OPEN

Artificial Intelligence for the Prediction of In-Hospital Clinical Deterioration: A Systematic Review

Lars I. Veldhuis, MD[1,2]

Nicky J. C. Woittiez, MD[3]

Prabath W. B. Nanayakkara, MD, PhD[4]

Jeroen Ludikhuize, MD, PhD[3,4]

OBJECTIVES: To analyze the available literature on the performance of artificial intelligence-generated clinical models for the prediction of serious life-threatening events in non-ICU adult patients and evaluate their potential clinical usage.

DATA SOURCES: The PubMed database was searched for relevant articles in English literature from January 1, 2000, to January 23, 2022. Search terms, including artificial intelligence, machine learning, deep learning, and deterioration, were both controlled terms and free-text terms.

STUDY SELECTION: We performed a systematic search reporting studies that showed performance of artificial intelligence-based models with outcome mortality and clinical deterioration.

DATA EXTRACTION: Two review authors independently performed study selection and data extraction. Studies with the same outcome were grouped, namely mortality and various forms of deterioration (including ICU admission, adverse events, and cardiac arrests). Meta-analysis was planned in case sufficient data would be extracted from each study and no considerable heterogeneity between studies was present.

DATA SYNTHESIS: In total, 45 articles were included for analysis, in which multiple methods of artificial intelligence were used. Twenty-four articles described models for the prediction of mortality and 21 for clinical deterioration. Due to heterogeneity of study characteristics (patient cohort, outcomes, and prediction models), meta-analysis could not be performed. The main reported measure of performance was the area under the receiver operating characteristic (AUROC) ($n = 38$), of which 33 (87%) had an AUROC greater than 0.8. The highest reported performance in a model predicting mortality had an AUROC of 0.935 and an area under the precision-recall curve of 0.96.

CONCLUSIONS: Currently, a growing number of studies develop and analyzes artificial intelligence-based prediction models to predict critical illness and deterioration. We show that artificial intelligence-based prediction models have an overall good performance in predicting deterioration of patients. However, external validation of existing models and its performance in a clinical setting is highly recommended.

KEY WORDS: critical care, intensive care, risk stratification models, artificial intelligence, clinical deterioration, Medical Emergency Team

Approximately 5–10% of hospitalized patients encounter a severe adverse event (SAE) during their hospital admission, including cardiac arrests (CAs) and ICU admission (1). Delay in recognition of deterioration leads to delayed diagnostic and therapeutic interventions resulting in increased morbidity and mortality (2, 3). Therefore, early recognition of deteriorating patients is pivotal to improve patient outcomes. To aid in this process,

DOI: 10.1097/CCE.0000000000000744

Early Warning Scores (EWSs) and Medical Emergency Teams (METs)/Rapid Response Teams (RRTs) were introduced in the late 1990s (4).

EWS are based on (abnormal) vital parameters that are frequently measured as input, resulting in an integer score. A higher score indicates a higher likelihood of clinical deterioration, generally defined as CAs, the need for ICU admission or death. Although EWS is a simple and easy-to-use tool, one of the main issues with EWS is the lack of specificity resulting in significant numbers of false-positive responses. This results in a poor positive predictive value of up to 5–10% (1, 5, 6).

To complicate matters further, the EWS is reliant on nurses taking vital signs. Frequency, (in)completeness and in some cases, incorrect calculation of the sum score can impact patient care significantly (7). Modification of the original EWS systems included adaptation of the thresholds of sum scores, including range of normality of vital signs and addition of blood measurements including lactate. These changes have led to enhanced performance. However, this still relies on timely awareness and interpretation of the staff by observing trends and absolute values.

These track and trigger systems were developed using variables that were often chosen because of accessibility and expected association with clinical deterioration. None have been fully validated and are mostly based on peer opinion (8). However, in recent years, medicine has witnessed the emergence of artificial intelligence (AI) as a tool to analyze large amounts of data (8–10). This offers the opportunity for unbiased analysis and uncovering hidden correlations in large datasets. To date, many AI-based algorithms have been published. Only a few have been tested in clinical practice, and scientific evaluation of these complex interventions and their clinical use are in its infancy (11).

