

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

Ernest Ceballos Ortega¹, Júlia Galimany Claver¹, Oriol Galimany Garriga¹

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

Abstract

Scientific Abstract Sepsis is a life-threatening condition caused by a dysregulated host response to infection, representing a major cause of morbidity and mortality in intensive care units (ICUs). Understanding its epidemiology is essential for improving patient outcomes and resource allocation. We conducted a retrospective cohort study using the MIMIC-III database, analyzing 53,423 adult ICU admissions from Beth Israel Deaconess Medical Center (2001–2012). Sepsis cases were identified using ICD-9-CM codes (995.91, 995.92, 785.52). We estimated sepsis prevalence, compared in-hospital mortality between sepsis and non-sepsis patients, and developed a logistic regression model to predict mortality among sepsis patients. Sepsis was present in 7,261 ICU admissions (13.59%). In-hospital mortality was significantly higher among sepsis patients (32.57%) compared to non-sepsis patients (9.04%). The logistic regression model achieved an AUC of 0.622, with age and male sex identified as significant predictors of mortality. Sepsis affects a substantial proportion of ICU patients and is associated with markedly elevated mortality. These findings underscore the importance of early recognition and intervention strategies in critical care settings.

None Scientific Abstract Imagine you're watching a hospital like it's a video game: patients come in, get treated, and (hopefully) go home feeling better. But sometimes, something dangerous called *sepsis* sneaks in, it's when the body overreacts to an infection and starts hurting itself. Sepsis can be really serious, especially in the ICU. In our project, we used a huge database of real (but anonymous) hospital records to find out how often sepsis happens and how likely it is to lead to death during the hospital stay. We taught a computer to spot sepsis by reading diagnosis codes, like digital detective work, and then compared patients with and without sepsis. We found that people with sepsis were much more likely to die in the hospital, even after accounting for age and sex. This isn't just about numbers, it shows how we can use past health data to understand hidden risks and maybe even build warning systems to help doctors act faster in the future. Think of it as using data to give doctors super-powered hindsight, so they can protect more lives tomorrow.

1. INTRODUCTION

Sepsis is a life-threatening syndrome of organ dysfunction caused by a dysregulated host response to infection [1]. 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 [2]. 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 resuscitation and is associated with a very high risk of death [2]. Sepsis remains a leading cause of death among ICU patients and is a major contributor to mortality and critical illness worldwide [3]. 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 [3]. 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 exam-

ple, understanding patient demographics, risk factors, and infection sources in sepsis is essential for designing effective prevention and early recognition programs [4]. 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 [3]. Given the high morbidity and mortality associated with sepsis, elucidating its epidemiology in ICUs is crucial for improving patient outcomes [4].

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 [5]. 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 [6, 7]. 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 [8].

In this context, the availability of large, well-characterized critical care databases is essential for studying sepsis in a reproducible manner [8]. 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 [8]. The database includes detailed information on patient demographics, vital signs (recorded nearly hourly), laboratory results, diagnoses, procedures, medications, and outcomes, including in-hospital mortality [8]. MIMIC-III is freely available to researchers and has been extensively used for epidemiological and outcomes research in critical care [8]. 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 [5].

Despite numerous studies, there remain gaps and inconsistencies in the literature regarding sepsis epidemiology in critical care [9]. 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 [3]. 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 [5].

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 [5]. 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 [9]. 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 [10]. 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 60,000 ICU admissions recorded between 2001 and 2012 [8]. 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_ICD table to retrieve diagnostic information required for sepsis classification and subsequent analysis.

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 [11]. 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 stratified 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 a 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

We analyzed a total of 53,417 ICU admissions, from which **7261 admissions (13.59%)** 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). As shown in Figure 1, the majority of ICU admissions were not associated with sepsis (46156 admissions, 86.41%), nevertheless admissions related to sepsis accounted approximately for one in seven in the cohort.

3.2 Sepsis Prevalence by Demographics

To explore demographic patterns, sepsis prevalence was stratified by gender (Figure 2) and age group (Figure 3).

