

Traumatic brain injury patients based on MIMIC-IV database Construction and validation of early epilepsy prediction model

Zepeng Qiu

Fujian Medical University

Guan Wei

76595825@qq.com

Fujian Medical University

Shuijin Wu

Fujian Medical University

Research Article

Keywords: traumatic brain injury, early epilepsy, prediction model, LASSO regression, mIMIC-IV database

Posted Date: September 25th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10573283/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Traumatic brain injury patients based on MIMIC-IV database

Construction and validation of early epilepsy prediction model

*Construction and Validation of a Prediction Model
for Early Post-Traumatic Seizures in Traumatic Brain Injury Patients
Based on the MIMIC-IV Database*

Study type : retrospective cohort study

Data sources: MIMIC-IV v3.1

Title: Traumatic brain injury patients based on MIMIC-IV database , Construction and validation of early epilepsy prediction model

Authors: Zepeng Qiu¹ , Guan Wei² , Shuijin Wu³

Affiliations: Emergency Department, Second Affiliated Hospital of Fujian Medical University, Fujian Medical University, Fujian, China

Corresponding author: Guan Wei

Email: 76595825@qq.com

Abstract

Background : Early epilepsy is a common complication after traumatic brain injury (TBI), which is more common. The purpose of this study is to try to establish a model with some conventional clinical indicators, and then to verify this model to see whether it can predict the possibility of early epilepsy in TBI patients.

Methods : The MIMIC-IV database provides data on TBI patients admitted from 2008 to 2019. These patients were selected according to stricter criteria, and then the data were thoroughly cleaned up. At the same time, epilepsy was detected in a hierarchical manner, so that a relatively high-quality data set can be prepared for the establishment of the model. Among the 92 possible variables, we first use the single factor method to screen, then check whether there is collinearity between them, and finally use LASSO regression to determine the important predictors. The whole process is divided into three stages. Next, the multivariate logistic regression is used to establish the prediction model, and then the Bootstrap method is used for internal verification. At the same time, a more comprehensive sensitivity analysis is made to verify the robustness of the model. The study was conducted in accordance with the Declaration of Helsinki.

Results : A total of 1293 patients were finally included, and the incidence of early epilepsy (within 7 days, high-quality evidence) was 8.58 % (111 cases). After three-step screening, 11 core predictive variables were obtained : GCS score (OR = 0.922), head AIS score (OR = 1.206), alcohol-related (OR = 1.677), body temperature (OR = 1.7667), oxygen saturation (OR = 1.069), blood glucose (OR = 1.006), number of operations before ICU (OR = 1.409), sedation at admission (OR = 2.094), platelets, potassium ions and imaging edema. The AUC of the model was 0.748 (95 % CI : 0.699-0.797), the calibration was good (Hosmer-Lemeshow $p = 0.740$), and the average AUC of Bootstrap validation was 0.759. The decision curve analysis showed that the model continued to produce positive net benefits within the clinically relevant risk threshold range (0.05-0.25). Sensitivity analysis showed that the model maintained stable performance under different time windows, different TBI severity subgroups and different outcome definitions.

Conclusion : The prediction model established in this study includes 11 variables. This model is still good in discriminant ability, and the calibration is also reliable. The overall performance is also relatively strong. These variables are indicators that will be collected in clinical routines, so it is more convenient and practical to obtain. It can be used to evaluate the risk of early epilepsy in TBI patients, and may also be helpful in formulating personalized treatment plans, which may have some improvement in the prognosis of patients.

Keywords : traumatic brain injury ; early epilepsy ; prediction model ; lASSO regression ; mIMIC-IV database

1. Introduction

Traumatic brain injury (TBI) is one of the leading causes of death and disability worldwide. [1][2][3], It affects about 69 million people every year. Early post-traumatic seizures (EPTS) is a common problem in patients with TBI, which is typically characterized by seizures within 7 days after injury, with an incidence of 4 % to 25 %.[4]. Early epilepsy increases the chance of secondary brain injury, prolongs hospital stay, and is also associated with poor long-term neurological outcomes and the development of late epilepsy. [5][6]. Therefore, accurate identification of high-risk patients and preventive antiepileptic treatment are essential to improve the prognosis of TBI patients.

At present, the clinical assessment of the risk of early epilepsy in patients with TBI mainly depends on the experience of doctors and a few known risk factors (such as skull fracture, subdural hematoma, low GCS score).[7], Lack of systematic and validated prediction tools [8]. Although previous studies have identified multiple risk factors for early epilepsy, problems such as limited sample size, unsystematic variable screening methods, and inconsistent definition of outcomes have limited its clinical application. In addition, most studies do not consider the impact of complex factors such as competitive risks (such as early death) and confounding factors (such as sedative drug use), which may lead to bias in the prediction model.

Medical Information Mart for Intensive Care IV (MIMIC-IV) is one of the largest intensive care databases. [9], It provides comprehensive clinical, laboratory, imaging and treatment information, and provides a good basis for establishing prediction models using real-world data. The purpose of this study is to apply the MIMIC-IV database, using rigorous methodological design and organized data processing. To establish and verify an accurate, reliable and clinically applicable early epilepsy prediction model for TBI patients, and to provide a basis for clinical decision-making and risk assessment. [10].

2. Research Method

2.1 Data sources

The data source used in this study is the MIMIC-IV v3.1 database, that is, Medical Information Mart for Intensive Care IV. This database is a relatively large public clinical database developed by the Computational Physiology Laboratory of the Massachusetts Institute of Technology and the Beth Israel Deaconess Medical Center. It collects de-identified medical records of patients admitted to the intensive care unit of Beth Israel Deaconess Medical Center from 2008 to 2019, including demographic information, vital signs, laboratory tests, imaging reports, medical records, medications, nursing records, clinical notes, and ICD diagnostic codes.[11], This study has been authorized to use the MIMIC-IV database, the certification number is 60908460, but also through the institutional review committee, that is, IRB exemption review.

2.2 Inclusion and exclusion criteria of research objects

2.2.1 Inclusion criteria

1. Adult patients between 18 and 90 years of age;
2. Conform to the diagnostic criteria of traumatic brain injury (TBI), ICD-10 : S06 / S02, ICD-9 : 85 / 80, including : (1) brain injury (concussion, brain edema, diffuse injury, focal injury, etc.) ; (2) Skull fracture (excluding nasal bone, orbital and facial fractures) ; (3) Intracranial hemorrhage (epidural hematoma, subdural hematoma, subarachnoid hemorrhage, etc.) ;
3. He was hospitalized in intensive care unit for more than one day.;
4. For the first time to enter the intensive care unit, but do not accept multiple entry.

2.2.2 Excluded criteria

1. There was a history of epilepsy before TBI.;
2. Drugs for epilepsy have been used before TBI.;
3. Intracranial disease, but not related to TBI.;
4. Before the occurrence of TBI, neurosurgery is useful for treatment.;
5. Brain lesions caused by poisoning or metabolism;
6. The data of GCS score, pupil reaction and CT imaging examination were missing.;
7. The data is incomplete and logically wrong.;

8. He had epilepsy before he entered the intensive care unit.;

According to the above criteria, a total of 1293 TBI patients were finally included in the study, and none of the subjects needed to be exclude(table1).

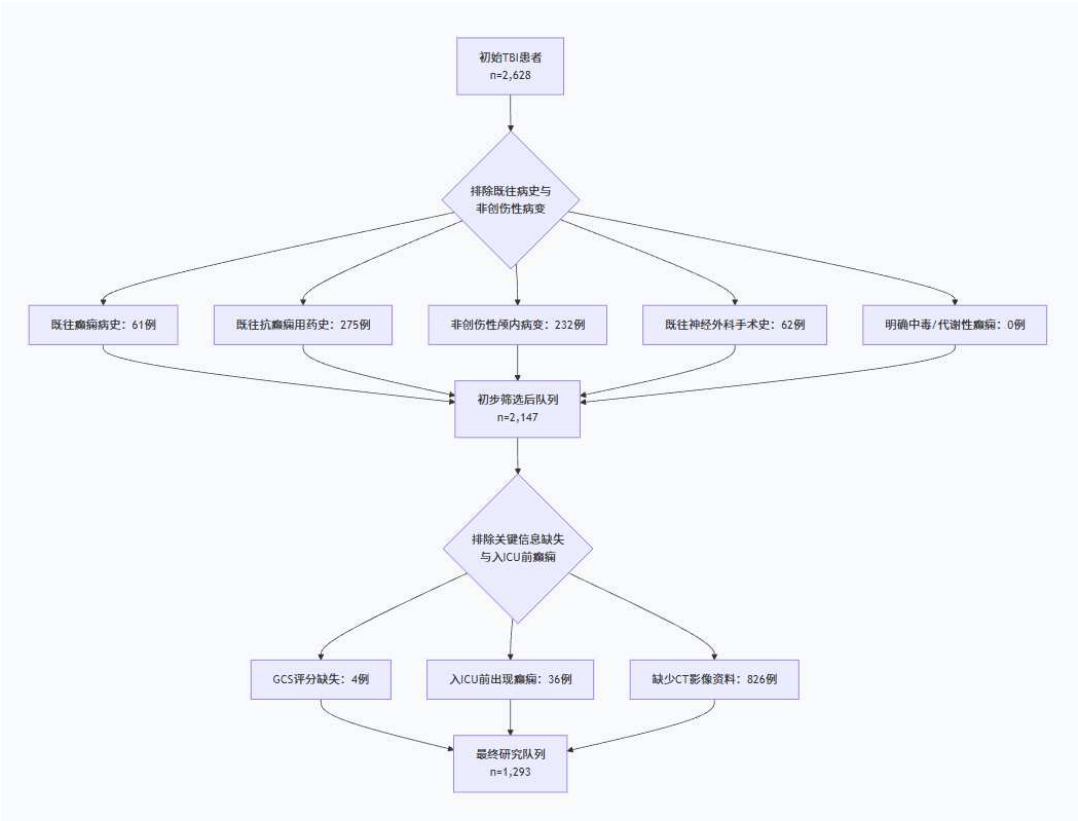


Table1

2.3 Early epilepsy detection methods and evidence grading

Because of the lack of EEG data in the MIMIC-IV database, this study integrated the data and established a reliable hierarchical evidence system to identify early epileptic seizures. Strict judgment rules are also set up to select the appropriate evidence.

