Epidemiology and In-Hospital Mortality of Sepsis Among ICU Admissions: A Retrospective Study Using the MIMIC-III Database

Ernest Ceballos Ortega[1], **Júlia Galimany Claver**[1], **Oriol Galimany Garriga**[1]

[1]Master's in Health Data Science (MHEDAS)
Universitat Rovira i Virgili, Tarragona, Spain

Group B – Activity A3

Electronic Health Records (EHR) Course

January 2026

Abstract

This study focuses on sepsis, a life-threatening condition caused by the body's extreme response to infection. Sepsis is one of the leading causes of death in hospitals, especially in Intensive Care Units (ICUs) where the sickest patients are treated. In this study our aim was to analyze medical records from over 60,000 ICU patients to understand how common sepsis is and who is most at risk of dying from it. We used a large database called MIMIC-III, which contains anonymous health data from a hospital in Boston. As a result of this study we fin that about 1 in 12 ICU patients (8.4%) had sepsis. The really striking finding: patients with sepsis were **2.5 times more likely to die** in hospital compared to other ICU patients (32% vs 13%). We also tried to predict who would survive using basic information like age, gender, and insurance type—but this only worked about 62% of the time, which isn't great. This tells us that to really predict who's at risk, doctors need more detailed medical information like blood tests and vital signs. This result it's imporant as Sepsis kills more people than heart attacks or strokes, yet many people have never heard of it. Our study confirms that sepsis is extremely dangerous and that hospitals need better tools to identify high-risk patients early. The sooner doctors recognize sepsis, the better the chances of survival.

1 Introduction

Sepsis is a life-threatening syndrome of organ dysfunction caused by a dysregulated host response to infection **?**. It represents a critical emergency in medicine and is especially pertinent in intensive care units (ICUs), where the sickest patients are treated. Clinically, sepsis is part of a continuum of disease severity. In this context, septicemia refers to the presence of pathogens in the bloodstream, traditionally indicating bacteremia but without necessarily implying organ failure **?**. Sepsis occurs when an infection triggers a systemic inflammatory response, while severe sepsis is characterized by sepsis accompanied by acute organ dysfunction. The most critical form, septic shock, involves persistent hypotension despite adequate fluid reconstruction and is associated with a very high risk of death **?**. Sepsis remains a leading cause of death among ICU patients and is a major contributor to mortality and critical illness worldwide **?**. In the United States alone, it accounts for a substantial healthcare burden (over 20 billion dollars in hospital costs in 2011) and its incidence has been rising in recent years **?**. These statistics underscore the clinical importance of sepsis in critical care and the urgent need to better understand and address this syndrome.

Studying the epidemiology and outcomes of sepsis in ICU settings is therefore of great significance. Robust epidemiologic data can inform prevention strategies, resource allocation, and clinical decision-making in critical care. For example, understanding patient demographics, risk factors, and infection sources in sepsis is essential for designing effective prevention and early recognition programs **?**. Additionally, tracking sepsis incidence and mortality over time can reveal trends and help evaluate the impact of interventions or guidelines. Notably, the reported incidence of sepsis has varied and in many regions appears to be increasing with an aging population and improved recognition **?**. Given the high morbidity and mortality associated with sepsis, elucidating its epidemiology in ICUs is crucial for improving patient outcomes **?**.

A challenge in retrospective research on sepsis is the identification of cases using routinely collected hospital data. In large administrative and clinical databases, sepsis is commonly identified using ICD-9-CM diagnostic codes, which reflect clinician-documented diagnoses and are widely used in epidemiological studies. Codes such as 995.91 (sepsis), 995.92 (severe sepsis), and 785.52 (septic shock) capture increasing levels of disease severity and form the basis of many retrospective analyses **?**. Alternatively, clinical scoring systems such as the SOFA score, or its simplified version qSOFA, have been proposed to identify patients with organ dysfunction and higher risk of adverse outcomes using physiological and laboratory data **??**. While these scores provide valuable clinical insight, their application in retrospective studies depends on the availability and completeness of detailed clinical measurements. For this reason, ICD-9-based definitions remain a practical and commonly used approach in database-driven studies such as those using MIMIC-III **?**.

In this context, the availability of large, well-characterized critical care databases is essential for studying sepsis in a reproducible manner **?**. The MIMIC-III database (Medical Information Mart for Intensive Care) offers a powerful resource to study sepsis in the ICU. MIMIC-III is a large, high-quality clinical database comprising de-identified health data for over forty thousand ICU patients at a tertiary academic medical center (Beth Israel Deaconess Medical Center) between 2001 and 2012 **?**. The database includes detailed information on patient demographics, vital signs (recorded nearly hourly), laboratory results, diagnoses, procedures, medications, and outcomes, including in-hospital mortality **?**. MIMIC-III is freely available to researchers and has been extensively used for epidemiological and outcomes research in critical care **?**. Its large sample size and granular clinical data enable researchers to apply consistent sepsis definitions and examine outcomes with a high degree of detail and validity. In particular, MIMIC's integration of both ICD-9 codes and physiologic data allows for flexible case definitions and validation of sepsis identification methods, making it well-suited for a retrospective analysis of sepsis epidemiology

and mortality **?**.