AI-based models have the potential to improve the care for in-hospital patients and especially those deteriorating on the ward. Appropriate escalation in case of deterioration is possible by integration of these models into clinical practice with the potential for real-time risk assessment (12). This would allow for more appropriate allocation of resources including staff as shortages are increasing and staffing ratios are directly correlated with outcome (13). However, AI-based prediction models have currently very limited penetration into clinical practice, and besides potential benefits, have inherent disadvantages. Overall, implementation may take significant time to fully get adopted into normal clinical routine. Also, the clinicians may be reluctant to use it due to the so called "Black Box Phenomenon" because they may find it unclear how decisions are made by the algorithm. In addition, the validity of a given model may vary depending on the applied patient population.

With the abundance of published models in the literature, our goal was to analyze the performance of these AI-generated clinical models for the prediction of serious life-threatening events in non-ICU adult in-hospital patients and evaluate their potential clinical usage.

METHODS

This review was performed according to methods detailed in the systematic review protocol and is reported in line with the Preferred Reporting Items for Systematic Reviews and Meta-Analysis statement (14). The study protocol was registered and approved in the International Prospective Register of Systematic Reviews, reference number CRD42021267982.

Search Strategy

The PubMed database was searched for relevant articles in English literature from January 1, 2000, to January 23, 2022. Search terms included controlled terms as well as free-text terms. The following terms were used: AI, machine learning, deep learning, and deterioration. The complete search strategy can be found in **Additional File 3** (http://links.lww.com/CCX/B47). The World Health Organization International Clinical Trials Registry Platform (www.who.int/ictrp/en/) and U.S. National Institutes of Health (NIH) (https://clinicaltrials.gov/) were searched to identify any unpublished ongoing trials. Reference lists of studies that met the eligibility criteria were hand-searched to identify any potentially missed studies.

Two review authors (L.I.V., N.J.C.W.) independently performed the title-abstract and full-text screening. Disagreements were resolved by a third person (J.L.). For the full-text screening, reasons for exclusion per article were recorded. Relevant data were extracted by the authors (L.I.V., N.J.C.W., J.L.).

Eligibility Criteria and Study Selection

Primary outcome was the model performance for the prediction of mortality and/or clinical deterioration.

Therefore, peer-reviewed studies focusing on the prediction of mortality and clinical deterioration for hospital admitted, non-ICU patients were included. Mortality was defined as all types of mortality (30-d mortality and in-hospital mortality). Clinical deterioration was defined as unplanned ICU admissions, emergency surgery, consultation by MET/RRT, hemodynamic instability, respiratory distress, and life-threatening events.

Studies including mixed-patient populations (i.e., ward and intensive care/intermediate care) were excluded unless there was a separate analysis for ward patients. COVID-19 as a specific disease entity was excluded from this analysis. As COVID-19 did account for significant deterioration at the ward level (2020–2021) resulting in ICU admissions and (unplanned) death. The current review focuses on future implications for modeling in the context of the deteriorating ward patient based on previous literature within this field. Although unclear about the future impact of COVID-19 in daily practice, we excluded COVID-19 to remain focus and correlate with previous work in this field of research (15).

Machine learning models were the index test of interest defined as any machine learning technique, deep learning, or other self-learning models. Primary outcome was expressed in area under the receiver operating characteristics (AUROCs), *C*-statistics, area under the precision-recall curve (AUPRC), and sensitivity and specificity for important thresholds.

Other items that were collected from the papers included the year of publication, study design, study population, input variables, whether cross-validation was performed and its number of folds.

Quality of Evidence and Risk of Bias

Risk of bias was assessed using the Risk Of Bias In Non-Randomized Studies—of Interventions (ROBINS-I) tool. The ROBINS-I is a tool developed to assess risk of bias in the results of nonrandomized studies that compare health effects of two or more interventions (16). The NIH quality assessment tool for observational cohort and cross-sectional studies was used to assess the methodology of each included study in case meta-analysis was performed.

Data Synthesis and Analysis

Studies with the same outcome were grouped, namely mortality and all forms of deterioration (i.e., ICU admission, adverse events, CAs). Meta-analysis was planned in case sufficient data would be extracted from each study and no considerable heterogeneity among studies was present.

Data were not pooled in case of heterogeneity that could not be explained by diversity of methodological or clinical input variables among the trials. Synthesis was performed using Review Manager, version 5.4. (RevMan 2020) (17).

RESULTS

Study Selection

After removal of duplicates and reference checking for additional studies, abstracts of 613 articles were screened for eligibility (**Fig. 1**). Eventually, the full text of 150 studies were obtained and assessed by two reviewers. The main reason for exclusion of articles after full-text screening was a mixed-patient population (including intensive care admitted patients without separate analysis) and studies with focus on COVID-19 infected patients. From these papers, 45 studies met our inclusion criteria for synthesis. Of the 45 included articles, 31 articles were published in the last 2 years.

