving prognostic models/risk scoring system | 91 (18.4%) | 80 (87.9%) | 7 (7.7%) | 3 (3.3%) | 1 (1.1%) | 0 (0%) | 0 (0%) |
| Classifying sub-populations | 58 (11.7%) | 53 (91.4%) | 1 (1.7%) | 4 (6.9%) | 0 (0%) | 0 (0%) | 0 (0%) |
| Determining physiological thresholds | 24 (4.9%) | 21 (87.5%) | 1 (4.2%) | 2 (8.3%) | 0 (0%) | 0 (0%) | 0 (0%) |
| Predicting length of stay | 22 (4.4%) | 22 (100%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) |
| Alarm reduction | 21 (4.3%) | 20 (95.2%) | 1 (4.8%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) |
| Predicting medication administration | 19 (3.8%) | 14 (73.7%) | 1 (5.3%) | 1 (5.3%) | 0 (0%) | 2 (10.5%) | 1 (5.3%) |
| Improving mechanical ventilation | 16 (3.2%) | 13 (81.3%) | 0 (0%) | 0 (0%) | 1 (6.3%) | 1 (6.3%) | 1 (6.3%) |
| Assessing clinical notes | 13 (2.6%) | 9 (69.2%) | 1 (7.7%) | 1 (7.7%) | 0 (0%) | 0 (0%) | 2 (15.4%) |
| Predicting readmissions | 12 (2.4%) | 11 (91.7%) | 1 (8.3%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) |
| Predicting relevance of clinical information | 8 (1.6%) | 5 (62.5%) | 1 (12.5%) | 2 (25%) | 0 (0%) | 0 (0%) | 0 (0%) |
| Assessing videos and images | 7 (1.4%) | 6 (85.7%) | 0 (0%) | 0 (0%) | 1 (14.3%) | 0 (0%) | 0 (0%) |
| Detecting spurious recorded values | 6 (1.2%) | 6 (100%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) |
| Predicting health improvement | 5 (1%) | 5 (100%) | 0 (0%) | 0 (0%) | 0 (0%) | 0(0%) | 0 (0%) |
| Predicting unnecessary lab tests | 3 (0.6%) | 3 (100%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) |
| Total (accounting for duplicates) | 494 | 421 (85.2%) | 35 (7.1%) | 20 (4%) | 8 (1.6%) | 5 (1%) | 5 (1%) |

¥ Where studies had more than one aim, all aims were recorded, thus percentages may exceed 100. *Retrospective studies were stratified according to their level of validation (e.g. internal, external and no reported validation)

within the testing and prototyping environment. Over time, the direction of the maturation of ICU-AI is mainly horizontal rather than diagonal, i.e. expansion of retrospective models instead of moving towards the clinical implementation (Fig. 3).

Considering the modest number of clinical trials (ten), AI does not have considerable impact on clinical decision-making in the ICU at this moment. Although there were only a few, the clinical trials included in this review nicely demonstrate how AI might actually benefit patient outcomes. A good example is the sepsis prediction model called 'InSight', which was tested by McCoy et al. and Shimabukuro et al. (Table 2), and which successfully moved from the testing and prototyping environment to the validation stage, and finally, the clinical evaluation stage [28, 33]. However, based on the findings of our review, we cannot conclude why this model was deployed successfully, as opposed to the majority of other ICU-AI models.

Generally, a possible explanation for the current direction of the maturation of AI may be that the development of prediction models in retrospective proof-of-concept studies is relatively 'easy' compared to leveraging AI to generate actually relevant information in clinical decision-making. Furthermore, interventions in the ICU are in general subject to the extreme complexity of the environment and the variation across sub-populations and in local practice. The medical AI community has achieved a major milestone in 2016, with the first AI model that has received legal approval by the United States Food and Drug Administration (FDA) [34]. However, the maturity of ICU AI has unfortunately not evolved much since then (Fig. 3).

The current review focused on the use of AI to analyze data gathered in the ICU, and thus no other clinical prediction models were included. Nonetheless, there are more examples of computerized decision support systems that can analyze such complex data, for example systems to improve weaning from mechanical ventilation [35]. However, many were knowledge-based systems that act upon preprogrammed rules and therefore did not meet the inclusion criteria. Furthermore, when it comes to clinical prediction models there is a variety of applied statistical methodologies, ranging from classical methodologies (e.g. logistic regression) to deep learning algorithms (e.g. deep neural networks). However, none of the ten clinical studies were based on deep learning algorithms. Future studies could delve into the comparison between AI models and models using classical methodologies.

