Delirium Prediction Models in Intensive Care Unit Patients Using Nursing Data

Mihui Kim

The Graduate School
Yonsei University
Department of Nursing

Delirium Prediction Models in Intensive Care Unit Patients Using Nursing Data

A Dissertation
Submitted to the Department of Nursing
and the Graduate School of Yonsei University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Mihui Kim

July 2022

This certifies that the dissertation
of Mihui Kim is approved.

_____
Thesis Supervisor: Mona Choi

_____
Eui Geum Oh: Thesis Committee Member

_____
Joungyoun Kim: Thesis Committee Member

_____
Dukyong Yoon: Thesis Committee Member

_____
Min-jeoung Kang: Thesis Committee Member

The Graduate School
Yonsei University
July 2022

ACKNOWLEDGMENTS

I would like to thank all the people who helped me a lot during the long doctoral journey. I hesitated about entering the Ph.D. program after graduating from the master's course for a long time, and finally could enter the Ph.D. course and finish. I would not have gone through this process without the guidance, prayers, and help of all the people involved with me.

First of all, I would like to sincerely appreciate my thesis supervisor, Professor Mona Choi. Thanks to your belief and consideration during my course, I could grow as a researcher. I could realize that waiting and encouraging with a "one word" rather than saying a hundred words have a strong power. I thank Professor Eui Geum Oh, who has taught and led me through my master's and Ph.D. course. Your insightful advice enabled me to clarify the phenomenon I was concerned about and think about the clinical adaptability of my research. I also thank Professor Joungyoun Kim, who gave me statistical advice when I was suffering. I could have a new perspective on statistical analysis through your guidance and finally finish my dissertation. Thanks also for Professor. Dukyong Yoon. I had a good opportunity to look back on the meaning of the dissertation, thanks to your valuable advice. I could rethink the value of research processes, not only of good results. Lastly, I thank Professor. Min-jeoung Kang, inspired

me to start my dissertation. Your studies' results motivated me to start my research and your cheers before the start of my research gave me the driving force to reach the end of the process of my research and dissertation.

I am deeply thankful to Professor Mona Choi and our research team (Yong Sook, Giwook, Yesol, Changhwan). We were able to positively influence each other to grow altogether during the research period. I was fortunate to have you all as my research colleague, and I was happy all the time with you all. Even though we all physically fall from each other, we will always be close at heart. In addition, I would like to especially thank my dear friend, Ohyoung, who prayed, cried, and laughed with me. Your advice gave me big energy to keep myself grounded and move forward in my academic journey. I also really appreciate Professor. Sujin Kim and Seongmi, who mentally supported me. Also, I would like to send my appreciation to my fellow classmates and Ph.D. course colleagues (Ohyoung, Soyoon, Eunkyoung, Eunjeoung, Sooyeon, Young Hee, Hyeyeon, and Mikyung) for being a constant source of strength as we worked on our dissertations together. Lastly, my lovely research room 402 fellows (Yesol, Seongmi, Sang Hwa, Won Jin, Ocksim, and Ho kon), who worked alongside me and laughed and cried with me as we sat next to the sunny window in Lab 402. Also, my thanks also go to the MIMIC team and the department of digital health of Severance Hospital for making my dissertation research possible.

I have devoted myself tirelessly but happily to my three and a half years of Ph.D. course without resting as I had wanted to do for a long time. I would not have been able to wholly devote myself to this program without the love of my family. I could not have undertaken this journey without my beloved husband, Yonsung's help. You have supported my decision and did so much on my behalf. I give all my love to my dear eldest son, Doyeop, for understanding, loving, and even being proud of his busy mother and to my youngest, adorable daughter Doa, who always makes me smile. I am thankful they have grown up so well. I thank my mother-in-law, who always understood and supported me. Lastly, my mom, Park Jeungim (Stelra), I deeply appreciate the time you had spent with my children with love when I was unavailable. Your prayers and sacrificial love made all this possible. I love and thank you all.

The past three and a half years allowed me to not only grow academically but also help me shape my future and make lasting relationships along the way. It was an opportunity to meet precious people who will be in my future. I will try to become a researcher who grows together and shares everything learned from the lessons and guidance from the journey of the Ph.D. program.

July 2022

Mihui Kim

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

ABSTRACT

Delirium Prediction Models in Intensive Care Unit Patients Using Nursing Data

Kim, Mihui
Dept. of Nursing
The Graduate School
Yonsei University

Introduction: Delirium frequently occurs among patients in intensive care units (ICU), leading to prolonged ICU stays, increased mortality, and higher healthcare costs. Nursing data include information related to nurses' observations and clinical judgments about patients' conditions; these data can be valuable indicators for predicting clinical deterioration in patients with rapidly changing clinical status. This study aimed to develop and validate machine learning-based delirium prediction models for ICU patients using nursing data that reflects time variation in electronic medical records (EMR).

Methods: This retrospective cohort study was performed using the Medical Information Mart for Intensive Care (MIMIC) database and the EMR from a single tertiary hospital in Seoul, South Korea. The cohorts included patients aged 18 years or older with a delirium

screening tool record admitted to the ICU for at least 24 hours. Patients who received hospice or palliative care were excluded from the study. EMR data included in the model predictors were extracted as forms of predisposing and precipitating factors, nursing assessments, and the frequency and intervention patterns of nursing documentation. Patient data were extracted for 24 hours before the occurrence of delirium based on the sliding window method. As a class imbalance, this study used combination sampling to obtain a balanced dataset and trained with five repetitions of stratified 10-fold cross-validation. Logistic regression, support vector machine, random forest, and neural network were used.

Results: The development and external validation cohorts included 9491 and 2629 admissions, among whom, delirium occurred in 17.0% and 8.4% of cases, respectively. The mean duration of delirium onset was 2.6 days in the development cohort. The best model performance of the Model I (40 predictors) was observed in the random forest method; and the area under receiver operating characteristics (AUROCs) and 95% confidence intervals (CIs) of the internal and external validation cohorts were 0.975 [0.967, 0.982] and 0.770 [0.733, 0.808], respectively. The Model II (31 predictors) used candidate predictors, excluding the Glasgow Coma Scale (GCS) and Richmond Agitation-Sedation Scale (RASS) among Model I predictors, and the random forest model exhibited the best performance (AUROC and 95% CI: 0.951 [0.940, 0.962]). Among the important predictors of the developed models, GCS, RASS, pain, and nursing

documentation frequency were ranked higher than the precipitating and predisposing factors.

Conclusion: This study used nursing data that reflected time variation to develop and validate machine learning-based delirium prediction models for ICU patients. Nursing data, including nurses' judgments, were important data resources in the delirium prediction models. The developed models were validated internally and externally with acceptable performance in two hospitals' EMR environments. The models developed in the current study may be used as fundamental resources for developing the clinical decision support algorithms in EMR for predicting delirium in ICUs.

Key words: delirium, electronic medical records, intensive care units, nursing assessment, nursing records, machine learning, prediction model

I. INTRODUCTION

1.1. Background

The intensive care unit (ICU) is where the physiological status of critically ill patients hospitalized with severe diseases or rapidly changing conditions are continuously supervised using monitoring devices (Awad et al., 2017). Over the past 20 years, with advances in healthcare technology and strategies to reduce the prevalence of complications, ICU care has increased patients' survival rates and decreased the duration of hospitalization (Vincent et al., 2018; Chiarici et al., 2019). Recently, ICU care has focused on fostering recovery with fast, safe, and effective therapy (Chiarici et al., 2019) which are priorities for ICU patient management (Connolly et al., 2015).

Delirium is a neurocognitive disorder that frequently occurs among ICU patients, characterized by inappropriate behaviors and acute and fluctuating changes in thinking, cognition, and sensation (Collet et al., 2018). It occurs in up to 55% of critically ill patients (Rood et al., 2018). Delirium can be categorized into hyperactive, hypoactive, and mixed according to its motoric presentation (Krewulak et al., 2018). Those with delirium have affected clinical outcomes, leading to prolonged ICU and hospital stays, increased mortality (Hshieh et al., 2015; Krewulak et al., 2020), higher healthcare costs, and additional workload for healthcare providers (Witlox et al., 2010; Hshieh et al., 2015).

Delirium screening is periodically performed in ICUs for the early detection and clinical intervention of delirious patients (Barr et al., 2013; Hshieh et al., 2015; Rood et al., 2018). Delirium screening is conducted using tools such as the Confusion Assessment Method for the Intensive Care Unit (CAM-ICU) (Ely et al., 2001) or the Intensive Care Delirium Screening Checklist (ICDSC) (Bergeron et al., 2001). However, implementing delirium screening and providing clinical interventions to prevent delirium among all patients requires substantial resources and staffing (Wassenaar et al., 2018). In addition, healthcare providers who are not specially trained in delirium screening methods reduce the sensitivity of screening (Aldecoa et al., 2017). Therefore, providing early clinical intervention through timely and accurate delirium prediction without additional assessments may help to improve resource management in the ICU and patient outcomes.

Nurses are often the first healthcare providers to recognize a patient's deterioration through their surveillance and monitoring activities in clinical practice (Douw et al., 2015). Nursing data in the electronic medical record (EMR) represent nursing activities to deliver holistic patient-centered care (Lewis et al., 2017) and consists of information related to nurses' observations and clinical judgments of patients' conditions (Odell et al., 2009; Cho et al., 2019; Kang et al., 2020). In EMR, nursing assessments combined with physiological data may facilitate the early detection of patient deterioration (Capan et al., 2017). An increase in nursing documentation frequency in the EMR may be evidence of nurses' concern about a patient's deterioration (Collins & Vawdrey, 2012; Collins et al., 2013). The optional documentation features of nursing data may reflect a nurse's concern

about a patient's risk of deterioration (Collins et al., 2013; Kang et al., 2020). Therefore, nursing data, including physiological data, nursing assessments, and patterns of both, may be valuable indicators to predict patients' deterioration with time variation (Rossetti et al., 2019).

Delirium prediction models have been developed based on patients' medical history, physiological data, and laboratory test results (van den Boogaard et al., 2012; Wassenaar et al., 2015; Hur et al., 2021). By extracting clinical features within the first 4 to 24 hours of ICU admission, these models are used to predict delirium during the first 24 hours or the entire ICU stay. They were useful models for predicting the first-time occurrence of delirium during ICU stay (Marra et al., 2018). However, these models have limited capacity to capture time-varying daily risk factors and changes in dynamic status among ICU patients and do not consider nursing data such as nursing assessments and nursing documentation patterns that reflect time variation. Nursing documentation patterns were associated with patient deterioration and consisted of documentation frequency and intervention found in nursing flowsheets and medication administration records (Schnock et al., 2021). The Healthcare Process Modeling Framework to Phenotype Clinician Behaviors for Exploiting the Signal Gain of Clinical Expertise (HPM-ExpertSignals), proposed by Rossetti et al. (2021), is a framework for capturing important features in the EMR and uses data patterns to predict associated outcomes. This framework reflects clinical decision-making processes and clinical concerns about a patient's condition (Rossetti et al., 2019; Rossetti et al., 2021).

Based on the HPM-ExpertSignals framework, this study used nursing data, including observations and clinical judgments, to develop and validate delirium prediction models for ICU patients that reflect time variability in a patient's conditions. The results may be used to identify patients at risk of delirium without the need for additional delirium screening. Using the developed delirium prediction models in clinical practice may ultimately improve patient outcomes and reduce nurses' workloads. In addition, the results of this study may emphasize the importance of nursing data that reflect nurses' observations and clinical judgments about patients' deterioration in clinical practice.

1.2. Purpose

The purpose of this study was to develop and validate machine learning-based delirium prediction models for ICU patients using nursing data in EMRs. The specific aims of this study are as follows:

First, to develop delirium risk prediction models that reflect time variation using nursing data in ICU patients;

Second, to validate the developed prediction models' performance in an internal validation cohort;

Third, to identify important variables and assess the reliability of the developed prediction models;

Fourth, to identify the most optimal models by comparing the developed prediction models; and

Fifth, to validate the optimal models' performance in an external validation cohort.

1.3. Definitions

1.3.1. Nursing data

1) Theoretical definition: Nursing data include nurses' observations of patients' clinical conditions and nurses' clinical judgments based on their interpretation of patients' subjective and objective signs and symptoms (Odell et al., 2009; Cho et al., 2019; Kang et al., 2020). These data contain information about the complexity, context, and richness of patients' nursing care (Dykes et al., 2009; De Georgia et al., 2015).

2) Operational definition: In the current study, nursing data recorded by nurses were defined as information about nurses' observations and clinical judgments, specifically including nursing assessments and nursing documentation patterns.

1.3.2. Delirium

1) Theoretical definition: Delirium is an acute confusion state related to a disturbance in attention and awareness that occurs concurrently with severe illness (American Psychiatric Association, 2013).

2) Operational definition: Delirium was defined as at least one positive assessment using CAM-ICU or ICDSC, or administration of haloperidol or lorazepam (van den Boogaard et al., 2012; Wassenaar et al., 2015) during ICU stay from 24 hours after ICU admission. If more than one delirium event occurred during an ICU stay, the first delirium event was defined as the primary outcome time.

II. LITERATURE REVIEW

This literature review was performed to explore delirium risk factors, delirium prediction models for critically ill patients, and predictors used in these models. In addition, the prediction models using nursing data as predictors were identified, and the types of nursing data were investigated.

2.1. Delirium risk factors in the intensive care unit

Delirium is caused by complex interactions between predisposing factors related to a patient's vulnerability at admission and precipitating factors associated with ICU admission (Inouye & Charpentier, 1996). Precipitating factors are caused by the environmental changes during hospitalization and are potentially modifiable (Green et al., 2019). Therefore, identifying modifiable factors among the precipitating factors associated with ICU admission and establishing related intervention strategies may be a way to minimize the occurrence of delirium.

A systematic review by Zaal et al. (2015) analyzed 33 studies investigating the risk factors for delirium in adult patients admitted to ICUs and derived a total of 11 risk factors (Zaal et al., 2015). It reported that the predisposing factors for delirium were age, history of dementia, delirium, and hypertension; and the precipitating factors for delirium were polytrauma, emergency surgery before ICU admission, Acute Physiology and

Chronic Health Evaluation-II (APACHE-II) score, sedation-induced coma, mechanical ventilation, and metabolic acidosis. Notably, among the precipitating factors, sedation-induced coma and mechanical ventilation were modifiable factors that may influence delirium occurrence. The use of dexmedetomidine in the ICU was a factor that reduced the risk of delirium.

In the clinical practice guidelines for pain, agitation, and delirium among adult patients in the ICU, risk factors for increasing delirium were classified as ‘strong’, ‘moderate’, or ‘inconclusive’ according to the quality of evidence (Devlin et al., 2018). Strong evidence of delirium risk factors were advancing age, dementia, altered cognitive function, emergency surgery, trauma, increased APACHE and American Society of Anesthesiologists (ASA) scores, use of benzodiazepines, and blood transfusions. Moderate evidence of delirium risk factors were history of hypertension, hospitalization for neurological disease, and use of antipsychotics. In contrast, nicotine use, smoking, ventilator use, history of respiratory disease, admission to an internal medicine ward, dialysis, and Glasgow Coma Scale (GCS) were not related to the occurrence of delirium. The use of benzodiazepines and blood transfusion were qualified as a form of strong evidence and modifiable factors. It is crucial to minimize these factors to reduce the risk of delirium occurrence. A systematic review reported that age and APACHE-II score were associated with delirium subtypes, delirium duration, length of hospital stays, and mortality (Krewulak et al., 2020).

Delirium in ICU patients is caused by the interactions between predisposing and precipitating factors, and the detailed factors are diverse. In summary, the reported predisposing factors for delirium are related to a patient's age, cognition (dementia, cognitive decline), history of hypertension and delirium, and ASA score. The reported precipitating factors for delirium are APACHE-II score, sedation-induced coma, emergency surgery or emergency admission, neurological disease, metabolic acidosis, polytrauma, mechanical ventilation, sedatives use, blood transfusion, multiple organ failure, infection, and blood urea nitrogen (BUN) elevation. It is necessary to understand these risk factors to predict delirium accurately.

2.2. Delirium prediction models in the intensive care unit

Delirium prediction models are developed for predicting patients at risk of delirium based on risk factors (Green et al., 2019). These models enable healthcare providers to actively monitor the high-risk patients for delirium, strengthen delirium prevention efforts (pharmacological and non-pharmacological interventions), and detect and diagnose delirium early (Halladay et al., 2018). A systematic review of 21 studies investigating delirium prediction models for ICU patients (Chen et al., 2020) reported that all studies used CAM-ICU or ICDSC for delirium screening and that candidate predictors were diverse. For model development, among the machine learning methods, 14 studies used

logistic regression. The area under the receiver operating characteristic (AUROC) of the developed models in the literature was between 0.73 and 0.91.

A literature review was performed in the current study to identify the delirium prediction models and the predictors used in model development. This review included studies that developed delirium prediction models for adult patients admitted to the ICUs and reported the developed models' predictive performance. The review excluded validation-only studies and those conducted only in limited ICUs or patients with specific diseases. The literature review included a previous systematic review (Chen et al., 2020) and identified new studies via seven electronic databases, including PubMed, CINAHL, Cochrane CENTRAL, EMBASE, IEEE Xplore Library, Web of Science, and Scopus. Search terms were (patient* AND ("intensive care unit*" OR "intensive care" OR "critical care")) AND ("prediction model*" OR "machine learning" OR (sensitivity AND specificity) OR "ROC curve")) AND (delirium OR "ICU psychosis" OR "ICU syndrome). Finally, eight studies met the eligibility criteria and were included in the review. Table 1 summarizes the outcome of the literature review on the delirium prediction models for ICU patients. The following elements of the selected studies were extracted: author, year of publication, model name, the timing of predictor measurement, follow-up duration, delirium measurement, predictors used in model development, machine learning method, and AUROC of developed models.

Table 1 Delirium prediction models for intensive care unit patients

Author (year)/ Model name	Timing of predictor measurement	Follow-up duration	Delirium measurement	Predictors used in model development	Machine learning method	AUROC [95% CI]
van den Boogaard et al. (2012)/ PRE-DELIRIC model	Within 24 hours of ICU admission	During ICU stay from 24 hours after ICU admission	CAM-ICU or treated with haloperidol	・Predisposing factors: age ・Disease-related factors: APACHE-II score, admission category, urgent admission, infection ・Precipitating factors: coma, use of sedatives and morphine, metabolic acidosis, BUN	Logistic regression (scoring formula)	・Development set: 0.87 [0.85, 0.89] ・Internal validation: 0.84 [0.82, 0.87]
Chen et al. (2017)	Within 24 hours after ICU admission	During ICU stay from 24 hours after ICU admission	CAM-ICU	・Predisposing factors: age, history of hypertension, delirium, and dementia ・Disease-related factors: APACHE-II score, coma, emergency operation, multiple traumas ・Precipitating factors: mechanical ventilation, use of dexmedetomidine hydrochloride, metabolic acidosis	Logistic regression (scoring formula)	・Internal validation: 0.78 [0.72, 0.83]
Oh et al. (2018)	Within 24 hours after ICU admission	During ICU stay	DSM-V and CAM-ICU	・Disease-related factors: heart rate variability	Support vector machine	—[a]

Table 1 Delirium risk prediction models for intensive care unit patients (continued)

Author (year)/ Model name	Timing of predictor measurement	Follow-up duration	Delirium measurement	Predictors used in model development	Machine learning method	AUROC [95% CI]
Cherak et al. (2020)	Within 24 hours after ICU admission	During ICU stay	Intensive Care Delirium Screening Checklist (ICDSC)	・Predisposing factors: age, sex, pre-existing neuropsychiatric disorder ・Disease-related factors: APACHE II score, Sequential Organ Failure Assessment score, Charlson Comorbidity Index, continuous renal replacement therapy ・Precipitating factors: vasoactive medication use, invasive mechanical ventilation ・Nursing data: Glasgow Coma Scale	Logistic regression	・Entire cohort set: 0.76 [0.75, 0.77]
Wassenaar et al. (2015)/ E-PRE-DELIRIC model	At the time of ICU admission	During ICU stay	CAM-ICU or treat with other anti-psychotics	・Predisposing factors: age, history of cognitive impairment, history of alcohol abuse ・Disease-related factors: admission category, urgent admission, respiratory failure ・Precipitating factors: corticosteroids use, BUN ・Nursing data: mean arterial blood pressure at the time of ICU admission	Logistic regression (scoring formula)	・Development set: 0.76 [0.73, 0.77] ・Internal validation: 0.75 [0.71, 0.79]

T**able 1** Delirium prediction models for intensive care unit patients (continued)

Author (year)/ Model name	Timing of predictor measurement	Follow-up duration	Delirium measurement	Predictors used in model development	Machine learning method	AUROC [95% CI]
Hur et al. (2021)/PRIDE	Within four hours after ICU admission	Within four to 24 hours after ICU admission	CAM-ICU	・Predisposing factors: age, sex ・Disease-related factors: admission category, reason for ICU admission, Charlson Comorbidity Index ・Precipitating factors: invasive mechanical ventilation, medications, laboratory test ・Nursing data: vital signs, Glasgow Coma Scale (eye, verbal, and motor)	Random forest [b]	・Internal validation: 0.92 [0.91, 0.92) ・External validation: 0.72 [0.71, 0.72]
Fan et al. (2019)/ DYNAMIC-ICU	Time-varying	During ICU stay	CAM-ICU	・Predisposing factors: chronic diseases history, hearing deficits ・Disease-related factors: infection, APACHE II score ・Precipitating factors: sedatives and analgesics use, indwelling catheter use, sleep disturbance (environmental element)	Logistic regression (scoring formula)	・Internal validation: 0.90 [0.86, 0.94]

T**able 1** Delirium prediction models for intensive care unit patients (continued)

Author (year)/ Model name	Timing of predictor measurement	Follow-up duration	Delirium measurement	Predictors used in model development	Machine learning method	AUROC [95% CI]
Moon et al. (2018)/Auto-DelRAS	During ICU stay (first, last, and maximum)	During ICU stay	CAM-ICU	・Predisposing factors: age, education ・Disease-related factors: infection ・Precipitating factors: surgery, medical ICU admission, BUN, number of catheters, restraint use, psychopharmacology drug use ・Nursing data: level of consciousness, pulse rate, activity level	Logistic regression (scoring formula)	・Internal validation: 0.90 ・External validation: 0.72 ・Post-implementation 1-year: 0.85─0.88

Note. APACHE-II = Acute Physiology and Chronic Health Evaluation-II; AUROC = area under the receiver operating characteristic; BUN = blood urea nitrogen, CAM-ICU = Confusion Assessment Method for the intensive care unit; CI = confidence interval; DSM-V = Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition; ICU = intensive care unit.

[a] Best-balanced accuracy: 74.83%. [b] Best machine learning technique among developed models.

The predictors used in model development were analyzed and classified into predisposing factors, precipitating factors, disease-related factors, and nursing data. Out of the eight selected studies, three had retrospective designs (Moon et al., 2018; Cherak et al., 2020; Hur et al., 2021), and the remaining five had prospective designs. The studies were classified into three categories according to the timing of the predictor measurement used in model development; data obtained during the first 24 hours of ICU admission (van den Boogaard et al., 2012; Chen et al., 2017; Oh et al., 2018; Cherak et al., 2020); data obtained during the first four hours of ICU admission (Wassenaar et al., 2015; Hur et al., 2021); and daily time-varying data (Moon et al., 2018; Fan et al., 2019).

