Predicting Acute Kidney Injury in Septic Patients Using Logistic Regression with MIMIC-III Data

Antonio García, Daniel Girbes, Rahul Sandhu

Master in Health Data Science — MHEDAS

January 23, 2025

Health
Co-funded by the European Union

UNIVERSITAT ROVIRA i VIRGILI
BIOINFORMATICS BARCELONA
UAB Universitat Autònoma de Barcelona
UNIVERSITAT DE BARCELONA
Universitat de Lleida
UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH
UNIVERSITAT DE VIC UNIVERSITAT CENTRAL DE CATALUNYA
Universitat de Girona
UGA Université Grenoble Alpes

Table of Contents

Introduction

- **Acute Kidney Injury (AKI):**
  - Sudden decline in kidney function.
  - Affects approximately 14% of hospitalized patients globally.
  - Higher prevalence in ICU.

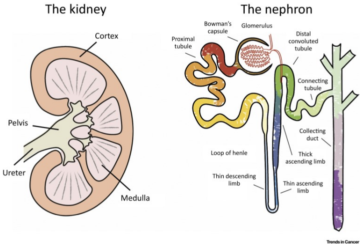


Figure: Kidney and nephron anatomy. Source: [1].

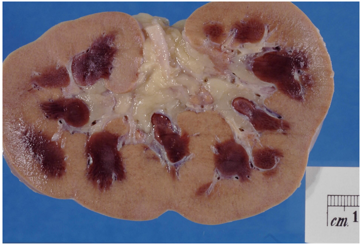


Figure: Kidney with AKI showing pale cortex and dark medullary tissue. Source: [2].

Cohort Building

Initial Cohort

`ICD9_CODE IN ('99591', '99592', '78552')`

Age: 18–89

LOS $\geq$ 48 hours

Gender, Ethnicity, Age

Comorbidities

Acute Kidney Injury (AKI), Chronic Kidney Disease (CKD), Coronary Artery Disease (CAD), Hypertension (HYP), Type 2 Diabetes Mellitus (DM2)

+

Laboratory Features

Albumin, Anion Gap, Bilirubin, Blood Urea Nitrogen (BUN), Chloride, Creatinine, estimated Glomerular Filtration Rate (eGFR), Glucose, Hematocrit, Hemoglobin, International Normalized Ratio (INR), Lactate, Partial Thromboplastin Time (PTT), Platelet Count, Potassium, Sodium

Chartevent Features

Diastolic Blood Pressure (DBP), Heart Rate (HR), Height, Oxygen Saturation (SpO2), Systolic Blood Pressure (SBP), Temperature, Weight

Mechanical Ventilation

Vasopressor Use

Column Mappings and eGFR Calculation

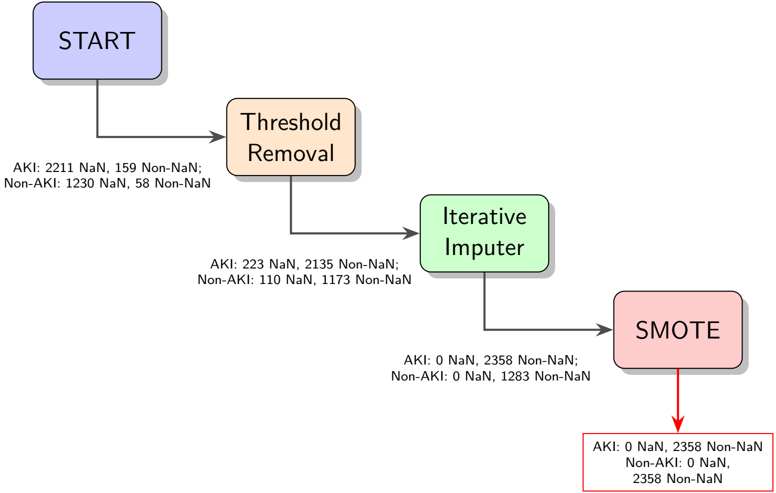
MHEDAS
Master Health Data Science

- Gender values = Male: **1**, Female: **0**.
- Ethnicities = **1**: Top1, **2**: Top2, **3**: Rest.
- All columns set to datatype **float** except for Gender, Ethnicities, and Comorbidities, which were **int**.
- eGFR calculated using a function provided in the **MIMIC-III GitHub** repository.
- Removed **Mechanical Ventilation** and **Vasopressor Use** columns as they were always 0 and provided no information.

```python
def egfr(creat: float, age: float, gender: int, ethnicity: float) -> float:
    """
    Calculate the estimated glomerular filtration rate (eGFR).

    Inputs:
        - creat (float): Creatinine level.
        - age (float): Age in years.
        - gender (int): Gender.
        - ethnicity (float): Ethnicity.

    Outputs:
        - float or None: The calculated eGFR value or none.
    """
    # Return None if inputs are not correct
    if pd.isnull(creat) or creat == 0.0 or age == 0.0:
        return None

    # Calculate eGFR
    factor_gender = 0.742 if gender == 0 else 1
    factor_ethnicity = 1.212 if ethnicity == 2 else 1
    eGFR = 175 * (creat**-1.154) * (age**-0.203) * factor_gender * factor_ethnicity

    return eGFR
```

Figure: eGFR function.