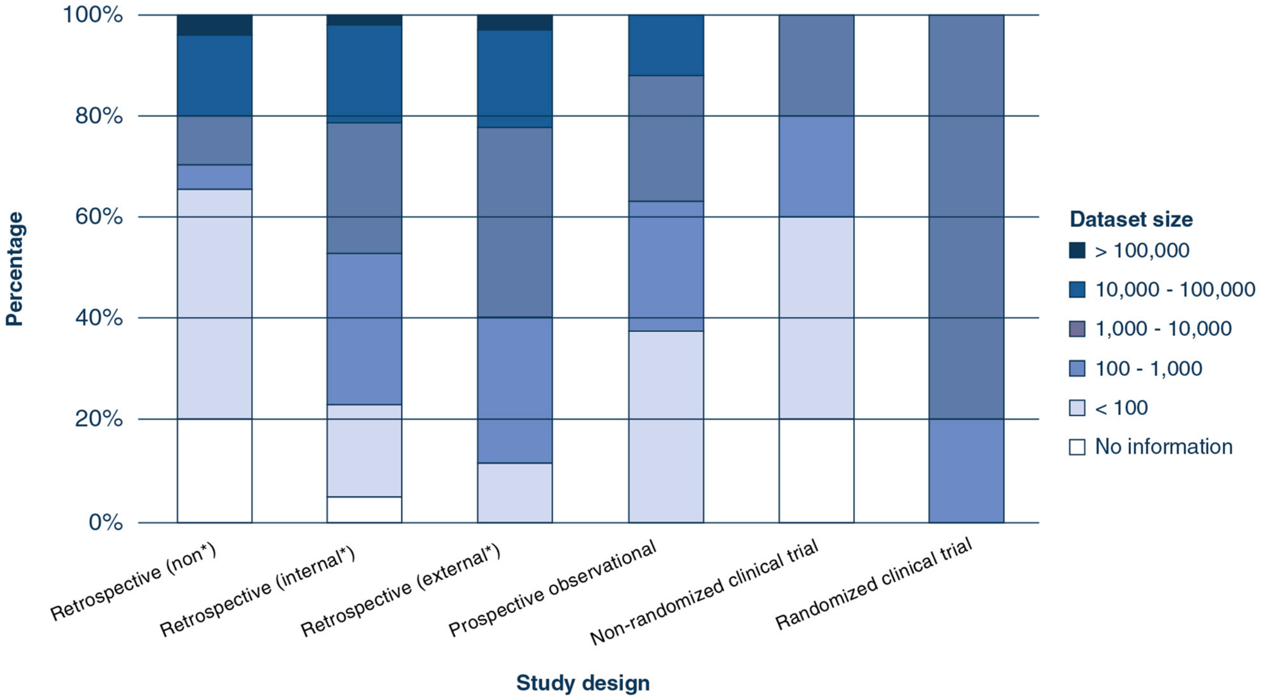

**Fig. 2** Proportion (%) of studies according to their design and the number of patients analyzed. *Studies with a retrospective design were stratified according to their level of validation (e.g. internal, external and no reported validation)