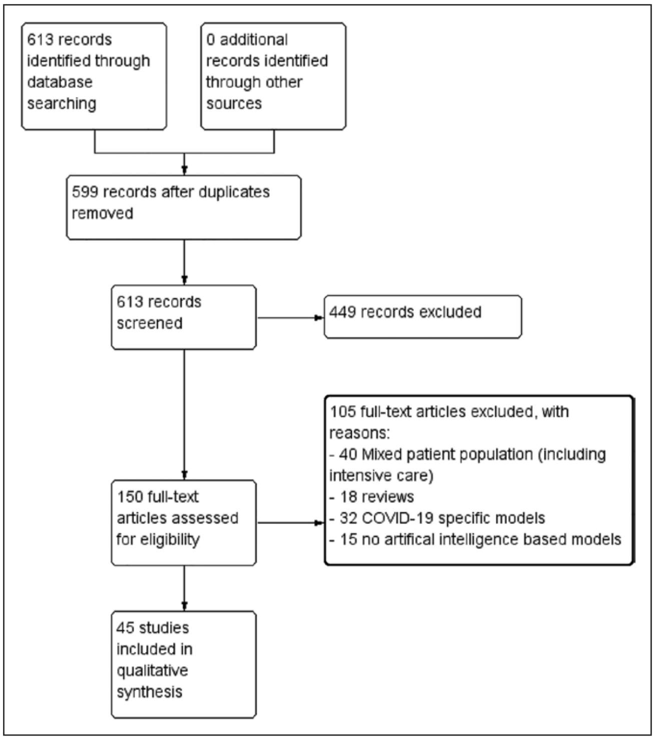

613 records identified through database searching
0 additional records identified through other sources
599 records after duplicates removed
613 records screened
449 records excluded
105 full-text articles excluded, with reasons:
- 40 Mixed patient population (including intensive care)
- 18 reviews
- 32 COVID-19 specific models
- 15 no artifical intelligence based models
150 full-text articles assessed for eligibility
45 studies included in qualitative synthesis

Figure 1. Study flow diagram.

Study Characteristics

Included studies could be subdivided according to their primary outcome; 24 with mortality (**Supplemental Content, Table 1**, http://links.lww.com/CCX/B47) and 21 with clinical deterioration as focus (**Supplemental Content, Table 2**, http://links.lww.com/CCX/B47). In total, five articles with mortality as endpoint had overlapping primary outcomes including also other forms of clinical deterioration. The 45 included studies developed in total 99 AI-models using 24 different methods with the three most used methods; random forest $n = 21$, gradient boosting $n = 20$, and deep neural networking $n = 10$ (Supplemental Files 1 and 2, http://links.lww.com/CCX/B47). Study populations consisted of Emergency Department (ED) patients (with several subpopulations) $n = 33$ and patients admitted to the ward (with several subpopulations) $n = 12$. The three most investigated outcomes were hospital mortality $n = 10$ and among the "group of clinical deterioration," ICU admission $n = 16$ and CA $n = 10$. All studies were based on retrospectively collected data.

Of the included studies, 42 used (among other variables) vital signs as their input. Other input variables consisted of laboratory values and patient characteristics (including age and gender). Due to the heterogeneity among studies, such as study population, types of machine learning models, and outcome parameters, meta-analysis could not be performed.