The DYNAMIC-ICU model (Fan et al., 2019) was developed by accumulating information from the point of ICU admission until the point of ICU discharge, to reflect the time-varying health status of ICU patients compared to other static models. This model obtained data on predisposing factors at admission, disease-related factors at the time of enrollment in the study, and precipitating and environmental factors on a daily basis during the ICU stay. Additionally, the data of time variation variables used the most abnormal values as the potential risk factors. The model emphasized delirium prediction rules by using cumulative data to represent the patients' dynamic conditions and stratified risk of delirium into low-, moderate-, and high-risk.

The Auto-DelRAS model was an automatic delirium risk scoring algorithm with potential predictors (Moon et al., 2018). The predicted delirium score was calculated by including data about age, education level, level of consciousness, pulse rate, activity level,

admission to a medical department, BUN level, presence/absence of infection, total number of catheters, use of physical restraints, and use of psychopharmacological drugs. The scores, including five default variables, were categorized into a high-, moderate-, and low-risk group. The Auto-DelRAS model was applied to clinical practice for 1 year and its predictive validity was evaluated using the first, last, and maximum values; the first values were extracted at ICU admission, and the last and maximum values were extracted before delirium onset or discharge. The results of the Auto-DelRAS models' performance were similar regardless of the extraction time points of the predictors.

All studies included in the literature review defined delirium using the CAM-ICU or ICDSC. In addition to the delirium screening tools, two studies defined delirium by the administration of antipsychotic medications (van den Boogaard et al., 2012; Wassenaar et al., 2015), and one defined delirium using the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-V) diagnosis (Oh et al., 2018).

The predictors used in the model developments varied between the included studies. Predisposing, precipitating, and disease-related factors were mainly used for model development. Four studies additionally used nursing data, such as GCS scores, activity level, and vital signs (Wassenaar et al., 2015; Moon et al., 2018; Cherak et al., 2020; Hur et al., 2021). Five models created a formula using the score assigned to each predictor based on the odds ratio values from the logistic regression (van den Boogaard et al., 2012; Wassenaar et al., 2015; Chen et al., 2017; Moon et al., 2018; Fan et al., 2019). The

models' AUROCs were 0.70 or higher, regardless of the time of delirium onset or predictors.

In summary, delirium prediction models were developed using predictors based on previously identified risk factors for delirium. Most were static models that did not reflect time variation. Some studies used additional predictors such as GCS scores, activity levels, and vital signs, whereas nursing documentation patterns were not used as potential predictors. As such, the use of nursing data for the development of delirium prediction models is limited, despite its clinical importance. The condition of patients in the ICUs can change rapidly (Awad et al., 2017), and physiological indicators alone have limited capacity to predict clinical prognosis. Therefore, developing delirium prediction models using nursing data that reflect nurses' clinical judgments and time variation is necessary to accurately predict delirium.

2.3. Prediction models using nursing data

Through continuous monitoring and direct contact with patients, nurses are often the first healthcare providers in the ICU to recognize patient deterioration (Douw et al., 2015). Therefore, nursing data represent patients' conditions, including nursing assessment records on objective and subjective patient conditions and documentation patterns on nurses' clinical judgments through the nursing process. Nursing assessments are recorded regularly during hospitalization, so patients' deterioration can be detected early through

the nursing records (Capan et al., 2017; Lewis et al., 2017). Using nursing flowsheets or records of medication administration, a previous study identified nursing documentation patterns associated with ICU patients' deterioration and classified nursing documentation patterns by frequency or intervention (Rossetti et al., 2021; Schnock et al., 2021). They reported that the frequency patterns indicated an increase in the documentation frequency of a single vital sign, a complete set of vital signs, and comments. Intervention patterns included the Pro re nata (PRN) medication administration and the withholding scheduled medications.

A systematic review including 170 studies investigated signs and symptoms with the potential to trigger nurses' concern about patients' clinical conditions. One-hundred and seventy symptoms and signs were extracted and classified into the following ten general indicators: changes in breathing; circulation; and mentation, temperature (rigors), agitation, pain; lack of progress, patient (indicates) not feeling well, subjective nurse observation, and subject knowledge (Douw et al., 2015). Six of the ten indicators can be identified through nursing assessment records in EMRs. Four indicators may or may not be present in EMRs, such as documentation patterns, and can be determined indirectly. Nurses increase the frequency of their surveillance activities due to concerns about patients (Schnock et al., 2021). Therefore, identifying documentation patterns in nursing records may be an important signal which predicts worsening patient deterioration (Romero-Brufau et al., 2019; Rossetti et al., 2021).

Douw et al. (2015) conducted a retrospective cohort study to compare model

performance using the vital sign-based Early Warning Score (EWS) and nurses' concern indicators (Douw et al., 2016). They reported that when the EWS and nurses' concern indicators were added as predictors, the AUROC for unplanned ICU admissions or unexpected mortality was higher (0.91) than EWS alone (0.86). This finding is in line with the European Resuscitation Council Guidelines 2021, which stated that hospitals should authorize healthcare providers to seek help based on their concerns and patients' vital signs (Soar et al., 2021).

A study used the Medical Information Mart for Intensive Care (MIMIC)-III database to develop machine learning-based prediction models using nursing notes and clinician notes to predict the length of stay and mortality (Huang et al., 2021). The models achieved AUROCs of 0.826 (nursing notes) and 0.796 (clinician notes), indicating that nursing notes showed greater predictive power than clinician notes. Nursing records contain clinically rich features that can be used to predict patients' outcomes. MIMIC databases containing information for ICU patients have been used to predict ICU readmission (Desautels et al., 2017; Rojas et al., 2018) and sepsis (Han et al., 2018; Garcia-Gallo et al., 2020) using machine learning techniques, and compare treatment effects and clinical outcomes (Song et al., 2019; Baker et al., 2020).

The current study performed a literature review to identify the machine learning-based prediction models using nursing data, and the delirium prediction models presented in Table 1 were excluded (Table 2). Studies that used only nursing notes or validation-only studies were excluded. Through the literature selection process, 11 studies were

included in this current review (Rothman et al., 2013; Zadravecz et al., 2015; Capan et al., 2017; Horng et al., 2017; Wellner et al., 2017; Beauchet et al., 2018; Rojas et al., 2018; Fu et al., 2021; Heyming et al., 2021; Song et al., 2021; Xu et al., 2022). The predictive purposes varied between studies and included the following: unplanned ICU admission, mortality, pressure injury, fall, and infection. Nursing data used in model development were categorized as nursing assessments and nursing documentation patterns.

For model development, one study used nursing documentation patterns (Fu et al., 2021), and the remaining ten studies used nursing assessments. Nursing assessments consisted of documentation of mainly head-to-toe examination, GCS, Braden scale score, fall, pain, and vital signs. Nursing documentation patterns included increased entries of vital signs and nursing flowsheet comments related to vital signs.

The results of the literature review demonstrated that patients' clinical outcomes could be predicted based on machine learning methods using nursing data in the form of nursing assessments and nursing documentation patterns as significant predictors. In summary, various types of nursing data have been used to predict clinical outcomes via machine learning methods. To develop delirium prediction models using nursing data, suitable types of nursing data must be selected after a review of potential predictors, and the predictive performance of the developed models should be validated.

Table 2 Summary of nursing data used in prediction models

Author (year)	Population	Prediction goal	Nursing data
Rojas et al. (2018)	ICU patients	Unplanned ICU readmission	・Nursing assessments: Braden scale score, Morse fall score, abdominal physical exam, cardiac rhythm
Wellner et al. (2017)	Inpatients	Unplanned ICU readmission	・Nursing assessments: GCS (eye, verbal), pupil reaction, level of consciousness, orientation, Braden risk, activity, mobility, retraction, moisture, skin, friction sheer, pain score, nutrition, perfusion cap refill/color/temperature, cough, FLAAC (Faces, Legs, Activity, Cry, Consolability), fluid balance, cardiovascular, pulse rate, heart rhythm, brachial/femoral/peripheral pulse, neurological, neurovascular check, respirations, work of breathing, secretion/sputum color
Fu et al. (2021)	ICU patients	Mortality, cardiac arrest, and rapid response team calls	・Nursing assessments: nursing notes ・Nursing documentation patterns: documentation frequency of vital signs and related comments
Capan et al. (2017)	Adult inpatients	Rapid Response Team activation, Code Blue activation, readmission, mortality	・Nursing assessments: neurological, respiratory, food (intake, swallowing), gastrointestinal, musculoskeletal, and genitourinary assessments, Braden scale score, Schmid score (patient safety)
Heyming et al. (2021)	Emergency department patients	Disposition of patients	・Nursing assessments: use of oxygen device, capillary refill, general appearance, level of consciousness, skin
Xu et al. (2022)	ICU patients	Pressure injury	・Nursing assessments: Braden scale score, GCS
Song et al. (2021)	Inpatients	Pressure injury	・Nursing assessments: GCS, level of consciousness, gait/transferring, activity

Table 2 Summary of nursing data used in prediction models (continued)

Author (year)	Population	Prediction goal	Nursing data
Beauchet et al. (2018)	Older inpatients	Fall	・Nursing assessments: history of falls during the past six months, mobility, five-times-sit-to-stand test, cognitive impairment, use of formal or informal home and social services
Horng et al. (2017)	Emergency department patients	Infection	・Nursing assessments: pain scale, free text on nursing assessment
Zadravecz et al. (2015)	Inpatients	Mortality	・Nursing assessments: GCS (total, Eye, Verbal, and Motor), Richmond Agitation-Sedation Scale
Rothman et al. (2013)	Inpatients	Mortality	・Nursing assessments: cardiac, food/nutrition, gastrointestinal, genitourinary, musculoskeletal, neurological, peripheral-vascular, psychosocial, respiratory, safety/fall risk, and skin/tissue assessments

Note. ICU = intensive care unit; GCS = Glasgow Coma Scale.

III. CONCEPTUAL FRAMEWORK

3.1. Theoretical framework

The HPM-ExpertSignals model (Rossetti et al., 2021), which is based on Donabedian's structure-process-outcome model (Donabedian, 1966), is driven by the clinician's knowledge-based behaviors and focuses on information generated by the clinical decision making process (Figure 1). In this process, clinicians increase the frequency of their surveillance based on their concern about the perceived potential risk signals of patients. As a result, clinicians increase their documentation frequency or documentation entry at uncommon times, and this information appears as a temporary pattern in the EMR. This information is affected by environmental and individual modifiers. Environmental modifiers include hospital setting, standards of care, and hospital policy, while individual modifiers include patient characteristics, physiological and disease prognosis, and clinician characteristics. This model emphasizes identifying, interpreting, and utilizing information on clinical behavior patterns represented in EMR data for clinical outcome prediction (Rossetti et al., 2021).

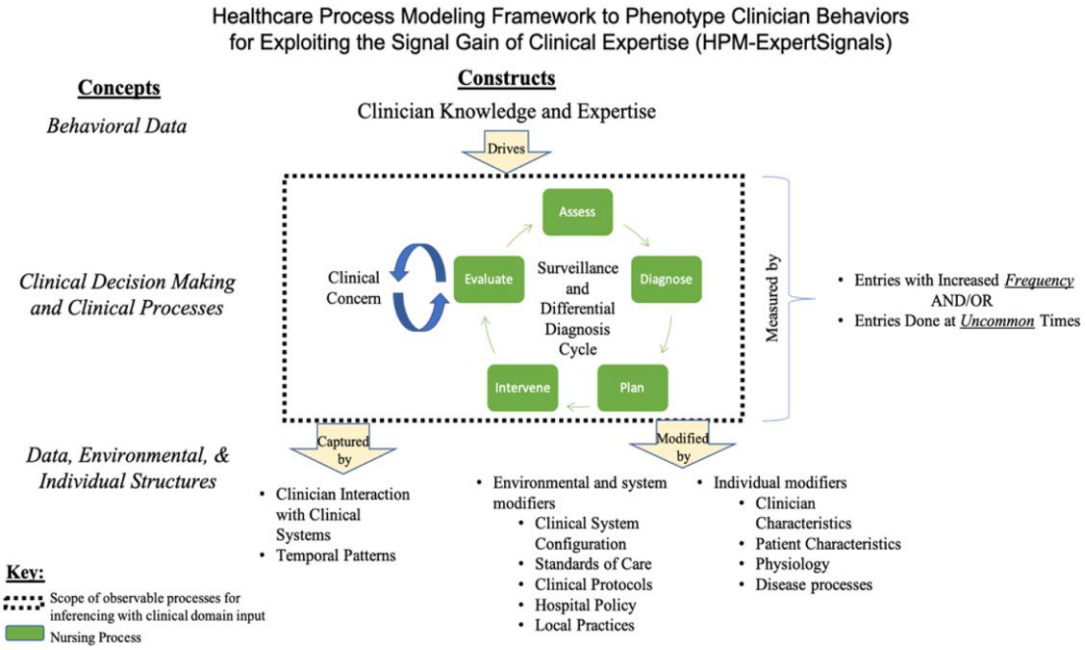


Figure 1 HPM-ExpertSignals Model

Note. Healthcare Process Modeling Framework to Phenotype Clinician Behaviors for Exploiting the Signal Gain of Clinical Expertise (HPM-ExpertSignals); Source: Rossetti et al., (2021).

3.2. Conceptual framework of this study

The conceptual framework of the delirium prediction models presented in the current study used a modified framework based on the HPM-ExpertSignals model (Figure 2). This framework assumes that nurses conduct surveillance activities based on their clinical concerns focused on delirium and record the information related to their surveillance activities in the EMR. Its information patterns recorded vary for each patient according to the nurses' concerns and environmental and individual modifiers. Therefore, the data patterns captured in the EMR represent the patient's condition.

The nursing data consisted of nursing assessments and nursing documentation patterns. Nursing assessments included pain, GCS, pressure injury, Richmond Agitation-Sedation Scale (RASS), and vital signs (blood pressure, pulse rate, respiratory rate, body temperature, and oxygen saturation). The nursing documentation patterns, which were identified in the nursing flowsheet and records of medication administration in the EMR, included frequency and intervention. Frequency patterns included entries in the nursing data related to vital signs, GCS, and RASS. The intervention patterns of medication administration, such as PRN/stat medication and withholding scheduled medication, were associated with temporal activities with changes in patients' conditions.

Environmental modifiers in the framework were defined as precipitating factors for delirium, and individual modifiers in the framework were defined as predisposing factors for delirium. The precipitating factors were related to the hospital environment (Inouye & Charpentier, 1996; Green et al., 2019), such as operation, use of mechanical ventilation,

foley catheter, physical restraints, blood transfusion, and serum creatinine. The predisposing factors refer to a vulnerability that a patient had at admission (Inouye & Charpentier, 1996), such as age, gender, history of dementia and hypertension, visual/hearing deficits, and history of falls.

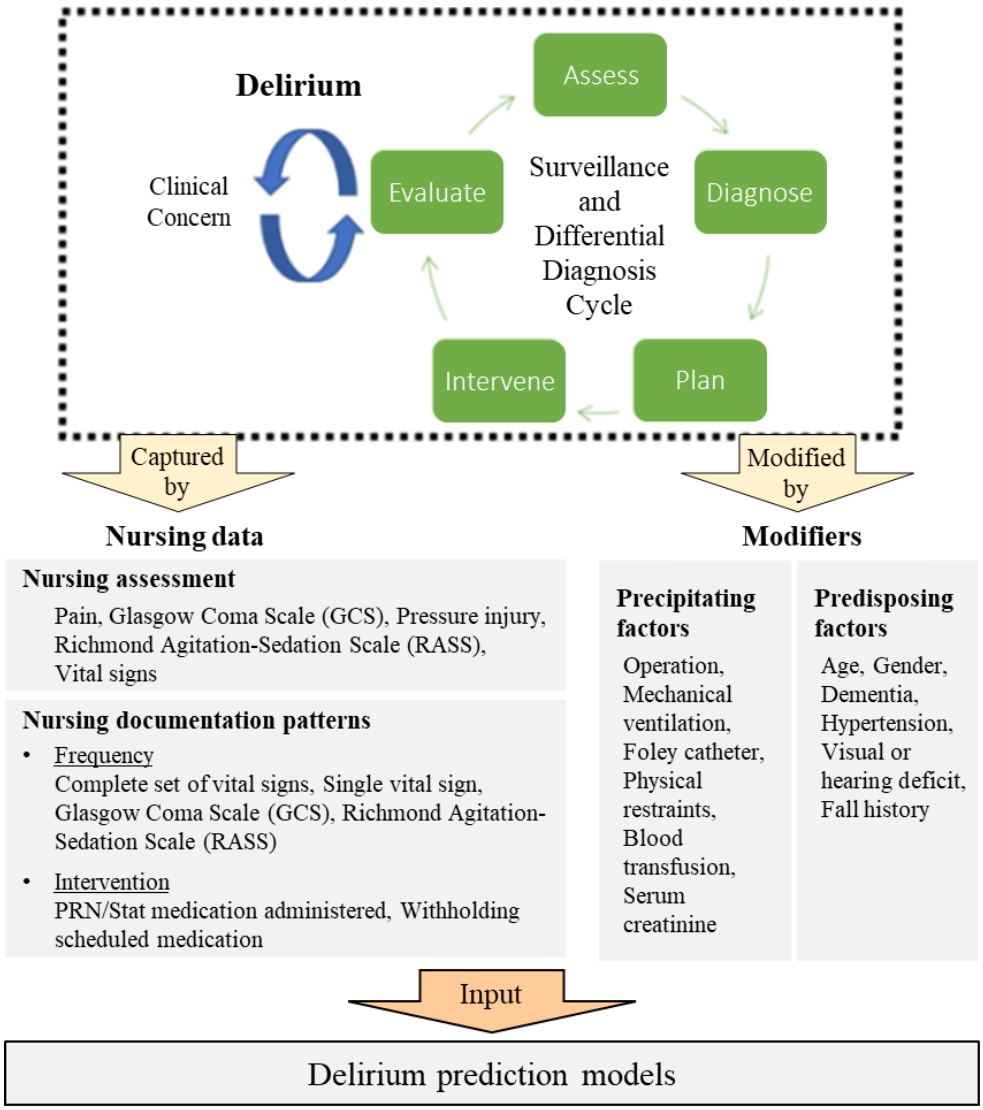


Figure 2 Conceptual framework for delirium prediction

IV. METHODS

4.1. Study design

This study was a retrospective cohort study that used nursing data from EMRs to develop machine learning-based delirium prediction models with time variation for ICU patients. Reporting followed the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) reporting guideline (Collins et al., 2015) (Appendix 1).

4.2. Data sources

This study used the MIMIC-IV database (version 1.0) and an EMR from a single tertiary hospital in Seoul, South Korea. The MIMIC-IV database was used for model development and internal validation. Meanwhile, the EMR was used for external validation to evaluate the performance of the developed models and explore their domestic applicability.

4.2.1. Development and internal validation dataset

The MIMIC-IV database includes 256,878 patients (523,740 admissions) admitted from January 2008 to December 2019 at the Beth Israel Deaconess Medical Center (BIDMC) in Boston, Massachusetts, USA (Goldberger et al., 2000; Johnson et al., 2021). It is a large publicly available electronic health record (EHR) grouped into five modules (core, hosp, icu, ed, cxr, and note), of which three modules (core, hosp, and icu) were selected considering the aim of this study (Appendix 2). The selected modules consisted of 27 tables linked by unique identifiers, such as SUBJECT_ID. All information was deidentified, and all dates were shifted to the future.

MIMIC-IV version 0.3 is an updated version of MIMIC-III, released in August 2020, that changed the modular approach to improving data usability, while MIMIC-IV version 1.0 is the current version released in March 2021.

4.2.2. External validation dataset

The external validation dataset used the EMR of patients admitted to the ICU in a single tertiary hospital in Seoul, South Korea. This EMR contains the clinical records of inpatients, outpatients, and emergency patients. Patients admitted to medical, surgical, and neurologic ICUs from March 2018 to August 2021 were selected, considering when the ICDSC was first used in each ICU. All information was deidentified to protect patient privacy by the Department of Digital Health of Severance Hospital in South Korea.

4.2.3. Ethical consideration

The MIMIC database was anonymized to be Health Insurance Portability and Accountability Act (HIPPA) compliant by the Massachusetts Institute of Technology (MIT). Access to the MIMIC-IV database was obtained by PhysioNet (http://physionet.org) after fulfilling the Collaborative Institutional Training Initiative (CITI) program (Appendix 3). This study received additional ethical approval from the Institutional Review Board of the Yonsei University Health System (approval number: 4-2021-1212) and the Data utilization Review Board (DRB; approval number, 2022100100) (Appendix 4). Finally, this study was conducted in accordance with the Declaration of Helsinki and ethical principles for medical research involving human subjects.

4.3. Cohort selection

The eligibility criteria for cohort selection is presented in Table 3. Patients who were aged between 18 and 89 years and admitted to the ICU for at least 24 hours were included. As the MIMIC database obscures the actual age of patients over 89 years (Johnson et al., 2021), these patients were excluded. Patients who received hospice or palliative care, in which clinicians focused more on patient comfort than clinical outcomes, were excluded (Fu et al., 2021). Patients with RASS score of -4 (deep sedation) or -5 (unarousable) or patients with delirium onset during the first 24 hours after ICU admission were excluded.

Table 3 Eligibility criteria

Inclusion criteria	Exclusion criteria
• 18 ≤ Age < 89 years old • ICU Length of stay ≥ 24 hours • Delirium screening tools [a] record ≥ 1 during the first 24 hours after ICU admission	• Hospice or palliative care patients • RASS score of -4 or -5 during the first 24 hours after ICU admission • Delirium during the first 24 hours after ICU admission

Note. ICU = intensive care unit; RASS = Richmond Agitation-Sedation Scale.

[a] Confusion Assessment Method for the intensive care unit (CAM-ICU) or the Intensive Care Delirium Screening Checklist (ICDSC).

Given that each patient could have multiple ICU admissions during the same hospitalization period, the first ICU admission was included in the cohort. The eligibility criteria for the development and internal validation cohorts (hereinafter, development cohort) were also applied to the external validation cohort. The development dataset was randomly split into a training set (80%) and a test set (20%). The training set was used for model development, while the test set was used for internal validation only (Figure 3).