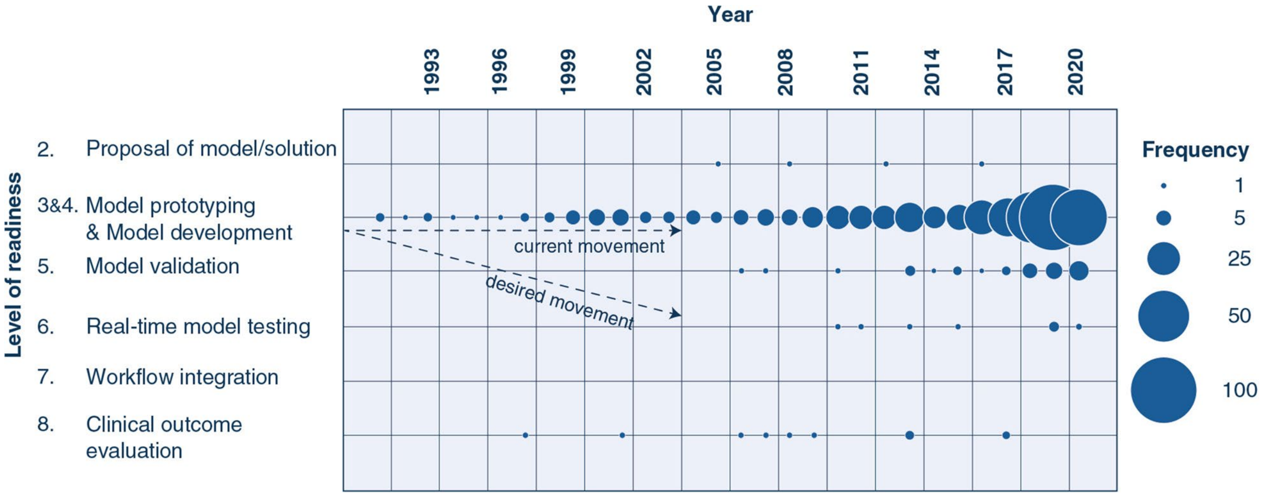

**Fig. 3** Number of studies published according to their level of readiness and year of publication. The total number of studies reporting on model development and prototyping (level 3 and 4), increased rapidly from 30 studies per year in 2017 to 92 studies per year in 2019. Furthermore, the number of studies per year reporting on external validation (level 5) increased from two in 2017 to seven in 2019. The current movement is mainly horizontal whereas the desired movement is diagonal, i.e. towards clinical evaluation

Implementation of AI is generally associated with barriers concerning data management, the development of models or the implementation in the (clinical) workflow. For instance, well-known barriers on the level of data and model development are: privacy/data sharing, regulation, and model generalizability [36–38]. Large amounts of