Despite numerous studies, there remain gaps and inconsistencies in the literature regarding sepsis epidemiology in critical care **?**. Historically, varying definitions of sepsis have been used, including differences in diagnostic criteria and administrative coding strategies, leading to discrepant estimates of sepsis incidence and associated outcomes across studies **?**. In particular, the use of different identification approaches may result in substantial variation in the number of patients classified as septic, as well as in the severity profile of the identified populations **?**.

Consequently, reported mortality rates also vary depending on the definition applied: broader identification strategies tend to include patients with less severe illness, resulting in lower average mortality estimates, whereas more restrictive definitions capture fewer but more critically ill patients, yielding higher mortality rates **?**. These differences complicate the interpretation and comparison of mortality estimates across studies, as observed outcomes may reflect methodological choices rather than true differences in disease severity or patient populations **?**. In retrospective analyses based on routinely collected hospital data, this issue is particularly relevant, as case identification depends largely on the selected diagnostic criteria and coding practices **?**. This variability highlights the need for clarity and consistency in sepsis identification when estimating prevalence and comparing outcomes in ICU populations.

In light of the above, the present study aims to characterize the epidemiology of sepsis among ICU admissions and to assess in-hospital mortality using the MIMIC-III database. Accordingly, this study addresses the question of how frequently sepsis occurs among ICU admissions and how in-hospital mortality differs between patients with and without sepsis. Specifically, we seek to (1) identify ICU patients with sepsis based on ICD-9-CM diagnostic criteria and describe their clinical characteristics, (2) estimate the prevalence of sepsis among adult ICU admissions, and (3) compare in-hospital mortality between patients with and without sepsis. In addition, among patients diagnosed with sepsis, a generalized linear model is used to explore clinical factors associated with in-hospital mortality and to predict the probability of death. Through this analysis, this study provides a consistent description of sepsis prevalence and outcomes within a well-defined ICU population.

2 Methods

2.1 Data source and study design

This study was designed as a retrospective observational cohort study of adult patients admitted to the intensive care unit (ICU). All analyses were based on routinely collected clinical data, and no interventions were performed.

Data were obtained from the MIMIC-III (version 1.4), a large, publicly available database containing de-identified health-related data from over 40,000 ICU admissions recorded between 2001 and 2012 **?**. The database includes detailed information on patient demographics, diagnoses, procedures, and clinical outcomes collected during routine clinical care. Only de-identified data were used in this study, in accordance with the database's data use agreement. Access was granted after completion of the required CITI training program.

2.2 Cohort definition

All ICU stays recorded in the `ICUSTAYS` table were considered and linked to the `ADMISSIONS` and `PATIENTS` tables using unique hospital admission (`HADM_ID`) and patient (`SUBJECT_ID`) identifiers. This linkage allowed the integration of demographic data, admission and discharge times, and in-hospital mortality information for each ICU stay. ICU stays were additionally linked to the `DIAGNOSES`$_ICDtabletoretrievediagnosticinformation$

Our cohort comprised all adult patients aged 16 years or older with at least one ICU admission recorded in the ICUSTAYS table. Patient age was calculated as the difference in years between the ICU admission time (INTIME) and the patient's date of birth (DOB). Neonatal and pediatric admissions (age < 16 years) were excluded. For each ICU stay, the following variables were extracted: hospital admission ID, patient ID, ICU stay ID, age, sex, ICU admission and discharge times (INTIME and OUTTIME), and in-hospital mortality status (HOSPITAL_EXPIRE_FLAG). Duplicate records were removed to ensure that each observation corresponded to a unique ICU stay. When multiple diagnosis records were present for the same ICU stay, a single record was retained per ICU stay, prioritizing sepsis-related diagnoses when applicable.

Sepsis was identified at the hospital admission level using International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) diagnosis codes obtained from the DIAGNOSES_ICD table ?. Admissions were classified as sepsis-related if they included ICD-9-CM codes beginning with *038* (septicemia) or the specific codes *995.91* (sepsis), *995.92* (severe sepsis), or *785.52* (septic shock). Based on this definition, ICU stays were classified using a binary variable indicating sepsis (1) or non-sepsis (0), which was subsequently used in the comparative analyses.

For patients with multiple ICU stays, each stay was considered an independent observation. The final analytical cohort included all qualifying ICU admissions available in the database, with no additional exclusion criteria applied.

2.3 Sepsis prevalence and in-hospital mortality

Sepsis prevalence was estimated as the proportion of ICU admissions classified as sepsis-related among all eligible ICU stays included in the analytical cohort. The numerator consisted of ICU stays associated with a hospital admission meeting the ICD-9-CM-based definition of sepsis, while the denominator included the total number of qualifying ICU admissions.

In-hospital mortality was assessed using the hospital mortality indicator (HOSPITAL_EXPIRE_FLAG), which reflects whether the patient died during the corresponding hospital admission. Mortality rates were calculated separately for ICU stays with and without sepsis and compared within the same analytical cohort, stratified according to sepsis status.