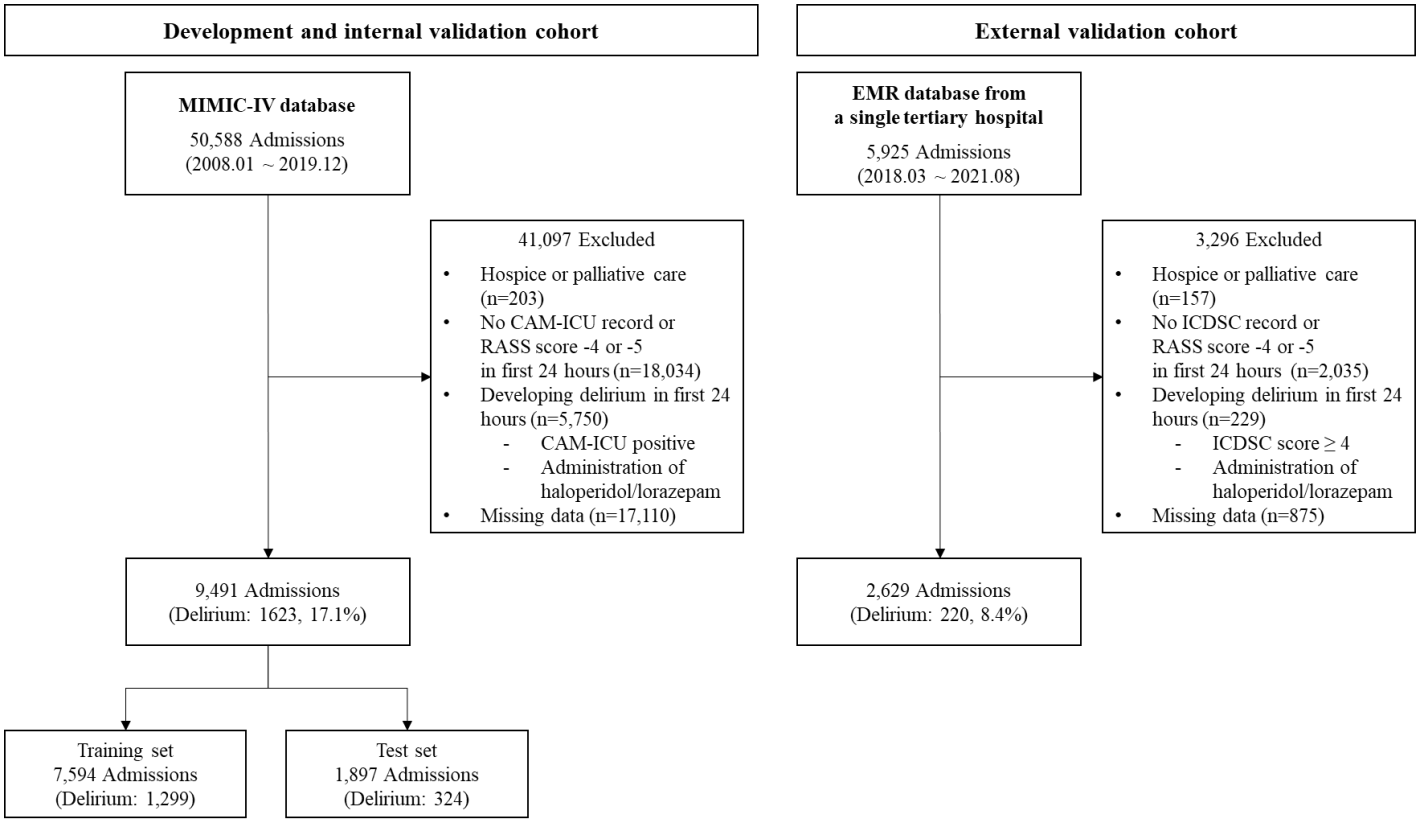


Figure 3 Process of cohort selection

Note. CAM-ICU = confusion assessment method for the intensive care unit; EMR = electronic medical record; ICDSC = intensive care delirium screening checklist; RASS = Richmond Agitation-Sedation Scale.

4.4. Delirium

The outcome of this study was delirium occurrence during ICU stay from 24 hours after ICU admission. The development and external validation cohorts used the CAM-ICU and ICDSC to screen delirium, respectively. Both CAM-ICU and ICDSC have validated ICU delirium screening tools and are recommended for use in the ICU despite their methodological differences in delirium screening (Gusmao-Flores et al., 2012; Chen et al., 2021) (Table 4).

The CAM-ICU categorizes four features: acute changes or fluctuations in mental status (Feature 1), inattention (Feature 2), disorganized thinking (Feature 3), and an altered level of consciousness (Feature 4). The diagnosis of delirium is positive if the patient manifests Features 1 and 2, along with Features 3 or 4 (Ely et al., 2001). The ICDSC consists of eight items based on DSM criteria. Each item is given zero or one point (score range: 0–8), and a total score of 4 or more confirms delirium (Bergeron et al., 2001). The CAM-ICU or ICDSC was not assessed when the RASS score was -4 or -5, which indicates deep sedation. Assessment time and frequency vary according to each hospital protocol.

Table 4 Comparison of delirium screening tools

Tools	CAM-ICU	ICDSC
Features or items	1. Acute changes or fluctuations in mental status 2. Inattention 3. Disorganized thinking 4. Altered level of consciousness	1. Altered level of consciousness 2. Inattention 3. Disorientation 4. Hallucination, delusion, or psychosis 5. Psychomotor agitation or retardation 6. Inappropriate speech or mood 7. Sleep/wake cycle disturbance 8. Symptom fluctuation
Delirium positive or cut-off score	Feature 1 and Feature 2 and either Feature 3 or Feature 4 are presented	Total score ≥ 4
Model performance	Sensitivity: 95–100% Specificity: 89–93% Accuracy: 95–96%	Sensitivity: 99% Specificity: 64% AUROC: 0.902

Note. AUROC = area under the receiver operating characteristic; CAM-ICU = confusion assessment method for the intensive care unit; ICDSC = intensive care delirium screening checklist. Source: Bergeron et al., (2001) and Ely et al., (2001).

In the external validation cohort, the ICDSC was evaluated every eight hours through continuous observation and recorded at the end of each shift. Furthermore, delirium assessment is performed during the ICU stay, except during the admission shift. Therefore, at least 16–24 hours are required to evaluate delirium using assessment tools after admission. The CAM-ICU is assessed at a one-time point based on the observation and evaluated two or three times a day.

Most delirium prediction models used the CAM-ICU (van den Boogaard et al., 2012; Wassenaar et al., 2015; Chen et al., 2017; Oh et al., 2018; Fan et al., 2019; Hur et al., 2021) or the ICDSC (Cherak et al., 2020) to define delirium occurrence. Note that ICU patients are routinely assessed using delirium screening tools during their ICU stays, and haloperidol or lorazepam are only used for the treatment of delirium in the ICU (Wassenaar et al., 2015; Kim et al., 2017; Collet et al., 2018). Therefore, delirium occurrence in the current study was defined as a CAM-ICU positive result, an ICDSC score of four or greater, or the administration of haloperidol or lorazepam.

4.5. Potential predictors

Predictors were selected based on the literature review, including nursing assessments, nursing documentation patterns, precipitating factors (clinical characteristics and laboratory test), and predisposing factors (demographics and clinical history) (Table 5).

In nursing assessments, pressure injury was recorded using the Braden scale in the development and external validation cohorts. Vital signs included blood pressure, pulse rate, respiratory rate, body temperature, and oxygen saturation. Among predisposing factors, history of dementia (Steinmetz et al., 2013) and hypertension (Yang et al., 2010) are defined by the International Classification of Diseases (ICD) codes, which were diagnosed in previous hospitalizations at the same hospital (Appendix 5). The final dataset in both cohorts included all these predictors.

Table 5 Summary of potential predictors

Category	Potential predictors	Type of data	MIMIC-IV tables
Nursing data			
Nursing assessments	Pain	Continuous	chartevents
	GCS	Continuous	chartevents
	Pressure injury	Continuous	chartevents
	RASS	Continuous	chartevents
	Vital signs [a]	Continuous	chartevents
Nursing documentation patterns	Frequency	Discrete	emar
	• Complete set of vital signs [a]		
	• Sigle vital sign		
	• GCS, RASS		
	Intervention	Continuous	chartevents
	• PRN/stat medication		
	• Withholding scheduled medication		
Modifiers			
Precipitating factors	Clinical characteristics		
	• Operation	Discrete	procedureevents
	• Mechanical ventilation	Discrete	chartevents
	• Foley catheter	Discrete	outputevents, datetimeevents
	• Physical restraints	Discrete	chartevents
	• Blood transfusion	Discrete	Inputevents, chartevents
	Laboratory test		
	• Serum creatinine	Continuous	labevents
Predisposing factors	Demographics		
	• Age	Continuous	patients
	• Gender	Discrete	patients
	Clinical history		
	• History of dementia	Discrete	diagnoses_icd
	• History of hypertension		
	• Visual or hearing deficit	Discrete	chartevents
	• History of fall		

Note. GCS = Glasgow Coma Scale; PRN = Pro re nata; RASS = Richmond Agitation-Sedation Scale.

[a] blood pressure, pulse rate, respiratory rate, body temperature, and oxygen saturation.

4.6. Sample size calculation

Larger sample size and sufficient data quality lead to the development of more robust prediction models (Riley et al., 2020). The sample size calculation of a rule of thumb for the binary outcome suggested ensuring at least ten events per variable (EPV) (Peduzzi et al., 1996) or ten events per predictor parameter (EPP) using all candidate predictors before any variable selection for model development (Riley et al., 2019). These methods have been widely used owing to their simplicity. However, these methods did not reflect actual context characteristics such as the total number of participants, outcome incidence, and the expected predictive performance of the model (Riley et al., 2020).

Therefore, the current study used the sample size calculation method proposed by Riley et al. (2020) to minimize overfitting and ensure precision for developing robust delirium risk prediction models for ICU patients. The sample size was calculated using the overall outcome proportion (0.17), the number of potential predictors (40), the shrinkage (default 0.9), and the target small expected optimism in the apparent R^2 (0.05). The R package *pmsampsize* was used for calculating the minimum sample size (Table 6). The results indicated that at least 6999 samples were required, corresponding to 1253 delirium occurrences and an EPP of 31.32.

Table 6 Calculation of the minimum sample size in R program

```
pmsampsize (type = "b", prevalence = 0.17, parameters = 40, shrinkage = 0.9,
                rsquared = 0.05)
```

In the current study, the development cohort consisted of 9491 admissions after cohort selection; this fulfilled the minimum sample size.

4.7. Data preprocessing

Data preprocessing involved cleaning the dataset by removing incomplete records to improve its quality (García et al., 2016; Mufti et al., 2019). This process is essential because a preprocessed final dataset allows the machine learning algorithm to operate stably and appropriately. In the current study, the data preprocessing steps were performed through outlier detection, handling of missing data, data balancing, and feature selection. The same data preprocessing steps were performed in both the development and the external validation cohorts. Complete datasets were reliable and suitable for model development and validation.

4.7.1. Outlier detection

An outlier is an unusual datum inconsistent with the remaining dataset (Barnett & Lewis, 1984). Outliers may decrease model performance and increase error variance; therefore, outlier identification is critical before statistical analyses. Outliers originate

from equipment errors, human errors, and natural variations within the patients (Salgado et al., 2016b).

This study assessed scale scores, such as pain, GCS, pressure injury, and RASS, based on the score range of each scale to detect outlier values. The actual value of vital signs used scientifically valid values (positive) (Kuhn & Johnson, 2013) and inter-quartile range (IQR; 25–75 percentile) to detect outliers (Steyerberg, 2009). Therefore, negative values were removed for vital signs, followed by the upper and lower extreme quartiles (Q1, Q3). In contrast, outliers were retained for potential predictors, which used documentation frequency rather than actual values, considering various entry errors.

4.7.2. Handling missing values

Most machine learning operates only on complete data, and it is necessary to handle missing values (Salgado et al., 2016a). Because most values affect each other, missing values should be determined by considering the relationship between variables (Little & Rubin, 2019). This study used the methods of dropping null values for vital signs and serum creatinine and assigning specific values for scale scores, such as pain, GCS, pressure injury, and RASS.

A total of 17110 admissions did not contain any actual values of vital signs or serum creatinine; therefore, these cases were excluded. In clinical practice, nursing assessment tools may not be measured when the patient is in a good or stable condition; therefore, the

null value can be replaced with the normal or best value. When pain and RASS scores were not recorded in the dataset, zero was assigned to replace the null value. The pressure injury score was measured using the Braden scale (range: 6–23), and a value of 23 was assigned to replace the null value. GCS total, GCS eye, GCS verbal, and GCS motor were assigned scores of 15, 4, 5, and 6, respectively, to replace the null value.

4.7.3. Data balancing

Class imbalanced data distribution is a common problem that yields biased results because classification algorithms are optimized towards the majority class (López et al., 2013). Imbalanced class modification techniques are used to alleviate the bias of algorithms (Mufti et al., 2019) by adding new or removing existing samples (Longadge & Dongre, 2013). The main sampling techniques are oversampling, undersampling, or combination sampling (Longadge & Dongre, 2013). Oversampling involves artificially replicating the number of cases of the minority class, undersampling involves decreasing the number of cases of the majority class, and combination sampling involves over and under sampling techniques (Longadge & Dongre, 2013; El-Rashidy et al., 2020). However, oversampling approaches may induce overfitting, while undersampling approaches may have the risk of losing valuable data (García et al., 2016). Other studies using the MIMIC database also used class-balancing techniques to resolve the class imbalance problem (Sundararaman et al., 2018; Xu et al., 2019; Tsiklidis et al., 2022).

In the development cohort, the delirium class distribution was imbalanced (17.0% of the patients experienced delirium). Therefore, the current study used a combination of the Synthetic Minority Over-sampling Technique (SMOTE) with Tomek links (Batista et al., 2004). First, the SMOTE (oversampling) technique was applied to create certain numbers in the minority class (Chawla et al., 2002). Thus, the minority class boundaries were spread into the majority class space to avoid overfitting. Then, Tomek-links (under sampling) was applied to remove overlapping samples. If the samples belonging to different classes are close in distance (Tomek-link), the samples belonging to the majority class are considered noisy and eliminated. Therefore, Tomek-links solve the problem of overlapping between classes that appear as samples generated without considering the distribution of the majority class of samples in SMOTE (Batista et al., 2004).

After cohort selection, 7594 admissions were included in the training set, from which 17.0% of the patients experienced delirium. Combination sampling was applied to maintain data balancing. The detailed process is shown in Figure 4. A total of 9752 cases were sampled for the final model training, from which 6495 cases experienced delirium.

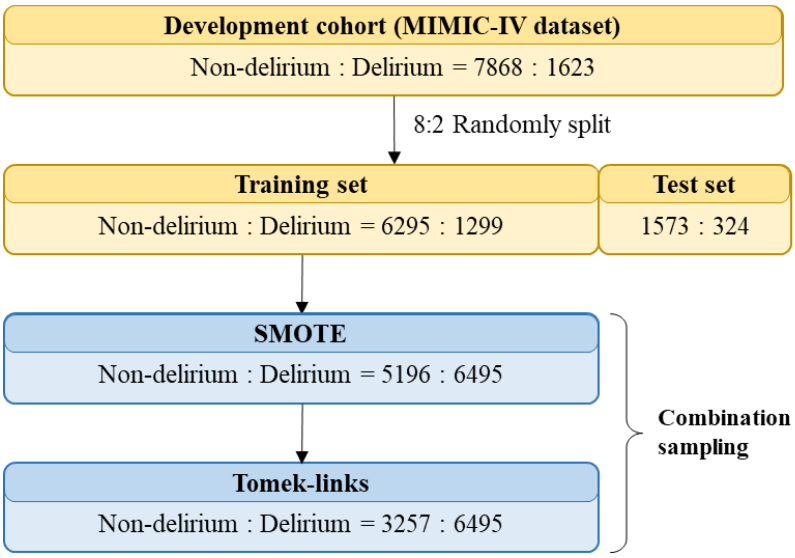


Figure 4 SMOTE with Tomek links technique

4.7.4. Feature selection

Feature selection step is an important component of machine learning and involves selecting the most critical features for model prediction. Consequently, unnecessary variables are removed to reduce the data and the training time (Stevens et al., 2020). Additionally, the risk of overfitting can be reduced through this process (García et al., 2016).

This study identified 40 clinical variables among potential predictors through a literature review, the researcher's clinical knowledge and experience, and available data in both cohorts. It includes nursing assessments, nursing documentation patterns, precipitating factors (clinical characteristics and laboratory test), and predisposing factors (demographics and clinical history) (Table 7).

Table 7 Summary of selected features

Category	Predictor (range)	Possible value
Nursing data		
Nursing assessments	Pain (0–10)	Maximum and median values
	GCS Total (3–15)	First, last, maximum, and minimum values
	GCS Eye response (1–4)	Minimum value
	GCS Verbal response (1–5)	Maximum and minimum values
	GCS Motor response (1–6)	Minimum value
	Pressure injury (6–23)	Minimum value
	RASS (-5 – +4)	Minimum value
	Systolic blood pressure, body temperature, respiratory rate, Pulse rate	Maximum value Maximum and median values
Nursing documentation patterns	Frequency • Complete set of vital signs • Sigle vital signs: systolic blood pressure, pulse rate, body temperature, respiratory rate, oxygen saturation • GCS, RASS	Frequency/24hours
	Intervention • PRN medication administered • Stat medication administered • Withholding scheduled medication	Yes/No
Modifiers		
Precipitating factors	Clinical characteristics • Operation • Mechanical ventilation • Foley catheter • Physical restraints • Blood transfusion	Yes/No
	Laboratory test • Serum creatinine, mg/dL	Maximum value
Predisposing factors	Demographics • Age, years	
	• Gender	Men/Women
	Clinical history • History of dementia, hypertension, fall, visual or hearing deficit	Yes/No

Note. GCS = Glasgow Coma Scale; PRN = Pro re nata; RASS = Richmond Agitation-Sedation Scale.

Each patient's observation window for the selected features during their ICU stay was based on the sliding window method. A sliding window model is the most frequently used processing method for data streaming (Zazzaro et al., 2021) that makes decisions based on recently observed data elements by appropriately reflecting constantly changing data (Datar et al., 2002). A sliding window moves along the time axis and rearranges the dataset focused on the last time-series points (Cho et al., 2019; Zazzaro et al., 2021).

The current study used a time variation-based sliding window of 24 hours to predict delirium during ICU stay, considering delirium assessment time. Therefore, the delirium group used data for 24 hours before delirium occurred, and the non-delirium group used data for all ICU stays to extract potential predictors (Figure 5). Exceptionally, predisposing factors used data at admission in both groups.

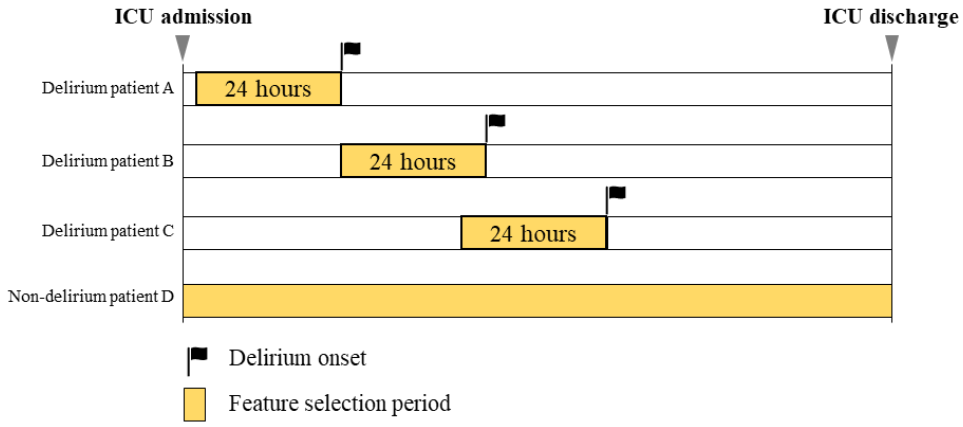


Figure 5 Examples of feature extraction using sliding window method
Note. ICU = intensive care unit.

Further, to compare the performance of the developed prediction models according to the feature selection time of the non-delirium group, features were extracted from the data for the first 24 hours after ICU admission and the last 24 hours before ICU discharge, respectively.

4.8. Model development and external validation

In ICUs, the patient's level of consciousness is periodically evaluated and recorded using GCS and RASS. Compared to other variables, this data can be easily extracted without missing values. Recently, prediction models have used the level of consciousness as a potential predictor or eligibility criteria (Wellner et al., 2017; Song et al., 2021; Xu et al., 2022). Although the level of consciousness was used as an important predictor by several delirium prediction models (Moon et al., 2018; Cherak et al., 2020; Hur et al., 2021), GCS and RASS values have been reported to be inconsistent evidence of delirium risk (Devlin et al., 2018). Therefore, Models I (40 predictors) and II (31 predictors) were developed to compare differences in model performance depending on whether the actual GCS and RASS values were included as predictors or not.

In classification modeling, it is common to separate the dataset into training, validation, and test sets to prevent overfitting of the developed models (Witten & Witten, 2017; Mufti et al., 2019). The training set was primarily used to develop the prediction models. The validation set was used to select the optimized model parameters. The test

set is an independent dataset that is not used in the formation of the algorithms but is used for the internal validation of the developed models (Witten & Witten, 2017). Here, the generated development cohort was randomly divided into a training set (80%) and a test set (20%) (Figure 6).

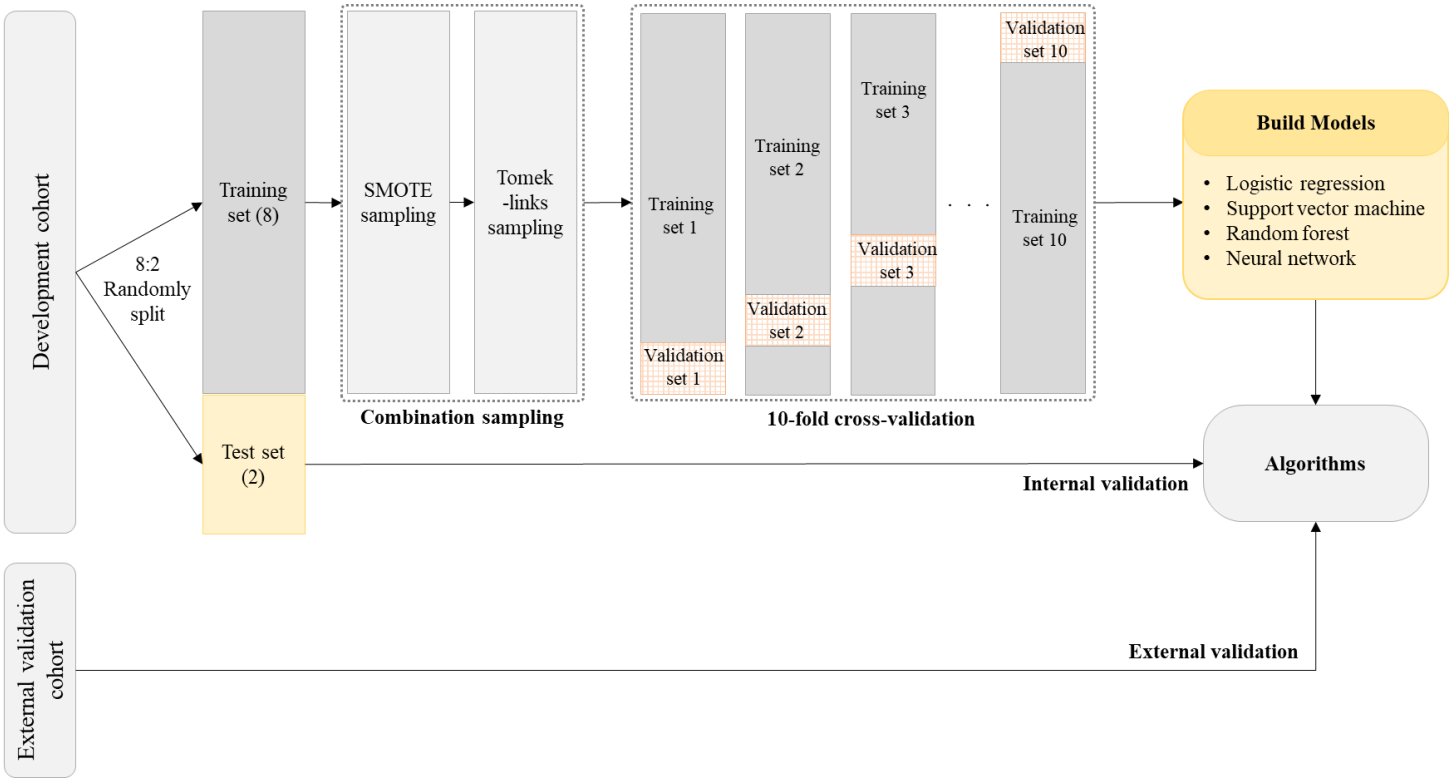


Figure 6 Model development and validation process

For the training set obtained by combination sampling, stratified 10-fold cross-validation was used for training algorithms. The training set was randomly divided into ten parts (training sets 1–10) of the same proportions (training set:validation set = 9:1). The learning procedures were executed ten times on different training sets, and this process was repeated five times. This study employed logistic regression, support vector machine, random forest, and neural network to develop delirium prediction models. The hyperparameters for each model were optimized using five repetitions of 10-fold cross-validation.