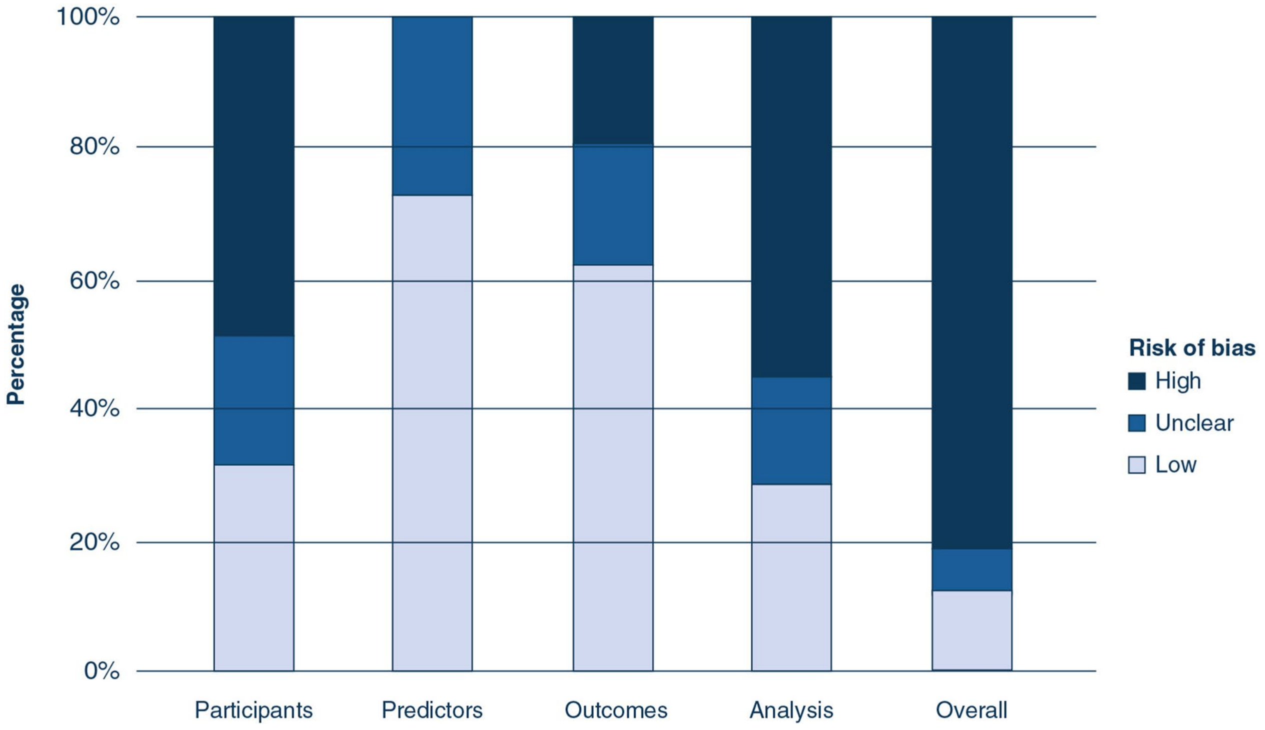

**Fig. 4** Percentage risk of bias according to the domain of assessment for all studies with a retrospective study design; assessed using PROBAST [23]

patient data are required to 'train' a new AI model. Ideally, ICU data are shared across institutions to construct large and diverse datasets. However, when using such sensitive information you have to comply with regulations like the General Data Protection and Regulation (GDPR) that has been issued by the European Union. Herein, the challenge is to balance the privacy and regulatory requirements with the desire to collect large and diverse datasets. When translating a model to a different institution you may encounter technical differences (for instance: differences in equipment, frequencies of data collection, and EHR systems) variations in local practices or variations in patient characteristics and as of a result, the model will tend to poorly generalize.

In this review, we have also identified specific barriers in the progress of AI in the ICU from model development to clinical implementation. First, 80% of the included retrospective studies were overall at high risk of bias. This is an indication that many studies may have been of poor quality or at least insufficient to serve as a starting point for successful maturity. The risk of bias was particularly high in the 'participants' PROBAST domain which implies that the quality of the used data may be poor. Frequently, the analyzed data were extracted directly from hospitals' electronic health record systems without proper validation. In general, obtaining high-quality data are a known challenge in the development of AI models [38]. Especially raw data, collected through continuous patient monitoring in the ICU environment, is prone to measurement errors [10, 39]. Several methods have been proposed to overcome this barrier, for instance using moving average models or signal estimators [40, 41]. It is not likely that the data will become entirely noise-free. Nonetheless, it is crucial to keep this in mind when developing an AI model.

Second, in most development studies, the size of the dataset was too small to exploit the full power of AI technologies. Deficiencies regarding the 'analysis' PROBAST domain most commonly related to Sect. 4.1, which means that studies did not use a reasonable number of patients relative to the number of predictor variables included in the AI model ($\geq$ 20 patients with the outcome of interest per predictor variable included in the model was considered to be reasonable [23]). This is a key issue for many uses of AI [42]. It is generally accepted that the more data an AI model gets access to, the more it can excel at its' predefined tasks [43]. To overcome this barrier, a solution may be to calculate the required sample size following the method proposed by Riley et al. [44].

Third, in 25% of the included studies in our review, it was unclear which variables were eventually used by the AI model. In addition, researchers in the field of AI commonly use terminology that is not familiar to clinicians and other researchers. Moreover, AI studies are