2.3.1 Multi-level evidence sources

(1) Chart Event Records

Chart event records come from real-time bedside records of ICU nurses and are the most reliable source of evidence. By searching the value field in the chartevents table, regular expressions were used to match epilepsy-related terms (seizure, convulsion, fitting, seizure, tonic-clonic, epilepsy, status epilepticus, etc.). According to the degree of detail and certainty of the record, the ChartEvents evidence is divided into three levels:

- Level 1-Definite eyewitness (DEFINITE) : Contains explicit descriptions such as ' seizure ', ' observed seizure ', ' active seizure ', ' seizure activity ', ' generalized tonic-clonic ', ' GTC ', ' grand mal ', etc., excluding negative statements. Confidence score : 95 %
- Level 2-probably (PROBABLE) : including ' seizure ', ' convulsion ', ' status epilepticus ', ' postictal ' and other terms, excluding negative, suspected or preventive expressions (such as ' no seizure ', ' possible seizure ', ' seizure prophylaxis '). Confidence score : 85 %
- Level 3-possibility (POSSIBLE) : Contains uncertain expressions such as ' possible seizure ', ' suspected seizure ', ' seizure-like ', and ' questionable '. Confidence rating : 70 %

In addition, preventive records (Level 4-EXCLUDED), such as ' seizure-free ', ' no seizure ', ' seizure prophylaxis ', ' seizure prophylaxis ', etc., were explicitly excluded. ChartEvents records contain accurate timestamps (charttime), which can accurately locate the time of seizures.

(2) Discharge Notes

- The clinical notes come from the discharge summary written by the doctor, and provide epilepsy diagnosis information through retrospective summary. Using natural language processing technology to analyze the text content in the discharge table, it is divided into three levels according to the strength of evidence :
- Level 2A-Definitive Diagnosis (DEFINITE) : Contains explicit expressions such as ' seizure ', ' observed seizure ', ' EEG... seizure ' (EEG confirmed), ' seizure... confirmed ', ' documented seizure ', etc. Confidence score : 90 %
- Level 2B-Physician diagnosis (PROBABLE) : Contains the expressions of ' post-traumatic seizure ', ' early seizure ', ' seizure... (occurred | developed | experienced) ' and other doctors ' definite diagnosis, excluding negative or preventive records. Confidence score : 80 %
- Level 3-possibly or suspected (POSSIBLE) : contains uncertain expressions such as ' possible seizure ', ' suspected seizure ', ' seizure-like ', ' questionable ', etc. Confidence score : 65 %

Clinical notes lack accurate timestamps, so a time inference method based on text cues is used :

- ICU 0-24 小时 (icu_0_24h_inferred) : "first day"、"day 1"、"initial"、"early"、"within...24 hour"
- ICU 24-72 小时 (icu_24_72h_inferred) : "second day"、"third day"、"day 2/3"、"within...48/72 hour"
- ICU period is not specified: "during ICU"、"in ICU"、"ICU stay"、"ICU course"
- Time unknown: There is no clear time clue

(3) ICD Diagnosis

The ICD diagnostic code is derived from the diagnoses _ icd table, which provides a formal diagnosis at the hospital level. Although ICD coding has diagnostic authority, its timeliness and specificity are relatively low. According to the specificity of coding, ICD evidence is divided into three levels :

- Tier A-status epilepticus (STATUS _ EPILEPTICUS) : ICD-10 G41.x (status epilepticus) or ICD-9 345.3. This is the most serious and definitive acute epileptic event with a confidence score of 90 %.
- Tier B-Acute Epilepsy / Convulsion (ACUTE _ SEIZURE) : ICD-10 R56.1 (febrile seizures), G40.901 (unspecified seizures) or ICD-9 780.39 (convulsive NOS). These codes suggest acute epileptic events, confidence score : 70 %.
- Tier C-non-specific seizures (NONSPECIFIC) : ICD-10 R56.9 (not designated seizures). This type of coding is highly non-specific and requires other evidence support. Confidence score : 50 %.

The diagnosis of chronic epilepsy (ICD-10 G40.x or ICD-9 345.x series, except G41 and 345.3) usually does not represent an acute event and is classified as CHRONIC _ EPILEPSY and does not include early epilepsy in the main analysis. ICD coding lacks accurate time, and time inference is performed using diagnostic order (seq _ num) and text cues.

(4) Antiepileptic drugs in acute phase

Antiepileptic drug records were from the inputevents table to retrieve newly initiated intravenous antiepileptic drugs during ICU hospitalization. Drug evidence is relatively indirect, and it is necessary to distinguish between therapeutic drugs and preventive drugs. Based on the mode of medication, the drug evidence is divided into the following categories :

Antiepileptic drugs monitored include : Levetiracetam、Phenytoi、Fosphenytoin、Valproate 、Lacosamide 、Phenobarbital and Benzodiazepines : Lorazepam 、Midazolam、Diazepam. Drug records contain precise administration time (starttime) , It can be used for time positioning.

drug-use mode	recognition standard	degree of confidence
load dosage (ACUTE_LOADING_DOSE)	Single large dose (such as Phenytoin $\geq$ 1000mg, Levetiracetam $\geq$ 1000mg, Valproate $\geq$ 800mg).	65% (Possible treatment)

Benzodiazepine rescue (ACUTE_RESCUE_BENZO)	Intravenous injection of any dose of Lorazepam, Midazolam or Diazepam.	75% (It is likely to be treated)
Drug upgrade mode (ACUTE_ESCALATION)	Upgrade from benzodiazepines to Phenytoin, Levetiracetam, or Valproate within 60 minutes	75% (It is likely to be treated)
High dose sedation (ACUTE_STATUS_CONTROL)	Propofol > 50 µg / kg / min (suggesting status epilepticus control).	65% (Possible treatment)
Preventive maintenance (PROPHYLAXIS_MAINTENANCE)	Standard dose of Phenytoin, Levetiracetam or Valproate, no loading dose or upgrade mode	0% (removal)

2.3.2 Evidence Integration and Priority Rule

In the case of multiple epilepsy-related records in the same patient, the following priority determination rules are used to screen the best evidence :

The core principles on which the priority determination rule is based :

- The evidence level of real-time recording is higher than that of retrospective summary, that is, the priority of ChartEvents (bedside real-time recording) is higher than that of Discharge Notes (discharge retrospective summary);
- • The evidence obtained from the definitive diagnosis is superior to the suspected diagnosis, and the seizures confirmed by eyewitness records or EEG results have the highest priority ;
- The priority of direct evidence is higher than that of indirect evidence, that is, clinical records have higher priority than ICD coding and drug use records;
- • When there are multiple records in the same evidence priority level, the judgment is based on time accuracy, that is, the earliest recorded seizure time is selected as the first seizure time;
- The final confidence score is required to be no less than 0.60 (60 %) in the main analysis, and a more stringent definition standard is used in the sensitivity analysis to increase the threshold to 0.80 (80 %).

priority	source of evidence	Evidence level coding
1 (maximum)	ChartEvents Level 1	ChartEvents_Level1
2	ChartEvents Level 2	ChartEvents_Level2
3	Notes Level 2A	Notes_Level2A
4	Notes Level 2B	Notes_Level2B
5	ICD Tier A	ICD_TierA
6	Medication Acute	Medication_Acute
7	ICD Tier B	ICD_TierB
8	ChartEvents Level 3	ChartEvents_Level3
9 (minimum)	Notes Level 3	Notes_Level3

2.3.3 Definition of early epilepsy

Based on evidence integration and time window judgment, this study adopted the following definition of early epilepsy:

- Main definition (strictly defined for main analysis) : Detection priority 1-4 (ChartEvents Level 1-2, Notes Level 2A-2B) and confidence ≥ 0.80 , onset time ≤ 7 days (168 hours) after ICU admission.;
- Secondary definition (medium definition, for sensitivity analysis) : detection priority 1-7 (including ICD Tier A-B, Medication Acute) and confidence ≥ 0.60 , onset time ≤ 7 days after ICU admission.;
- Loose definition (for exploratory analysis) : detection priority 1-9 (including ChartEvents Level 3, Notes Level 3) and confidence ≥ 0.50 , onset time ≤ 7 days after ICU admission..

In addition, early epilepsy with 24-hour and 72-hour time windows were defined for time window sensitivity analysis.

2.3.4 Treatment of patients without clear evidence

For patients with only antiepileptic drug use records but no other substantial clinical evidence, that is, primary _ evidence _ source is labeled as 'No _ Strong _ evidence ', these patients are excluded from the epilepsy group during the preliminary analysis, but they are still considered in the sensitivity analysis. These patients may also include:

- Only preventive medication, such as anti-epileptic drugs routinely given after surgery, but the patient did not have a real seizure record;
- Medication with unclear treatment intention, that is, its medication mode does not meet the standard characteristics of acute epileptic seizures, such as no load dose, no rescue medication, and no step-by-step intensive treatment upgrade process;

- Incomplete recording information, that is, although there are relevant medication records, there is no mention of epilepsy-related diagnosis in clinical course records and diagnostic codes;
- For other indications, that is, antiepileptic drugs are not for epilepsy treatment, but for other clinical purposes, such as emotional stability or pain management, etc.

2.3.5 Evidence quality control

In order to ensure the accuracy and reliability of detecting epileptic seizures, the following methods are used:

- Multi-Source Validation Principle : Prioritize seizures supported by multiple evidences;
- Time consistency check : For evidence with a precise timestamp, it is necessary to check whether the occurrence time is within the set research time.;
- Negative exclusion : Exclusion of evidence with negative meaning;
- Preventive exclusion : Exclude clear preventive records;
- Chronic Epilepsy Exclusion : Only the coding of acute epilepsy is retained for analysis;
- Manual spot check : The evidence with high priority (priority 1-4) is checked manually and checked at a ratio of 10 %.