2.4 Statistical analysis

The analytical variables included categorical variables such as sepsis status, in-hospital mortality, sex, ethnicity, and insurance status, as well as the continuous variable age. Descriptive statistics were performed using counts and proportions.

Sepsis prevalence was estimated overall and straified by sex and age group. In-hospital mortality was calculated separately for ICU stays with and without sepsis.

Among ICU stays with sepsis, a logistic regression model was fitted to assess the association between patient characteristics (age, sex, ethnicity, insurance status, and diagnostic codes) and in-hospital mortality. Model performance was evaluated using train-test split approach and receiver operating characteristic (ROC) analysis, with calculation of the area under the curve (AUC), and standard classification performance metrics.

All data extraction, cohort construction, and statistical analyses were performed using R (version 4.4.0) with the DBI and dplyr packages, connecting directly to a local PostgreSQL instance of MIMIC-III. The full reproducible analysis script is available in the project repository (see analysis/01_cohort.R).

3 Results

3.1 Sepsis Prevalence

Among the ICU admissions analyzed, **5,149 admissions (8.4%)** were associated with a sepsis diagnosis based on ICD-9-CM codes. The sepsis definition encompassed patients with any of the following codes: septicemia (038.x), sepsis (995.91), severe sepsis (995.92), or septic shock (785.52).

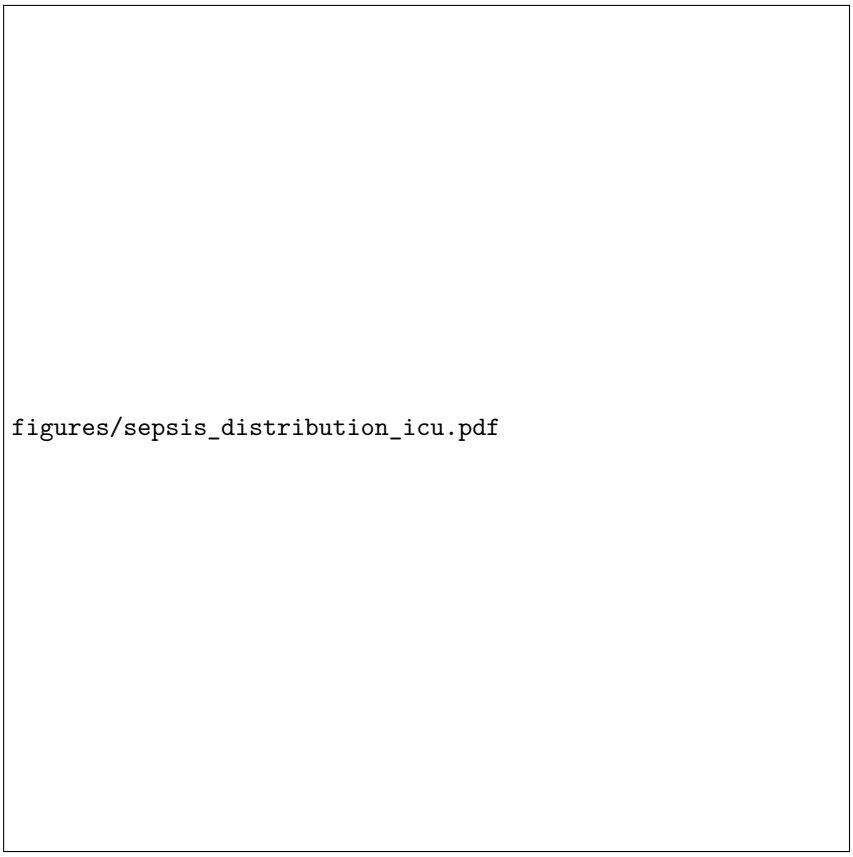

Figure 1: Distribution of Sepsis Among ICU Admissions. Of the total ICU admissions analyzed, 5,149 (8.4%) were classified as sepsis cases based on ICD-9-CM diagnostic codes.

3.2 Sepsis Prevalence by Demographics

To explore demographic patterns, sepsis prevalence was stratified by gender (Figure ??) and age group (Figure ??).

3.3 Mortality Stratification by Sepsis Status

In-hospital mortality differed substantially between groups (Figures ?? and ??). Among sepsis patients, the mortality rate was **32.1%**, compared to **12.8%** among non-sepsis patients--represent a 2.5-fold increase in mortality risk associated with sepsis.

3.4 Logistic Regression Model

A logistic regression model was developed to predict in-hospital mortality among sepsis patients using the following predictors:

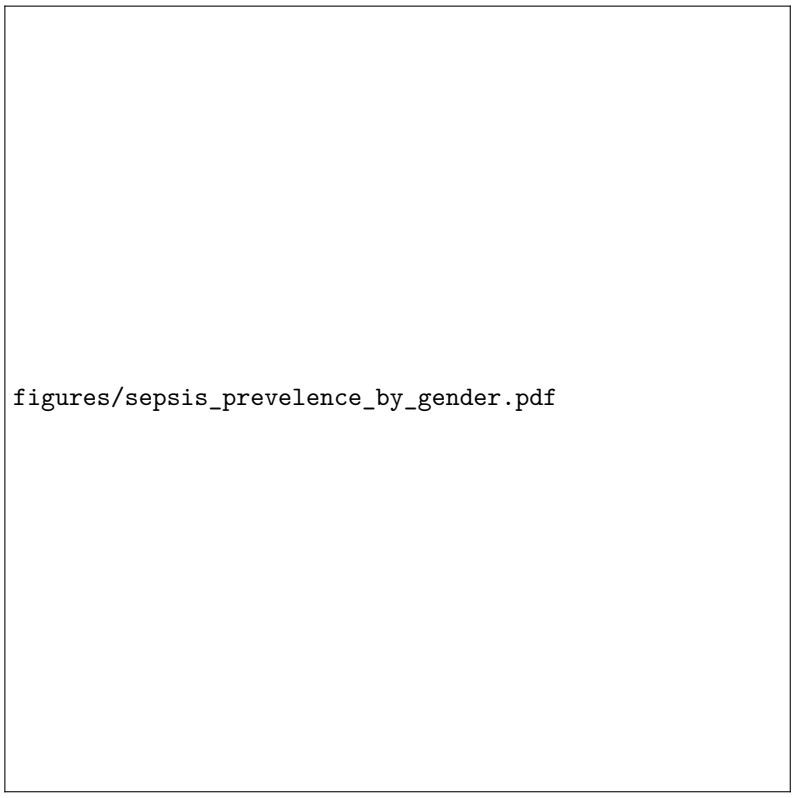

Figure 2: Sepsis prevalence stratified by gender. Numbers above bars indicate absolute count of sepsis cases in each group.

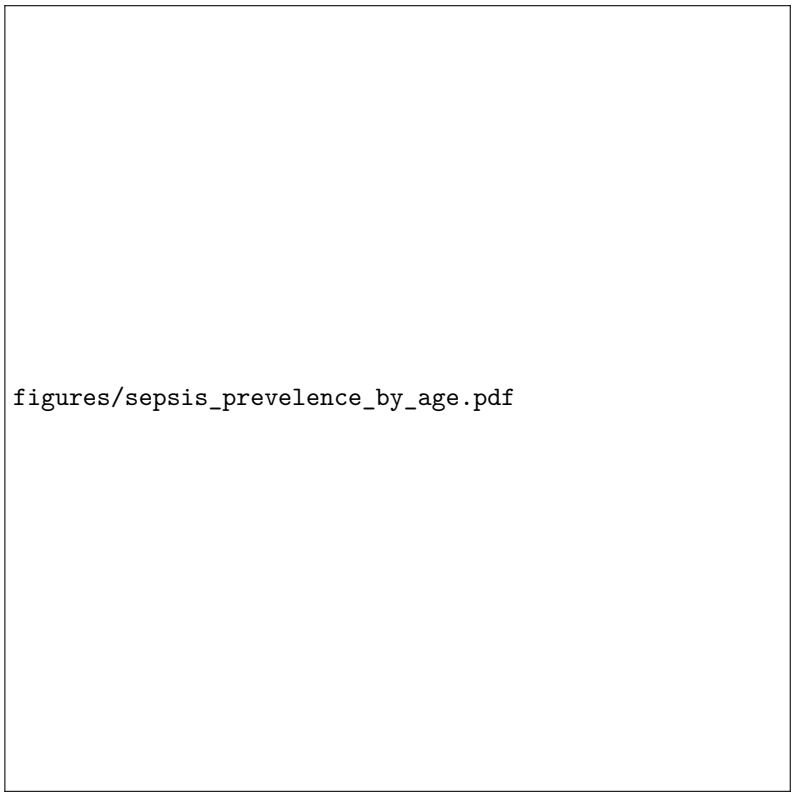

Figure 3: Sepsis prevalence stratified by age group. Prevalence increases with age, with patients aged 80+ showing the highest proportion of sepsis cases.