**Table 2 Characteristics and outcomes of clinical ICU-AI studies**

| Study author and year | Study aim | Number of patients | AI intervention | Effect on patient outcomes |
|---|---|---|---|---|
| Non-randomized clinical trial | | | | |
| Dojat et al. 1997 [24] | Predicting outcomes of weaning trials | No information | Ventilator settings were adapted to the patients' respiratory state | No information |
| Haddad et al. 2007 [25] | Optimizing sedative dosing | No information | A automated closed-loop infusion system was used | No information |
| Ross et al. 2009 [26] | Controlling systolic blood pressure in cardiac surgery patients | 3 | The clinicians were provided with predicted optimal treatment strategies (administration of fluid, commencing a drug, altering the drug infusion rate) | Successful prediction of drug therapies and appropriate infusion rates (without further details) |
| Cho et al. 2013 [27] | Predicting hospital acquired pressure ulcers | 866 | The nursing staff was provided with a daily predicted ulcer probability | Significantly reduced prevalence of HAPU (OR = 0.1; $p < 0.0001$) and reduced length of stay (OR = 0.67 [$p < 0.001$]) |
| McCoy et al. 2017 [28] | Predicting severe sepsis | 1328 | Sepsis onset was predicted in advance in real-time | Sepsis-related in hospital mortality rate decreased by 60.24% ($p < 0.001$) |
| Randomized clinical trial | | | | |
| Dazzi et al. 2001 [29] | Optimizing glycemic control | 40 | Blood glucose was controlled by automated insulin infusion rates | Improved blood glucose control (without further details) |
| Meystre et al. 2006 [30] | Automated problem list generation | 105 | The medical staff was provided with an automatic proposed list of clinically relevant problems for each patient | Significantly increased sensitivity of problem lists (from 8.9% to 41%) |
| Meystre et al. 2008 [31] | Automated problem list generation | 247 | The medical staff was provided with an automatic proposed list of clinically relevant problems for each patient | Significantly increased sensitivity of problem lists (from 9 to 41%) |
| Hsu et al. 2013 [32] | Predicting outcomes of weaning trials | 380 | Weaning trials were initiated when the AI model predicted that the patient could successfully be weaned from mechanical ventilation | Significantly reduced days on mechanical ventilation (intervention group 38.41 days ± 3.35 vs. control group 43.69 days ± 14.89, with a difference on average of 5.2 days [$p < 0.01$]) |
| Shimabukuro et al. 2017 [33] | Predicting severe sepsis | 142 | Sepsis onset was predicted in advance in real-time | Significantly reduced average LOS (2.7 days on average; $p = 0.042$) and reduced in-hospital mortality [12.4% ($p = 0.018$)] |

*AI* artificial intelligence, *HAPU* hospital acquired pressure ulcers, *LOS* length of stay, *OR* odds ratio, *p* p- value (two-tailed). ± plus and minus interquartile range often published in specific journals that are not familiar to clinicians [42]. The use of reporting standards, as used for conventional multivariable prediction models (TRIPOD), may enhance transparency and thus promote the progress of ICU-AI from development to validation studies [45]. Since existing reporting standards are tailored to conventional prediction models, the introduction of an ML-specific standard has been announced lately [46]. The use of reporting standards will enhance transparency and completeness when incorporating AI in a clinical trial protocol (SPIRIT-AI) and when reporting AI related results of a clinical trial (CONSORT-AI) [47, 48]. These are extensions of the conventional SPIRIT and CONSORT guidelines tailored to AI and can help both researchers and editors to better grasp the relevance of AI models [49, 50].

Finally, the concern for patient safety may impede the progress of AI [42]. At the moment, much remains unclear in the development and safe delivery of AI to patients. For example, if an AI model is inaccurate, poorly calibrated, or used in a biased way, and still used for decision support, this could lead to wrong clinical decisions [51, 52]. The FDA has proposed a regulatory framework for 'good machine learning practice' [53]. More recently, they came up with an action plan to update the previously published regulatory framework and to advance real-world clinical trials in order to provide information on what a real-world evidence generation program could look like [54]. This is an iterative process that will evolve as more clinical trials have been conducted. However, we believe that establishing a uniform and structured approach for the implementation of AI models is paramount to enable safe development and delivery of AI to patients in the ICU. The current findings and the idea for a uniform structured approach are in line with a previous position paper by Cosgriff et al. [55].

The last systematic review on the use of AI and ML in the ICU was published in 2019 [16]. In particular, it concluded that different AI/ML statistical methods (e.g. decision trees, neural networks, and random forest) were used over time and was restricted to the use of data routinely collected in the ICU. A more recent narrative review explained the different types of ML (e.g. supervised learning and unsupervised learning) in relation to specific clinical problems in the ICU (e.g. sepsis, length of stay, and mortality) [7]. Our review is the first to assess the current level of AI maturity, research methods, and the risk of bias in these studies.

Some limitations must be acknowledged. First, in our review, we have specifically focused on the development of AI models in the ICU. However, there may exist AI models which can be translated from other specialties to the ICU environment using similar aims and variables. Second, the guideline we used to assess the risk of bias (PROBAST) was designed for conventional prediction models rather than AI-based models, and so our findings should be interpreted in this context. Third, although comprehensive, the literature search might have missed studies that could have been included. However, to ensure that we did not underestimate the level of clinical readiness or missed FDA approved AI models, we additionally searched the online open access database for models that have been approved by the FDA, but did not identify additional AI models for use in the ICU [34]. Fourth, the PROBAST method was used to assess the risk of bias for prediction model studies. As such, other important determinants of model performance such as discrimination, calibration, and algorithmic bias were not taken into account. Last, we did not take racial and minority bias into account and did not gather such data. Future studies should consider gathering such information.