2.3.6 Methodological advantages and limitations

The multi-source evidence integration method of this study has the following advantages:

- Make up for the lack of EEG : in the absence of EEG data, identify epilepsy by sorting out clinical records, drug use and other information;
- Establish a hierarchical evidence system : establish a priority system based on the quality and reliability of the evidence to ensure that the selected evidence is credible;
- Precise time : Prioritize evidence with accurate time stamps, such as ChartEvents and Medication, to improve the reliability of time positioning;
- Multi-angle test : mutual verification of evidence from different sources can improve the credibility and reliability of the results;
- Reproducibility : This method is based on structured data and clear decision rules, allowing other researchers to perform repeated verification and result reproduction;

However, this method has the following limitations:

- The lack of EEG gold standard : Because of the lack of systematic EEG data, there may be omissions, such as non-convulsive seizures without obvious clinical manifestations;
- Deviation of retrospective data : Because this study mainly relies on previous records to judge, epilepsy events that are not fully recorded may be excluded;
- Time inference uncertainty : The lack of evidence of accurate time stamps, such as clinical notes or ICD codes, may result in a certain degree of error;
- Indirectness of drug evidence : because drug use records cannot be directly used to prove seizures, there will be misjudgment of preventive medication as seizures;
- Lack of evidence : Some patients may be classified as ' no clear evidence ' due to incomplete records;

2.3.7 Distribution and Statistics of Evidence Sources

A total of 1293 TBI patients were included in this study, and epilepsy-related records were identified and judged by integrating evidence from various sources. Then, according to the rule of evidence priority, only one source of evidence with the highest quality is retained for each patient. For early epilepsy, it is defined as an epileptic event that occurs within 7 days after admission. At the same time, a stricter criterion is adopted, that is, the priority of evidence is required to reach 1-4 levels and the confidence score is not less than 0.80 ; finally, 111 eligible cases were identified, accounting for about 8.58 % of all subjects. The specific distribution of various sources of evidence is as follows:

priority	source of evidence	number of patients (n)	percentage (%)	Incorporate strict definitions
1	ChartEvents Level 1 (Clearly witness epilepsy)	0	0.0	✓
2	ChartEvents Level 2 (Epilepsy is likely)	69	5.3	✓
3	Notes Level 2A (diagnosed definitely)	118	9.1	✓

4	Notes Level 2B (doctors' diagnoses)	5	0.4	✓
5	ICD Tier A (status epilepticus)	18	1.4	×
6	Medication Acute (Acute treatment medication)	714	55.2	×
7	ICD Tier B (Acute epilepsy / convulsion)	24	1.9	×
8	ChartEvents Level 3 (Possible epilepsy)	20	1.5	×
9	Notes Level 3 (Suspected epilepsy)	292	22.6	×
—	No Strong Evidence (There is no clear evidence)	33	2.6	×

Table 1. Distribution of evidence sources for early epilepsy detection (N=1293)

Note : In the strict definition (confidence ≥ 0.80), the evidence of priority 1 ~ 4 (chart event 1 ~ 2, note 2A ~ 2B) was included. In 192 records (0 + 69 + 118 + 5), 111 patients met the time window and confidence criteria, accounting for 8.58 %. ChartEvents Level 1 represents the highest level of evidence, but it shows zero patients, indicating that there is no systematic EEG monitoring record in the MIMIC-IV database. In the main analysis, evidence with a priority level of 5-9 was excluded from the epilepsy group due to low confidence or significant indirectness, but was considered in the sensitivity analysis. There is no strong evidence involving patients who have a drug record but lack other clinical evidence.

Data Extraction and Variable Definition

Extract the following data from the MIMIC-IV database via structured SQL queries:

Demographic and baseline characteristics:

Age, gender, race, admission time, ICU stay time, etc. The mean age of the cohort was 57.8 ± 21.1 years, with 65.1 % of males and 57.8 % of whites.

TBI characteristics:

ICD-9 and ICD-10 codes were used to identify injury types, including subdural hematoma (n = 508, 39.2 %), subarachnoid hemorrhage (n = 369, 12.9 %), brain contusion (n = 158, 12.1 %), epidural hematoma (n = 45, 3.48 %) and diffuse axonal injury (n = 38, 2.94 %). Injury mechanisms include motor vehicle accidents, falls, blow injuries, etc.

severity score:

Glasgow coma score (GCS, first recorded) : mild (13-15 points, n = 577, 44.6 %), moderate (9-12 points, n = 202, 15.6 %), severe (3-8 points, n = 514, 35.7 %) ; aPACHE II score, SOFA score, head concise injury scale (Head AIS).

Physiological indicators and laboratory tests:

After entering the intensive care unit, the first recorded vital signs (body temperature, heart rate, respiratory rate, blood pressure, etc.), pupil reaction, laboratory examination (blood routine, renal function, liver function, arterial blood gas analysis, etc.) were collected.

therapeutic intervention:

The treatment of patients (mechanical ventilation, sedation, analgesia, hypertension, hypotension and neurosurgery, etc.) was collected.

outcome index:

The clinical outcomes of the patients were collected. The average length of stay in the intensive care unit was 4.95 ± 6.12 days, the total length of stay was 11.7 ± 12.0 days, the mortality rate during hospitalization was 11.3 %, the mortality rate within 7 days was 7.79 %, the mortality rate within 72 hours was 4.33 %, and the seizures within 7 days and 72 hours.

Overview of queue characteristics

Among the 1293 patients with traumatic brain injury, 597 patients had seizures within 7 days, accounting for 46.1 %. 558 cases occurred within 72 hours, accounting for 43.2 %, 465 cases occurred within 24 hours, accounting for 43.9 %. From the time distribution of seizures, most cases occurred within 24 hours after entering the intensive

care unit (n = 465). There were fewer patients with epilepsy at 24-72 hours (n = 94) and 72 hours to 7 days (n = 40).

Intracranial hemorrhage was the most common type of traumatic brain injury, with a total of 716 cases, accounting for 55.4 %, of which subdural hematoma was the most common. There were 369 cases of brain injury, accounting for 28.5 %, and 336 cases of skull fracture, accounting for 26.0 %. In addition, multiple injuries were also common, with a total of 221 cases (17.1 %). Some patients also had other injuries such as intracranial hemorrhage, brain injury and skull fracture.

This study reflects the typical clinical characteristics of patients with traumatic brain injury : mainly middle-aged and elderly male patients, the possibility of early epilepsy is higher, which further suggests the significance of prevention. Through the strategy of integrating multi-source data, seizures in clinical practice are collected to a large extent, which lays a reliable foundation for subsequent research.

- According to the above strict criteria, 111 patients with early epilepsy were finally identified, accounting for 8.58 % of all subjects. The distribution of the evidence sources of these 111 patients is :ChartEvents Level 2: 69 例 (62.2%)
- Notes Level 2A: 37 cases (33.3%)
- Notes Level 2B: 5 cases (4.5%)

Note : Of the 118 Notes Level 2A records, only 37 met the time window (≤ 7 days) and confidence (≥ 0.80) requirements, so 37 were included in the main analysis instead of 118.

classification	Number of cases (percentage)
Early epilepsy (positive)	111 (8.58%)
No early epilepsy (negative).	1, 182 (91.42%)
grand total	1, 293 (100.00%)

4. Cleaning Results

4.1 Necessity of data cleaning

The data cleaning process of this study also shows that it is more complicated and important to do quality control in real world data. Although the original data of MIMIC-IV itself has certain reliability, there are still many problems in data quality. For example, first, some data missing rates are relatively high, some variables are missing by more than 30 %, and even some variables have no data at all. Second, 131 variables are related to different clinical fields, sometimes there will be duplication or collinearity ; third, the significance of some outcomes is not clear, the quality of evidence related to epilepsy is also uneven, and the definition of time window is relatively vague. If these problems are not solved well, it may seriously affect the validity and reliability of the prediction model.

4.2 Data cleaning

4.2.1 High missing rate variable processing

- In order to ensure the robustness of the model and reduce the interpolation deviation, the variables with high missing rate are first screened and processed. The missing threshold is set to 30 %, and the predictive variables whose missing rate is higher than this threshold are deleted. A total of 19 variables were deleted, including :
- Completely missing data, such as `ddimer _ first`;
- Blood gas analysis related data, such as bicarbonate, oxygen partial pressure and carbon dioxide partial pressure, etc;
- Liver function-related data, such as alanine aminotransferase, aspartate aminotransferase, and albumin;
- Some coagulation function related data and metabolic data.

The principle of retention : When the missing rate of patient identification (ID) variables and clinical outcome variables exceeds 30 %, they are still retained because of their irreplaceability.

4.2.2 Interpolation of low missing rate variable in

For data with a missing rate of less than 30 %, different strategies (single-valued interpolation) are used according to the degree of missing.:

The median is used for continuous variables and the mode is used for categorical variables.

- There are 5 medium missing variables with a missing rate of 5 % -30 %, which are interpolated with median and mode.

- There are 24 variables with a missing rate of less than 5 %, which are interpolated with median and mode.

Processing results : Through the above processing, all the predictive variables used have become complete, which has established a good foundation for the next analysis.

4.2.3 Near-zero variance variable screening

- In order to avoid the bad influence on the stability and fitting efficiency of the model, some variables whose variance is close to zero, the data distribution is particularly unbalanced or the amount of information is very low are eliminated. This threshold is set at 99 %. That is to say, if a certain value in a variable appears more than 99 %, it is considered that this variable does not provide any useful information. Finally, 8 of these variables are removed:
- cvd (99.1%)
- any_vasopressor_at_admission (99.4%)
- brain_herniation (99.8%)
- mass_effect (99.5%)
- cistern_compressed (99.8%)
- cistern_absent (99.7%)
- ventricle_compressed (99.6%)
- hydrocephalus (99.1%)

4.2.4 Near-zero variance variable screening

Newly defined early epilepsy outcome:

- early_seizure_24h (Strict 24 hours : 87 cases, 6.7%)
- early_seizure_72h (Strict 72 hours : 106 cases, 8.2%)
- early_seizure_strict (Strict 7 days : 111 cases, 8.6%)
- early_seizure_moderate (Medium 7 days : 597 cases, 46.2%)

At this point, the TBI patient group in the MIMIC-IV database was effectively converted into a high-quality modeling data set through a thorough data cleaning process. The initial data set of 1293 patients and 131 variables was reduced to a comprehensive data set of 1293 patients and 107 variables, including 90 predictors.

variable type	quantity	explanation
ID variable	3	subject_id, hadm_id, stay_id
Outcome variables	12	It includes 4 definitions of early epilepsy and original epilepsy variables.

prognosis variate	90	Demographics, clinical evaluation, imaging, treatment, etc.
grand total	107	complete data set

5. Statistical Analysis Method

5.1 Variable screening strategy

The number of predictors in this study is relatively large, a total of 90, but the number of positive events is small, only 111 cases, resulting in a relatively low EPV, only 1.21. Therefore, a three-step screening strategy is used to construct a more streamlined and more robust prediction model.

5.1.1 single factor screening

In the early stages of the study, all variables were subjected to univariate logistic regression analysis, and variables with a P value of less than 0.10 were retained as the criteria for screening data. [12]. This standard is relatively loose, and the main purpose is to avoid prematurely excluding variables that may be important in the model ; for each variable, its odds ratio and 95 % confidence interval are calculated, and finally the analysis results are ranked according to the P value.