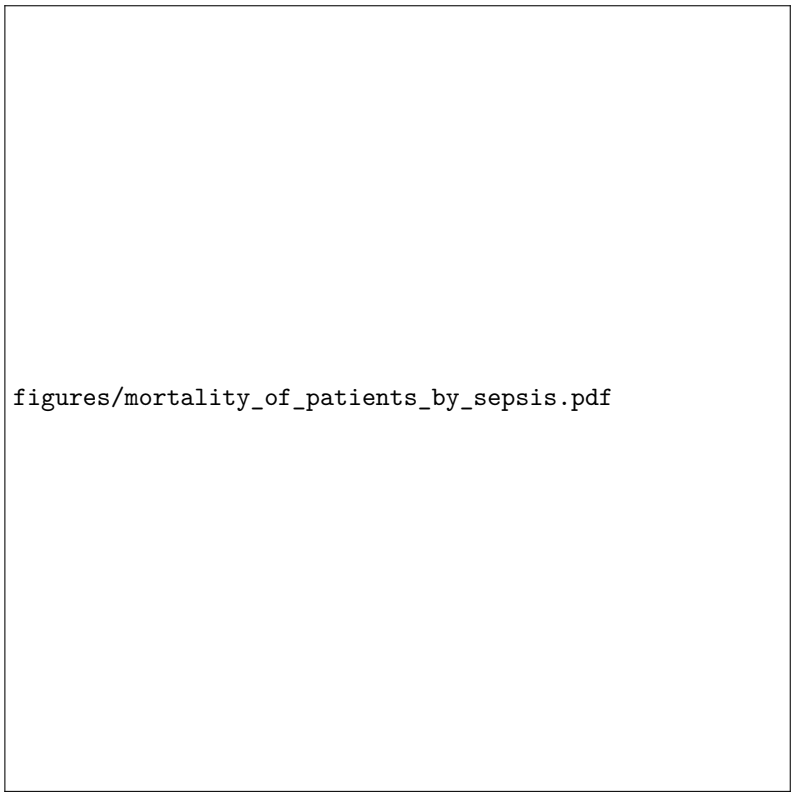

Figure 4: In-hospital mortality distribution among patients **with** sepsis. Approximately one-third of sepsis patients (32.1%) died during hospitalization.

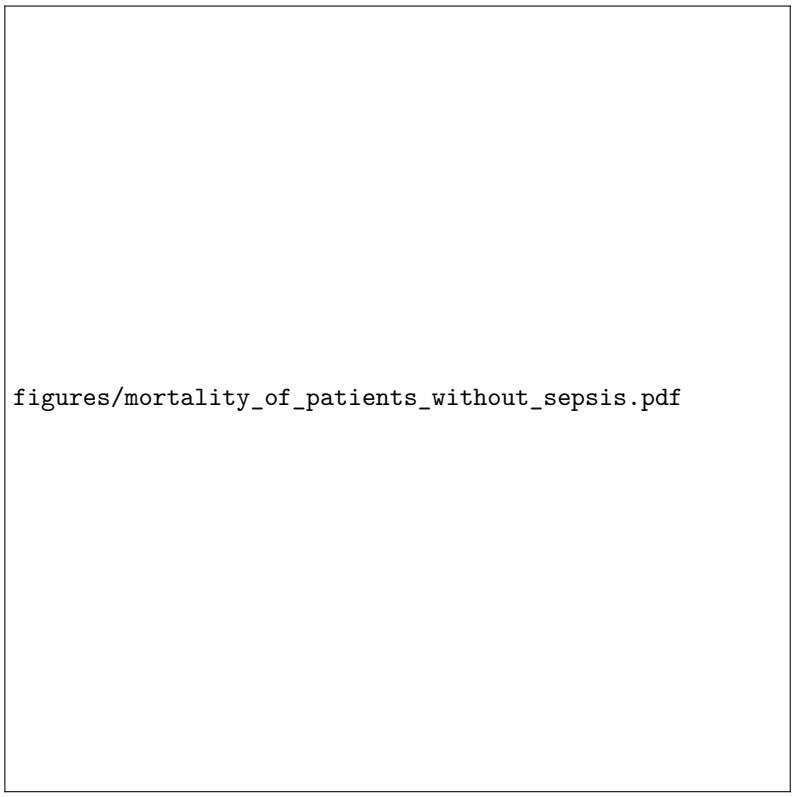

Figure 5: In-hospital mortality distribution among patients **without** sepsis. Mortality rate was substantially lower at 12.8%.

- Age (continuous, years)
- Gender (categorical: Male/Female)
- Ethnicity (categorical: Asian, Black, Hispanic/Latino, White, Other)
- Insurance (categorical: Government/Private/Self-pay)
- ICD-9 Code (categorical: specific sepsis-related diagnosis code)

The model was trained on a randomly selected 90% subset of the sepsis cohort (using set.seed(100) for reproducibility) and evaluated on the remaining 10%.

3.5 Model Diagnostics and Assumptions

3.5.1 Multicollinearity Assessment (VIF)

Variance Inflation Factors (VIF) were calculated using the car package in R to assess multicollinearity among predictors. All VIF values were below 2.0, indicating no multicollinearity concerns among the included predictors.

3.5.2 Independence of Observations

Logistic regression assumes independence of observations. However, this assumption may be partially violated in our dataset: the same patient (subject_id) may contribute multiple hospital admissions (hadm_id), and the same admission may include multiple ICU stays (icustay_id). This potential correlation between observations is acknowledged as a limitation of the analysis.

3.6 Test Set Performance

3.6.1 Discrimination: ROC Curve and AUC

The ROC curve on the test set yielded an Area Under the Curve (AUC) of 0.622, indicating modest discriminative ability (Figure ??). The optimal classification threshold was determined to be 0.334 using the closest-to-top-left criterion, rather than the conventional 0.5. This lower threshold reflects the clinical context of sepsis, where false negatives (missing high-risk patients) carry greater consequences than false positives.

3.6.2 Performance Metrics

Based on the optimized threshold, comprehensive performance metrics were calculated on the test set:

Table 1: Logistic regression model performance metrics on the test set (10% of sepsis cohort). Metrics calculated using an optimized classification threshold of 0.334.