## Conclusion

AI is an innovative and quickly evolving field of research with the potential to improve clinical outcomes for ICU patients. The vast majority of currently developed ICU-AI models remain within the testing and prototyping environment and are not tested at the bedside. There are a number of barriers to overcome to move towards clinical evaluation and eventually implementation of AI to aid clinical decision-making. Obviously more research is needed to gain more insight into this process. This can be supported by a structured approach to develop AI models and to ensure safe delivery to patients.

### Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1007/s00134-021-06446-7.

### Abbreviations

AI: Artificial intelligence; FDA: Food and Drug Administration; GDPR: General data protection and regulation; HAPU: Hospital acquired pressure ulcers; ICU: Intensive care unit; ML: Machine learning; NASA: National Aeronautics and Space Administration; PROBAST: Prediction model Risk Of Bias Assessment Tool; PRISMA: Preferred Reporting Items for Systematic reviews and Meta-analysis; RCT: Randomized controlled trial; SPIRIT-AI: Reporting guidelines for clinical trial protocols for interventions involving artificial intelligence; CONSORT-AI: Reporting guidelines for clinical trial reports for interventions involving artificial intelligence; TRIPOD: Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis.

### Author details

[1] Department of Adult Intensive Care, Erasmus MC University Medical Center, Room Ne-413, Doctor Molewaterplein 40, 3015 GD Rotterdam, The Netherlands. [2] SAS Institute, Health Care Analytics, Huizen, The Netherlands.

### Acknowledgements

The authors wish to thank Christa Niehot from the Erasmus MC Medical Library for developing and updating the search strategies.

### Author contributions

DvdS designed the study, collected the data, analyzed and interpreted the data and drafted the manuscript. MvG participated in the study design, interpreted the data and drafted the manuscript. JH participated in the study design, interpreted the data and drafted the manuscript. DG conceived the study, participated in its design and coordination sign, and reviewed the manuscript. JvB conceived the study, participated in its design and coordination sign, and reviewed the manuscript. All authors read and approved the final manuscript for publication.

### Funding

No funding fees.

### Availability of data and materials

The dataset analyzed during the current study is available from the corresponding author on request.

### Declarations

### Conflicts of interest

The authors declare that they have no conflicts of interest. DG has received speakers' fees and travel expenses from Dräger, GE Healthcare (medical advisory board 2009–2012), Maquet, and Novalung (medical advisory board 2015–2018). All other authors declare no competing interests.