NaN Removal


START
Threshold Removal
Iterative Imputer
SMOTE
AKI: 2211 NaN, 159 Non-NaN;
Non-AKI: 1230 NaN, 58 Non-NaN
AKI: 223 NaN, 2135 Non-NaN;
Non-AKI: 110 NaN, 1173 Non-NaN
AKI: 0 NaN, 2358 Non-NaN;
Non-AKI: 0 NaN, 1283 Non-NaN
AKI: 0 NaN, 2358 Non-NaN
Non-AKI: 0 NaN,
2358 Non-NaN

Feature Selection

- Features were normalized to the range $[0, 1]$ using the Max-Min function:

$$X_{\text{scaled}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

- Variance Inflation Factor (VIF) was calculated for all features using the formula:

$$\text{VIF}_i = \frac{1}{1 - R_i^2}$$

- Features with VIF > 10 were removed iteratively. The process continued until all features had VIF < 10.
- The t-test was used to compare the means of quantitative features between the AKI and Non-AKI cohorts.
- χ^2-test to assess associations between categorical features and AKI status.

Feature Selection

MHEDAS Master Health Data Science

Table: Comparison of features between AKI and Non-AKI patients. Quantitative features are summarized by mean and standard deviation (SD), while categorical features are presented as counts for each group. A feature is considered statistically significant if the p-value is ≤ 0.05.

Feature (Unit)	AKI Mean (SD) / Count	Non-AKI Mean (SD) / Count	p-value
Age (years)	65.457 (14.986)	62.623 (15.045)	1.002×10^{-10}
Minimum Creatinine (mg/dL)	1.930 (1.513)	1.387 (1.753)	1.441×10^{-29}
Maximum Anion gap (mmol/L)	16.749 (4.605)	14.624 (3.604)	1.476×10^{-67}
Minimum Anion gap (mmol/L)	14.186 (3.753)	12.650 (3.070)	4.015×10^{-52}
Maximum Chloride (mmol/L)	108.420 (7.428)	107.488 (6.011)	2.234×10^{-6}
Minimum Chloride (mmol/L)	105.122 (7.557)	104.669 (5.856)	2.144×10^{-2}
Maximum Glucose (mg/dL)	174.921 (92.680)	155.957 (67.823)	1.341×10^{-15}
Maximum Platelet Count ($10^3/\mu L$)	225.550 (137.671)	237.127 (142.619)	4.588×10^{-3}
Maximum Potassium (mmol/L)	4.504 (0.822)	4.243 (0.683)	4.955×10^{-32}
Minimum Potassium (mmol/L)	3.862 (0.636)	3.707 (0.527)	1.777×10^{-19}
Maximum Sodium (mmol/L)	140.210 (6.018)	139.444 (4.198)	4.048×10^{-7}
Minimum Sodium (mmol/L)	137.386 (5.812)	137.058 (4.498)	3.030×10^{-2}
Minimum Hematocrit (%)	29.201 (5.483)	29.885 (4.839)	5.719×10^{-6}
Maximum Hemoglobin (g/dL)	10.647 (1.869)	10.867 (1.668)	2.086×10^{-5}
Minimum eGFR (mL/min/1.730 m^2)	54.405 (43.681)	97.813 (80.100)	5.736×10^{-112}
Minimum BUN (mg/dL)	39.019 (26.021)	20.502 (14.097)	2.487×10^{-185}
Minimum SBP (mmHg)	84.855 (14.873)	86.624 (13.538)	1.976×10^{-5}
Minimum DBP (mmHg)	41.174 (10.049)	42.601 (9.722)	7.452×10^{-7}
Maximum Temperature (°C)	37.611 (0.981)	37.858 (0.925)	7.728×10^{-19}
Minimum Temperature (°C)	36.077 (0.856)	36.221 (0.781)	1.383×10^{-9}
Gender	{0: 962, 1: 1396}	{0: 1172, 1: 1186}	9.689×10^{-10}
Ethnicity	{1: 1693, 2: 260, 3: 405}	{1: 1651, 2: 349, 3: 358}	2.707×10^{-4}
DM2	{0: 1382, 1: 976}	{0: 1597, 1: 761}	1.044×10^{-10}
CAD	{0: 1691, 1: 667}	{0: 1853, 1: 505}	5.794×10^{-8}
CKD	{0: 1593, 1: 765}	{0: 1972, 1: 386}	1.355×10^{-37}
HYP	{0: 1209, 1: 1149}	{0: 1451, 1: 907}	1.473×10^{-12}

Model Performance Metrics

Table: Performance metrics for training and testing datasets.