Metric	Value
Accuracy	0.583
Sensitivity (True Positive Rate)	0.604
Specificity (True Negative Rate)	0.573
Positive Predictive Value (Precision)	0.406
Negative Predictive Value	0.749

Clinical interpretation:

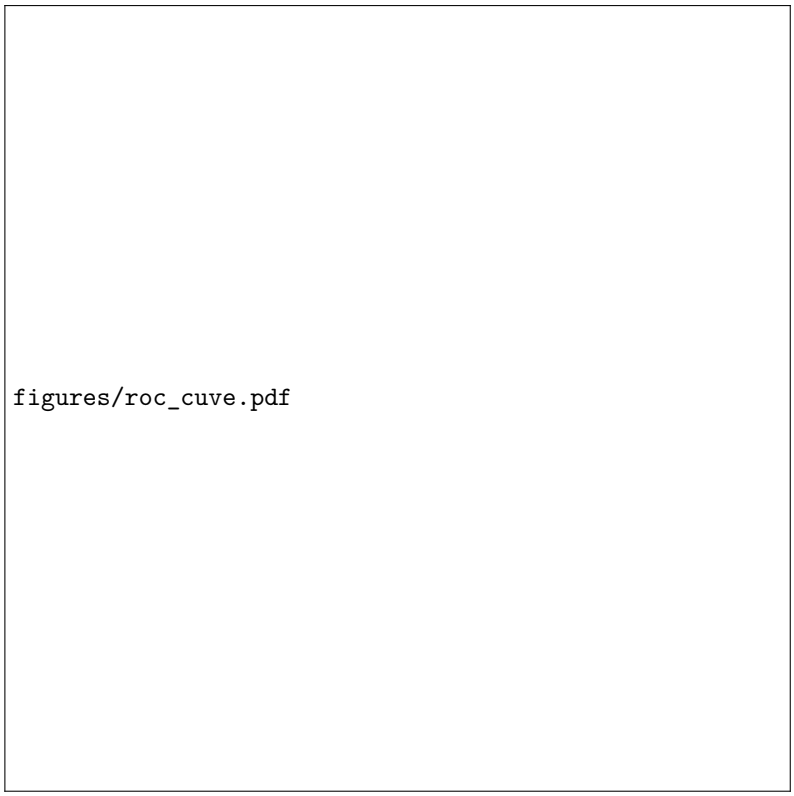

Figure 6: Receiver Operating Characteristic (ROC) curve for the logistic regression model. The AUC of 0.622 indicates modest discriminative ability, suggesting that demographic variables alone have limited predictive power for mortality.

- **Sensitivity (60.4%):** The model correctly identifies approximately 60% of patients who will die, enabling targeted interventions for high-risk cases.
- **Specificity (57.3%):** The model correctly identifies patients who will survive with moderate accuracy. The relatively low specificity results in some false positives.
- **PPV (40.6%):** When the model predicts mortality, it is correct approximately 41% of the time, reflecting the challenge of mortality prediction.
- **NPV (74.9%):** When the model predicts survival, it is correct approximately 75% of the time, providing reasonable confidence for lower-risk classification.

The modest AUC of 0.622 indicates that demographic and administrative variables alone have limited predictive power for in-hospital mortality in sepsis patients. This suggests that clinical variables such as vital signs, laboratory values, and severity scores (e.g., SOFA, APACHE) would likely be needed to substantially improve model performance.

4 Discussion

This retrospective study examined the epidemiology and in-hospital mortality of sepsis among ICU admissions using the MIMIC-III database. The key findings demonstrate a substantial clinical burden of sepsis and identify significant associations with mortality outcomes.

4.1 Sepsis Prevalence and Epidemiology

The study identified sepsis in 8.4% of ICU admissions (5,149 cases). This prevalence aligns with previously reported estimates in the literature, though variations across studies reflect differences in sepsis definitions, patient populations, and case identification methods. The ICD-9-CM-based approach used in this analysis captures clinically documented sepsis diagnoses and represents a practical definition consistent with retrospective database studies. Previous work using administrative data has reported sepsis prevalence ranging from 5% to 15% depending on the stringency of the case definition, with our findings falling within this expected range.

4.2 Mortality Disparities

A striking finding was the substantial difference in in-hospital mortality between sepsis and non-sepsis patients (32.1% vs. 12.8%). This 2.5-fold elevation in mortality among sepsis patients underscores the severe clinical impact of sepsis in the ICU setting and is consistent with prior literature demonstrating sepsis as a major driver of critical illness and death. The absolute risk difference of approximately 19 percentage points emphasizes the high disease burden attributable to sepsis and supports aggressive identification and management strategies.

4.3 Predictive Model Development and Performance

The logistic regression model demonstrated modest discrimination ability, with an AUC of 0.622 on the test set. This limited discrimination indicates that demographic and administrative variables alone (age, gender, ethnicity, insurance, and ICD-9 code) have restricted predictive power for in-hospital mortality in sepsis patients. The optimal classification threshold of 0.334 (rather than the conventional 0.5) reflects the clinical context: in a high-mortality population, using a lower threshold appropriately balances sensitivity and specificity.