5.1.2 Examination of collinearity

After completing the single factor screening, the collinearity test of the relevant variables was further carried out. The correlation degree of variables is evaluated by calculating the Pearson correlation coefficient between all numerical variables. When the absolute value of the correlation coefficient of the two variables exceeds 0.8, it is judged that there is a high degree of collinearity. For highly collinear variable combinations, variables with smaller P values in the univariate analysis are preferentially retained, and variables with relatively large P values are deleted.

5.1.3 LASSO regression screening

LASSO regression is used to make the final selection of variables. It automatically selects variables by adding an L1 penalty term, which can help to identify the most important predictor variables in high-dimensional data. In order to find the most suitable regularization parameter λ , 10-fold cross-validation is used in the study, and then the λ value that maximizes the cross-validation AUC is selected, that is, λ_{1se} , which is used to obtain a more streamlined model. Finally, the variables with non-zero coefficients in LASSO regression are retained as the final modeling variables.

5.2 Model Construction

Multivariate logistic regression was used to construct an early epilepsy prediction model.

The model form is :

$$\text{logit}(P) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

Among them, P is the probability of early epilepsy, X_1 to X_k is the selected k predictive variables, β_0 is the intercept, and β_1 to β_k is the regression coefficient. The adjusted OR value, 95 % CI and Wald test p value were calculated for each predictor variable.

5.3 Model Performance

5.3.1 distinguished ability

The area under the receiver operating characteristic (ROC) curve (AUC) was used to evaluate the discriminant ability of the model. The AUC value ranged from 0.5 to 1.0, of which 0.5 indicated no discriminant ability, 0.7-0.8 was acceptable, 0.8-0.9 was excellent, and > 0.9 was excellent. The 95 % CI of AUC was calculated using the DeLong method. The optimal classification threshold was determined by Youden index (sensitivity + specificity-1), and the sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and accuracy under the threshold were calculated.

5.3.2 calibration

In order to evaluate the calibration of the model, the Hosmer-Lemeshow goodness-of-fit test was used in the study, and then the patients were divided into 10 small intervals according to the predicted probability, and then the observed events in each interval were compared with the predicted events. Whether it is consistent, this model is considered to be well calibrated, and the P value is greater than 0.05, which indicates that there is no significant difference between the predicted probability and the actual result.

5.3.3 internal verification

In order to judge the stability of the model and whether there is over-fitting, the Bootstrap method is used to verify the model internally. This verification is based on 1000 retrievable extractions in the original dataset, and the corresponding AUC values are calculated each time. The average AUC and its 95 % confidence interval were calculated by Bootstrap results. If the AUC obtained by Bootstrap verification is close to the result of the original model, it can be proved that the model has better robustness. If the AUC is significantly lower than the original AUC, it indicates that the model has over-fitting problems.

5.4 Clinical Utility Evaluation

The net benefit of the model in clinical use can be estimated by decision curve analysis. [13]. Decision curve analysis determines the actual value of the model in real clinical activities by analyzing the different results of real positive and false positives under different risks. The net benefit calculation formula is :

$$NB = (TP/n) - (FP/n) \times [P/(1-P_t)]$$

In the formula, TP is the number of true positive people, FP is the number of false positive people, n is the number of total samples, and P corresponds to the risk threshold ; then the model is compared with the net benefits obtained in the two extreme cases of ' Treat All ' and ' Treat None '. If the curve obtained by the model is above the above two baselines, it shows that the model has certain application value in clinical activities.

5.5 Sensitivity Analysis

In order to test the robustness of the model, the following sensitivity analysis is carried out.:

- Time window analysis : 24 hours, 72 hours and 7 days were used as time windows to define early epilepsy, and then analyzed. Finally, the results were compared to compare the performance of the model under different time windows.

- Severity stratification analysis : patients were graded with GCS score, 3 ~ 8 points were severe, 9 ~ 12 points were moderate, 13 ~ 15 points were mild, and then analyzed to evaluate the performance of the model in different degrees of injury ;
- Competitive risk analysis : the data of early death is regarded as a competitive event, and then the patients who died within 7 days and the patients who died within 72 hours are excluded, and the model is verified.
- Outcome definition comparison : The prediction performance of the model under the two early epilepsy criteria of strict definition (priority 2-4) and medium definition (priority 2-4,6,9) was further compared, and the stability of the OR values of each predictor was evaluated.

5.6 Statistical Software

All statistical analyses were performed using R software (version 4.3.0). The main R packages used in the research process include dplyr (data processing), glmnet (LASSO regression analysis), pROC (ROC analysis), rms (model calibration evaluation), boot (Bootstrap verification) and ggplot2 (data visualization). Statistical tests were performed using a two-sided test method, and $P < 0.05$ was used as a statistically significant criterion for the difference. [14].

6. Modeling Results

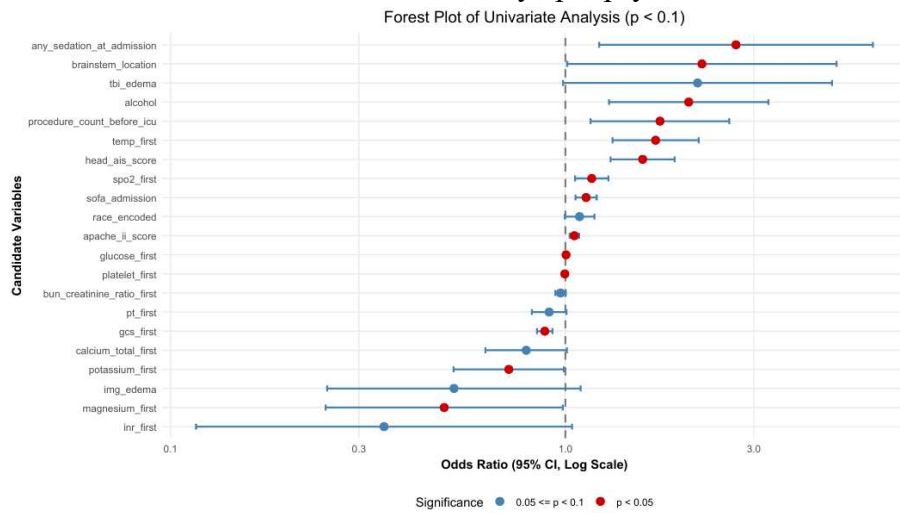
6.1 Variable Screening Results

6.1.1 single factor screening

Univariate logistic regression analysis was performed on 90 candidate predictors. A total of 21 potential predictors (23.3 %) were screened with $p < 0.10$ as the standard, and the EPV was increased to 5.28 after screening.

Univariate analysis showed that GCS score (OR = 0.81, 95 % CI : 0.77-0.86, $p < 0.001$), head AIS score (OR = 1.50, 95 % CI : 1.32-1.70, $p < 0.001$), alcohol-related (OR = 2.08, 95 % CI : 1.35-3.21, $p = 0.001$) and other variables were significantly

associated with early epilepsy.

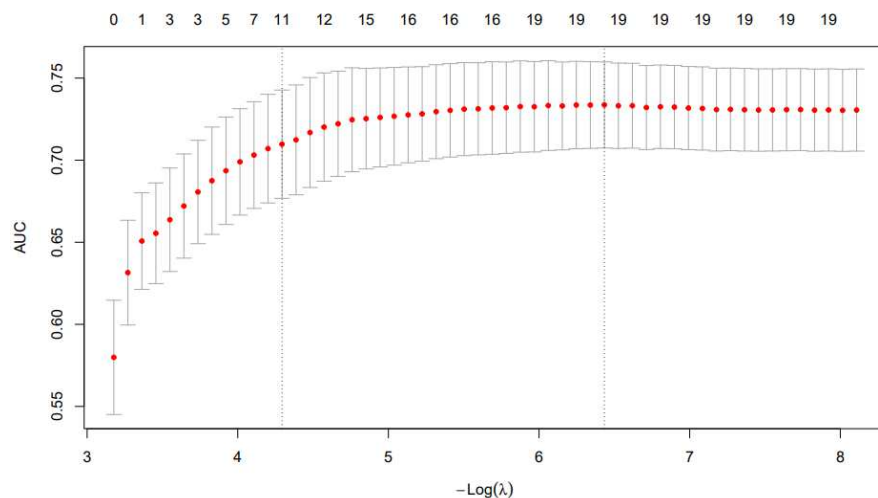


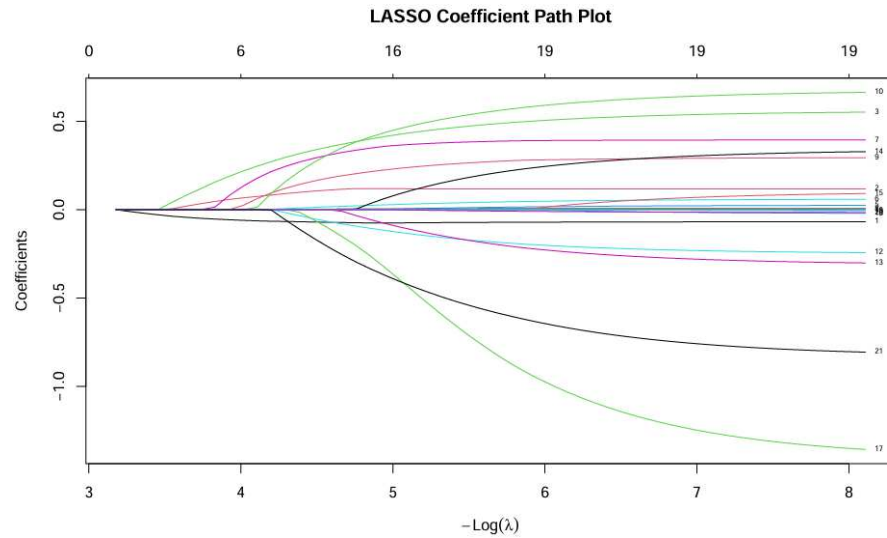
6.1.2 collinearity processing

Highly collinear variables were not detected in 21 screening variables (correlation coefficient > 0.8).