The four machine learning methods were mainly used to predict binary outcomes. The logistic regression model was used to predict a binary outcome due to its simplicity and flexibility (Steyerberg, 2009; Han et al., 2016; Géron, 2019). Support vector machine is a powerful and highly flexible modeling technique that constructs optimal decision boundaries for classification cases of different class labels (Steyerberg, 2009; Kuhn & Johnson, 2013). Random forest is a tree-based model that improves model performance by constructing multiple trees through permutation and resampling approaches to reduce the variance of predicted values (Breiman, 2001; Kuhn & Johnson, 2013; Géron, 2019). In the current study, the random forest was applied with ntree = 500 and mtry = 1:10 (representing the number of variables considered in each node split in the tree). The neural network is modeled by invisible layers such as hidden units (Goodfellow et al., 2017). The neural network model used in this study applies a multilayer perceptron with four hidden layers between the input and output layers.

The performance of the developed prediction models was validated using the test set and external validation cohort. For internal validation, the developed models were fitted to the test set, and the models' predictive performances were calculated. In addition, the same statistical methods as in the internal validation were used in the external validation cohort.

4.9. Data analysis

Data preprocessing, model development, and validation of the machine learning algorithms were performed using R version 4.1.2 (R Core Team, 2021). The method for data analysis was as follows.

First, descriptive statistics of the development and external validation cohorts were analyzed and represented as mean and standard deviation (SD) or frequency and percentage. The differences between the groups (non-delirium or delirium) according to general and admission-related characteristics, predisposing factors, precipitating factors, and nursing data were analyzed using *t*-test or chi-square test.

Second, the developed prediction models were evaluated for internal validation using the test set in the development cohort. The model performances were evaluated were sensitivity, specificity/recall, positive predictive value (PPV)/precision, negative predictive value (NPV), F_1 score, accuracy, Youden index, and AUROC (Figure 7). Sensitivity is the proportion of actual positives, and specificity/recall is the proportion of

actual negatives. PPV/precision is the proportion of patients with positive delirium results who had delirium, and NPV is the proportion of patients with negative delirium results who did not have delirium. The F_1 score is the harmonic mean of precision and recall (Witten, 2017). Youden index is calculated by sensitivity and specificity, and a score close to 1 indicates higher predictive capability (Kallner, 2018). Accuracy is defined as the proportion of correct predictions, and the AUROC curve plots the sensitivity and 1-specificity (false positive rate, FPR) (Irizarry, 2020).

Third, variable importance between the predictors was identified using the Gini index, and the reliability was assessed through calibration plots related to goodness-of-fit. Variable importance was measured by the Mean Decrease Gini (MDG) which was derived from the training procedures of the random forest classifier (Menze et al., 2009). The Gini importance score provides a relative ranking of features (Kuhn & Johnson, 2013). The calibration plot reflects the agreement between the observed and predicted risks, to ensure the model's reliability (Han et al., 2016).

Fourth, the developed models were compared using the DeLong test, and the most optimal model was selected. The DeLong test is used to determine statistical differences when comparing two correlated models derived from the same dataset (DeLong et al., 1988; Han et al., 2016).

Fifth, the performances of the optimal prediction models were evaluated using the external validation cohort.

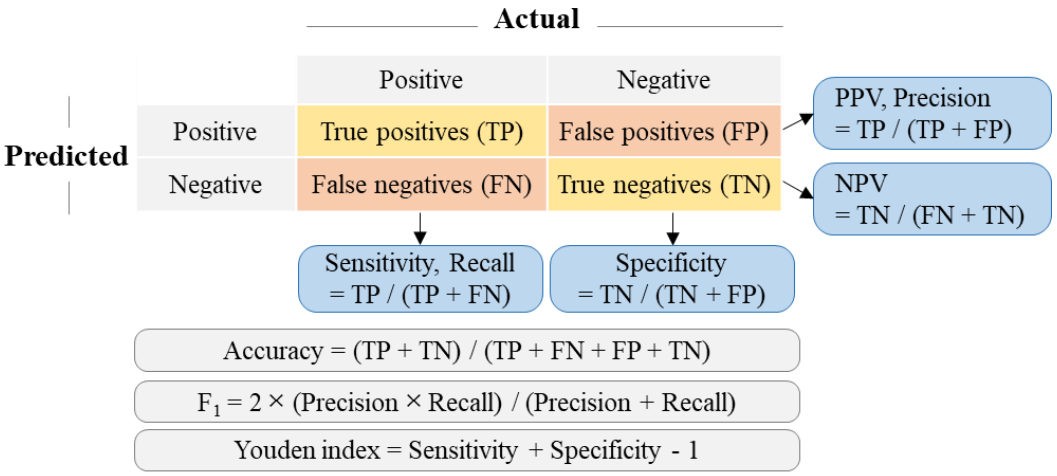


Figure 7 Evaluation of prediction model

Note. FN = false negative; FP = false positive; NPV = negative predictive value; PPV = positive predictive value; TN = true negative; TP = true positive.

V. RESULTS

5.1. Cohort development

This study utilized the MIMIC-IV dataset for model development and internal validation, and the EMR dataset of a single tertiary hospital for external validation. A total of 50588 and 5925 admissions were included in the MIMIC-IV and the EMR datasets, respectively. During the cohort selection process, 9491 admissions of 8696 patients who met the eligibility criteria were included in the development cohort, and 2629 admissions of 2596 patients who met the eligibility criteria were included in the external validation cohort. Delirium was detected in 17.0% and 8.4% admissions in the development and external validation cohorts, respectively. The general characteristics of the cohorts are shown in Table 8.

Table 8 Characteristics of the cohorts

	Development cohort (N = 9491)								External validation cohort (N = 2629)			
	Development (n=7594)				Internal validation (n=1897)				Non-delirium	Delirium		
	Non-delirium	Delirium	t/χ^2	p	Non-delirium	Delirium	t/χ^2	p			t/χ^2	p
	n (%) or $M \pm SD$				n (%) or $M \pm SD$				n (%) or $M \pm SD$			
Patients	5602 (81.78)	1248 (18.22)			1523 (82.50)	323 (17.50)			2376 (91.53)	220 (8.47)		
Admissions	6295 (82.89)	1299 (17.11)			1573 (82.92)	324 (17.08)			2409 (91.63)	220 (8.37)		
Demographics												
Age, years	62.49 ± 15.51	63.72 ± 15.13	-2.60	.009	62.13 ± 15.25	64.28 ± 15.45	-2.31	.021	59.60 ± 14.53	67.13 ± 13.49	-7.40	<.001
Gender, men	3632 (57.69)	726 (55.89)	1.37	.242	868 (55.18)	175 (54.01)	0.10	.746	1210 (50.23)	129 (58.64)	5.37	.020
Type of ICU												
Combined medical/surgical	1180 (18.75)	239 (18.40)	0.12	.942	312 (19.83)	50 (15.43)	3.43	.180	1329 (55.17)	90 (40.91)	81.62	<.001
Medical	1965 (31.22)	404 (31.10)			492 (31.28)	109 (33.64)			337 (13.99)	82 (37.27)		
Surgical	3150 (50.04)	656 (50.50)			769 (48.89)	165 (50.93)			743 (30.84)	48 (21.82)		
ICU length of stay, days	2.45 ± 1.92	8.01 ± 7.60	-26.22	<.001	2.47 ± 2.06	7.14 ± 6.47	-12.86	<.001	1.54 ± 0.57	6.90 ± 7.69	-10.34	<.001
Predisposing factors												
Dementia	112 (1.78)	61 (4.70)	39.85	<.001	23 (1.46)	18 (5.56)	19.40	<.001	38 (1.58)	14 (6.36)	21.41	<.001
Hypertension	3614 (57.41)	599 (46.11)	55.19	<.001	873 (55.50)	157 (48.46)	5.09	.024	741 (30.76)	101 (45.91)	20.56	<.001
Visual or hearing deficit	1503 (23.88)	55 (4.23)	253.56	<.001	352 (22.38)	12 (3.70)	59.22	<.001	413 (17.14)	62 (28.18)	15.85	<.001
Fall history	1486 (23.60)	428 (32.95)	49.36	<.001	373 (23.71)	110 (33.95)	14.30	<.001	86 (3.57)	22 (10.00)	19.56	<.001
Precipitating factors												
Operation	1368 (21.73)	51 (3.93)	223.51	<.001	335 (21.30)	19 (5.86)	41.15	<.001	1784 (74.06)	24 (10.91)	371.35	<.001
Mechanical ventilation	2120 (33.68)	471 (36.26)	3.08	.079	537 (34.14)	127 (39.20)	2.80	.094	473 (19.63)	49 (22.27)	0.72	.395
Foley catheter	539 (8.56)	46 (3.54)	37.48	<.001	143 (9.09)	11 (3.40)	10.93	<.001	1446 (60.02)	7 (3.18)	261.18	<.001
Physical restraints	1983 (31.50)	692 (53.27)	222.72	<.001	439 (27.91)	169 (52.16)	50.34	<.001	1057 (43.88)	105 (47.73)	1.06	.303
Transfusion	1328 (21.10)	221 (17.01)	10.81	.001	328 (20.85)	58 (17.90)	1.27	.260	468 (19.43)	57 (25.91)	4.90	.027
Serum creatinine	1.53 ± 1.71	1.65 ± 1.72	-2.23	.025	1.49 ± 1.67	1.50 ± 1.46	-0.11	.915	0.97 ± 1.15	1.24 ± 1.25	-3.33	<.001

Table 8 Characteristics of the cohorts (continued)

	Development cohort (N = 9491)								External validation cohort (N = 2629)			
	Development (n=7594)				Internal validation (n=1897)				Non-delirium	Delirium		
	Non-delirium	Delirium	t/χ^2	p	Non-delirium	Delirium	t/χ^2	p			t/χ^2	p
	n (%) or $M \pm SD$				n (%) or $M \pm SD$				n (%) or $M \pm SD$			
Nursing assessments												
GCS Total, first	12.04 ± 4.65	11.60 ± 3.67	3.69	<.001	12.02 ± 4.71	11.61 ± 3.74	1.66	.097	14.73 ± 1.57	14.34 ± 2.01	2.83	.005
GCS Total, last	14.65 ± 1.53	11.79 ± 3.23	31.14	<.001	14.65 ± 1.58	11.76 ± 3.27	15.55	<.001	14.93 ± 0.77	14.41 ± 1.61	4.74	<.001
GCS Total, max	14.89 ± 0.70	12.87 ± 2.73	26.52	<.001	14.91 ± 0.62	12.84 ± 2.77	13.37	<.001	14.98 ± 0.27	14.67 ± 1.11	4.14	<.001
GCS Total, min	11.13 ± 4.82	10.31 ± 3.96	6.56	<.001	11.11 ± 4.85	10.26 ± 4.04	3.34	<.001	14.66 ± 1.76	14.15 ± 2.14	3.50	<.001
GCS Eye, min	2.90 ±1.22	2.76 ± 1.14	4.10	<.001	2.88 ± 1.22	2.80 ± 1.10	1.22	.221	3.90 ± 0.48	3.82 ± 0.59	1.96	.051
GCS Verbal, max	4.92 ± 0.49	3.50 ± 1.80	28.26	<.001	4.94 ± 0.42	3.52 ± 1.80	14.12	<.001	4.99 ± 0.15	4.81 ± 0.60	4.53	<.001
GCS Verbal, min	3.45 ± 1.86	2.60 ± 1.75	15.75	<.001	3.47 ± 1.86	2.57 ± 1.76	8.00	<.001	4.88 ± 0.66	4.55 ± 0.99	4.82	<.001
GCS Motor, min	4.74 ± 2.06	4.88 ± 1.66	-2.65	.008	4.71 ± 2.07	4.83 ± 1.81	-0.99	.320	5.13 ± 0.71	5.15 ± 0.78	-0.38	.707
RASS, min	-0.56 ± 1.33	-0.54 ± 1.05	-0.49	.623	-0.54 ± 1.34	-0.60 ± 1.16	0.82	.413	-0.99 ± 1.67	-0.52 ± 1.52	-3.95	<.001
Pain, max	3.54 ± 3.78	1.73 ± 3.07	18.56	<.001	3.74 ± 3.75	1.87 ± 3.15	9.41	<.001	3.33 ± 3.07	1.66 ± 2.76	7.79	<.001
Pain, median	0.40 ± 1.06	0.48 ± 1.36	-1.95	.051	0.44 ± 1.17	0.48 ± 1.49	-0.46	.644	0.01 ± 0.09	0.04 ± 0.28	-1.67	.096
Pressure injury, min	17.20 ± 4.42	16.69 ± 4.74	3.51	<.001	17.12 ± 4.37	16.62 ± 4.73	1.85	.064	21.40 ± 2.70	19.16 ± 2.78	11.76	<.001
SBP, max	151.37 ± 20.70	146.43 ± 22.10	7.41	<.001	150.89 ± 20.5	148.36 ± 22.8	1.99	.047	153.35 ± 20.74	155.45 ± 21.31	-1.43	.152
PR, max	106.54 ± 19.40	105.44 ± 19.86	1.84	.065	106.32 ± 18.9	106.27 ± 19.6	0.05	.963	104.79 ± 20.17	109.98 ± 20.72	-3.65	<.001
PR, median	83.15 ± 14.40	86.41 ± 16.41	-6.64	<.001	83.02 ± 14.3	87.47 ± 16.9	-4.42	<.001	79.85 ± 14.68	88.72 ± 17.69	-7.21	<.001
BT, max	37.35 ± 0.44	37.35 ± 0.49	0.22	.829	37.36 ± 0.44	37.38 ± 0.49	-0.66	.512	37.72 ± 0.49	37.61 ± 0.55	2.80	.005
RR, max	28.85 ± 4.52	27.49 ± 4.90	9.24	<.001	28.97 ± 4.48	27.40 ± 4.87	5.69	<.001	26.00 ± 4.21	27.14 ±4.85	-3.37	<.001

Table 8 Characteristics of the cohorts (continued)

	Development cohort (N = 9491)								External validation cohort (N = 2629)			
	Development (n=7594)				Internal validation (n=1897)							
	Non-delirium	Delirium	t/χ^2	p	Non-delirium	Delirium	t/χ^2	p	Non-delirium	Delirium	t/χ^2	p
	n (%) or $M \pm SD$				n (%) or $M \pm SD$				n (%) or $M \pm SD$			
Nursing documentation patterns												
Frequency												
Complete set of vital signs [a]	3.94 ± 3.85	4.40 ± 4.83	-3.23	.001	3.89 ± 3.80	4.37 ± 5.47	-1.51	.132	17.43 ± 9.42	13.86 ± 9.67	5.37	<.001
SBP [a]	25.54 ± 6.08	27.03 ± 9.02	-5.67	<.001	25.41 ± 5.95	26.92 ± 8.44	-3.07	.002	31.70 ± 17.28	30.15 ± 18.67	1.26	.207
PR [a]	25.84 ± 4.35	26.70 ± 8.02	-3.77	<.001	25.82 ± 4.30	26.61 ± 6.48	-2.10	.037	40.68 ± 12.68	39.19 ± 13.60	1.66	.097
BT [a]	7.07 ± 3.60	7.69 ± 4.72	-4.47	<.001	7.01 ± 3.43	7.91 ± 5.22	-3.00	.003	29.21 ± 3.67	28.40 ± 5.77	2.05	.042
RR [a]	25.65 ± 4.37	26.54 ± 7.74	-4.00	<.001	25.63 ± 4.45	26.38 ± 6.64	-1.94	.053	39.49 ± 12.53	37.79 ± 13.42	1.92	.056
SpO2 [a]	25.25 ± 4.49	26.29 ± 7.94	-4.57	<.001	25.24 ± 4.48	26.23 ± 6.40	-2.64	.009	24.95 ± 3.08	20.54 ± 7.20	9.01	<.001
GCS [a]	6.03 ± 3.71	6.78 ± 4.65	-5.52	<.001	5.96 ± 3.63	6.89 ± 4.71	-3.33	<.001	0.87 ± 1.67	1.34 ± 1.93	-3.46	<.001
RASS [a]	3.00 ± 2.43	3.31 ± 2.84	-3.66	<.001	2.99 ± 2.39	3.32 ± 2.89	-1.96	.051	0.68 ± 1.71	0.77 ± 1.99	-0.67	.501
Intervention												
PRN medication	1678 (26.66)	374 (28.79)	2.38	.123	423 (26.89)	100 (30.86)	1.93	.165	1824 (75.72)	159 (72.27)	1.11	.292
Stat medication	2215 (35.19)	518 (39.88)	10.08	.001	572 (36.36)	121 (37.34)	0.07	.786	2392 (99.29)	210 (95.45)	25.59	<.001
Withholding scheduled medication	369 (5.86)	63 (4.85)	1.87	.171	103 (6.55)	13 (4.01)	2.58	.108	2391 (99.25)	215 (97.72)	3.79	.051

Note. BT = body temperature; GCS = Glasgow Coma Scale; PR = pulse rate; ICU = intensive care unit; max = maximum score; min = minimum score; PRN = Pro re nata; RASS = Richmond Agitation-Sedation Scale; RR = respiratory rate; SBP = systolic blood pressure; SpO2 = oxygen saturation.

[a] Documentation frequency per 24 hours.

The mean age for both the development and external validation cohorts was 59–67 years, and more than half were men (55.6%). Approximately 50% of the patients in the development cohort were admitted to a surgical ICU. In the external validation cohort, most patients were admitted to a combined ICU. The length of ICU stay was longer in the delirium group than in the non-delirium group for both cohorts (p <.001).

More than half of the patients in both cohorts (52.2%) suffered from hypertension. The non-delirium group in the development cohort and the delirium group in the external validation cohort had higher proportions of patients with hypertension or visual or hearing deficits (P <.05). History of dementia and fall incidence within the last six months was higher in patients with delirium than in those without delirium (p <.001). The proportions of use of mechanical ventilation or physical restraints were higher in the delirium groups than in the non-delirium groups. The proportions of patients who underwent surgery or used foley catheter were lower in the delirium groups than in the non-delirium groups (p <.001).

The GCS, pressure injury, and pain scores were lower in the delirium groups than in the non-delirium groups. Vital signs values did not exhibit clear data patterns, except for pulse rate in both cohorts. The median pulse rate was higher in the delirium groups than in the non-delirium groups (p <.001).

Frequency patterns in the nursing data showed that vital signs, GCS, and RASS were more frequently recorded in the delirium group than in the non-delirium group in the development cohort. For the external validation cohort, GCS and RASS were more

frequently documented in the delirium group than in the non-delirium group. Intervention patterns related to medication administration were higher in the external validation cohort; in particular, over 95% of patients experienced stat medication administration and withholding scheduled medication.

The mean duration of delirium onset after ICU admission was 2.6 days (*SD*: 2.3), regardless of onset time. Among all delirium cases, 95.3% occurred within the first week of admission (Figure 8). In the development cohort, delirium events mostly occurred during the day shift (48.7%; 6 AM–2 PM), followed by the evening (34.9%; 2 PM–10 PM) and night shifts (16.4%; 10 PM–6 AM). In the external validation cohort, delirium events mostly occurred during the night (39.5%) and day (35.0%) shifts (Appendix 6).

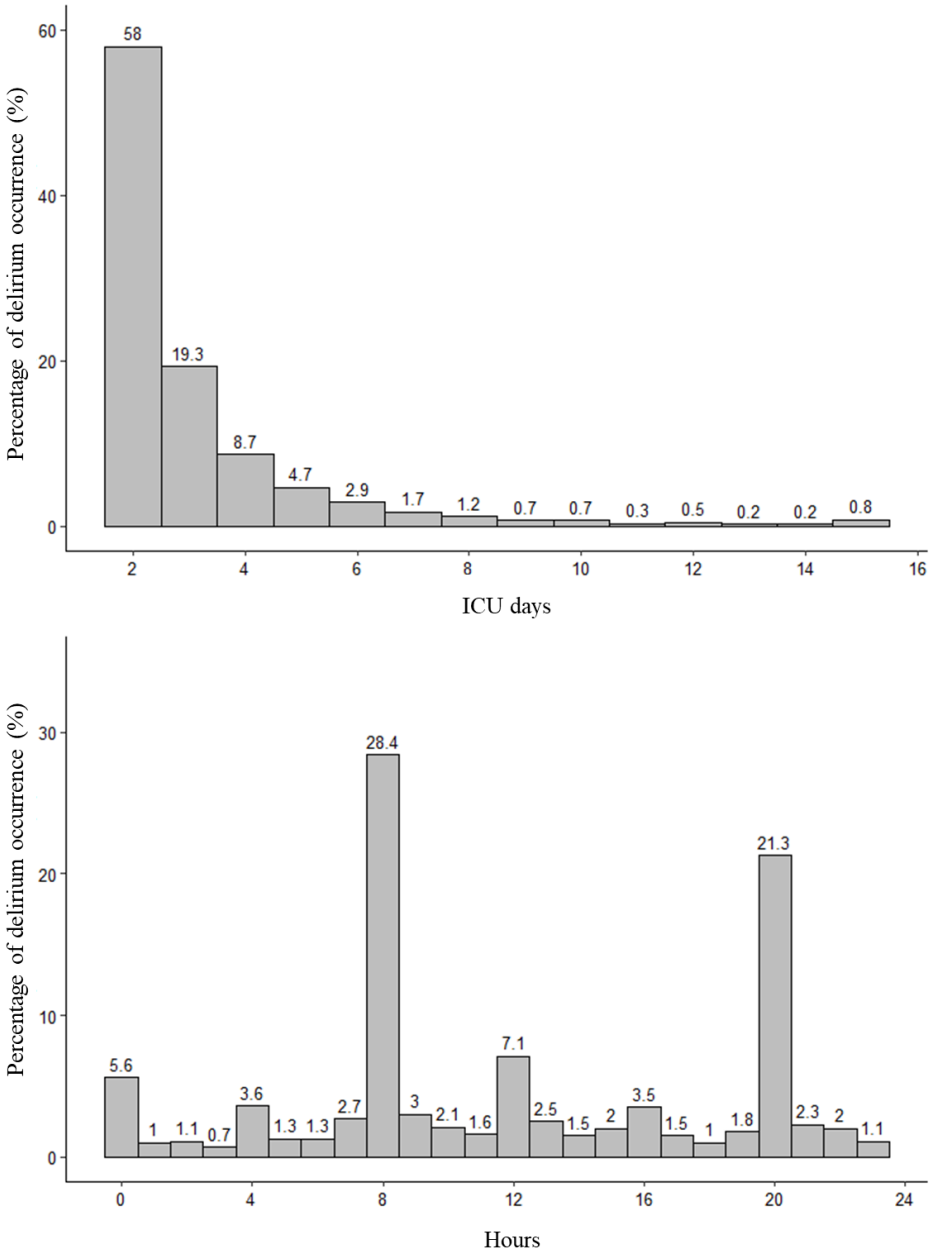


Figure 8 Delirium occurrence based on day and time of onset

Note. ICU = intensive care unit.