### Open Access

## Publisher's Note

Received: 24 March 2021 Accepted: 23 May 2021
Published online: 5 June 2021

### References

1. Adhikari NK, Fowler RA, Bhagwanjee S, Rubenfeld GD (2010) Critical care and the global burden of critical illness in adults. Lancet 376:1339–1346
2. Adhikari NK, Rubenfeld GD (2011) Worldwide demand for critical care. Curr Opin Crit Care 17:620–625
3. Citerio G, Park S, Schmidt JM, Moberg R, Suarez JI, Le Roux PD (2015) Data collection and interpretation. Neurocrit Care 22:360–368
4. Roshdy A (2019) Admission to the intensive care unit: the need to study complexity and solutions. Ann Intensive Care 9:14
5. Fuhrmann V, Weber T, Roedl K, Motaabbed J, Tariparast A, Jarczak D, de Garibay APR, Kluwe J, Boenisch O, Herkner H, Kellum JA, Kluge S (2020) Advanced organ support (ADVOS) in the critically ill: first clinical experience in patients with multiple organ failure. Ann Intensive Care 10:1
6. Vincent JL, Lefrant JY, Kotfis K, Nanchal R, Martin-Loeches I, Wittebole X, Sakka SG, Pickkers P, Moreno R, Sakr Y, Investigators IS (2018) Comparison of European ICU patients in 2012 (ICON) versus 2002 (SOAP). Intens Care Med 44:337–344
7. Gutierrez G (2020) Artificial intelligence in the intensive care unit. Crit Care 24:101
8. Morris AH (2018) Human cognitive limitations. Broad, consistent, clinical application of physiological principles will require decision support. Ann Am Thorac Soc 15:S53–S56
9. Hastie T, Tibshirani R, Friedman J (2017) The elements of statistical learning, data mining, inference and prediction. Springer, Berlin
10. Johnson AE, Ghassemi MM, Nemati S, Niehaus KE, Clifton DA, Clifford GD (2016) Machine learning and decision support in critical care. Proc IEEE Inst Electr Electron Eng 104:444–466
11. Lindsey R, Daluiski A, Chopra S, Lachapelle A, Mozer M, Sicular S, Hanel D, Gardner M, Gupta A, Hotchkiss R, Potter H (2018) Deep neural network improves fracture detection by clinicians. Proc Natl Acad Sci USA 115:11591–11596
12. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams T, Cuadros J, Kim R, Raman R, Nelson PC, Mega JL, Webster DR (2016) Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 316:2402–2410
13. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature 542:115–118
14. Safavi KC, Khaniyev T, Copenhaver M, Seelen M, Zenteno Langle AC, Zanger J, Daily B, Levi R, Dunn P (2019) Development and Validation of a machine learning model to aid discharge processes for inpatient surgical care. JAMA Netw Open 2:e1917221
15. Murdoch TB, Detsky AS (2013) The inevitable application of big data to health care. JAMA 309:1351–1352
16. Shillan D, Sterne JAC, Champneys A, Gibbison B (2019) Use of machine learning to analyse routinely collected intensive care unit data: a systematic review. Crit Care 23:284
17. Fleuren LM, Klausch TLT, Zwager CL, Schoonmade LJ, Guo T, Roggeveen LF, Swart EL, Girbes ARJ, Thoral P, Ercole A, Hoogendoorn M, Elbers PWG (2020) Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. Intensive Care Med 46:383–400
18. Keane PA, Topol EJ (2018) With an eye to AI and autonomous diagnosis. NPJ Digit Med 1:40
19. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P (2009) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ 339:b2535
20. Van de Sande DGMM, Van Genderen ME, Huiskens J, Gommers DAMPJ, Van Bommel J (2020) Moving from bytes to bedsides: a systematic review on the use of artificial intelligence in daily intensive care unit clinical practice. PROSPERO 2020 CRD42020199863. Available from: https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42020199863
21. Mankins JC (1995) Technology readiness levels. http://www.artemisinnovation.com/images/TRL_White_Paper_2004-Edited.pdf
22. Fleuren LM, Thoral P, Shillan D, Ercole A, Elbers PWG, Right Data Right Now C (2020) Machine learning in intensive care medicine: ready for take-off? Intensive Care Med 46:1486–1488
23. Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, Reitsma JB, Kleijnen J, Mallett S (2019) PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med 170:W1–W33
24. Dojat M, Pachet F, Guessoum Z, Touchard D, Harf A, Brochard L (1997) NeoGanesh: a working system for the automated control of assisted ventilation in ICUs. Artif Intell Med 11:97–117
25. Haddad WM, Bailey JM, Hayakawa T, Hovakimyan N (2007) Neural network adaptive output feedback control for intensive care unit sedation and intraoperative anesthesia. IEEE Trans Neural Netw 18:1049–1066
26. Ross JJ, Denai MA, Mahfouf M (2009) A hybrid hierarchical decision support system for cardiac surgical intensive care patients. Part II. Clinical implementation and evaluation. Artif Intell Med 45:53–62
27. Cho I, Park I, Kim E, Lee E, Bates DW (2013) Using EHR data to predict hospital-acquired pressure ulcers: a prospective study of a Bayesian Network model. Int J Med Inform 82:1059–1067
28. McCoy A, Das R (2017) Reducing patient mortality, length of stay and readmissions through machine learning-based sepsis prediction in the emergency department, intensive care unit and hospital floor units. BMJ Open Qual 6:e000158
29. Dazzi D, Taddei F, Gavarini A, Uggeri E, Negro R, Pezzarossa A (2001) The control of blood glucose in the critical diabetic patient: a neuro-fuzzy method. J Diabetes Complicat 15:80–87