6.1.3 LASSO screening

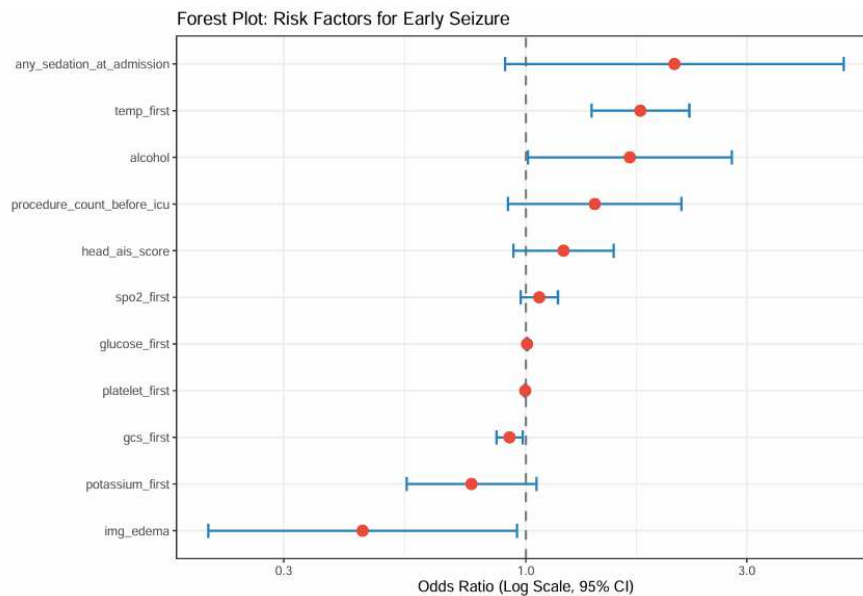
Through the 10-fold cross-validation LASSO regression, the parameter lambda.1 se was used to determine 11 key predictors. The final calculated EPV was 10.09, that is, 111 divided by 11. This value satisfies the robustness standard of traditional logistic regression, and EPV is greater than or equal to 10.





The model includes 11 predictors:

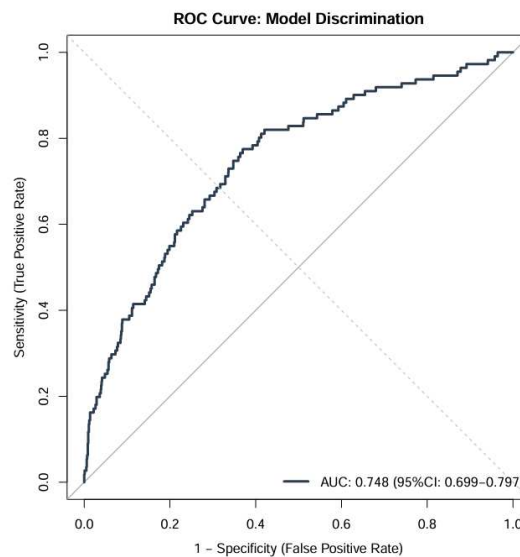
- First GCS score
- Head AIS score
- First body temperature
- First blood oxygen saturation
- Alcohol-related
- First blood glucose
- Number of operations before ICU
- Sedation on admission
- First platelet
- First potassium ion
- Imaging edema



6.2 Model Performance

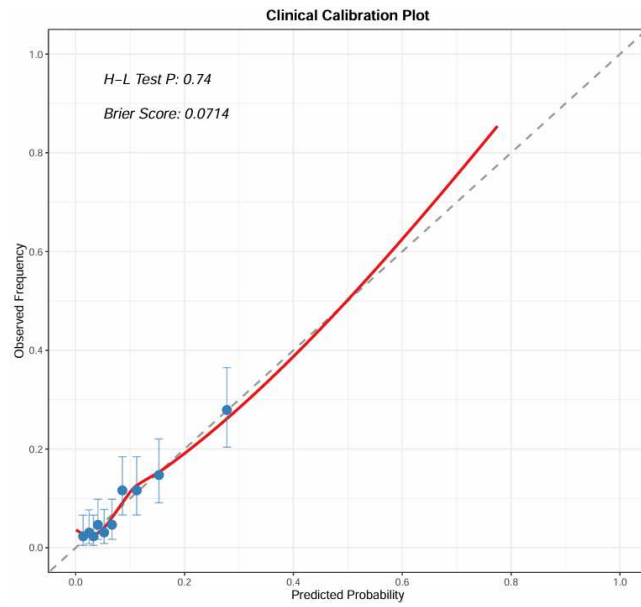
6.2.1 distinguished ability

The area under the ROC curve of the model was 0.748 (95 % CI : 0.699-0.797), indicating that the model had acceptable discriminant ability. Using the Youden index, 0.087 was found to be the best classification threshold. The performance of the model at this level was : sensitivity was 77.5 %, specificity was 63.0 %, positive predictive value was 16.4 %, negative predictive value was 96.8 %, and accuracy was 64.3 %. The negative predictive value of the model was 96.8 %, which could well exclude low-risk patients.



6.2.2 calibration

The Hosmer-Lemeshow goodness-of-fit test showed that $\chi^2 = 5.162$, $p = 0.740$ ($p > 0.05$), indicating that the model was well calibrated and the predicted probability was highly consistent with the actual incidence.

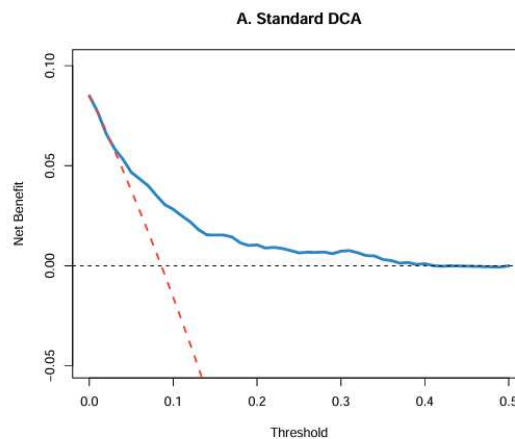


6.2.3 internal verification

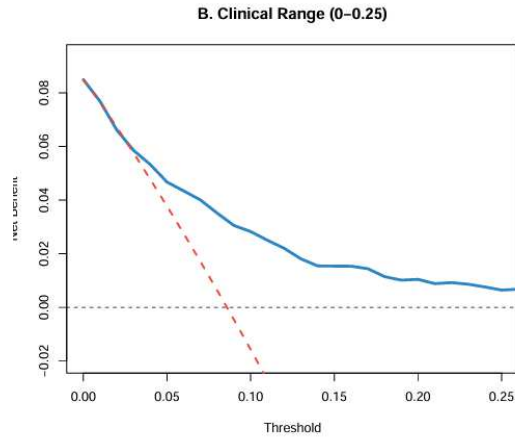
Bootstrap 1000 re-sampling validation showed that the average AUC was 0.759 (95 % CI : 0.708-0.806), which was very close to the original AUC (0.748), and the AUC difference was only 0.011 (1.5 %), indicating that the model had good stability and no obvious over-fitting phenomenon.

6.3 Clinical Effect

In the range of threshold probability p_t of 0-0.50, the net benefit (NB) of the three strategies of ' Model-guided intervention (Model) ', ' Treat-all ' and ' Treat-none ' were compared (Figure X).



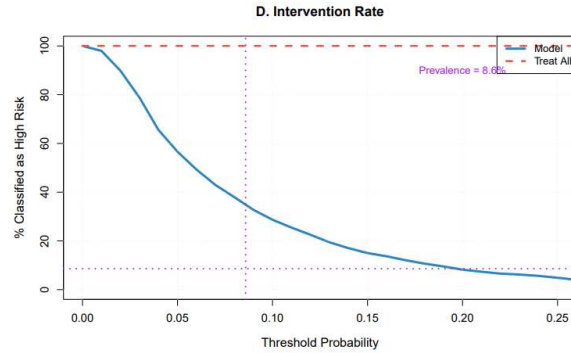
Within the clinically operable threshold range (e.g., $p_t = 0.05 - 0.25$), the net benefit of the model strategy is generally higher than that of Treat-all and Treat-none, suggesting that the model has potential clinical utility.



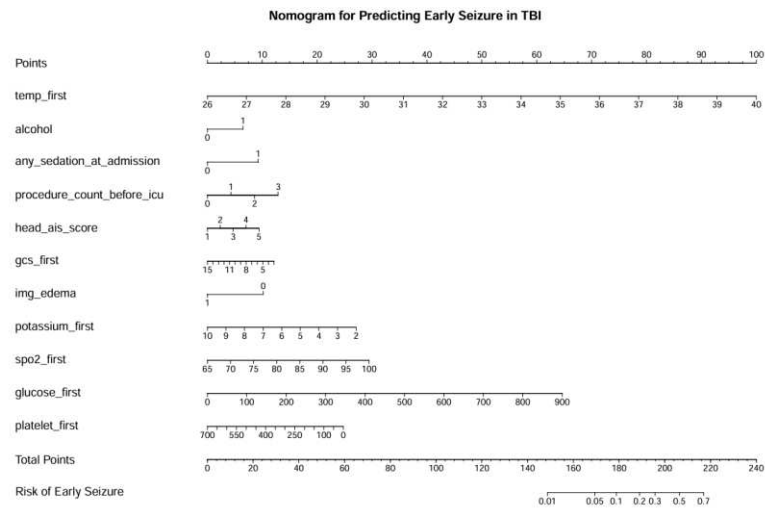
The decomposition analysis at the key threshold point shows (Table X) : when $p_t = 0.05$, the model predicts a high risk of 56.6 % (732 / 1293) and a net benefit of 0.0467, higher than Treat-all 's 0.0377 ; when $p_t \geq 0.10$, the net benefit of Treat-all becomes negative (e.g., $NB = -0.0157$ when $p_t = 0.10$), while the model still maintains a positive net benefit ($NB = 0.0283$), and as the threshold increases, the net benefit advantage of the model relative to Treat-all further expands (Table X).

A	B	C	D	E
Threshold	Net_Benefit_Model	Net_Benefit_All	Advantage_vs_All	Intervention_Rate
5%	0.0467	0.0377	0.009	56.60%
10%	0.0283	-0.0157	0.044	28.70%
15%	0.0154	-0.0755	0.0909	15%
20%	0.0104	-0.1427	0.1531	8.20%

The intervention rate of the model decreased while the threshold gradually increased : when $p_t = 0.10$, about 28.7 % of the patients needed intervention or intensive treatment ; when $p_t = 0.20$, only about 8.2 % of patients needed intervention (Fig.X-D). Then we use Number Needed to Intervene to quantify resource efficiency : in Treat-all 's strategy, NNI is 11.6 (1293 / 111), but in the model 's strategy, NNI is 7.8, 5.3, 4.2 and 3.3 at $p_t = 0.05, 0.10, 0.15, 0.20$, respectively, and the corresponding improvement is about 33.1 % -71.6 % (Table X). According to the above results, compared with ' all interventions ', the model can reduce unnecessary interventions and improve the effectiveness of unit interventions at the same time.