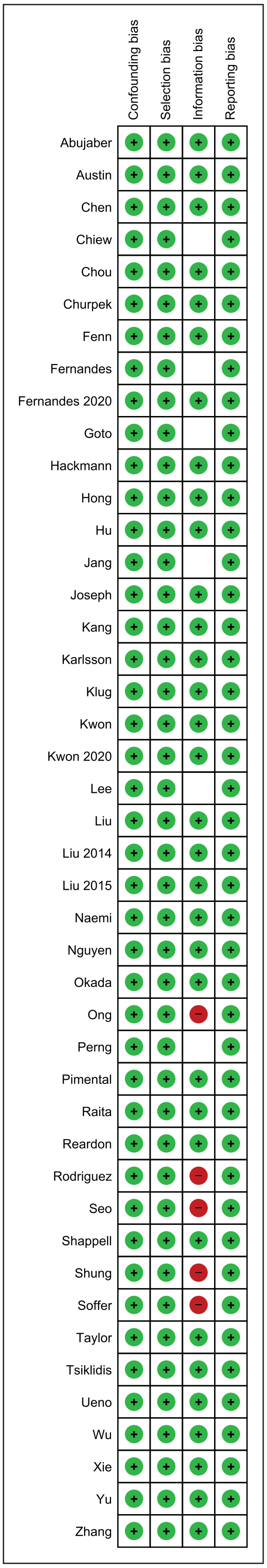
Model Performance

Out of 45 articles, primary outcome was mainly defined using AUROC/area under the curve $n = 38$. Other used performance indicators were AUPRC, sensitivity and specificity, positive predictive value, and F1. Of 38 articles reporting AUROC, 33 (87%) had a model performance greater than 0.8.

All included studies developed a model tailor to their collected data. Of these 45 studies, 37 performed internal validation with a subset of their patients. External validation was only performed by three studies.

Risk of Bias Assessment

Using the ROBINS-I tool, bias was rated high among five studies and unknown for six studies. This was related to potential information bias in all these circumstances (**Fig. 2**).



	Confounding bias	Selection bias	Information bias	Reporting bias
Abujaber	+	+	+	+
Austin	+	+	+	+
Chen	+	+	+	+
Chiew	+	+		+
Chou	+	+	+	+
Churpek	+	+	+	+
Fenn	+	+	+	+
Fernandes	+	+		+
Fernandes 2020	+	+	+	+
Goto	+	+		+
Hackmann	+	+	+	+
Hong	+	+	+	+
Hu	+	+	+	+
Jang	+	+		+
Joseph	+	+	+	+
Kang	+	+	+	+
Karlsson	+	+	+	+
Klug	+	+	+	+
Kwon	+	+	+	+
Kwon 2020	+	+	+	+
Lee	+	+		+
Liu	+	+	+	+
Liu 2014	+	+	+	+
Liu 2015	+	+	+	+
Naemi	+	+	+	+
Nguyen	+	+	+	+
Okada	+	+	+	+
Ong	+	+	−	+
Perng	+	+		+
Pimental	+	+	+	+
Raita	+	+	+	+
Reardon	+	+	+	+
Rodriguez	+	+	−	+
Seo	+	+	−	+
Shappell	+	+	+	+
Shung	+	+	−	+
Soffer	+	+	−	+
Taylor	+	+	+	+
Tsiklidis	+	+	+	+
Ueno	+	+	+	+
Wu	+	+	+	+
Xie	+	+	+	+
Yu	+	+	+	+
Zhang	+	+	+	+

Figure 2. Risk of bias summary: review authors' evaluation of risk of bias for each included study. Symbols: *green circle* = low bias risk, *no circle* = unknown risk, *red circle* = high bias risk.

DISCUSSION

This is the first study to systematically analyze studies investigating machine learning models and prediction of deterioration in ED and hospitalized adult patients. It is a field of science that is evolving rapidly with studies primarily being published in the last few years. In total, 45 articles were identified employing different methods of AI to develop risk stratification models. Most of the articles use vital parameters as part of their prediction models, which is in concordance with the "classic" EWS. However, many parameters that are not integrated in EWS, such as laboratory values and clinical characteristics, were also frequently used in the described models. Overall, performance of the included AI-based models is high and promising but external validity and performance in the clinical domain remains a focus for future (validation) studies.

Clinical Relevance

This review included studies focusing on the detection of deteriorating hospitalized, non-ICU patients. As mentioned before, adequate recognition of these patients is highly relevant as unpredicted ICU admissions and other SAEs are directly correlated with increased morbidity and mortality (2, 3). This specifically applies to patients admitted to nursing wards where nurse-patient ratio is lower compared with the ICU. This makes ward patients more vulnerable for delayed detection of clinical deterioration and thereby increased mortality (18). To achieve early adequate treatment and to reduce the number of SAEs, clinical prediction models, that is, the EWS, were developed. As part of Rapid Response Systems, well-integrated EWS with RRT have shown to reduce mortality and predominantly in-hospital CA (19). However, simply introducing EWS and other track and trigger-based systems alone, have not shown to significantly reduce SAEs nor mortality (20). This may be explained by the relatively low specificity of EWS, leading to a high false-positive rate, thus resulting in alarm fatigue (21). However, evidence regarding improved patient outcomes with the introduction of AI-based decision-making is lacking as implementation in clinical practice is scarce/absent. It is therefore highly relevant to investigate the effect of clinical implementation of AI-based risk stratification on patient outcome. Interesting is that none of the included articles discusses concrete plans to implement their developed model for clinical usage to improve patient care. Furthermore, only three studies in our review have externally validated their model as a first step toward this goal. To improve internal validity and to facilitate comparison between models, universal reporting of the diagnostic performance should be listed. Specific use of certain parameters such as the AUPRC is recommended because the primary outcome is a relatively rare event. Thus, the model performance using the AUPRC may be more accurate compared with the AUROC. Further studies therefore should use the AUPRC rather than the AUROC.