30. Meystre S, Haug P (2006) Improving the sensitivity of the problem list in an intensive care unit by using natural languageprocessing. In: AMIA annual symposium proceedings/AMIA symposium, pp 554–558
31. Meystre SM, Haug PJ (2008) Randomized controlled trial of an automated problem list with improved sensitivity. Int J Med Inform 77:602–612
32. Hsu JC, Chen YF, Chung WS, Tan TH, Chen TS, Chiang JY (2013) Clinical verification of a clinical decision support system for ventilator weaning. Biomed Eng Online 12:S4
33. Shimabukuro DW, Barton CW, Feldman MD, Mataraso SJ, Das R (2017) Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. BMJ Open Respir Res 4:e000234
34. Benjamens S, Dhunnoo P, Mesko B (2020) The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digit Med 3:118
35. Burns KEA, Lellouche F, Nisenbaum R, Lessard MR, Friedrich JO (2014) Automated weaning and SBT systems versus non-automated weaning strategies for weaning time in invasively ventilated critically ill adults. Cochrane Database Syst Rev. https://doi.org/10.1002/14651858.CD008638.pub2
36. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D (2019) Key challenges for delivering clinical impact with artificial intelligence. BMC Med 17:195
37. Char DS, Shah NH, Magnus D (2018) Implementing machine learning in health care—addressing ethical challenges. N Engl J Med 378:981–983
38. Rajkomar A, Dean J, Kohane I (2019) Machine learning in medicine. N Engl J Med 380:1347–1358
39. Maslove DM, Dubin JA, Shrivats A, Lee J (2016) Errors, omissions, and outliers in hourly vital signs measurements in intensive care. Crit Care Med 44:e1021–e1030
40. Imhoff M, Bauer M, Gather U, Lohlein D (1998) Statistical pattern detection in univariate time series of intensive care on-line monitoring data. Intensive Care Med 24:1305–1314
41. Becker C, Gather U (2001) The largest nonidentifiable outlier: a comparison of multivariate simultaneous outlier identification rules. Comput Stat Data Anal 36:119–127
42. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K (2019) The practical implementation of artificial intelligence technologies in medicine. Nat Med 25:30–36
43. Mesko B, Gorog M (2020) A short guide for medical professionals in the era of artificial intelligence. NPJ Digit Med 3:126
44. Riley RD, Ensor J, Snell KIE, Harrell FE Jr, Martin GP, Reitsma JB, Moons KGM, Collins G, van Smeden M (2020) Calculating the sample size required for developing a clinical prediction model. BMJ 368:441
45. Collins GS, Reitsma JB, Altman DG, Moons KG (2015) Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ 350:7594
46. Collins GS, Moons KGM (2019) Reporting of artificial intelligence prediction models. Lancet 393:1577–1579
47. Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ, Spirit AI, Group C-AW, Group C-AS (2020) Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med 26:1351–1363
48. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, Spirit AI, Group C-AW (2020) Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med 26:1364–1374
49. Chan AW, Tetzlaff JM, Altman DG, Laupacis A, Gotzsche PC, Krleza-Jeric K, Hrobjartsson A, Mann H, Dickersin K, Berlin JA, Dore CJ, Parulekar WR, Summerskill WS, Groves T, Schulz KF, Sox HC, Rockhold FW, Rennie D, Moher D (2013) SPIRIT 2013 statement: defining standard protocol items for clinical trials. Ann Intern Med 158:200–207
50. Schulz KF, Altman DG, Moher D (2010) CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ 340:c332
51. Liu VX (2020) The future of AI in critical care is augmented, not artificial, intelligence. Crit Care 24:673. https://doi.org/10.1186/s13054-020-03404-5
52. Colak E, Moreland R, Ghassemi M (2021) Five principles for the intelligent use of AI in medical imaging. Intens Care Med 47:154–156
53. Administration FaD (2019) Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD)—discussion paper and request for feedback. Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD)—discussion paper and request for feedback. Food and Drug Administration, Silver Spring. Available from: https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf
54. Administration FaD (2021) Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD)action plan. Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) action plan. Food andDrug Administration, Silver Spring. Available from: https://www.fda.gov/media/145022/download
55. Cosgriff CV, Stone DJ, Weissman G et al (2020) The clinical artificial intelligence department: a prerequisite for success. BMJ Health Care Inform 27:e100183. https://doi.org/10.1136/bmjhci-2020-100183