A	B	C	D
Threshold	NNI_Model	NNI_All_Strategy	Efficiency_Improvement_Pct
5%	7.8	11.6	32.80%
10%	5.3	11.6	54.30%
15%	4.2	11.6	63.80%
20%	3.3	11.6	71.60%



6.4 Sensitivity Analysis Results

6.4.1 Different time windows

Under two different conditions, the model still maintains excellent stability. The AUC of the 72-hour time window was 0.740, and the AUC of the 7-day time window was 0.748. The AUC difference between the two is only 0.008, indicating that regardless of any change in the time window, the prediction performance of the model is not affected.

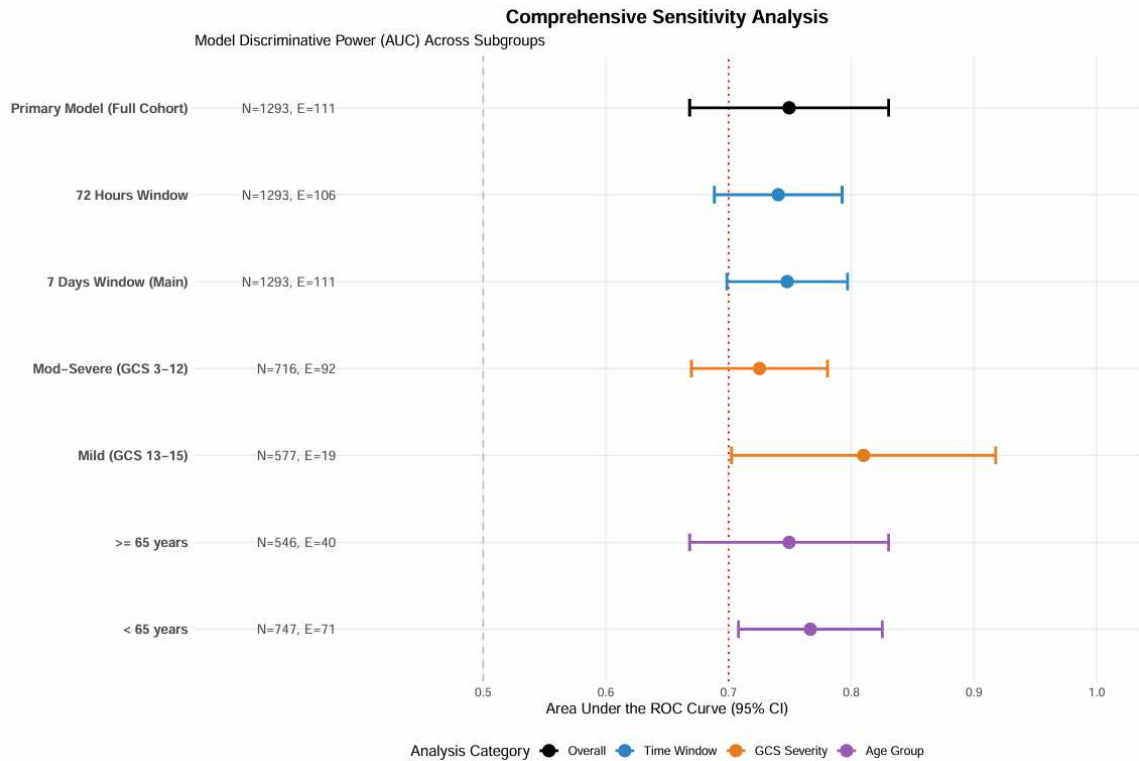
6.4.2 TBI severity stratification

In patients with severe TBI (GCS 3-12, $n = 716$), the AUC of the model was 0.725 (95 % CI : 0.670-0.781). In patients with mild TBI (GCS 13-15, $n = 577$), the AUC of the model was 0.810 (95 % CI : 0.702-0.918), which was similar to the AUC of the whole

population (0.748), indicating that the model still maintained good performance in patients with different severity.

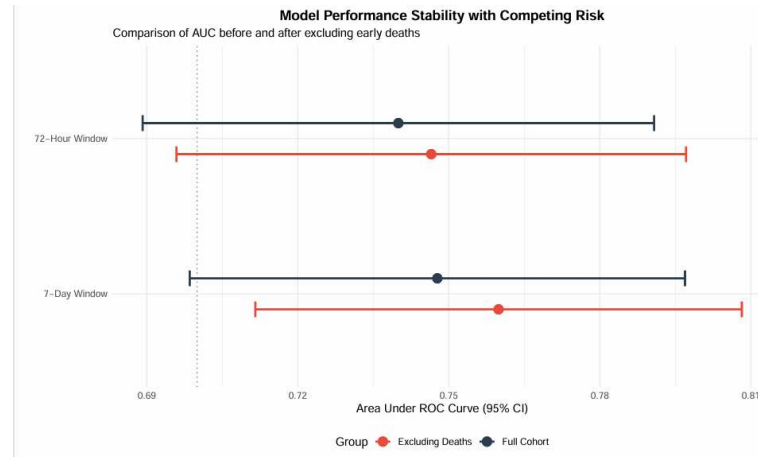
6.4.3 Stratification of different ages

For the elderly group, that is, patients aged 65 and above, there were a total of 546 cases. The AUC of this model was 0.749, and the 95 % confidence interval was 0.668 to 0.830. This value and the AUC of the whole population, that is, 0.748, indicating that this model still maintains a relatively strong predictive performance in high-risk elderly patients.



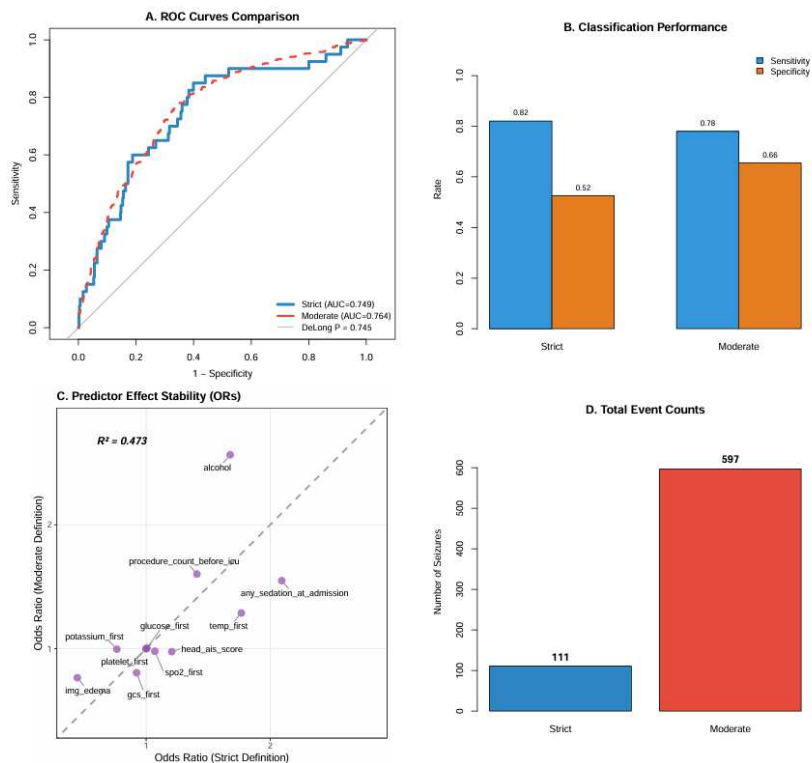
6.4.4 competing risk analysis

When the patients who died within 7 days after admission were excluded, that is, 1191 cases, the AUC of the model was 0.762, which only changed by 0.015 compared with the original. If the patients who died within 72 hours after admission were excluded, that is, 1237 cases, the AUC was 0.745, and the change was only 0.005, which indicated that early death as a competitive event had little effect on the performance of the model.



6.4.5 Comparison of outcome definitions

The strictly defined model, a total of 111 cases, AUC is 0.749, while the moderately defined model has 597 cases, AUC is 0.764, the OR value of the predictor is basically consistent under the two definitions, the average OR value is 1.069, which indicates that the model is not sensitive to the change of the definition of epilepsy.

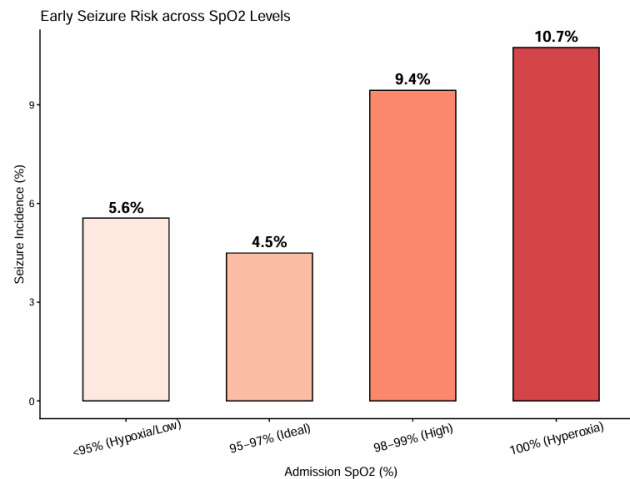


Complementary analysis 1 : Nonlinear relationship test of blood oxygen saturation

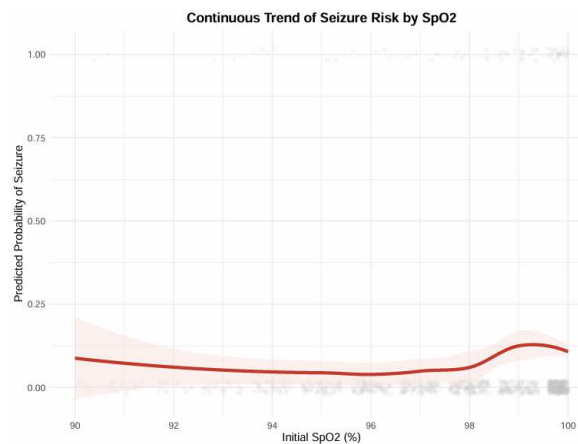
1.1 Background-Avoiding hyperoxia poisoning

Multivariate analysis showed that oxygen saturation, that is, SpO₂, was positively correlated with the risk of epilepsy. The OR was 1.07, and the 95 % confidence interval was 0.98 to 1.17. In addition, the trend analysis showed that with the different levels of SpO₂, the risk had a more obvious gradient. The incidence of epilepsy was about 2.7 % in patients with SpO₂ less than 90 % at admission, while the incidence of epilepsy increased significantly to 10.7 % in patients with SpO₂ of 100 %.

o



A continuous loess regression model showed that the risk of early epilepsy increased significantly when SpO₂ exceeded the 97 % threshold, indicating that hyperoxia contributed significantly to cortical hyperexcitability in the acute phase of traumatic brain injury compared to normoxia.