Limitations and Potential Pitfalls of AI

As briefly discussed in the introduction, potential drawbacks in relation to AI do exist (22). The main issue of AI-based risk stratification models is that they are not completely flawless. Implementing these models may lead to overreliance of these nonfaultless systems when medical personnel solely rely on these models. This could lead to SAEs or even death. However, a similar risk also exists with the use of EWS systems but has thus far not been reported to our knowledge.

As AI slowly finds its way into the clinical arena, methodological issues regarding model development still exist. Currently, models are developed based on relatively small datasets from mostly single centers and data selection does often not consider external factors, such as possible correlation with seasons. Also, models trained on data derived from single centers may only reflect good performance within this specific center, and extrapolation and usage in other hospitals has not been analyzed to date due to case-mix differences.

While most of included studies in this review are published recently, only three out of 45 articles externally validated their developed model. Prognostic models that aim to improve prediction of clinical events, require external validation and clinical testing (23). This is especially important, as prediction models tend to have a poorer performance in external validation cohorts compared with the original study population (24).

Future Opportunities for AI

A potential benefit of AI is the use of real-time or integrated data to continuously update the model and provide flexibility for changes in patient population, seasonality, and treatment effects. In medicine, this has not been widely employed yet, as described earlier. However, performance and effectiveness of the models based on real-time data should be further investigated.

Strengths and Limitations of This Review Study

This review is the first to systematically analyze all available literature from the past 20 years regarding AI for the prediction of critical illness in hospitalized, non-ICU, patients. It provides a comprehensive up-to-date overview of current pitfalls and highlights the plurality of available AI-based models and the current lack in clinical applicability. Our review shows that research efforts have been made as the number of models is booming in recent years. This provides promising avenues for future research and potential improvement in clinical care for this patient group.

Despite applying strict selection criteria, a meta-analysis could not be performed due to considerable heterogeneity in both study populations, input and output variables and methods used for model derivation. This deprives the opportunity to directly compare published AI-based models and their performances.

Another important limitation is that study outcomes including “hemodynamic instability” and “respiratory distress” are important clinical entities, but by nature lack a clearly defined and generally accepted definition, as is the case with mortality and in hospital cardiac arrests. Studying (and comparing) models based on these “vague” parameters requires clear definitions that are often absent at this time. This makes it more difficult to analyze, comprehend and compare consequences of our (and future) studies, and let alone to extrapolate into clinical practice.

CONCLUSIONS

A rapidly growing number of studies uses AI-based prediction models to identify critical illness and deterioration. This review shows that, despite heterogeneity among studies, AI-based prediction models have an overall good performance in predicting clinical deterioration of hospitalized non-ICU patients. However, standardized reporting of outcome variables for diagnostic models and subsequent external validation of these models are first steps to promote validity of the models. Second, investigation of clinical applicability and performance after implementation are major aspects for future studies to ultimately assess clinical utility.

1 Department of Anesthesiology, Amsterdam UMC, Location Academic Medical Center, Amsterdam, The Netherlands.

2 Department of Intensive Care, Amsterdam UMC, Location VU University Medical Centre, Amsterdam, The Netherlands.

3 Department of Intensive Care, Haga Teaching Hospital, Den Haag, The Netherlands.

4 Section General Internal Medicine, Department of Internal Medicine, Amsterdam Public Health Research Institute, Amsterdam UMC, Location VU University Medical Centre, Amsterdam, The Netherlands.

Drs. Veldhuis and Woittiez are both primary authors and contributed equally to this study.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal’s website (http://journals.lww.com/ccejournal).

All contributed to the study concept and design, acquisition and interpretation of data, and critically revised the article for important intellectual content.

The authors have disclosed that they do not have any potential conflicts of interest.

Data is available upon request.

Registered and approved on the International Prospective Register of Systematic Reviews (reference number CRD42021267982).