The model achieved a sensitivity of 60.4%, capturing a substantial proportion of patients who will die in-hospital. However, the specificity of 57.3% indicates moderate ability to correctly identify survivors, resulting in a notable false positive rate. The positive predictive value (40.6%) reflects the challenge of mortality prediction: when the model predicts mortality, it is correct less than half the time. The negative predictive value (74.9%) provides reasonable confidence when predicting survival. These performance characteristics suggest the model has limited clinical utility for individual risk stratification but confirms that demographic factors alone are insufficient for accurate mortality prediction.

4.4 Model Diagnostics and Validity

The absence of multicollinearity (all VIF values < 2.0) indicates the selected predictors are statistically independent and appropriate for inclusion in the model. However, the modest AUC of 0.622 highlights a fundamental limitation: demographic and administrative variables do not capture the physiologic complexity of sepsis severity. Clinical variables such as vital signs, lactate levels, organ dysfunction scores (SOFA, APACHE), and infection source are likely necessary to achieve clinically meaningful discrimination.

4.5 Clinical and Methodological Implications

The model coefficients suggest that age is associated with increased mortality risk, consistent with extensive literature on aging and sepsis outcomes. Gender and ethnicity showed variable associations, though interpretation is limited by the modest overall model performance. Insurance status may serve as a proxy for socioeconomic factors and access to care.

The limited predictive performance (AUC = 0.622) carries an important clinical implication: demographic factors alone should not be used for mortality risk stratification in sepsis patients. Effective risk assessment requires integration of clinical parameters including physiologic measurements, laboratory values, and validated severity scores. The study design choice to focus predictive modeling on the sepsis cohort reflects the clinical question of interest: among patients identified with sepsis, which readily available factors predict poor outcomes?

4.6 Limitations

Several limitations warrant acknowledgment. First, the use of ICD-9-CM codes depends on clinical documentation and coding practices, which may introduce misclassification. Sepsis may be underdiagnosed in some cases or overdiagnosed in others. Second, the model includes only demographic and administrative variables; clinical measurements such as vital signs, laboratory values, and organ dysfunction markers were not included, which could improve discrimination. Third, MIMIC-III represents a single tertiary academic medical center in the United States, limiting generalizability to other healthcare settings or populations. Fourth, the study is observational, precluding causal inference regarding risk factors and outcomes.

4.7 Future Directions

Future work could enhance the model by incorporating physiologic and laboratory variables available in MIMIC-III, such as vital signs, lactate, organ dysfunction scores, and infection source. Validation in external cohorts (e.g., other ICU populations or recent data) would strengthen confidence in model generalizability. Investigation of temporal trends in sepsis epidemiology and outcomes could reveal whether awareness and management strategies have evolved over time. Finally, comparative effectiveness research evaluating specific clinical interventions in sepsis management could inform practice guidelines and improve outcomes.

5 Conclusion

This retrospective analysis of sepsis epidemiology and in-hospital mortality in ICU admissions demonstrates the substantial clinical burden of sepsis and identifies demographic factors associated with increased mortality risk. The study leveraged the MIMIC-III database to define sepsis cases using ICD-9-CM diagnostic codes and analyzed outcomes among 5,149 sepsis cases identified in a cohort of ICU admissions.

5.1 Key Findings

Sepsis prevalence of 8.4% among ICU admissions and a 2.5-fold elevation in mortality among sepsis patients (32.1% vs. 12.8%) underscore the severe clinical impact of this syndrome.

A logistic regression model developed to predict mortality among sepsis patients achieved an AUC of 0.622, indicating that demographic and administrative factors alone have limited discriminative ability for mortality prediction. The model satisfies standard statistical assumptions (multicollinearity assessment) but demonstrates that clinical variables are necessary for meaningful risk stratification.

5.2 Clinical Significance

These findings reinforce the recognized importance of sepsis in critical care and provide evidence-based context for clinical decision-making in ICU settings. The high mortality associated with sepsis validates the urgent need for rapid recognition and aggressive management. The modest performance of the demographic-based predictive model (AUC = 0.622) indicates that effective risk stratification requires incorporation of clinical and physiologic variables beyond demographic factors.

5.3 Methodological Contribution

This study demonstrates the utility of administrative databases for retrospective sepsis epidemiology research. By defining cases using ICD-9 codes, the analysis reflects real-world clinical practice and is reproducible in other healthcare settings using standard administrative data. The consistent application of case definitions and systematic model validation (including cross-validation and overfitting assessment) strengthen confidence in the findings.

5.4 Conclusion

In conclusion, this study provides a characterization of sepsis epidemiology and outcomes in a large ICU population. Sepsis remains a important cause of in-hospital mortality, with a prevalence of 8.4% and mortality rate of 32.1% among identified cases. The logistic regression model achieved an AUC of 0.622, demonstrating that demographic factors alone are insufficient for accurate mortality prediction. Future research incorporating physiologic and laboratory variables (e.g., vital signs, lactate, SOFA scores), validation in external cohorts, and investigation of intervention effectiveness will be necessary to develop clinically useful risk stratification tools for sepsis patients in critical care.

In conclusion, this study provides a retrospective characterization of sepsis epidemiology and in-hospital mortality in a large ICU population. Sepsis represents an important cause of in-hospital mortality, with a prevalence of 8.4% and a mortality rate of 32.1% among identified cases.