1.2 Clinical Explanation and Conclusion

Our study found that there is a significant positive correlation between high oxygen saturation levels and the risk of early epilepsy, which is a challenge to the traditional focus on the prevention of hypoxia. Ensuring sufficient oxygen has always been the basis of TBI treatment. However, from our research results, compared with relatively low oxygen levels, acute hyperoxia, that is, $\text{SpO}_2 \geq 98\%$, is closely related to an increase in the incidence of epilepsy by about four times.

This association is biologically reasonable and may involve a variety of different mechanisms ; on the one hand, hyperoxia promotes the excessive production of reactive oxygen species (ROS) and further induces lipid peroxidation of neuronal cell membranes. These oxidative stress-related factors can directly interfere with the GABA-mediated inhibitory neurotransmission process, thereby reducing the seizure threshold. On the other hand, higher levels of oxygen partial pressure may also cause cerebral vasoconstriction, which in turn reduces cerebral blood flow perfusion, especially in some more sensitive marginal brain regions ; in this case, secondary ischemic injury may be further aggravated and promote the occurrence of focal seizures.

In addition, the higher incidence of epilepsy (10.7 %) in the $\text{SpO}_2 = 100\%$ group may reflect a stronger intensity of resuscitation intervention to some extent. In this study cohort, patients with lower GCS scores were more likely to have higher SpO_2 levels, which may be related to the use of higher concentrations of inhaled oxygen (FiO_2) during acute tracheal intubation and early life support.

Based on the results described above, it is supported that the use of a more ' conservative oxygenation strategy ' in neurocritical care may be beneficial. It is not only helpful to reduce the risk of oxidative lung injury, but also may reduce the occurrence of early post-traumatic epilepsy.

Complementary analysis 2 : Paradoxical association between imaging brain edema and early epilepsy

2.1 Background and Paradox Description-Effect of Treatment Factors on Prediction Models

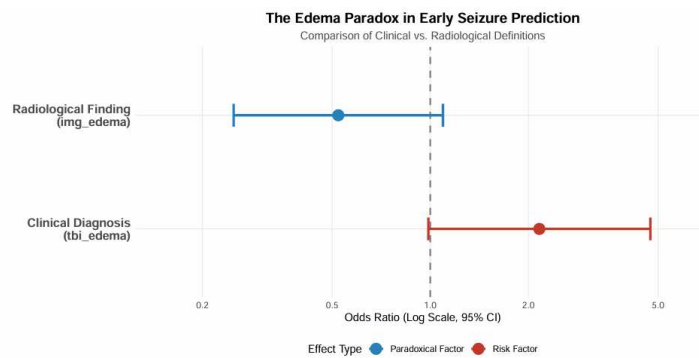
In the final model, there is a negative correlation with imaging brain edema, OR is 0.444, 95 % confidence interval is 0.206-0.957, P value is 0.038, but this result is different from the previous cognition. Brain edema is often regarded as a sign of severe craniocerebral injury, which should increase the risk of epilepsy. However, the model results show that the incidence of epilepsy is lower in patients with brain edema. This ' conclusion ' needs to be further studied.

2.2 Paradox Verification Analysis

2.2.1 descriptive analysis

Our analysis shows that the effect of brain edema on the risk of early epilepsy is significantly different between clinical definition and imaging definition. Clinically diagnosed brain edema, that is, tbi _ edema, is a relatively strong risk factor, and its incidence of early epilepsy is about 16.32 %. On the contrary, imaging evidence suggests that brain edema, that is, img _ edema, shows a seemingly contradictory protective relationship. Its unadjusted odds ratio OR is 0.52, and the 95 % confidence interval is 0.25 to 1.09. The incidence of epilepsy in patients with imaging brain edema was significantly lower than that in patients without imaging brain edema, 4.97 % vs 9.10 %, respectively.

	A	B	C	D	E	F
1	Definition	Total_N	Seizure_Events	Incidence_Rate	Unadjusted_OR	CI_95
2	Clinical (tbi_edema)	49	8	16.33%	2.161	(0.987-4.734)
3	Radiological (img_edema)	161	8	4.97%	0.522	(0.249-1.094)



In the subgroups of GCS 6 to 8, that is, moderate to severe, the risk of epilepsy was most significantly reduced. When the admission CT showed imaging brain edema, the incidence of early epilepsy decreased from 16.2 % to 6.52 %, which was more obvious. Further subgroup analysis also strengthened this 'paradox' phenomenon. The SOFA score of patients with imaging brain edema was significantly higher than that of patients without brain edema, which was 5.26 vs 4.34, respectively. The statistical difference was significant. The mechanical ventilation rate, that is, as an alternative indicator of high-intensity intervention, showed an upward trend in patients with brain edema. It was 30.4 % vs. 23.2 %, and the difference was not statistically significant. However, combined with the significant difference in SOFA score, this shows that patients with 'cerebral edema and coma' will adopt more active respiratory protection and deep sedation strategies. This higher proportion of medical intervention is likely to be the direct cause of their 'paradoxical' decline in the incidence of epilepsy.

In patients with mild traumatic brain injury, that is, GCS 13 to 15, the incidence of early epilepsy in patients with imaging brain edema was 0 %, while that in patients without imaging brain edema was 3.75 %. Mechanism verification analysis showed that

compared with patients without imaging brain edema, patients with img _ edema were more likely to receive some active cluster interventions in the early stage of admission, including hypertonic therapy and deep sedation.

With the model from the single factor, that is, the unadjusted model, the gradual addition of sedative drugs to adjust, coupled with the adjustment of the severity of the injury, that is, the GCS adjustment, the OR value of the imaging brain edema has been basically kept below 1.0, from 0.52 Slowly dropped to 0.49.

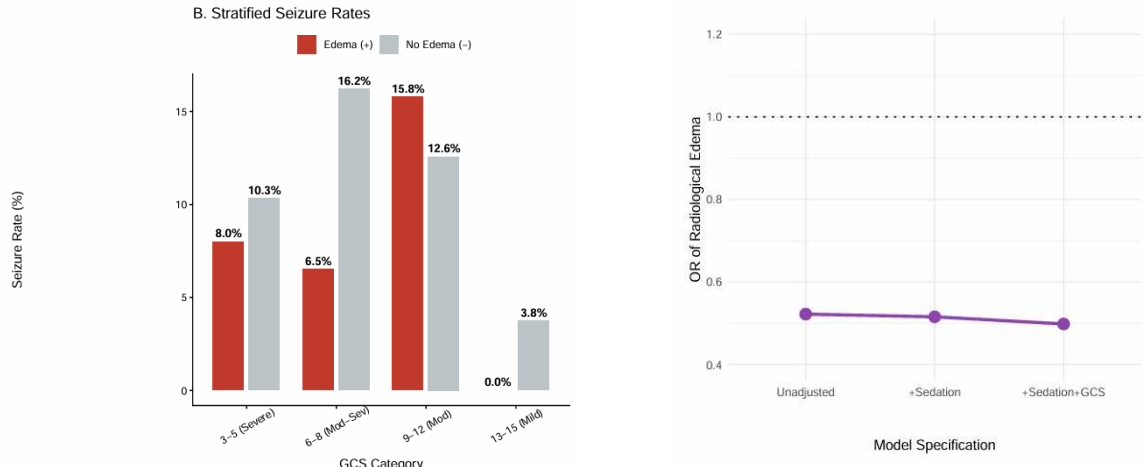


Table s2

Note : Table S2, subgroup analysis of the incidence of early epilepsy according to imaging brain edema status.

Our subgroup analysis, which is shown in Table S2, provides some empirical evidence that there is a 'protection paradox' in imaging brain edema. The results show that although GCS and SOFA scores are similar to those of patients without brain edema, there are a total of 152 patients with brain edema on imaging. The incidence of early epilepsy is significantly lower than that of patients without brain edema, which is 4.6 % vs.8.8 %, respectively, which provides some evidence that radiological findings may be used as a 'sentinel signal' by initiating some more advanced clinical interventions. It effectively prevents the disease from developing into obvious clinical symptoms, such as seizures. On the contrary, if the disease develops to the stage of clinically diagnosed brain edema, the risk of seizures in patients will surge to 17.5 %, suggesting that brain damage at that time may have exceeded the range that conventional intervention can control.

	Both	Clinical Only	None	Radiological Only	p
n	9	40	1092	152	
age (mean (SD))	51.70 (17.59)	52.50 (19.78)	58.33 (21.08)	55.91 (21.80)	0.159
gcs_first (median (IQR))	7.00 (6.00, 7.00)	6.50 (3.00, 10.00)	10.00 (7.00, 15.00)	10.00 (8.00, 15.00)	<0.001
sofa_admission (median (IQR))	8.00 (6.00, 8.00)	4.00 (2.75, 7.00)	3.00 (1.00, 5.00)	3.00 (1.00, 5.00)	<0.001
mechanical_vent_at_admission = 1 (%)	2 (22.2)	6 (15.0)	162 (13.9)	24 (15.8)	0.83
early_seizure_strict = 1 (%)	1 (11.1)	7 (17.5)	96 (8.8)	7 (4.6)	0.064
early_death_7d = 1 (%)	2 (22.2)	10 (25.0)	80 (7.3)	10 (6.6)	<0.001

2.3 Comprehensive Interpretation And Clinical Significance

One of the main findings of this study is the so-called 'brain edema paradox', that is, brain edema seen on imaging seems to be associated with a low risk of early epilepsy. However, we believe that this phenomenon is not a real biological benefit, but more likely a 'treatment-related masking' phenomenon that is common in neurocritical settings.