5.2. Model development and internal validation

In the current study, Models I (40 predictors) and II (31 predictors) were developed to compare differences in model performance depending on whether the actual GCS and RASS scores were included as predictors. Based on the sliding window method, potential predictors were extracted from the data 24 hours before delirium occurred for patients with delirium, while all data were used for non-delirium patients during their ICU stay. Furthermore, to compare the performance of the developed prediction models according to the feature selection time in non-delirium patients, features were additionally extracted from the data for the first 24 hours after ICU admission (control A) and the last 24 hours before ICU discharge (control B).

Models I and II employed logistic regression, support vector machine, random forest, and neural network to develop delirium prediction models. The confusion matrix for the internal validation is shown in Table 9. The accuracy of Models I and II across the four machine learning methods was 0.840–0.931 and 0.703–0.863, respectively. Random forest method among developed models had the highest accuracy (Model I: 0.931, [95% CI; 0.919, 0.942]; Model II: 0.863, [95% CI; 0.847, 0.878]).

Table 9 Confusion matrix of internal validation

Model	Prediction modeling method	Predicted	Actual		Accuracy [95% CI]
			Delirium	Non-delirium	
Model I	Logistic regression	Delirium	270	250	0.840
		Non-delirium	54	1323	[.822, .856]
	Support vector machine	Delirium	245	156	0.876
		Non-delirium	79	1417	[.860, .891]
	Random Forest	Delirium	288	94	0.931
		Non-delirium	36	1479	[.919, .942]
	Neural network	Delirium	264	216	0.855
		Non-delirium	60	1357	[.838, .870]
Mode II [a]	Logistic regression	Delirium	273	512	0.703
		Non-delirium	51	1061	[.682, .724]
	Support vector machine	Delirium	248	359	0.771
		Non-delirium	76	1214	[.751, .790]
	Random Forest	Delirium	297	233	0.863
		Non-delirium	27	1340	[.847, .878]
	Neural network	Delirium	263	500	0.704
		Non-delirium	61	1073	[.683, .725]

Note. CI = confidence interval.

[a] The model used predictors excluding actual values of GCS and RASS among Model I predictors.

For the control A models (data used during the first 24 hours after ICU admission in the non-delirium group), the accuracies were 0.761–0.861 and 0.636–0.780 in Models I and II, respectively (Appendix 7). For the control B models (data used during the last 24 hours before ICU discharge), the accuracies were 0.837–0.908 and 0.745–0.840 in Models I and II, respectively (Appendix 8). The control B models were more accurate than control A models, but slightly less accurate than in basic models (all data used during ICU stay in the non-delirium group). Random forest was the most accurate method between the developed control A and B models.

The predictive performance of the developed models is shown in Table 10. The receiver operating characteristic (ROC) curves are shown in Figure 9. Random forest showed the best performance for Model I, with an AUROC and 95% CI of 0.975 [0.967, 0.982], and an estimated out-of-bag (OOB) error rate of 2.43%. This model's sensitivity was 0.889, specificity was 0.940, PPV was 0.755, NPV was 0.974, F_1 score was 0.816, and Youden index was 0.829. Random forest showed the best performance for Model II as well, with an AUROC and 95% CI of 0.951 [0.940, 0.962], and an estimated OOB error rate of 3.80%. This model's sensitivity was 0.917, specificity was 0.852, PPV was 0.556, NPV was 0.975, F_1 score was 0.696, and Youden index was 0.769.

Table 10 Model performance of each model in the internal validation

Model	AUROC [95% CI]	Sensitivity	Specificity	PPV	NPV	F_1 score	Youden index
Model I							
Logistic regression	0.915 [.897, .932]	0.833	0.841	0.528	0.960	0.640	0.674
Support vector machine	0.829 [.804, .853]	0.756	0.901	0.624	0.954	0.676	0.657
Random forest	0.975 [.967, .982]	0.889	0.940	0.755	0.974	0.816	0.829
Neural Network	0.839 [.900, .934]	0.815	0.863	0.556	0.963	0.657	0.677
Model II [a]							
Logistic regression	0.836 [.811, .860]	0.843	0.675	0.374	0.954	0.492	0.517
Support vector machine	0.769 [.743, .794]	0.765	0.772	0.436	0.942	0.533	0.537
Random forest	0.951 [.940, .962]	0.917	0.852	0.556	0.975	0.696	0.769
Neural Network	0.747 [.802, .852]	0.812	0.682	0.370	0.952	0.484	0.494

Note. AUROC = area under the receiver operating characteristic; CI = confidence interval; NPV = negative predictive value; PPV = positive predictive value.

[a] The model used predictors excluding the actual values of GCS and RASS among Model I predictors.

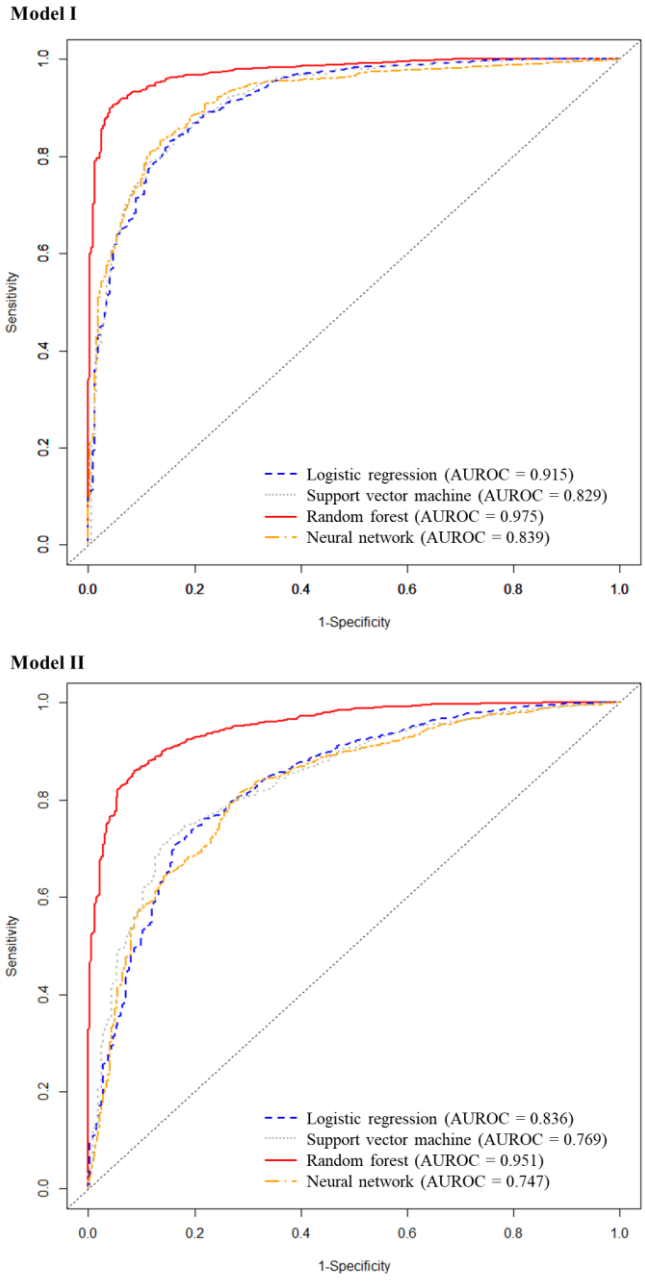


Figure 9 Receiver operating characteristic curves of prediction models

Note. AUROC = area under the receiver operating characteristic.

In control A models, random forest achieved the best predictive performance, with AUROCs and 95% CIs for Models I and II of 0.896 [0.877, 0.914] and 0.846 [0.823, 0.868], respectively (Appendix 9). Random forest achieved the best predictive performance in control B models. Control B models' AUROCs and 95% CIs for Models I and II were 0.926 [0.940, 0.962] and 0.897 [0.878, 0.916], respectively (Appendix 10), which was better than the predictive performance of control A models, but slightly lower than that of basic models. In the random forest method, control A and B models were lower in sensitivity and similar in specificity than the basic model, but the AUROCs were 0.80 or higher.

5.3. Variable importance

All variables included in the developed model were grouped as predisposing factors, precipitating factors, nursing assessments, and nursing documentation patterns. The top 20 important predictors for the random forest model are shown in Figure 10. The most significant variable was the last GCS in Model I and the maximum pain score in Mode II. Among the important variables, GCS, RASS, pain, and documentation frequency patterns (single vital sign, RASS, and GCS) were ranked higher than the precipitating and predisposing factors and actual values of vital signs in both models. These predictors were extracted from nursing assessments and nursing flowsheets in EMR.

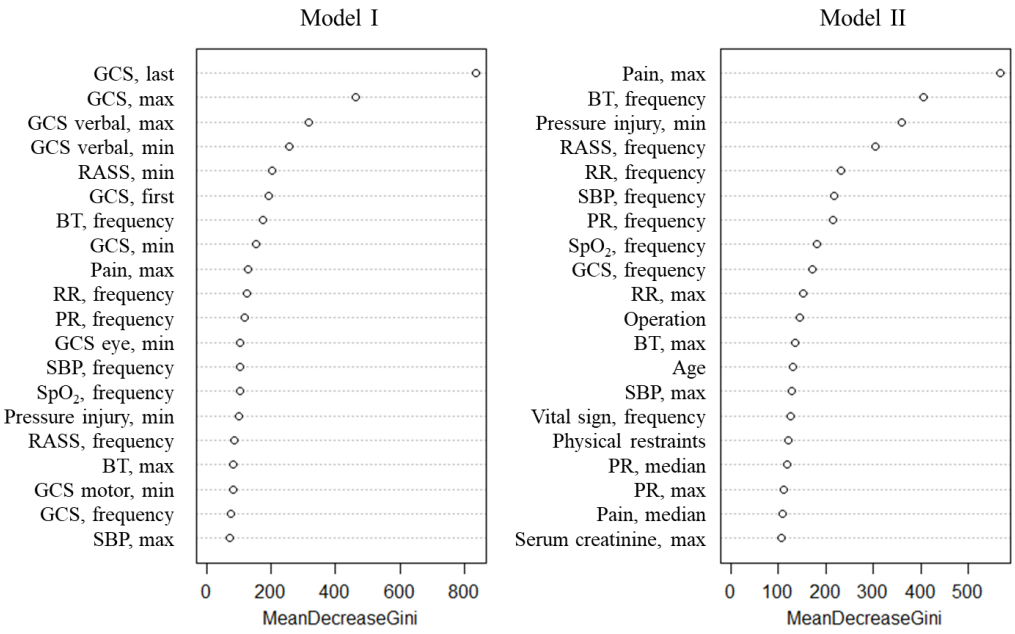


Figure 10 Top 20 important predictors in the random forest models

Note. BT = body temperature; GCS = Glasgow Coma Scale; PR = pulse rate; RASS = Richmond Agitation-Sedation Scale; RR = respiratory rate; SBP = systolic blood pressure; SpO_2 = oxygen saturation.

5.4. Calibration plot

A calibration plot of the random forest models for delirium prediction is shown in Figure 11. Both classification models had similar AUROCs with 95% CIs (Model I: 0.975 [0.967, 0.982], Model II: 0.951 [0.940, 0.962]). The calibration plot shows that the class probabilities of the two models were similar in the test set, but Model I was slightly dominant.

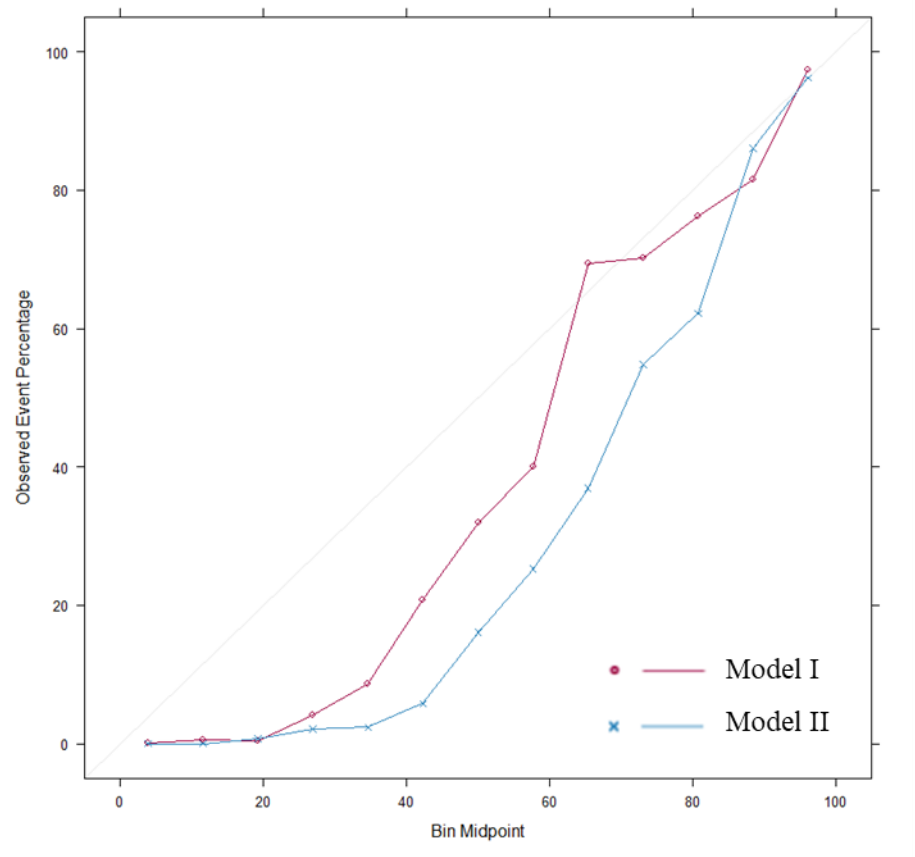


Figure 11 Calibration curves for random forest models

5.5. Comparison between machine learning models

The random forest method showed the best predictive performance in the development cohort. The DeLong test was performed to compare the AUROCs between the models (p <.05 indicated statistical significance) (Figure 12). The AUROCs with 95% CIs of Models I and II were 0.975 [0.967, 0.982] and 0.951 [0.940, 0.962], respectively. The best delirium prediction models' performance, with or without GCS and RASS values as predictors, was over 0.950 (AUROC). The DeLong test showed that Model I had a statistically higher predictive performance than Model II (z = 4.530, p <.001).

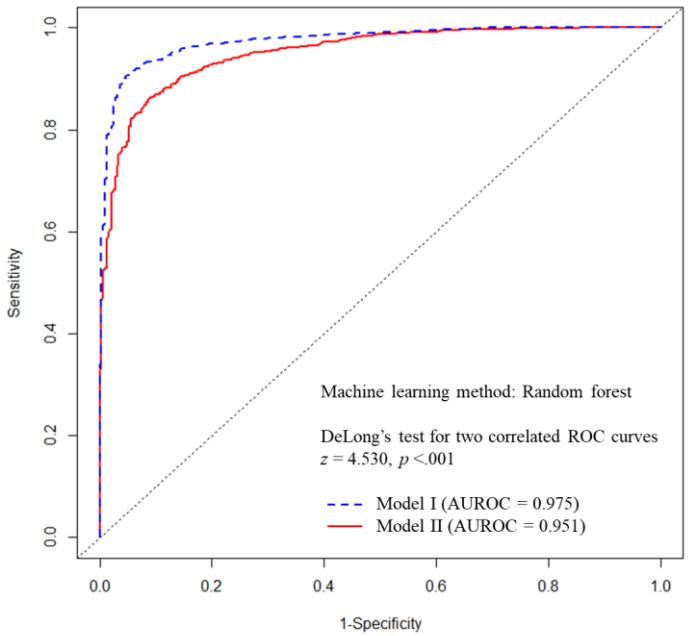


Figure 12 Comparison of random forest models

Note. AUROC = area under the receiver operating characteristic.

5.6. External validation

External validation was conducted to determine the generalizability of the developed models. The confusion matrix for the external validation is presented in Table 11. Neural network and logistic regression showed the best accuracies for Models I (0.832) and II (0.749). The predictive performance of Model I was in the order of neural network, logistic regression, and random forest, with AUROCs and 95% CIs of 0.825 [0.796, 0.853], 0.799 [0.768, 0.830], and 0.770 [0.733, 0.808], respectively (Table 12). Meanwhile, for Model II, logistic regression had the best predictive performance, followed by random forest, with AUROCs and 95% CIs of 0.833 [0.806, 0.860] and 0.791 [0.764, 0.818], respectively. The developed delirium prediction models' performance in external validation, with or without actual GCS and RASS values as predictors, was over 0.770 (AUROC) in logistic regression and random forest methods.

The best predictive performance model for external validation was the logistic regression method in Model II, with an AUROC of 0.833 (95% CI [0.806, 0.860]). This model's sensitivity was 0.736, specificity was 0.750, PPV was 0.148, NPV was 0.990, F_1 score was 0.329, and Youden index was 0.486.

Table 11 Confusion matrix of external validation

Model	Prediction modeling method	Predicted	Actual		Accuracy [95% CI]
			Delirium	Non-delirium	
Model I	Logistic regression	Delirium	133	470	0.788
		Non-delirium	87	1939	[.772, .804]
	Support vector machine	Delirium	132	1230	0.499
		Non-delirium	88	1179	[.479, .518]
	Random Forest	Delirium	167	1124	0.552
		Non-delirium	53	1285	[.533, .571]
	Neural network	Delirium	124	346	0.832
		Non-delirium	96	2063	[.817, .846]
Mode II [a]	Logistic regression	Delirium	162	603	0.749
		Non-delirium	58	1806	[.732, .765]
	Support vector machine	Delirium	179	1193	0.530
		Non-delirium	41	1216	[.511, .550]
	Random Forest	Delirium	207	1327	0.490
		Non-delirium	13	1082	[.471, .510]
	Neural network	Delirium	209	1602	0.387
		Non-delirium	11	807	[.368, .405]

Note. CI = confidence interval.

[a] The model used predictors excluding the actual values of GCS and RASS among Model I predictors.

Table 12 Model performance of each model in the external validation

Model	AUROC [95% CI]	Sensitivity	Specificity	PPV	NPV	F_1 score	Youden index
Model I							
Logistic regression	0.799 [.768, .830]	0.605	0.805	0.145	0.979	0.323	0.409
Support vector machine	0.545 [.511, .579]	0.559	0.537	0.108	0.950	0.169	0.096
Random forest	0.770 [.733, .808]	0.759	0.533	0.129	0.960	0.221	0.293
Neural Network	0.825 [.796, .853]	0.564	0.856	0.161	0.974	0.359	0.420
Model II [a]							
Logistic regression	0.833 [.806, .860]	0.736	0.750	0.148	0.990	0.329	0.486
Support vector machine	0.659 [.632, .687]	0.814	0.505	0.130	0.967	0.225	0.318
Random forest	0.791 [.764, .818]	0.941	0.449	0.124	0.986	0.236	0.390
Neural Network	0.642 [.803, .853]	0.950	0.335	0.135	0.992	0.206	0.285

Note. AUROC = area under the receiver operating characteristic; CI = confidence interval; NPV = negative predictive value; PPV = positive predictive value.

[a] The model used predictors excluding the actual values of GCS and RASS among Model I predictors.

VI. DISCUSSION

This study was a retrospective cohort study that used nursing data that reflected time variation in EMR to develop and validate machine learning-based delirium prediction models for ICU patients. The MIMIC-IV database was used for model development and internal validation. An EMR database from a single tertiary hospital in Seoul, South Korea, was used for external validation to evaluate the predictive performance of the developed models. This chapter discusses the delirium incidence and general characteristics of delirium group in cohorts as well as the models' development and predictive performance of developed models. Finally, the chapter concludes with a discussion of the study's significance, limitations, and suggestions for future research.

6.1. Delirium incidence and general characteristics of delirium group

In the development and external validation cohorts, the delirium incidence 24 hours after ICU admission were 17.0% and 8.4%, respectively. A previous study reported that delirium was detected in 21.1–30.8% of patients in a mixed ICU and 32.1% in a cardiac ICU (van den Boogaard et al., 2012; Guenther et al., 2013; van den Boogaard et al., 2014; Wassenaar et al., 2015; Moon et al., 2018; Hur et al., 2021), which was more prevalent than reported in the current study. The previous studies were conducted prospectively and assessed delirium by trained researchers using CAM-ICU at the bedside. The incidence of

delirium may be under reported if the healthcare provider did not write delirium-related records in the EMR or did not recognize hypoactive delirium, which occurs more frequently than hyperactive (Barr & Pandharipande, 2013; Krewulak et al., 2018). In addition, while other studies excluded patients with neurologic or psychiatric diseases, severe visual or hearing disorders, and aphasia (Wassenaar et al., 2015; Moon et al., 2018; Zhao & Luo, 2021), the current study did not limit patients' diagnoses, whether or not they underwent surgery, or ICU type. Furthermore, the current study excluded patients with delirium within 24 hours of ICU admission. Therefore, the delirium incidence reported here may be lower compared to previous studies.

In the current study, the delirium incidence was lower in the external validation cohort (8.4%) than in the development cohort (17.0%). These results were similar to a meta-analysis on ICU delirium conducted in 18 studies which reported that the overall pooled delirium incidence were 4% (hyperactive), 11% (hypoactive), and 7% (mixed) (Krewulak et al., 2018). On the other hand, two studies conducted in South Korea reported that the incidences of delirium were 21.1% and 30.8% (Moon et al., 2018; Hur et al., 2021), which was higher than that of the current external validation cohort. The higher reported incidences may be owing to differences in the delirium assessment tools (previous studies used CAM-ICU and external validation cohort used ICDSC for delirium screening) or the longer ICU stays. The CAM-ICU and ICDSC are well-validated tools, but CAM-ICU has higher specificity than ICDSC (Barr et al., 2013) and is superior in detecting delirium in patients who use mechanical ventilation (Chen et al., 2021).

The characteristics of delirium patients in the current study are consistent with previously reported studies, such as older age, history of dementia, mechanical ventilation, physical restraints, length of ICU stay, and fall history (Barr et al., 2013; Zaal et al., 2015; Krewulak et al., 2020). History of hypertension was reported as inconsistent or moderate evidence predictors by other studies (Devlin et al., 2018; Mufti et al., 2019; Hur et al., 2021). Similarly, the current study showed inconsistent results; the proportion of hypertensive patients was higher in the non-delirium group in the development cohort and the delirium group in the external validation cohort than in the other group.

Pain is reported as a risk factor for postoperative delirium, and untreated pain is related to agitation or anxiety rather than delirium (Barr & Pandharipande, 2013; Aldecoa et al., 2017). In the current study, the maximum pain score and the proportion of patients who underwent surgery were lower in the delirium group than in the non-delirium group. The cohorts of this study were selected regardless of surgical history, and therefore, the pain score may be lower in patients with delirium who did not undergo surgery.