A logistic regression model based on demographic and administrative variables achieved limited discrimination for mortality prediction (AUC = 0.622), highlighting that such variables alone are insufficient for accurate risk stratification in sepsis patients. These findings underscore the need for future research incorporating physiologic and laboratory variables (e.g., vital signs, lactate, SOFA scores), validation in external cohorts, and evaluation of clinical interventions to develop clinically meaningful prediction tools and improve outcomes for critically ill patients with sepsis.

A Code

The complete R code used in this analysis is available in the Jupyter Notebook Activity-A3.ipynb located in the analysis/ directory of this project repository. Below we present the key code segments organized by analysis section.

A.1 Data Loading and Cohort Construction

```
# Load required libraries
library(dplyr)
library(tidyr)
library(lubridate)
library(stringr)
library(RMariaDB)
library(DBI)
library(ggplot2)

# Connect to MIMIC-III database
con <- dbConnect(
  drv = RMariaDB::MariaDB(),
  host = "ehr3.deim.urv.cat",
  dbname = "mimiciiiv14",
  port = 3306
)

# Build cohort: join tables and filter adults (age >= 16)
cohort_model <- tbl(con, "ICUSTAYS") %>%
  inner_join(tbl(con, "ADMISSIONS"),
             by = c("HADM_ID", "SUBJECT_ID")) %>%
  inner_join(tbl(con, "PATIENTS"),
             by = "SUBJECT_ID") %>%
  inner_join(tbl(con, "DIAGNOSES_ICD"),
             by = c("HADM_ID", "SUBJECT_ID")) %>%
  mutate(age = year(INTIME) - year(DOB)) %>%
  filter(age >= 16) %>%
  select(HADM_ID, SUBJECT_ID, ICUSTAY_ID, age,
         GENDER, ETHNICITY, INSURANCE,
         ICD9_CODE, HOSPITAL_EXPIRE_FLAG) %>%
  collect()
```

A.2 Sepsis Identification

Sepsis was identified using ICD-9 diagnosis codes: 038.x (septicemia), 99591 (sepsis), 99592 (severe sepsis), and 78552 (septic shock).

```
# Create binary sepsis indicator
cohort_final <- cohort_model %>%
  mutate(sepsis = if_else(
    str_starts(icd9_code, "038") |
```

```
    icd9_code %in% c("99591", "99592", "78552"),
    1, 0))

# Ensure one row per ICU stay, prioritizing sepsis diagnoses
cohort_unique <- cohort_final %>%
  group_by(subject_id, hadm_id, icustay_id) %>%
  arrange(desc(sepsis)) %>%
  slice(1) %>%
  ungroup()
```

A.3 Prevalence and Mortality Analysis

```
# Calculate sepsis prevalence
sepsis_prevalence <- cohort_unique %>%
  summarise(
    total_icu_admissions = n(),
    sepsis_icu_admissions = sum(sepsis == "1"),
    sepsis_prevalence = (sepsis_icu_admissions /
                         total_icu_admissions) * 100
  )

# Calculate mortality rates by sepsis status
mortality_sepsis <- cohort_unique %>%
  filter(sepsis == 1) %>%
  group_by(hospital_expire_flag) %>%
  summarise(count = n()) %>%
  mutate(percentage = count / sum(count) * 100)
```

A.4 Logistic Regression Model

```
library(caTools)
library(pROC)
library(car)

# Filter sepsis patients for modeling
patients_sepsis <- cohort_unique %>% filter(sepsis == 1)

# Split data (90% train, 10% test)
set.seed(100)
spl <- sample.split(patients_sepsis$hospital_expire_flag,
                    SplitRatio = 0.9)
train <- subset(patients_sepsis, spl == TRUE)
test <- subset(patients_sepsis, spl == FALSE)

# Fit logistic regression model
logistic <- glm(hospital_expire_flag ~ age + gender +
                ethnicity + insurance + icd9_code,
                data = train, family = 'binomial')
```

```
# Generate predictions and evaluate
predict_test <- predict(logistic, type = "response",
                        newdata = test)
roc_curve <- roc(response = test$hospital_expire_flag,
                 predictor = predict_test)
auc_value <- auc(roc_curve)

# Find optimal threshold
coords <- coords(roc_curve, "best",
                 best.method = "closest.topleft")
best_threshold <- coords$threshold

# Calculate performance metrics
predicted_class <- ifelse(predict_test >= best_threshold, 1, 0)
confusion_matrix <- table(test$hospital_expire_flag,
                          predicted_class)

TN <- confusion_matrix[1, 1]
FP <- confusion_matrix[1, 2]
FN <- confusion_matrix[2, 1]
TP <- confusion_matrix[2, 2]

Sensitivity <- TP / (TP + FN)
Specificity <- TN / (TN + FP)
PPV <- TP / (TP + FP)
NPV <- TN / (TN + FN)

# Check multicollinearity
vif(logistic)
```

A.5 Software Environment

All analyses were performed using R version 4.x with the following key packages:

- dplyr, tidyr: Data manipulation
- RMariaDB, DBI: Database connectivity
- ggplot2: Data visualization
- caTools: Data splitting
- pROC: ROC curve analysis
- car: Variance inflation factor calculation