If the imaging at the time of admission indicates obvious cerebral edema, the patient will be treated as a clinical high-risk group, and then neuroprotection-related treatment will be carried out early, such as hypertonic treatment and enhanced sedation management to prevent fluctuations in intracranial pressure as much as possible. Under these early sedation treatments, the seizure threshold may increase and the excitability of the cortex may decrease, thereby reducing the possibility of seizures. According to the analysis of the results, in the group with GCS 6 ~ 8, if CT showed cerebral edema when entering the hospital, the risk of epilepsy would be significantly reduced, and the incidence of early epilepsy also showed a decreasing trend. In clinical practice, brain edema is often used as a 'sentinel signal', and strategies are often used to stabilize neurons, such as hypertonic therapy.

At the same time, brain edema still showed OR > 1.0 results in the model, showing a significant independent risk effect, which indicated that brain edema may need to reach a certain threshold to have a significant impact, and the promotion of epilepsy will be greater than the inhibition brought by treatment. It is worth noting that even if the factors of sedative drugs are corrected, the OR of brain edema is 0.498, indicating that this 'masking effect' may not only be due to sedative drugs, but also be related to the overall neuroprotective treatment strategy, which is often used in patients with imaging brain edema.

In addition, in mild craniocerebral injury, the incidence of early epilepsy in patients with imaging brain edema is 0 %, which shows that when the degree of injury is relatively mild, brain edema may be a transient inflammatory response and cannot induce epilepsy.

In summary, just seeing the imaging findings is not equivalent to the clinical pathophysiological pro-epileptic state. The 'cerebral edema paradox' highlights a more fundamental challenge in predictive modeling, that is, the so-called 'treatment as a confounding factor' bias problem. Correct identification of this phenomenon helps to avoid clinicians misinterpreting imaging cerebral edema as a low-risk signal, but should be regarded as a sign of high-risk patients who have received high-intensity monitoring and intervention, but still need continuous close monitoring.

5. Discuss

5.1 Main Findings

Using the MIMIC-IV database of 1293 TBI patients, this study developed an early epilepsy prediction model containing 11 clinical variables. This model has good discriminant ability, AUC is 0.748, and the correction ability is also reliable. The p-value of the Hosmer-Lemeshow test is 0.740, and the internal validation shows that the Bootstrap AUC is 0.759. This model includes 11 variables covering different aspects, such as neurological assessment, injury severity, physiological markers, metabolic factors, and treatment strategies. These are routinely collected clinical indicators and do not require additional testing, so they are relatively easy to obtain in a clinical setting.

5.2 Comparison With Previous Studies

Previous studies on early epilepsy prediction in TBI patients were mostly single-center and small-sample studies, and the construction methods and performance of prediction models were quite different. The classic study of Temkin et al. identified risk factors such as depressed skull fracture, subdural hematoma, and GCS30 minutes, but did not construct a quantitative prediction model. In recent years, some studies have attempted to establish early epilepsy prediction models, and the AUC values reported are between 0.65 and 0.85. However, most studies have a sample size of < 500 cases, a variable screening method is not systematic, and there is a lack of external verification. Nowadays, the prediction model of machine learning methods is gradually applied to TBI-related epilepsy risk assessment, but its clinical feasibility and generalization ability are still controversial[15].

The innovation of this study has several aspects. First, the definition of outcome is relatively strict, and the quality of epilepsy evidence is considered, that is, the priority of detection, and the time window is limited to 7 days, so as to avoid the common definition ambiguity in previous studies. Second, we used a systematic variable screening strategy. Through the three steps of single factor screening, collinearity examination and LASSO regression, 11 core predictors were selected from 92 candidate variables, which increased the EPV from 1.21 to 10.09. Third, a comprehensive sensitivity analysis was performed. The robustness of the model was fully evaluated by covering the four aspects of time window, severity stratification, competitive risk and outcome definition. Fourthly, the mechanism was discussed in depth, including the non-linear relationship of SpO₂ and the paradoxical relationship of cerebral edema. Systematic verification was also made to improve the interpretability of the model in clinical practice.

5.3 Clinical Value

The prediction model constructed in this study has many clinical application values.:

- Risk stratification : The model can divide TBI patients into low, medium and high risk groups to guide individualized monitoring and prevention strategies;
- Treatment decision : High NPV (96.8 %) makes the model particularly suitable for excluding low-risk patients and avoiding unnecessary prophylactic antiepileptic drug therapy and its adverse reactions;
- Resource optimization : According to the results of the decision curve analysis, the model can improve medical efficiency and is very important for the lack of medical resources;
- Early warning : The variables of the model are obtained when the patient enters the intensive care unit or within 24 hours after entering the intensive care unit, which can provide early warning of early risks;
- Clinical accessibility : All variables are routinely collected in clinical practice, which is more convenient for clinical promotion and use.

5.4 Research Limitations

- Single data source : Although the MIMIC-IV database samples are of high quality and large quantity, they all come from the same medical center - Beth Israel Deacon Medical Center, so there may be bias in the region, individual and medical methods, which may limit the promotion and use of the research results;
- Retrospective design : Some interference factors are not fully controlled and may be affected by early epilepsy risk perception;
- Lack of external verification : the model only uses Bootstrap for internal verification, and there is a lack of external data to test whether the model is universal;
- Epilepsy detection bias : Retrospection relies on existing clinical records, so there may be seizures that are not fully recorded that are missed;
- Low EPV : The EPV was raised to 10.09 by LASSO, but it is still at the edge of the traditional standard and may have an impact on the stability of the model;
- Lack of neuroimaging quantitative indicators : No further inclusion of CT or MRI indicators may limit the ability of the model.

5.5 Future Research Direction

- External validation : Using data from different regions or different medical institutions to test the performance and ability of the model;
- Prospective study : In order to evaluate the actual ability of the model, and to identify seizures more accurately, a prospective study was conducted;
- Model upgrade : include imaging quantitative indicators, continuous EEG data, etc., to further improve the performance of the model;

- Clinical decision support system : Combining the model with the electronic medical record system to develop a tool that can automatically assess risks;
- Intervention study : a randomized controlled trial of the model to assess whether the strategy is effective and the cost is controllable;
- Machine learning : Use machine learning algorithms such as random forest, gradient boosting, and neural networks to find ways to improve model prediction performance.

6. Conclusion

In this study, through rigorous methodological design and systematic data processing, an early epilepsy prediction model for TBI patients based on 11 clinical routine variables was successfully constructed. This model has acceptable discriminant ability, AUC is 0.748, and the calibration degree is also good. The p value of the Hosmer-Lemeshow test is 0.740, and the internal verification shows that the Bootstrap AUC is 0.759, and a more comprehensive sensitivity analysis is performed. The results show that the model can maintain a relatively stable performance under different time windows, different patient subgroups and different outcome definitions. Decision curve analysis also shows that this model has certain clinical application value, which can provide some evidence-based basis for individualized early epilepsy risk assessment and preventive treatment decision-making of TBI patients.

The model contains a total of 11 variables, all of which are derived from conventional clinical collection indicators and do not require additional special examinations. Therefore, it is easy to obtain in practical clinical applications and has good practicability and operability. The negative predictive value (NPV) of the model was high, reaching 96.8 %, suggesting that it has a prominent application value in identifying low-risk patients and reducing unnecessary preventive treatment. However, it should be noted that this study is based on single-center retrospective data, and there are still some selective bias and extrapolation limitations. Therefore, the model still needs to further verify its stability and generalization ability in an independent external population. At the same time, it is necessary to evaluate its practical application effect in real-world clinical environment in future research, and further combine neuroimaging quantitative markers, biomarkers and machine learning methods to build a more perfect clinical decision support system, so as to promote the clinical transformation and application of the model.

7. Funding

The authors declare that no funding was received for this study.

8. Declarations

1、Ethics approval and consent to participate.

2、Consent to participate:

The requirement for informed consent was waived because this study used de-identified data from the MIMIC-IV database. No identifiable individual patient data were accessed or used in this study.

Reference

- [1] Centers for Disease Control and Prevention. Report to Congress: traumatic brain injury in the United States: epidemiology and rehabilitation. Atlanta (GA): US Department of Health and Human Services; 2015.
- [2] Annegers JF, Hauser WA, Coan SP, Rocca WA. A population-based study of seizures after traumatic brain injuries. *N Engl J Med*. 1998;338(1):20-24.
- [3] Christensen J, Pedersen MG, Pedersen CB, Sidenius P, Olsen J, Vestergaard M. Long-term risk of epilepsy after traumatic brain injury in children and young adults: a population-based cohort study. *Lancet*. 2009;373(9670):1105-1110.
- [4] Fordington S, Manford M. A review of seizures and epilepsy following traumatic brain injury. *J Neurol*. 2020;267(11):3105-3111.
- [5] Jennett WB, Lewin W. Traumatic epilepsy after closed head injuries. *J Neurol Neurosurg Psychiatry*. 1960;23(4):295-301.
- [6] Lee ST, Lui TN. Early seizures after mild closed head injury. *J Neurosurg*. 1992;76(3):435-439.
- [7] Temkin NR. Antiepileptogenesis and seizure prevention trials with antiepileptic drugs: meta-analysis of controlled trials. *Epilepsia*. 2001;42(4):515-524. doi:10.1046/j.1528-1157.2001.28900.x.
- [8] Schubert KM, Schmick A, Statmann M, Galovic M. Prognostic models for seizures and epilepsy after stroke, tumors and traumatic brain injury. *Clin Neurophysiol Pract*. 2025;10:116-128. doi:10.1016/j.cnp.2025.02.008.
- [9] 吴俊,曹冠柏,杨以平,等.血清 CRP、CKBB 和 MMP-2 联合检测对轻型创伤性脑损伤危险分层的价值[J].检验医学与临床,2020,17(20):2957-2960.
- [10] Malik H, Wolff MD, Teskey GC, Mychasiuk R. Electrographic seizures and brain hyperoxia may be key etiological factors for postconcussive deficits. *J Neurophysiol*. 2022;128(3):727-737. doi:10.1152/jn.00533.2021.
- [11] 李杨,杜雷雷,许飞,等.大数据与人工智能在医学领域的应用进展[J].协和医学杂志,2023,14(1):184-189.
- [12] Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J*. 2014;35(29):1925-1931.

- [13] Vickers AJ, Elkin EB, Gonen O. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 2008;28(5):623-634.
- [14] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ*. 2015;350:g7594.
- [15] Senders JT, Arnaout O, Karhade AV, Dasenbrock HH, Gormley WB, Broekman MLD, et al. Machine learning and neurosurgical outcome prediction. *Neurosurgery*. 2018;82(5):608-615. doi:10.1093/neuros/nyx054.