6.2. Development of delirium prediction models

This study developed delirium prediction models using nursing data that reflected time variability. The EMR data included in the developed models were demographic data, clinical history, clinical characteristics, laboratory test, nursing assessments, and nursing documentation patterns of frequency and intervention. The nursing assessment contains

actual values based on assessment tools and the measurement of vital signs. Nursing documentation patterns include frequency and intervention patterns extracted from nursing flowsheets and records of medication administration (Collins & Vawdrey, 2012; Collins et al., 2013; Schnock et al., 2021). The most important variables for delirium prediction were GCS, RASS, pain score, and documentation frequency patterns; these were ranked higher among the top 20 important predictors than well-known modifiers, such as age, operation, physical restraints, and serum creatinine level. According to the guideline, GCS is not related to delirium incidence (Devlin et al., 2018) or did not show an important feature (Wassenaar et al., 2015). However, GCS was identified as the most important predictor in the current study. Similarly, a recent study used GCS scores (eye, verbal, and motor) for the development of delirium prediction models and ranked second (GCS, verbal) and eighth (GCS, motor) among delirium predictors (Hur et al., 2021). In addition, a study used the level of consciousness score to develop delirium risk scoring algorithms (Moon et al., 2018).

It should be noted that there may be differences between the critical features identified by analyzing data patterns based on machine learning methods and previously known delirium risk factors. Delirium is an acute state of confusion assessed by healthcare providers. It may be associated with GCS and RASS scores. In the current study, the AUROCs of the developed models using the random forest method were 0.975 and 0.951, respectively, depending on whether GCS and RASS were included as predictors or not. The models achieved good predictive performance regardless of GCS

and RASS. GCS and RASS are valuable resources that can be used without missing values in predictive models because they are assessed and recorded periodically in ICUs. Therefore, GCS and RASS can be used as important predictors to predict delirium depending on the hospital environment.

Nursing documentation patterns were identified as frequency and intervention in the nursing data. These patterns have been previously used to predict mortality (Collins et al., 2013; Fu et al., 2021), cardiac arrest (Collins & Vawdrey, 2012; Fu et al., 2021), and patient deterioration (Schnock et al., 2021). A study reported that ICU patients with delirium risk had significantly increased all-cause mortality and prolonged durations of ICU and hospital stay (Fan et al., 2019). Therefore, using nursing documentation patterns as predictors in delirium prediction models may be acceptable. In the current study, approximately 95% of ICU patients in the external validation cohort experienced stat medication administration and withholding scheduled medication. Nursing documentation patterns may vary according to the hospital policy, guidelines, and the clinical environment. Therefore, intervention patterns from data on medication administration should be considered as a predictor according to each hospital's data patterns.

The selected predictors in the current study were slightly different from previously reported delirium risk factors. This study selected predictors considering available variables in two cohorts and focused on nursing data that represented time variation. As a

result, GCS, RASS, pain, and nursing documentation patterns were ranked higher than previously reported risk factors.

EMR data accumulate continuously, and their patterns can change according to patient conditions. Therefore, this study used a sliding window of 24 hours to represent the most recent patient condition before the onset of delirium. The sliding window methods used here required the instance at which data should be extracted in the non-delirium group during ICU stay to be determined. The DYNAMIC-ICU model was developed with accumulated information that reflected the time variation by risk factors (Fan et al., 2019). In contrast, the fall risk prediction model using a time-variant method extracted data on fallers within 24 hours before their fall, while all data were used for non-fallers (Cho et al., 2019). The current study extracted data on the delirium group within 24 hours before delirium onset, and all accumulated data were extracted in the non-delirium group.

Additionally, in the current study, data were extracted from the first 24 hours after ICU admission (control A) and the last 24 hours before ICU discharge (control B) to compare the predictive performance of the developed models according to the feature selection time of the non-delirium group. Thus, the developed delirium prediction models' AUROCs, regardless of the feature selection time in the non-delirium group, were over 0.846 in the random forest method. The control B model (relatively low severity condition during ICU stay in the non-delirium group) had a slightly better predictive performance than the control A model (relatively high severity condition in the non-

delirium group). These results are similar to a previous study that reported delirium predicted performance using first, last, and maximum data (Moon et al., 2018).

6.3. Predictive performance of the developed delirium models

This study developed four prediction models each of which depended on the predictors. In recent studies, the best performing delirium prediction models' methods were random forest with an AUROC and 95% CI of 0.92 [0.91, 0.92] (Hur et al., 2021) and logistic regression with AUROC and 95% CI of 0.90 [0.86, 0.94] (Fan et al., 2019). Meanwhile, in the current study, random forest and logistic regression models had AUROCs and 95% CIs of 0.975 [0.967, 0.982] and 0.915 [0.897, 0.932], respectively. Although direct comparisons are difficult due to differences in predictors and follow-up duration, the current study's models showed acceptable performance compared with the previous models.

The PRE-DELIRIC model showed similar performance in internal (AUROC: 0.87) and external validation (AUROC: 0.84) reported by a study conducted in five ICUs in the Netherlands (one hospital was used for model development and internal validation, and four were used for external validation) (van den Boogaard et al., 2012). The PRIDE model showed AUROCs of 0.92 in the internal validation (EMR data from South Korea) and 0.72 in the external validation (EMR data from Boston) (Hur et al., 2021). In the current study, the developed models showed the best AUROCs of 0.98 in the internal

validation (EMR data from Boston) and 0.77 in the external validation (EMR data from South Korea). In addition, the delirium group of both cohorts had different characteristics in the history of hypertension or visual/hearing deficits; and nursing documentation patterns (frequency and intervention). The non-delirium group in the development cohort and the delirium group in the external validation cohort had higher proportions of patients with hypertension or visual/hearing deficits. In nursing documentation patterns, the delirium group in the development cohort and the non-delirium group in the external validation cohort had higher documentation frequency of vital signs and administration of PRN/stat medication. Furthermore, the proportions of patients who underwent surgery or used foley catheter and administration of PRN/stat medication were remarkably higher in the external validation cohort than in the development cohort. The current study's results indicate that differences in documentation culture, practice patterns, the policies, and the guidelines of each hospital may have different data characteristics, which may influence internal and external validation predictive performance.

6.4. Significance of the study

6.4.1. Nursing theory

This study used the HPM-ExpertSignals framework to develop delirium prediction models. This framework provides evidence of the importance of information (data patterns) in EMR for the clinical outcome prediction models. The delirium prediction

models in the current study were developed using information such as modifiers and data patterns representing nurses' behaviors. Thus, the importance of nursing data in prediction models was identified based on the current study's conceptual framework. The results of the current study will add scientific knowledge to bridge the gap between theory and clinical practice and help develop delirium prediction models for use in clinical practice.

6.4.2. Nursing research

The current study used nursing data recorded for the changing conditions of ICU patients to develop delirium prediction models that reflect time variability. Nursing data consisting of nursing assessments and nursing documentation patterns were identified as important features in the developed models. This result demonstrates the reusability of nursing data that reflects the variations of patients' conditions over time in the clinical prediction model. The models developed in this study may be used as fundamental resources for developing the clinical decision support algorithms in EMR for predicting delirium in ICUs. Furthermore, the findings of the current study provide additional insights and evidence regarding developing various clinical outcome prediction models that apply various forms of nursing data as important predictors.

6.4.3. Nursing practice

This study provides a practical strategy for delirium screening of ICU patients by using delirium prediction models without the need for additional assessments. The current study identified important features of nursing data to predict delirium, and these models were validated using domestic EMR to confirm their applicability in Korean healthcare environments. Further studies, such as the development of the clinical decision support system in EMR for delirium prediction, may facilitate the application of the developed models to clinical practice. In addition, the importance of nursing data reflecting nurses' concerns about patients in the current study is significant in that it presented a direction to improve the usability of an EMR that can record and communicate nurses' clinical judgments about patients' conditions.

6.5. Limitations

This study has several limitations. First, it was conducted retrospectively using EMR data. CAM-ICU or ICDSC are assessed periodically in clinical practice, thus identifying patients with delirium at an early stage or may delay the recording than the actual occurrence of delirium. Therefore, a gap may occur between the actual delirium time of onset and the time it was recorded in the EMR.

Second, selection bias may have occurred as patients with missing data or outliers were excluded from the cohort. Retrospective EMR data may contain missing or incorrect

information. Therefore, data require preprocessing to create a suitable dataset to apply machine learning techniques.

Third, this study extracted structured information from the nursing data. Structured and unstructured data are stored in the EMR database of each hospital (Abhyankar et al., 2014). Notably, unstructured data contain richer information (Hashir & Sawhney, 2020) that can facilitate the discrimination of diseases or events. However, this study did not consider unstructured data because the two cohorts used different languages in the EMR as well as different EMR systems.

Fourth, even though the developed delirium prediction models were validated internally and externally, there may be limitations in applying them to clinical practice. This study focused more on the applicability and usability of nursing data for delirium prediction models but not on ease of use in clinical practice. Therefore, this current study did not create prediction algorithms or formula for scores assigned to each predictor that is easy to apply in clinical practice.

Finally, the developed model used data from the MIMIC database in Boston, USA, and external validation used EMR from a single tertiary hospital in Seoul, South Korea, to support its generalizability. However, since only two hospitals' EMRs were used, it is necessary to interpret the results carefully.

6.6. Suggestions for future studies

This study developed and validated delirium prediction models for ICU patients using nursing data. Some suggestions for future studies are as follows. First, the current study was performed retrospectively using EMR. The developed delirium prediction models may be applied to prospective longitudinal cohort studies for validation purposes. A prospective design would enable the measurement of all predictors identified in the developed models and delirium occurrences, which would minimize missing values and maintain the integrity of the data; this may also help examine the validity of the developed models. In addition, environmental factors that are important delirium risk factors but not recorded in EMR, such as sleep disturbance (Moon et al., 2018; Fan et al., 2019), may be observed during ICU stay.

Second, this study did not consider free-text data as a source of predictors. Further studies propose using and validating unstructured data, including the clinical context, to predict delirium accurately. In addition, to utilize unstructured data in prediction models, it is necessary to extract and validate vocabulary lists indicating delirium risk in EMR data through healthcare providers' knowledge and expertise.

Third, the findings of the current study may be used to facilitate the development of prediction algorithms or formulas, such as clinical decision support systems, for clinical practice applications. These formulas may be generated using the identified features of the developed delirium prediction models, and the developed formula should be validated using other datasets. In addition, such prediction formulas may be applied to EMR in the

form of clinical decision support systems to predict delirium.

Finally, several important features in the prediction models were derived from nursing documentation patterns. These patterns have been identified in prediction models of other clinical outcomes, such as cardiac arrest, mortality, and pressure injury. Future studies should also identify changes in delirium-specific nurses' behavior, such as increased surveillance activity and additional observations before delirium onset, and these should be extracted as documentation or intervention patterns stored in the EMR.

VII. CONCLUSION

This study developed and validated machine learning-based delirium prediction models for ICU patients using nursing data that reflected time variation in EMRs. The most important predictors of delirium prediction models were GCS, RASS, pain, and documentation frequency. The best machine learning method in developed models was random forest, and it was validated internally and externally in two different hospitals' EMR environments with acceptable performance. Therefore, the delirium prediction models developed using nursing data with predisposing and precipitating factors were able to discriminate patients at risk of delirium occurrence. In addition, the importance of nursing data that reflected nurses' clinical judgments and time variation in patients' conditions was identified. The evidence and insights revealed by the current study may be used for developing delirium prediction algorithms for application in clinical practice as a form of clinical decision support systems in EMR. Future research can utilize the result of the current study to develop prediction models associated with clinical outcomes.

REFERENCE

Abhyankar, S., Demner-Fushman, D., Callaghan, F. M., & McDonald, C. J. (2014). Combining structured and unstructured data to identify a cohort of ICU patients who received dialysis. *Journal of the American Medical Informatics Association : JAMIA, 21*(5), 801-807. https://doi.org/10.1136/amiajnl-2013-001915

Aldecoa, C., Bettelli, G., Bilotta, F., Sanders, R. D., Audisio, R., Borozdina, A., Cherubini, A., Jones, C., Kehlet, H., MacLullich, A., Radtke, F., Riese, F., Slooter, A. J. C., Veyckemans, F., Kramer, S., Neuner, B., & Weiss, B., & Spies, C. D. (2017). European Society of Anaesthesiology evidence-based and consensus-based guideline on postoperative delirium. *European Journal of Anaesthesiology, 34*(4), 192-214. https://doi.org/10.1097/EJA.0000000000000594

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). https://doi.org/10.1176/appi.books.9780890425596

Awad, A., Bader-El-Den, M., McNicholas, J., & Briggs, J. (2017). Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach. *International Journal of Medical Informatics, 108*, 185-195. https://doi.org/10.1016/j.ijmedinf.2017.10.002

Baker, L., Maley, J. H., Arevalo, A., DeMichele, F., 3rd, Mateo-Collado, R., Finkelstein, S., & Celi, L. A. (2020). Real-world characterization of blood glucose control and insulin use in the intensive care unit. *Scientific Reports, 10*(1), 10718. https://doi.org/10.1038/s41598-020-67864-z

Barnett, V., & Lewis, T. (1984). *Outliers in statistical data*. Wiley.

Barr, J., Fraser, G. L., Puntillo, K., Ely, E. W., Gelinas, C., Dasta, J. F., Davidson, J. E., Devlin, J. W., Kress, J. P., Joffe, A. M., Coursin, D. B., Herr, D. L., Tung, A., Robinson, B. R. H., Fontaine, D. K., Ramsay, M. A., Riker, R. R., Sessler, C. N., Pun, B., Skrobik, Y., & Jaeschke, R. (2013). Clinical practice guidelines for the

management of pain, agitation, and delirium in adult patients in the intensive care unit. *Critical Care Medicine, 41*(1), 263-306. https://doi.org/10.1097/CCM.0b013e3182783b72

Barr, J., & Pandharipande, P. P. (2013). The pain, agitation, and delirium care bundle: Synergistic benefits of implementing the 2013 pain, agitation, and delirium guidelines in an integrated and interdisciplinary fashion. *Critical Care Medicine, 41*(9 Suppl 1), S99-115. https://doi.org/10.1097/CCM.0b013e3182a16ff0

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter, 6*(1), 20-29. https://doi.org/10.1145/1007730.1007735

Beauchet, O., Noublanche, F., Simon, R., Sekhon, H., Chabot, J., Levinoff, E. J., Kabeshova, A., & Launay, C. P. (2018). Falls risk prediction for older inpatients in acute care medical wards: Is there an interest to combine an early nurse assessment and the artificial neural network analysis? *Journal of Nutrition, Health & Aging, 22*(1), 131-137. https://doi.org/10.1007/s12603-017-0950-z

Bergeron, N., Dubois, M. J., Dumont, M., Dial, S., & Skrobik, Y. (2001). Intensive Care Delirium Screening Checklist: Evaluation of a new screening tool. *Intensive Care Medicine, 27*(5), 859-864. https://doi.org/10.1007/s001340100909

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32. https://doi.org/10.1023/a:1010933404324

Capan, M., Wu, P., Campbell, M., Mascioli, S., & Jackson, E. V. (2017). Using electronic health records and nursing assessment to redesign clinical early recognition systems. *Health Systems, 6*(2), 112-121. https://doi.org/10.1057/hs.2015.19

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research, 16*, 321-357. https://doi.org/10.1613/jair.953

Chen, T. J., Chung, Y. W., Chang, H. R., Chen, P. Y., Wu, C. R., Hsieh, S. H., & Chiu, H. Y. (2021). Diagnostic accuracy of the CAM-ICU and ICDSC in detecting

intensive care unit delirium: A bivariate meta-analysis. *International Journal of Nursing Studies, 113*, 103782. https://doi.org/10.1016/j.ijnurstu.2020.103782

Chen, X., Lao, Y., Zhang, Y., Qiao, L., & Zhuang, Y. (2020). Risk predictive models for delirium in the intensive care unit: A systematic review and meta-analysis. *Annals of Palliative Medicine*. 10(*2*), 1467. https://doi.org/10.21037/apm-20-1183

Chen, Y., Du, H., Wei, B. H., Chang, X. N., & Dong, C. M. (2017). Development and validation of risk-stratification delirium prediction model for critically ill patients: A prospective, observational, single-center study. *Medicine (Baltimore), 96*(29), e7543. https://doi.org/10.1097/MD.0000000000007543

Cherak, S. J., Soo, A., Brown, K. N., Ely, E. W., Stelfox, H. T., & Fiest, K. M. (2020). Development and validation of delirium prediction model for critically ill adults parameterized to ICU admission acuity. *PLoS One, 15*(8), e0237639. https://doi.org/10.1371/journal.pone.0237639

Chiarici, A., Andrenelli, E., Serpilli, O., Andreolini, M., Tedesco, S., Pomponio, G., Gallo, M. M., Martini, C., Papa, R., & Ceravolo, M. G. (2019). An early tailored approach is the key to effective rehabilitation in the intensive care unit. *Archives of Physical Medicine and Rehabilitation, 100*(8), 1506-1514. https://doi.org/10.1016/j.apmr.2019.01.015

Cho, I., Boo, E. H., Chung, E., Bates, D. W., & Dykes, P. (2019). Novel approach to inpatient fall risk prediction and its cross-site validation using time-variant data. *Journal of Medical Internet Research, 21*(2), e11505. https://doi.org/10.2196/11505

Collet, M. O., Caballero, J., Sonneville, R., Bozza, F. A., Nydahl, P., Schandl, A., Wøien, H., Citerio, G., van den Boogaard, M., Hästbacka, J., Haenggi, M., Colpaert, K., Rose, L., Barbateskovic, M., Lange, T., Jensen, A., Krog, M. B., Egerod, I., Nibro, H. L.,. . . co-authors, A.-I. c. s. (2018). Prevalence and risk factors related to haloperidol use for delirium in adult intensive care patients: The multinational AID-ICU inception cohort study. *Intensive Care Medicine, 44*(7), 1081-1089.

https://doi.org/10.1007/s00134-018-5204-y

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. (2015). Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD statement. *British Journal of Surgery, 102*(3), 148-158. https://doi.org/10.1002/bjs.9736

Collins, S. A., Cato, K., Albers, D., Scott, K., Stetson, P. D., Bakken, S., & Vawdrey, D. K. (2013). Relationship between nursing documentation and patients' mortality. *American Journal of Critical Care, 22*(4), 306-313. https://doi.org/10.4037/ajcc2013426

Collins, S. A., & Vawdrey, D. K. (2012). "Reading between the lines" of flow sheet data: Nurses' optional documentation associated with cardiac arrest outcomes. *Applied Nursing Research, 25*(4), 251-257. https://doi.org/10.1016/j.apnr.2011.06.002

Connolly, B., O'Neill, B., Salisbury, L., McDowell, K., Blackwood, B., & Enhanced Recovery After Critical Illness Programme, G. (2015). Physical rehabilitation interventions for adult patients with critical illness across the continuum of recovery: An overview of systematic reviews protocol. *Systematic Reviews, 4*(1), 130. https://doi.org/10.1186/s13643-015-0119-y

Datar, M., Gionis, A., Indyk, P., & Motwani, R. (2002). Maintaining stream statistics over sliding windows. *SIAM Journal on Computing, 31*(6), 1794-1813. https://doi.org/10.1137/s0097539701398363

De Georgia, M. A., Kaffashi, F., Jacono, F. J., & Loparo, K. A. (2015). Information technology in critical care: review of monitoring and data acquisition systems for patient care and research. *TheScientificWorldJournal, 2015*, 727694. https://doi.org/10.1155/2015/727694

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics, 44*(3), 837-845. https://www.ncbi.nlm.nih.gov/pubmed/3203132

Desautels, T., Das, R., Calvert, J., Trivedi, M., Summers, C., Wales, D. J., & Ercole, A. (2017). Prediction of early unplanned intensive care unit readmission in a UK tertiary care hospital: A cross-sectional machine learning approach. *BMJ Open, 7*(9), e017199. https://doi.org/10.1136/bmjopen-2017-017199

Devlin, J. W., Skrobik, Y., Gelinas, C., Needham, D. M., Slooter, A. J. C., Pandharipande, P. P., Watson, P. L., Weinhouse, G. L, Nunnally, M. E., Rochwerg, B., Balas, M. C., van den Boogaard, M., Bosma, K. J., Brummel, N. E., Chanques, G., Denehy, L., Drouot, X., Fraser, G. L., Harris, J. E., . . . Alhazzani, W. (2018). Clinical practice guidelines for the prevention and management of pain, agitation/sedation, delirium, immobility, and sleep disruption in adult patients in the ICU. *Critical Care Medicine, 46*(9), e825-e873. https://doi.org/10.1097/CCM.0000000000003299

Donabedian, A. (1966). Evaluating the quality of medical care. *Milbank Memorial Fund Quarterly, 44*(3), Suppl:166-206. https://www.ncbi.nlm.nih.gov/pubmed/5338568

Douw, G., Huisman-de Waal, G., van Zanten, A. R., van der Hoeven, J. G., & Schoonhoven, L. (2016). Nurses' 'worry' as predictor of deteriorating surgical ward patients: A prospective cohort study of the Dutch-Early-Nurse-Worry-Indicator-Score. *International Journal of Nursing Studies, 59*, 134-140. https://doi.org/10.1016/j.ijnurstu.2016.04.006

Douw, G., Schoonhoven, L., Holwerda, T., Huisman-de Waal, G., van Zanten, A. R., van Achterberg, T., & van der Hoeven, J. G. (2015). Nurses' worry or concern and early recognition of deteriorating patients on general wards in acute care hospitals: A systematic review. *Critical Care, 19*(1), 230. https://doi.org/10.1186/s13054-015-0950-5

Dykes, P. C., Kim, H.-e., Goldsmith, D. M., Choi, J., Esumi, K., & Goldberg, H. S. (2009). The adequacy of ICNP version 1.0 as a representational model for electronic nursing assessment documentation. *Journal of the American Medical Informatics Association : JAMIA, 16*(2), 238-246. https://doi.org/10.1197/jamia.

M2956

El-Rashidy, N., El-Sappagh, S., Abuhmed, T., Abdelrazek, S., & El-Bakry, H. M. (2020). Intensive care unit mortality prediction: An improved patient-specific stacking ensemble model. *IEEE Access : practical Innovations, Open solutions, 8*, 133541-133564. https://doi.org/10.1109/access.2020.3010556

Ely, E. W., Margolin, R., Francis, J., May, L., Truman, B., Dittus, R., Speroff, T., Gautam, S., Bernard, G. R., & Inouye, S. K. (2001). Evaluation of delirium in critically ill patients: validation of the Confusion Assessment Method for the Intensive Care Unit (CAM-ICU). *Critical Care Medicine, 29*(7), 1370-1379. https://doi.org/10.1097/00003246-200107000-00012

Fan, H., Ji, M., Huang, J., Yue, P., Yang, X., Wang, C., & Ying, W. (2019). Development and validation of a dynamic delirium prediction rule in patients admitted to the intensive care units (DYNAMIC-ICU): A prospective cohort study. *International Journal of Nursing Studies, 93*, 64-73. https://doi.org/10.1016/j.ijnurstu.2018.10.008

Fu, L. H., Knaplund, C., Cato, K., Perotte, A., Kang, M. J., Dykes, P. C., Albers, D., & Collins Rossetti, S. (2021). Utilizing timestamps of longitudinal electronic health record data to classify clinical deterioration events. *Journal of the American Medical Informatics Association : JAMIA, 28*(9), 1955-1963. https://doi.org/10.1093/jamia/ocab111

Garcia-Gallo, J. E., Fonseca-Ruiz, N. J., Celi, L. A., & Duitama-Munoz, J. F. (2020). A machine learning-based model for 1-year mortality prediction in patients admitted to an intensive care unit with a diagnosis of sepsis. *Medicina Intensiva, 44*(3), 160-170. https://doi.org/10.1016/j.medin.2018.07.016

García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M., & Herrera, F. (2016). Big data preprocessing: Methods and prospects. *Big Data Analytics, 1*(1), 1-22. https://doi.org/10.1186/s41044-016-0014-0

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*.