Metric	Train	Test
Accuracy	0.737	0.741
Precision	0.751	0.770
Recall	0.716	0.673
F1-score	0.733	0.718

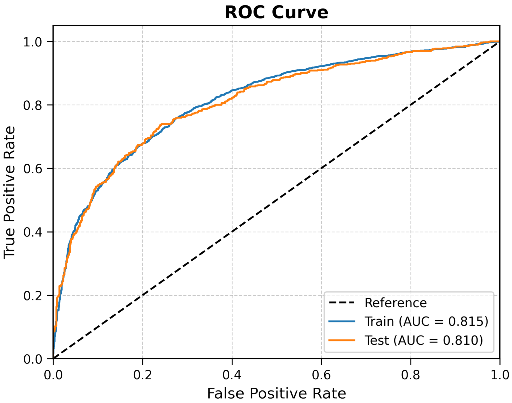


Figure: ROC curves for training and testing datasets.

Feature Contribution Analysis

- eGFR (most important marker for kidney function).
- Creatinine and BUN (indicators of renal health).
- Others: age, electrolytes, hemoglobin.

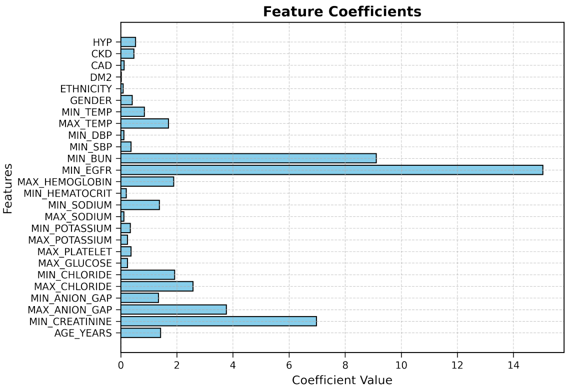


Figure: Feature importance in the logistic regression model.

Benchmarking Against Existing Studies

- Roknaldin et al. (2024): MIMIC-III database; 3301 ICU patients with sepsis; analyzed demographics, labs, vitals, and interventions. Multiple models.
- Malhotra et al. (2017): Multicenter dataset; 717 ICU patients; developed and validated an AKI risk score using logistic regression.
- Jiang et al. (2023): MIMIC-III database; 963 ICU patients with acute pancreatitis; developed a nomogram for AKI prediction using logistic regression.

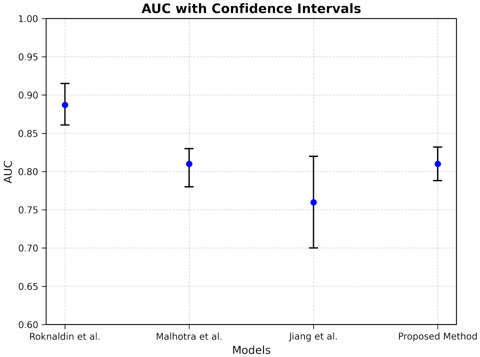


Figure: Comparison of AUC values across studies using logistic regression for AKI prediction.

Conclusions

MHEDAS Master Health Data Science

- **Study Findings:**
  - Developed a logistic regression model for AKI prediction.
  - Achieved AUC = 0.810 (95% CI: 0.788–0.832), accuracy = 74.06%, precision = 76.97%, recall = 67.34%, and F1-score = 71.83%.
  - Key predictors: eGFR, creatinine, BUN; secondary: age, electrolytes
- **Limitations:**
  - Single-source data limits generalizability.
  - Logistic regression may miss non-linear patterns.
  - Missing data imputation introduces uncertainty.
- **Future Directions:**
  - Add features like ICU interventions to improve accuracy.
  - Validate on external datasets for broader applicability.
  - Explore advanced models for capturing complex relationships.

References I

Helen C. Looker, Michael Mauer, and Robert G. Nelson.
Role of kidney biopsies for biomarker discovery in diabetic kidney disease.
Advances in Chronic Kidney Disease, 25(2):192–201, 2018.
Diabetic Kidney Disease (c. 2018).

Wikipedia contributors.
Acute kidney injury, 2025.
Accessed: 2025-01-20.

Esra Adiyeke, Yuanfang Ren, Shmuel Fogel, Parisa Rashidi, Mark Segal, Elizabeth A. Shenkman, Azra Bihorac, and Tezcan Ozrazgat-Baslanti.
Epidemiology, trajectories and outcomes of acute kidney injury among hospitalized patients: A retrospective multicenter large cohort study.
arXiv preprint arXiv:2403.08020, Mar 2024.