For information regarding this article, E-mail: L.I.Veldhuis@amsterdamumc.nl

REFERENCES

1. Churpek MM, Yuen TC, Park SY, et al: Using electronic health record data to develop and validate a prediction model for adverse outcomes in the wards*. *Crit Care Med* 2014; 42:841–848
2. Mardini L, Lipes J, Jayaraman D: Adverse outcomes associated with delayed intensive care consultation in medical and surgical inpatients. *J Crit Care* 2012; 27:688–693
3. Young MP, Gooder VJ, McBride K, et al: Inpatient transfers to the intensive care unit: Delays are associated with increased mortality and morbidity. *J Gen Intern Med* 2003; 18:77–83
4. Lee A, Bishop G, Hillman KM, et al: The medical emergency team. *Anaesth Intensive Care* 1995; 23:183–186
5. Smith GB, Prytherch DR, Meredith P, et al: The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. *Resuscitation* 2013; 84:465–470
6. Subbe CP, Kruger M, Rutherford P, et al: Validation of a modified Early Warning Score in medical admissions. *QJM* 2001; 94:521–526
7. Clifton DA, Clifton L, Sandu DM, et al: ‘Errors’ and omissions in paper-based early warning scores: The association with changes in vital signs–a database analysis. *BMJ Open* 2015; 5:e007376

8. Wuytack F, Meskell P, Conway A, et al: The effectiveness of physiologically based early warning or track and trigger systems after triage in adult patients presenting to emergency departments: A systematic review. *BMC Emerg Med* 2017; 17:38
9. Beam AL, Kohane IS: Big data and machine learning in health care. *JAMA* 2018; 319:1317–1318
10. Yu KH, Beam AL, Kohane IS: Artificial intelligence in healthcare. *Nat Biomed Eng* 2018; 2:719–731
11. Yin J, Ngiam KY, Teo HH: Role of artificial intelligence applications in real-life clinical practice: Systematic review. *J Med Internet Res* 2021; 23:e25759
12. Dahella SS, Briggs JS, Coombes P, et al: Implementing a system for the real-time risk assessment of patients considered for intensive care. *BMC Med Inform Decis Mak* 2020; 20:161
13. McHugh MD, Aiken LH, Sloane DM, et al: Effects of nurse-to-patient ratio legislation on nurse staffing and patient mortality, readmissions, and length of stay: A prospective study in a panel of hospitals. *Lancet* 2021; 397:1905–1913
14. Page MJ, McKenzie JE, Bossuyt PM, et al: The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* 2021; 372:n71
15. Jones DA, DeVita MA, Bellomo R: Rapid-response teams. *N Engl J Med* 2011; 365:139–146
16. Sterne JA, Hernán MA, Reeves BC, et al: ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* 2016; 355:i4919
17. ReviewManager: Review Manager (RevMan). Version 5.4. Accessed January 2023. Available at revman.cochrane.org
18. Chen J, Bellomo R, Flabouris A, et al; MERIT Study Investigators for the Simpson Centre; ANZICS Clinical Trials Group: The relationship between early emergency team calls and serious adverse events. *Crit Care Med* 2009; 37:148–153
19. Solomon RS, Corwin GS, Barclay DC, et al: Effectiveness of rapid response teams on rates of in-hospital cardiopulmonary arrest and mortality: A systematic review and meta-analysis. *J Hosp Med* 2016; 11:438–445
20. Alam N, Hobbelink EL, van Tienhoven AJ, et al: The impact of the use of the Early Warning Score (EWS) on patient outcomes: A systematic review. *Resuscitation* 2014; 85:587–594
21. Veldhuis LI, Nanayakkara PW: Early warning score to detect deterioration in the hospital: Role of staffing and alarm fatigue? *Neth J Crit Care* 2021; 29:3
22. Sunarti S, Fadzlul Rahman F, Naufal M, et al: Artificial intelligence in healthcare: Opportunities and risk for future. *Gac Sanit* 2021; 35(Suppl 1):S67–S70
23. Holdsworth LM, Kling SMR, Smith M, et al: Predicting and responding to clinical deterioration in hospitalized patients by using artificial intelligence: Protocol for a mixed methods, stepped wedge study. *JMIR Res Protoc* 2021; 10:e27532
24. Ramspek CL, Voskamp PW, Van Ittersum FJ, et al: Prediction models for the mortality risk in chronic dialysis patients: A systematic review and independent external validation study. *Clin Epidemiol* 2017; 9:451–464