O'Reilly Media, Inc.

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C-K., & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation, 101*(23), E215-220. https://doi.org/10.1161/01.cir.101.23.e215

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2017). *Deep learning* (Vol. 1). MIT press

Green, C., Bonavia, W., Toh, C., & Tiruvoipati, R. (2019). Prediction of ICU delirium: Validation of current delirium predictive models in routine clinical practice. *Critical Care Medicine, 47*(3), 428-435. https://doi.org/10.1097/CCM.0000000000003577

Guenther, U., Theuerkauf, N., Frommann, I., Brimmers, K., Malik, R., Stori, S., Scheidemann, M., Putensen, C., & Popp, J. (2013). Predisposing and precipitating factors of delirium after cardiac surgery: A prospective observational cohort study. *Annals of Surgery, 257*(6), 1160-1167. https://doi.org/10.1097/SLA.0b013e318281b01c

Gusmao-Flores, D., Salluh, J. I., Chalhub, R. A., & Quarantini, L. C. (2012). The confusion assessment method for the intensive care unit (CAM-ICU) and intensive care delirium screening checklist (ICDSC) for the diagnosis of delirium: A systematic review and meta-analysis of clinical studies. *Critical Care, 16*(4), R115. https://doi.org/10.1186/cc11407

Halladay, C. W., Sillner, A. Y., & Rudolph, J. L. (2018). Performance of electronic prediction rules for prevalent delirium at hospital admission. *JAMA Network Open, 1*(4), e181405. https://doi.org/10.1001/jamanetworkopen.2018.1405

Han, K., Song, K., & Choi, B. W. (2016). How to develop, validate, and compare clinical prediction models involving radiological parameters: Study design and statistical methods. *Korean Journal of Radiology, 17*(3), 339-350. https://doi.org/10.

3348/kjr.2016.17.3.339

Han, Y. Q., Zhang, L., Yan, L., Li, P., Ouyang, P. H., Lippi, G., & Hu, Z. D. (2018). Red blood cell distribution width predicts long-term outcomes in sepsis patients admitted to the intensive care unit. *Clinica Chimica Acta, 487*, 112-116. https://doi.org/10.1016/j.cca.2018.09.019

Hashir, M., & Sawhney, R. (2020). Towards unstructured mortality prediction with free-text clinical notes. *Journal of Biomedical Informatics, 108*, 103489. https://doi.org/10.1016/j.jbi.2020.103489

Heyming, T. W., Knudsen-Robbins, C., Feaster, W., & Ehwerhemuepha, L. (2021). Criticality index conducted in pediatric emergency department triage. *American Journal of Emergency Medicine, 48*, 209-217. https://doi.org/10.1016/j.ajem.2021.05.004

Horng, S., Sontag, D. A., Halpern, Y., Jernite, Y., Shapiro, N. I., & Nathanson, L. A. (2017). Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. *PLoS One, 12*(4), e0174708. https://doi.org/10.1371/journal.pone.0174708

Hshieh, T. T., Yue, J., Oh, E., Puelle, M., Dowal, S., Travison, T., & Inouye, S. K. (2015). Effectiveness of multicomponent nonpharmacological delirium interventions: A meta-analysis. *JAMA Internal Medicine, 175*(4), 512-520. https://doi.org/10.1001/jamainternmed.2014.7779

Huang, K., Gray, T. F., Romero-Brufau, S., Tulsky, J. A., & Lindvall, C. (2021). Using nursing notes to improve clinical outcome prediction in intensive care patients: A retrospective cohort study. *Journal of the American Medical Informatics Association : JAMIA, 28*(8), 1660-1666. https://doi.org/10.1093/jamia/ocab051

Hur, S., Ko, R. E., Yoo, J., Ha, J., Cha, W. C., & Chung, C. R. (2021). A machine learning-based algorithm for the prediction of intensive care unit delirium (PRIDE): Retrospective study. *JMIR Medical Informatics, 9*(7), e23401. https://doi.org/10.2196/23401

Inouye, S. K., & Charpentier, P. A. (1996). Precipitating factors for delirium in hospitalized elderly persons: Predictive model and interrelationship with baseline vulnerability. *JAMA, 275*(11), 852-857. https://www.ncbi.nlm.nih.gov/pubmed/8596223

Irizarry, R. A. (2020). *Introduction to data science: Data analysis and prediction algorithms with R.* CRC Press.

Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L. A., & Mark, R. (2021). MIMIC-IV (version 1.0). *PhysioNet*. https://doi.org/https://doi.org/10.13026/s6n6-xd98

Kallner, A. (2018). Formulas. In A. Kallner (Ed.), *Laboratory Statistics* (pp. 1-140): Elsevier. https://doi.org/10.1016/b978-0-12-814348-3.00001-0

Kang, M. J., Dykes, P. C., Korach, T. Z., Zhou, L., Schnock, K. O., Thate, J., Whalen, K., Jia, H., Schwartz, J., Garcia, J. P., Knaplund, C., Cato, K. D., & Rossetti, S. C. (2020). Identifying nurses' concern concepts about patient deterioration using a standard nursing terminology. *International Journal of Medical Informatics, 133*, 104016. https://doi.org/10.1016/j.ijmedinf.2019.104016

Kim, D. H., Lee, J., Kim, C. A., Huybrechts, K. F., Bateman, B. T., Patorno, E., & Marcantonio, E. R. (2017). Evaluation of algorithms to identify delirium in administrative claims and drug utilization database. *Pharmacoepidemiology and Drug Safety, 26*(8), 945-953. https://doi.org/10.1002/pds.4226

Krewulak, K. D., Stelfox, H. T., Ely, E. W., & Fiest, K. M. (2020). Risk factors and outcomes among delirium subtypes in adult ICUs: A systematic review. *Journal of Critical Care, 56*, 257-264. https://doi.org/10.1016/j.jcrc.2020.01.017

Krewulak, K. D., Stelfox, H. T., Leigh, J. P., Ely, E. W., & Fiest, K. M. (2018). Incidence and prevalence of delirium subtypes in an adult ICU: A systematic review and meta-analysis. *Critical Care Medicine, 46*(12), 2029-2035. https://doi.org/10.1097/ CCM.0000000000003402

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer.

Lewis, S. L., Bucher, L., Heitkemper, M. M., Harding, M. M., Kwong, J., & Roberts, D. (2017). *Medical-surgical nursing: Assessment and management of clinical problems* (10th ed.). Elsevier.

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793). John Wiley & Sons.

Longadge, R., & Dongre, S. (2013). Class imbalance problem in data mining review. *arXiv preprint arXiv:1305.1707*. https://doi.org/10.48550/arXiv.1305.1707

López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences, 250*, 113-141. https://doi.org/10.1016/j.ins.2013.07.007

Marra, A., Pandharipande, P. P., Shotwell, M. S., Chandrasekhar, R., Girard, T. D., Shintani, A. K., Peelen, L. M., Moons, K. G., Dittus, R. S., & Vasilevskis, E. E. (2018). Acute brain dysfunction: Development and validation of a daily prediction model. *Chest, 154*(2), 293-301. https://doi.org/10.1016/j.chest.2018.03.013

Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., & Hamprecht, F. A. (2009). A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics, 10*(1), 213. https://doi.org/10.1186/1471-2105-10-213

Moon, K. J., Jin, Y., Jin, T., & Lee, S. M. (2018). Development and validation of an automated delirium risk assessment system (Auto-DelRAS) implemented in the electronic health record system. *International Journal of Nursing Studies, 77*, 46-53. https://doi.org/10.1016/j.ijnurstu.2017.09.014

Mufti, H. N., Hirsch, G. M., Abidi, S. R., & Abidi, S. S. R. (2019). Exploiting machine learning algorithms and methods for the prediction of agitated delirium after cardiac surgery: Models development and validation Study. *JMIR Medical*

Informatics, 7(4), e14993. https://doi.org/10.2196/14993

Odell, M., Victor, C., & Oliver, D. (2009). Nurses' role in detecting deterioration in ward patients: Systematic literature review. *Journal of Advanced Nursing, 65*(10), 1992-2006. https://doi.org/10.1111/j.1365-2648.2009.05109.x

Oh, J., Cho, D., Park, J., Na, S. H., Kim, J., Heo, J., Shin, C. S., Kim, J-I., Park, J. Y., & Lee, B. (2018). Prediction and early detection of delirium in the intensive care unit by using heart rate variability and machine learning. *Physiological Measurement, 39*(3), 035004. https://doi.org/10.1088/1361-6579/aaab07

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology, 49*(12), 1373-1379. https://doi.org/10.1016/s0895-4356(96)00236-3

R Core Team. (2021). R: A language and environment for statistical computing. In. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Riley, R. D., Ensor, J., Snell, K. I. E., Harrell, F. E., Jr., Martin, G. P., Reitsma, J. B., Moons, K. G., Collins, G., & van Smeden, M. (2020). Calculating the sample size required for developing a clinical prediction model. *BMJ (Clinical Research Ed.), 368*, m441. https://doi.org/10.1136/bmj.m441

Riley, R. D., Snell, K. I., Ensor, J., Burke, D. L., Harrell Jr, F. E., Moons, K. G. M., & Collins, G. S. (2019). Minimum sample size for developing a multivariable prediction model: Part II-binary and time-to-event outcomes. *Statistics in Medicine, 38*(30), 5672. https://doi.org/10.1002/sim.7992

Rojas, J. C., Carey, K. A., Edelson, D. P., Venable, L. R., Howell, M. D., & Churpek, M. M. (2018). Predicting intensive care unit readmission with machine learning using electronic health record data. *Annals of the American Thoracic Society, 15*(7), 846-853. https://doi.org/10.1513/AnnalsATS.201710-787OC

Romero-Brufau, S., Gaines, K., Nicolas, C. T., Johnson, M. G., Hickman, J., &

Huddleston, J. M. (2019). The fifth vital sign? nurse worry predicts inpatient deterioration within 24 hours. *JAMIA Open, 2*(4), 465-470. https://doi.org/10.1093/jamiaopen/ooz033

Rood, P., Huisman-de Waal, G., Vermeulen, H., Schoonhoven, L., Pickkers, P., & van den Boogaard, M. (2018). Effect of organisational factors on the variation in incidence of delirium in intensive care unit patients: A systematic review and meta-regression analysis. *Australian Critical Care, 31*(3), 180-187. https://doi.org/10.1016/j.aucc.2018.02.002

Rossetti, S. C., Knaplund, C., Albers, D., Dykes, P. C., Kang, M. J., Korach, T. Z., Zhou, L., Schnock, K., Garcia, J., Schwartz, J., Fu, L-H., Klann, J. G., Lowenthal, G., & Cato, K. (2021). Healthcare Process Modeling to Phenotype Clinician Behaviors for Exploiting the Signal Gain of Clinical Expertise (HPM-ExpertSignals): Development and evaluation of a conceptual framework. *Journal of the American Medical Informatics Association : JAMIA, 28*(6), 1242-1251. https://doi.org/10.1093/jamia/ocab006

Rossetti, S. C., Knaplund, C., Albers, D., Tariq, A., Tang, K., Vawdrey, D., Yip, N. H., Dykes, P. C., Klann, J. G., & Kang, M. J. (2019). Leveraging clinical expertise as a feature-not an outcome-of predictive models: Evaluation of an early warning system use case. *AMIA Annual Symposium Proceedings, 2019*, 323-332.

Rothman, M. J., Rothman, S. I., & Beals, J. t. (2013). Development and validation of a continuous measure of patient condition using the Electronic Medical Record. *Journal of Biomedical Informatics, 46*(5), 837-848. https://doi.org/10.1016/j.jbi.2013.06.011

Salgado, C. M., Azevedo, C., Proenca, H., & Vieira, S. M. (2016a). Missing data. In M. I. T. C. Data (Ed.), *Secondary Analysis of Electronic Health Records* (pp. 143-162). Cham (CH): Springer International Publishing. https://doi.org/10.1007/978-3-319-43742-2_13

Salgado, C. M., Azevedo, C., Proenca, H., & Vieira, S. M. (2016b). Noise versus outliers.

In M. I. T. C. Data (Ed.), *Secondary Analysis of Electronic Health Records* (pp. 163-183). Cham (CH): Springer International Publishing. https://doi.org/10.1007/978-3-319-43742-2_14

Schnock, K. O., Kang, M. J., Rossetti, S. C., Garcia, J., Lowenthal, G., Knaplund, C., Chang, F., Albers, D., Korach, T. Z., Zhou, L., Klann, J. G., Cato, K., Bates, D., & Dykes, P. C. (2021). Identifying nursing documentation patterns associated with patient deterioration and recovery from deterioration in critical and acute care settings. *International Journal of Medical Informatics, 153*, 104525. https://doi.org/10.1016/j.ijmedinf.2021.104525

Soar, J., Bottiger, B. W., Carli, P., Couper, K., Deakin, C. D., Djarv, T., Lott, C., Olasveengen, T., Paal, P., Pellis, T., Perkins, G. D., Sandroni, C., & Nolan, J. P. (2021). European resuscitation council guidelines 2021: Adult advanced life support. *Resuscitation, 161*, 115-151. https://doi.org/10.1016/j.resuscitation.2021.02.010

Song, W., Kang, M. J., Zhang, L., Jung, W., Song, J., Bates, D. W., & Dykes, P. C. (2021). Predicting pressure injury using nursing assessment phenotypes and machine learning methods. *Journal of the American Medical Informatics Association : JAMIA, 28*(4), 759-765. https://doi.org/10.1093/jamia/ocaa336

Song, Y., Gao, S., Tan, W., Qiu, Z., Zhou, H., & Zhao, Y. (2019). Dexmedetomidine versus midazolam and propofol for sedation in critically ill patients: Mining the Medical Information Mart for Intensive Care data. *Annals of Translational Medicine, 7*(9), 197. https://doi.org/10.21037/atm.2019.04.14

Steinmetz, J., Siersma, V., Kessing, L. V., Rasmussen, L. S., & Group, I. (2013). Is postoperative cognitive dysfunction a risk factor for dementia? A cohort follow-up study. *British Journal of Anaesthesia, 110 Suppl 1*(suppl_1), i92-97. https://doi.org/10.1093/bja/aes466

Stevens, L. M., Mortazavi, B. J., Deo, R. C., Curtis, L., & Kao, D. P. (2020). Recommendations for reporting machine learning analyses in clinical research.

Circulation. Cardiovascular Quality and Outcomes, 13(10), e006556. https://doi.org/10.1161/CIRCOUTCOMES.120.006556

Steyerberg, E. W. (2009). *Clinical prediction models*. Springer.

Sundararaman, A., Valady Ramanathan, S., & Thati, R. (2018). Novel approach to predict hospital readmissions using feature selection from unstructured data with class imbalance. *Big Data Research, 13*, 65-75. https://doi.org/10.1016/j.bdr.2018.05.004

Tsiklidis, E. J., Sinno, T., & Diamond, S. L. (2022). Predicting risk for trauma patients using static and dynamic information from the MIMIC III database. *PLoS One, 17*(1), e0262523. https://doi.org/10.1371/journal.pone.0262523

van den Boogaard, M., Pickkers, P., Slooter, A. J., Kuiper, M. A., Spronk, P. E., van der Voort, P. H., Van Der Hoeven, J. G., Donders, R., van Achterberg, T., & Schoonhoven, L. (2012). Development and validation of PRE-DELIRIC (PREdiction of DELIRium in ICu patients) delirium prediction model for intensive care patients: Observational multicentre study. *BMJ (Clinical Research Ed.), 344*, e420. https://doi.org/10.1136/bmj.e420

van den Boogaard, M., Schoonhoven, L., Maseda, E., Plowright, C., Jones, C., Luetz, A., Sackey, P. V., Jorens, P. G., Aitken, L. M., van Haren, F. M. P., Donders, R., van der Hoeven, J. G., & Pickkers, P. (2014). Recalibration of the delirium prediction model for ICU patients (PRE-DELIRIC): A multinational observational study. *Intensive Care Medicine, 40*(3), 361-369. https://doi.org/10.1007/s00134-013-3202-7

Vincent, J. L., Lefrant, J. Y., Kotfis, K., Nanchal, R., Martin-Loeches, I., Wittebole, X., Sakka, S. G., Pickkers, P., Moreno, R., Sakr, Y., Pavlik, P., Manak, J., Kieslichova, E., Turek, R., Fischer, M., Valkova, R., Dadak, L., Dostal, P., Malaska, J., & investigators, S. (2018). Comparison of European ICU patients in 2012 (ICON) versus 2002 (SOAP). *Intensive Care Medicine, 44*(3), 337-344. https://doi.org/10.1007/s00134-017-5043-2

Wassenaar, A., Schoonhoven, L., Devlin, J. W., van Haren, F. M. P., Slooter, A. J. C., Jorens, P. G., van der Jagt, M., Simons, K. S., Egerod, I., Burry, L. D., Beishuizen, A., Matos, J., Donders, A. R. T., Pickkers, P., & van den Boogaard, M. (2018). Delirium prediction in the intensive care unit: Comparison of two delirium prediction models. *Critical Care, 22*(1), 114. https://doi.org/10.1186/s13054-018-2037-6

Wassenaar, A., van den Boogaard, M., van Achterberg, T., Slooter, A. J., Kuiper, M. A., Hoogendoorn, M. E., Simons, K. S., Maseda, E., Pinto, N., Jones, C., Luetz, A., Schandl, A., Verbrugghe, W., Aitken, L. M., van Haren, F. M. P., Donders, A. R. T., Schoonhoven, L., & Pickkers, P. (2015). Multinational development and validation of an early prediction model for delirium in ICU patients. *Intensive Care Medicine, 41*(6), 1048-1056. https://doi.org/10.1007/s00134-015-3777-2

Wellner, B., Grand, J., Canzone, E., Coarr, M., Brady, P. W., Simmons, J., Pinto, N., Jones, C., Luetz, A., Schandl, A., Verbrugghe, W., Aitken, L. M., van Haren, F. M. P., Donders, A. R. T., Schoonhoven, L., & Sylvester, P. (2017). Predicting unplanned transfers to the intensive care unit: A machine learning approach leveraging diverse clinical elements. *JMIR Medical Informatics, 5*(4), e45. https://doi.org/10.2196/medinform.8680

Witlox, J., Eurelings, L. S., de Jonghe, J. F., Kalisvaart, K. J., Eikelenboom, P., & van Gool, W. A. (2010). Delirium in elderly patients and the risk of postdischarge mortality, institutionalization, and dementia: A meta-analysis. *JAMA, 304*(4), 443-451. https://doi.org/10.1001/jama.2010.1013

Witten, I. H. (2017). *Data mining: practical machine learning tools and techniques* (4th ed.). Elsevier : Morgan Kaufmann.

Witten, I. H., & Witten, I. H. (2017). *Data mining*. Elsevier :Morgan Kaufmann.

Xu, J., Chen, D., Deng, X., Pan, X., Chen, Y., Zhuang, X., & Sun, C. (2022). Development and validation of a machine learning algorithm-based risk prediction model of pressure injury in the intensive care unit. *International*

Wound Journal. https://doi.org/10.1111/iwj.13764

Xu, Z., Feng, Y., Li, Y., Srivastava, A., Adekkanattu, P., Ancker, J. S., Jiang, G., Kiefer, R. C., Lee, K., Pacheco, J. A., Rasmussen, L. V., Pathak, J., Luo, Y., & Wang, F. (2019). Predictive modeling of the risk of acute kidney injury in critical care: A systematic investigation of the class imbalance problem. *AMIA Joint Summits on Translational Science Proceedings, 2019*, 809-818. https://www.ncbi.nlm.nih.gov/pubmed/31259038

Yang, S., Khang, Y. H., Harper, S., Davey Smith, G., Leon, D. A., & Lynch, J. (2010). Understanding the rapid increase in life expectancy in South Korea. *American Journal of Public Health, 100*(5), 896-903. https://doi.org/10.2105/AJPH.2009.160341

Zaal, I. J., Devlin, J. W., Peelen, L. M., & Slooter, A. J. (2015). A systematic review of risk factors for delirium in the ICU. *Critical Care Medicine, 43*(1), 40-47. https://doi.org/10.1097/CCM.0000000000000625

Zadravecz, F. J., Tien, L., Robertson-Dick, B. J., Yuen, T. C., Twu, N. M., Churpek, M. M., & Edelson, D. P. (2015). Comparison of mental-status scales for predicting mortality on the general wards. *Journal of Hospital Medicine, 10*(10), 658-663. https://doi.org/10.1002/jhm.2415

Zazzaro, G., Cuomo, S., Martone, A., Montaquila, R. V., Toraldo, G., & Pavone, L. (2021). EEG signal analysis for epileptic seizures detection by applying data mining techniques. *Internet of Things, 14*, 100048. https://doi.org/10.1016/j.iot.2019.03.002

Zhao, Y., & Luo, Y. (2021). Unsupervised learning to subphenotype delirium patients from electronic health records. *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2949-2961. https://doi.org/10.1109/BIBM52615.2021.9669806

APPENDICES

Appendix 1 TRIPOD Checklist: Model Development and Validation

Section/Topic	Item		Checklist Item	Page
Title and abstract				
Title	1	D;V	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	vii
Abstract	2	D;V	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	vii
Introduction				
Background and objectives	3a	D;V	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	1–4
	3b	D;V	Specify the objectives, including whether the study describes the development or validation of the model or both.	4–5
Methods				
Source of data	4a	D;V	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	28–29
	4b	D;V	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	28–29
Participants	5a	D;V	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	28–29
	5b	D;V	Describe eligibility criteria for participants.	31–32
	5c	D;V	Give details of treatments received, if relevant.	31–32
Outcome	6a	D;V	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	34–36
	6b	D;V	Report any actions to blind assessment of the outcome to be predicted.	–
Predictors	7a	D;V	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	36–37
	7b	D;V	Report any actions to blind assessment of predictors for the outcome and other predictors.	–
Sample size	8	D;V	Explain how the study size was arrived at.	38
Missing data	9	D;V	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	39–41
Statistical analysis methods	10a	D	Describe how predictors were handled in the analyses.	39–41
	10b	D	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	44–51
	10c	V	For validation, describe how the predictions were calculated.	51–53
	10d	D;V	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	51–53
	10e	V	Describe any model updating (e.g., recalibration) arising from the validation, if done.	51–31
Risk groups	11	D;V	Provide details on how risk groups were created, if done.	34–36
Development vs. validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	47–51

Results				
Participants	13a	D;V	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	54, Figure3
	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	54–59
	13c	V	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	54–59
Model development	14a	D	Specify the number of participants and outcome events in each analysis.	54–59
	14b	D	If done, report the unadjusted association between each candidate predictor and outcome.	Table 8
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	67
	15b	D	Explain how to the use the prediction model.	61–67
Model performance	16	D;V	Report performance measures (with CIs) for the prediction model.	64, 72
Model-updating	17	V	If done, report the results from any model updating (i.e., model specification, model performance).	–
Discussion				
Limitations	18	D;V	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	82–83
Interpretation	19a	V	For validation, discuss the results with reference to performance in the development data, and any other validation data.	75–80
	19b	D;V	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	73–80
Implications	20	D;V	Discuss the potential clinical use of the model and implications for future research.	80–82
Other information				
Supplementary information	21	D;V	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	–
Funding	22	D;V	Give the source of funding and the role of the funders for the present study.	–

Appendix 2 MIMIC-IV database

Module	Table	Definition
Core	admissions	Demographics for the patients
	patients	Record for each hospitalization
	transfers	Record for each ward stay within a hospitalization
hosp	d_hcpcs	Description of recorded in *hpcsevents*
	d_icd_diagnoses	Description of ICD billed diagnoses
	d_icd_procedures	Description of ICD billed procedures
	d_labitems	Description of all laboratory measurements
	diagnoses_icd	Billed ICD diagnoses for patients
	drgcodes	Billed diagnosis related groups (DRG) codes for patients
	emar	Barcode scanning of medications at the time of administration
	emar_detail	Supplementary information or recorded in *emar*
	hpcsevents	Billed events occurring during the hospitalization
	labevents	Contains all laboratory measurements for a given patient
	microbiologyevents	Contains microbiology cultures information
	pharmacy	Detailed information regarding prescriptions (formulary dose, route, frequency, dose, duration)
	poe	Orders made by providers relating to patient care
	poe_detail	Supplementary information for orders made by providers in the hospital
	prescriptions	Contains medication related order entries
	procedures_icd	Billed procedures for patients during their hospital stay
	services	Lists services that a patient was admitted/transferred under
ICU	d_items	Definition table for all items in the ICU module
	chartevents	Contains all charted data for all patients during the ICU stay
	datetimeevents	Contains all date formatted data
	ICU stays	Tracking information for each ICUSTAY_ID in the database
	inputevents	Input data for patients
	outputevents	Output data for patients
	procedureevents	Contains procedures during the ICU stay

Note. DRG = Diagnosis related group; ICD = international classification of disease; ICU = intensive care unit; MIMIC = Medical Information Mart for Intensive Care.