References II

Giuseppe Vagliano et al.
A systematic review of machine learning models for predicting acute kidney injury.
Clinical Kidney Journal, 15(12):2266–2277, 2022.

Jianjun Wei et al.
K-nearest neighbors algorithm for predicting 90-day renal recovery in severe aki patients.
BMC Medical Informatics and Decision Making, 2023.

Feng Yue et al.
Application of xgboost algorithm for predicting aki in critically ill septic patients.
BMC Translational Medicine, 2022.

References III

Alistair Johnson, Tom Pollard, and Roger Mark.
Mimic-iii clinical database (version 1.4).
PhysioNet, 2016.

MIT Laboratory for Computational Physiology.
Mimic code issue #608.
GitHub.
Accessed: 2025-01-23.

Stef Van Buuren and Karin Groothuis-Oudshoorn.
mice: Multivariate imputation by chained equations in r.
Journal of Statistical Software, 45(3):1–67, 2011.

References IV

György Kovács.
Smote-variants: A python implementation of 85 minority oversampling techniques.
Neurocomputing, 366:352–354, 2019.

User Community.
When using variance inflation factor (vif), should i do the removal recursively?
Cross Validated (Stack Exchange).
Accessed: 2025-01-23.

Built In.
T-test vs. chi-square: What's the difference?
Built In.
Accessed: 2025-01-23.

References V

- Scikit learn Developers.
  Logisticregression — scikit-learn 1.5 documentation.
  Scikit-learn Documentation.
  Accessed: 2025-01-23.
- User Community.
  Confidence interval auc with the bootstrap method.
  Stack Overflow.
  Accessed: 2025-01-23.
- Alan S Go, Glenn M Chertow, Diane Fan, Charles E McCulloch, and Chi-Yuan Hsu.
  Chronic kidney disease and the risks of death, cardiovascular events, and hospitalization.
  New England Journal of Medicine, 351(13):1296–1305, 2004.

References VI

Andrew S Levey and Josef Coresh.
Chronic kidney disease.
The Lancet, 379(9811):165–180, 2012.

Pierre Delanaye, Etienne Cavalier, and Hans Pottel.
Creatinine: not so simple!
Nephron Clinical Practice, 120(3):c194–c197, 2012.

HJ Carroll.
Blood urea nitrogen.
Journal of Clinical Pathology, 50(5):473–473, 1997.

J W Ter Avest, J P van Dijk, L D Dikkeschei, and T S van der Werf.
Chloride: the queen of electrolytes?
Clinical Chemistry, 46(9):1435–1438, 2000.

References VII

Kyle J Gunnerson and John A Kellum.
Clinical review: the meaning of acid–base abnormalities in the intensive care unit–effects on patient outcome.
Critical Care, 9(5):508, 2005.

Marcello Tonelli, Natasha Wiebe, Bruce Culleton, Andrew House, Christopher Rabbat, Marcus Fok, Finlay McAlister, and Amit X Garg.
Chronic kidney disease and mortality risk: a systematic review.
Journal of the American Society of Nephrology, 17(7):2034–2047, 2006.

Nicolaos E Madias.
Hyponatremia.
Kidney International, 52:S1–S8, 1997.

References VIII

John H Galla.
Disorders of potassium homeostasis: hypokalemia and hyperkalemia.
Critical Care Clinics, 16(2):311–328, 2000.

Aleyeh Roknaldin, Zehao Zhang, Jiayuan Xu, Kamiar Alaei, and Maryam Pishgar.
Utilizing machine learning models to predict acute kidney injury in septic patients from mimic-iii database.
arXiv preprint arXiv:2412.03737, 2023.

Ziming Jiang, Xiangyu An, Yueqian Li, Chen Xu, Haining Meng, and Yan Qu.
Construction and validation of a risk assessment model for acute kidney injury in patients with acute pancreatitis in the intensive care unit.
BMC Nephrology, 24:123, 2023.

References IX

MHEDAS Master Health Data Science

Rakesh Malhotra, Kianoush B. Kashani, Etienne Macedo, Jihoon Kim, Josee Bouchard, Susan Wynn, Guangxi Li, Lucila Ohno-Machado, and Ravindra Mehta.
A risk prediction score for acute kidney injury in the intensive care unit.
Nephrology Dialysis Transplantation, 32(5):814–822, 2017.

Predicting Acute Kidney Injury in Septic Patients Using Logistic Regression with MIMIC-III Data

Antonio García, Daniel Girbes, Rahul Sandhu

Master in Health Data Science — MHEDAS

January 23, 2025

Co-funded by the European Union