Appendix 3 Credentialing applications from PhysioNet

Appendix 4 Approval from the institutional review board

연세의료원 세브란스병원 연구심의위원회
Yonsei University Health System, Severance Hospital, Institutional Review Board
서울특별시 서대문구 연세로 50-1 (우) 03722
Tel.02 2228 0430~4, 0450~4 Fax.02 2227 7888~9 Email. irb@yuhs.ac

심 의 일 자	2021년 10 월 21 일
접 수 번 호	2021-3095-001
과제승인번호	4-2021-1212

세브란스병원 연구심의위원회의 심의 결과를 다음과 같이 알려 드립니다.

Protocol No.	
연 구 제 목	간호기록의 패턴을 활용한 중환자실 환자의 섬망 예측모델 개발
연 구 책 임 자	최모나 / 세브란스병원 간호학과
의 뢰 자	(학)연세대학교
연구예정기간	2021.10.21 ~ 2022.10.20
지속심의 빈도	면제
과 제 승 인 일	2021.10.21
위 험 수 준	Level I 최소위험
심 의 방 법	신속
심 의 유 형	신규과제
심 의 내 용	- 연구계획서 (국문) - 증례기록서 - 연구책임자 이력 및 경력에 관한 사항
심 의 위 원 회	제7위원회
참 석 위 원	제7위원회 신속심의자
심 의 결 과	승인, 대상자 동의 면제
심 의 의 견	-

※ 본 통보서에 기재된 사항은 세브란스병원 연구심의위원회의 기록된 내용과 일치함을 증명합니다.
※ 세브란스병원 연구심의위원회는 국제 임상시험 통일안(ICH-GCP), 임상시험 관리기준(KGCP), 생명윤리 및 안전에 관한 법률을 준수합니다.
※ 연구책임자 및 연구담당자가 IRB위원인 경우, 해당 위원은 위 연구의 심의과정에 참여하지 않았습니다.

연세의료원 세브란스병원

연구심의위원회 위원장

Appendix 4 Approval from the institutional review board (continued)

데이터 신청 내용 | IRB 신청 내용

데이터 대상 병원	신촌					의료자산활용대상여부	비상정 대상, 비보고 대상
데이터 대상 기간	2018-03-01~2021-08-31		데이터 보유 기간	2021-10-21~2021-08-31		데이터 수령 매체	연세의료원 NAS
영상데이터필요여부	불필요	영상데이터PACS정보		규정준수서약서	3/3등의	보안서약서	
연구 구역	원내 연구 (연세의료원 내부/연구용 클라우드 포함)	데이터 반출 기관		등록번호실명화여부	불필요	DRB연계동의여부	Y
데이터 반출 사유							
요청 데이터 설명 및 항목	2018.03.01~2021.08.31 까지 내과/외과/신경외과 중환자실에 입실한 만 19세 이상의 대상자로 추출이 필요한 데이터는 파일로 첨부하였음.						

제출서류현황 | 보완작업요청및현황 | DRB총괄심의현황 | 작업배정및현황 | 작업진행현황 | 작업확정및이력관리

2021-11차 DRB총괄 심의(2021-12-15~ 2021-12-15) - 심의결과 : 승인

Appendix 5 International Classification of Disease codes

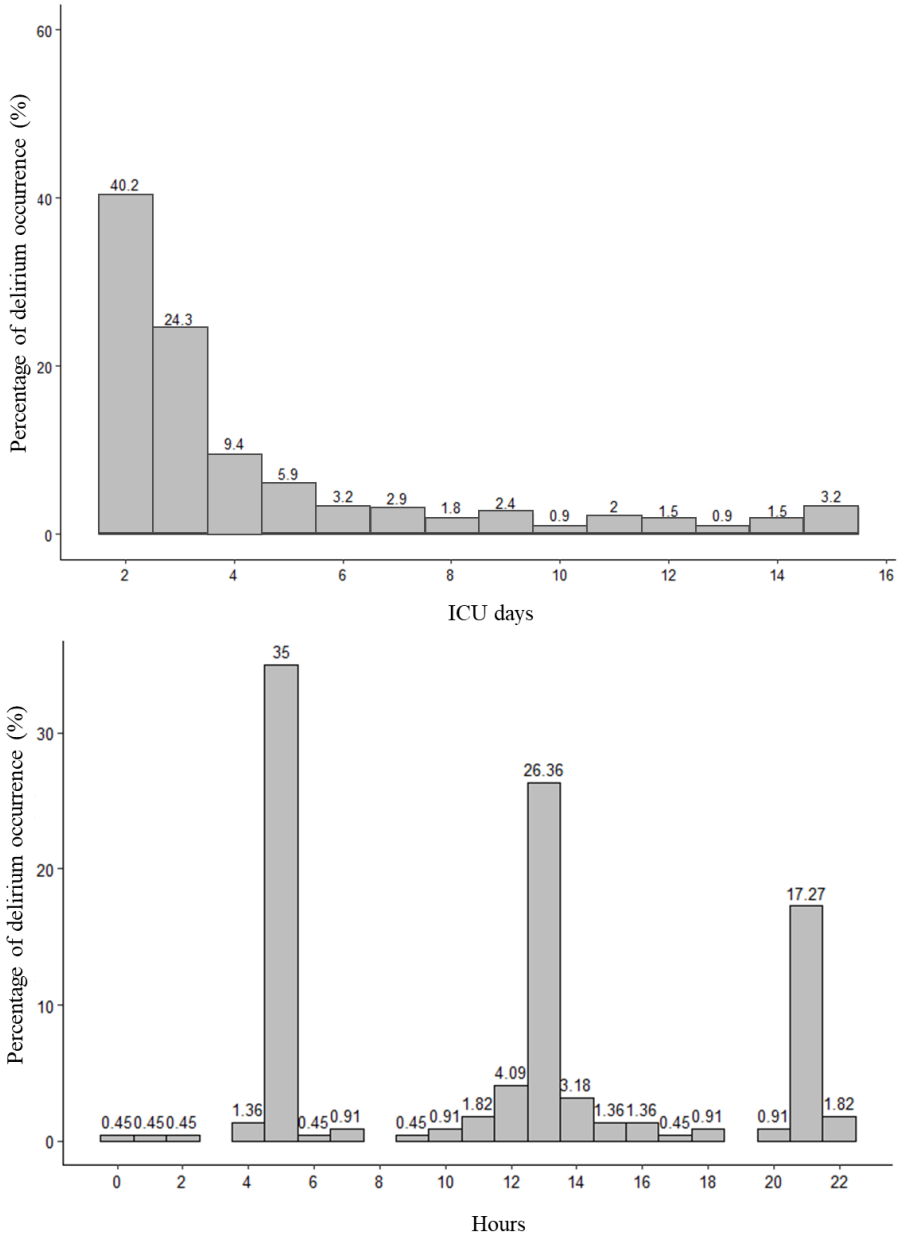
ICD code	ICD version	Code titles
Dementia		
G30	10	Alzheimer's disease
G300	10	Alzheimer's disease with early onset
G301	10	Alzheimer's disease with late onset
G308	10	Other Alzheimer's disease
G309	10	Alzheimer's disease, unspecified
F01	10	Vascular dementia
F015	10	Vascular dementia
F0150	10	Vascular dementia without behavioral disturbance
F0151	10	Vascular dementia with behavioral disturbance
F02	10	Dementia in other diseases classified elsewhere
F028	10	Dementia in other diseases classified elsewhere
F0280	10	Dementia in other diseases classified elsewhere without behavioral disturbance
F0281	10	Dementia in other diseases classified elsewhere with behavioral disturbance
F03	10	Unspecified dementia
F039	10	Unspecified dementia
F0390	10	Unspecified dementia without behavioral disturbance
F0391	10	Unspecified dementia with behavioral disturbance
29010	9	Presenile dementia, uncomplicated
29011	9	Presenile dementia with delirium
29012	9	Presenile dementia with delusional features
29013	9	Presenile dementia with depressive features
3310	9	Alzheimer's disease
29040	9	Vascular dementia, uncomplicated
29041	9	Vascular dementia, with delirium
29042	9	Vascular dementia, with delusions
29043	9	Vascular dementia, with depressed mood
29410	9	Dementia in conditions classified elsewhere without behavioral disturbance
29411	9	Dementia in conditions classified elsewhere with behavioral disturbance
2900	9	Senile dementia, uncomplicated
29020	9	Senile dementia with delusional features
29021	9	Senile dementia with depressive features
2908	9	Other specified senile psychotic conditions
Hypertension		
I10	10	Essential (primary) hypertension
I11	10	Hypertensive heart disease
I110	10	Hypertensive heart disease with heart failure
I119	10	Hypertensive heart disease without heart failure
I12	10	Hypertensive chronic kidney disease
I120	10	Hypertensive chronic kidney disease with stage 5 chronic kidney disease or end stage renal disease
I129	10	Hypertensive chronic kidney disease with stage 1 through stage 4 chronic kidney disease, or unspecified chronic kidney disease
I13	10	Hypertensive heart and chronic kidney disease

Appendix 5 International Classification of Disease codes (continued)

ICD code	ICD version	Code titles
I130	10	Hypertensive heart and chronic kidney disease with heart failure and stage 1 through stage 4 chronic kidney disease, or unspecified chronic kidney disease
I131	10	Hypertensive heart and chronic kidney disease without heart failure
I1310	10	Hypertensive heart and chronic kidney disease without heart failure, with stage 1 through stage 4 chronic kidney disease, or unspecified chronic kidney disease
I1311	10	Hypertensive heart and chronic kidney disease without heart failure, with stage 5 chronic kidney disease, or end stage renal disease
I132	10	Hypertensive heart and chronic kidney disease with heart failure and with stage 5 chronic kidney disease, or end stage renal disease
4010	9	Malignant essential hypertension
4011	9	Benign essential hypertension
4019	9	Unspecified essential hypertension
40200	9	Malignant hypertensive heart disease without heart failure
40201	9	Malignant hypertensive heart disease with heart failure
40210	9	Benign hypertensive heart disease without heart failure
40211	9	Benign hypertensive heart disease with heart failure
40290	9	Unspecified hypertensive heart disease without heart failure
40291	9	Unspecified hypertensive heart disease with heart failure
40400	9	Hypertensive heart and chronic kidney disease, malignant, without heart failure and with chronic kidney disease stage I through stage IV, or unspecified
40401	9	Hypertensive heart and chronic kidney disease, malignant, with heart failure and with chronic kidney disease stage I through stage IV, or unspecified
40402	9	Hypertensive heart and chronic kidney disease, malignant, without heart failure and with chronic kidney disease stage V or end stage renal disease
40403	9	Hypertensive heart and chronic kidney disease, malignant, with heart failure and with chronic kidney disease stage V or end stage renal disease
40410	9	Hypertensive heart and chronic kidney disease, benign, without heart failure and with chronic kidney disease stage I through stage IV, or unspecified
40411	9	Hypertensive heart and chronic kidney disease, benign, with heart failure and with chronic kidney disease stage I through stage IV, or unspecified
40412	9	Hypertensive heart and chronic kidney disease, benign, without heart failure and with chronic kidney disease stage V or end stage renal disease
40413	9	Hypertensive heart and chronic kidney disease, benign, with heart failure and chronic kidney disease stage V or end stage renal disease
40490	9	Hypertensive heart and chronic kidney disease, unspecified, without heart failure and with chronic kidney disease stage I through stage IV, or unspecified
40491	9	Hypertensive heart and chronic kidney disease, unspecified, with heart failure and with chronic kidney disease stage I through stage IV, or unspecified
40492	9	Hypertensive heart and chronic kidney disease, unspecified, without heart failure and with chronic kidney disease stage V or end stage renal disease
40493	9	Hypertensive heart and chronic kidney disease, unspecified, with heart failure and chronic kidney disease stage V or end stage renal disease

Appendix 6 Delirium occurrence based on day and time of onset

External validation cohort



Note. ICU = intensive care unit.

Appendix 7 Confusion matrix of control A models

Control A [a] model	Prediction modeling method	Predicted	Actual		Accuracy [95% CI]
			Delirium	Non-delirium	
Model I	Logistic regression	Delirium	264	395	0.761
		Non-delirium	60	1178	[.740, .779]
	Support vector machine	Delirium	228	246	0.820
		Non-delirium	96	1327	[.802, .837]
	Random Forest	Delirium	244	183	0.861
		Non-delirium	80	1390	[.844, .876]
	Neural network	Delirium	244	331	0.783
		Non-delirium	80	1242	[.764, .802]
Mode II [b]	Logistic regression	Delirium	263	630	0.636
		Non-delirium	61	943	[.614, .657]
	Support vector machine	Delirium	218	435	0.715
		Non-delirium	106	1138	[.694, .735]
	Random Forest	Delirium	238	332	0.780
		Non-delirium	86	1241	[.760, .798]
	Neural network	Delirium	244	537	0.675
		Non-delirium	80	1036	[.653, .696]

Note. CI = confidence interval.

[a] Data were extracted from the first 24 hours after ICU admission in the control group.

[b] Predictors excluding the actual values of GCS and RASS among Model I predictors

Appendix 8 Confusion matrix of control B models

Control B [a] model	Prediction modeling method	Predicted	Actual		Accuracy [95% CI]
			Delirium	Non-delirium	
Model I	Logistic regression	Delirium	251	237	0.837
		Non-delirium	73	1336	[.819, .853]
	Support vector machine	Delirium	247	188	0.860
		Non-delirium	77	1385	[.844, .876]
	Random Forest	Delirium	251	101	0.908
		Non-delirium	73	1472	[.894, .921]
	Neural network	Delirium	249	230	0.839
		Non-delirium	75	1343	[.822, .856]
Mode II [b]	Logistic regression	Delirium	258	409	0.750
		Non-delirium	66	1164	[.730, .769]
	Support vector machine	Delirium	244	311	0.794
		Non-delirium	80	1262	[.775, .812]
	Random Forest	Delirium	258	238	0.840
		Non-delirium	66	1335	[.823, .856]
	Neural network	Delirium	265	424	0.745
		Non-delirium	59	1149	[.725, .765]

Note. CI = confidence interval.

[a] Data were extracted from the last 24 hours before ICU discharge in the control group.

[b] Model used predictors excluding the actual values of GCS and RASS among Model I predictors

Appendix 9 Model performance of control A models

Control A [a] model	AUROC [95% CI]	Sensitivity	Specificity	PPV	NPV	F_1 score	Youden index
Model I							
Logistic regression	0.856 [.834, .878]	0.815	0.749	0.401	0.952	0.537	0.564
Support vector machine	0.774 [.747, .800]	0.704	0.844	0.481	0.933	0.571	0.547
Random forest	0.896 [.877, .914]	0.753	0.884	0.571	0.946	0.650	0.637
Neural Network	0.853 [.831, .875]	0.753	0.790	0.424	0.940	0.543	0.543
Model II [b]							
Logistic regression	0.781 [.753, .808]	0.812	0.599	0.295	0.939	0.432	0.411
Support vector machine	0.698 [.670, .726]	0.673	0.724	0.334	0.915	0.446	0.396
Random forest	0.846 [.823, .868]	0.735	0.789	0.418	0.935	0.532	0.524
Neural Network	0.790 [.762, .818]	0.753	0.659	0.312	0.928	0.442	0.412

Note. AUROC = area under the receiver operating characteristic; CI = confidence interval; NPV = negative predictive value; PPV = positive predictive value.

[a] Data were extracted from the first 24 hours after ICU admission in the control group.

[b] Model used predictors excluding the actual values of GCS and RASS among Model I predictors

Appendix 10 Model performance of control B models

Control B [a] model	AUROC [95% CI]	Sensitivity	Specificity	PPV	NPV	F_1 score	Youden index
Model I							
Logistic regression	0.886 [.863, .908]	0.775	0.849	0.514	0.948	0.618	0.624
Support vector machine	0.821 [.797, .846]	0.762	0.881	0.568	0.947	0.651	0.643
Random forest	0.926 [.940, .962]	0.775	0.936	0.713	0.953	0.743	0.710
Neural Network	0.858 [.832, .884]	0.769	0.854	0.520	0.947	0.620	0.622
Model II [b]							
Logistic regression	0.843 [.818, .869]	0.796	0.740	0.387	0.946	0.521	0.536
Support vector machine	0.779 [.752, .803]	0.753	0.802	0.440	0.940	0.555	0.555
Random forest	0.897 [.878, .916]	0.796	0.849	0.520	0.953	0.629	0.645
Neural Network	0.834 [.809, .859]	0.818	0.730	0.385	0.951	0.523	0.548

Note. AUROC = area under the receiver operating characteristic; CI = confidence interval; NPV = negative predictive value; PPV = positive predictive value.

[a] Data were extracted from the last 24 hours before ICU discharge in the control group.

[b] Model used predictors excluding the actual values of GCS and RASS among Model I predictors

KOREAN ABSTRACT

간호데이터를 이용한 중환자실 환자의 섬망 예측모델

김 미 희

연세대학교 대학원 간호학과

섬망은 중환자실에서 빈번히 발생하는 신경 정신장애로 부적절한 행동과 생각, 인지 및 감각의 급격한 변화를 나타낸다. 중환자실 환자에서 섬망 발생은 환자의 예후에 부정적인 영향을 미쳐 재원 기간을 연장하고, 사망률을 높이며, 의료비용과 의료진의 업무 부담을 증가시킨다. 중환자실에서는 섬망을 조기에 선별하고, 관리하기 위한 전략으로 섬망 선별도구를 이용하여 섬망을 주기적으로 사정하고, 이에 따라 중재를 제공하고 있지만, 모든 환자를 대상으로 한 섬망 선별과 예방 중재는 많은 자원과 인력의 투입이 필요하다. 따라서 섬망을 정확하게 예측하고, 고위험 대상자에게 맞춤형 중재를 제공하는 것이 환자 예후와 의료자원 관리를 위해 중요하다. 전자의무기록에서 간호데이터는 환자 상태에 대한 간호사의 관찰 및 임상 판단과 관련된 정보를 포함하고 있어 빠르게 변화하는 환자의 상태를 예측하는데 중요한 지표로 활용될 수 있다. 따라서 본 연구는 전자의무기록에서 시간의 변동성이 반영된 간호데이터를 이용하여 기계학습 방법을 기반으로 중환자실 환자의 섬망 예측모델을 개발하고 검증하는데 있다.

본 연구는 전자의무기록을 이용한 후향적 연구로 섬망 예측모델 개발 및 내부 타당도 검증을 위해 미국 중환자실 데이터인 Medical Information Mart for Intensive Care (MIMIC) 데이터베이스를 이용하였고, 외부 타당도 검증을 위해 국내 상급종합병원의 전자의무기록을 이용하였다. 본 연구에서는 중환자실에서 24시간 이상 치료를 받은 18세 이상의 대상자 중 입실 24시간 이내에 섬망 선별도구를 이용한 기록이 1회 이상인 환자를 대상으로 하였고, 호스피스나 완화치료를 받고 있거나 입실 24시간 이내에 섬망이 발생한 대상자는 제외하였다. 모델 개발에 포함된 예측인자는 섬망 발생 직전 24시간 데이터를 추출하여 대상자별로 소인요인, 촉진요인, 간호사정 기록, 간호기록의 빈도와 중재 패턴을 선별하였다. 최종 생성된 개발 코호트는 9,491명, 외부 코호트는 2,629명의 대상자가 포함되었고, 이중 섬망은 17.0%, 8.4%에서 발생하였다. 개발 코호트는 섬망 클래스에 대한 데이터 불균형을 해결하기 위해 결합 샘플링 방법을 이용하여 데이터의 균형을 유지하였다. 본 연구에서는 기계학습 방법 중 로지스틱 회귀분석, 서포트 벡터 머신, 랜덤 포레스트, 신경망을 기반으로 10배 교차검증을 수행하여 예측모델을 개발하였고, 개발된 모델의 예측성능을 확인하기 위해 내부 검증과 외부 코호트 데이터를 이용하여 외부 검증을 수행하였다.

40개의 예측인자로 개발된 모델 I에서는 랜덤 포레스트 기반 알고리즘 모델이 가장 높은 예측 성능을 보였고, 곡선 아래 면적 (Area under the receiver operating characteristic, AUROC)은 내부 검증에서 0.975, 외부 검증에서 0.770으로 확인되었다. 개발된 모델의 예측변수 중 Glasgow Coma Scale (GCS), Richmond Agitation-Sedation scale (RASS), 통증, 간호기록의 빈도 패턴은 변수 중요도

지수에서 소인이나 촉진요인보다 더 높은 중요도에 위치함을 확인하였다.

본 연구는 간호데이터를 이용하여 중환자실 환자의 섬망 예측모델을 개발하고, 개발된 모델의 성능을 내부 및 외부 데이터를 이용하여 평가하였다. 또한, 간호사의 임상적 판단이 포함된 간호데이터는 섬망 예측의 중요한 변수로 확인되었다. 본 연구 결과는 섬망 예측모델을 임상 현장에 적용하기 위해 전자의무기록의 의사결정지원 시스템의 형태로 예측 알고리즘을 개발하는데 기초자료로 활용될 수 있을 것이다.

핵심되는 말: 간호기록, 간호사정, 머신러닝, 섬망, 예측모델, 전자의무기록, 중환자실