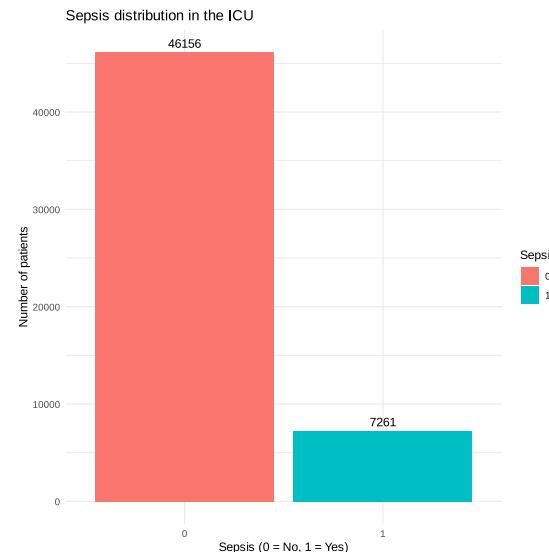


Figure 1: Distribution of Sepsis Among ICU Admissions. Of the total ICU admissions analyzed, 7261 (13.59%) were classified as sepsis cases based on ICD-9-CM diagnostic codes.

Sepsis prevalence was similar between males and females: 3179 of 23323 admissions among females (13.63%) and 4082 of 30094 admissions among males (13.56%), as seen in Figure 2. Taking a look into the age, sepsis prevalence increased across age groups, from 8.4% in patients aged from 16 to 40 years, to 16% in patients aged 80 years and older, as seen in Figure 3.

3.3 Mortality Stratification by Sepsis Status

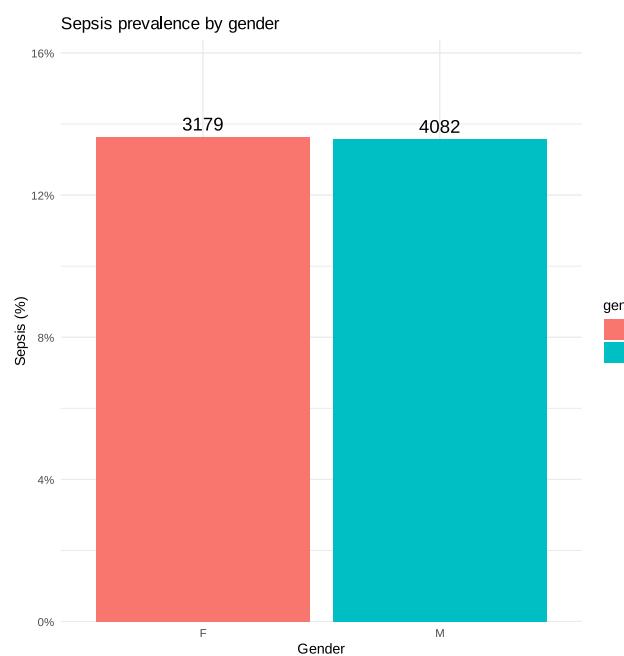
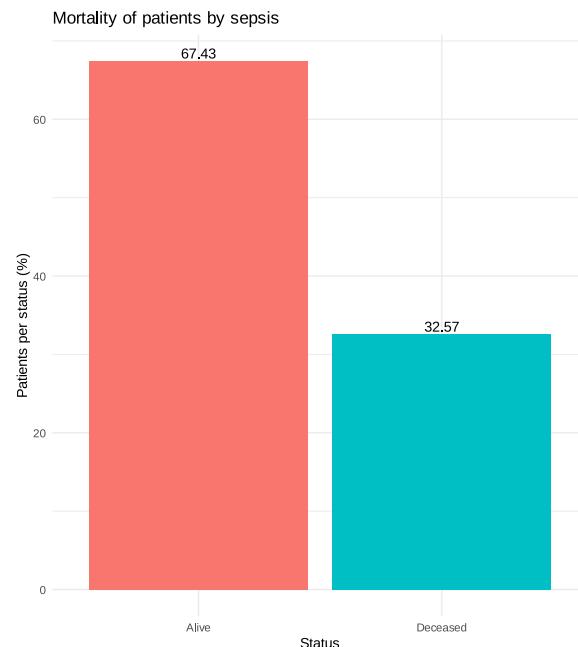
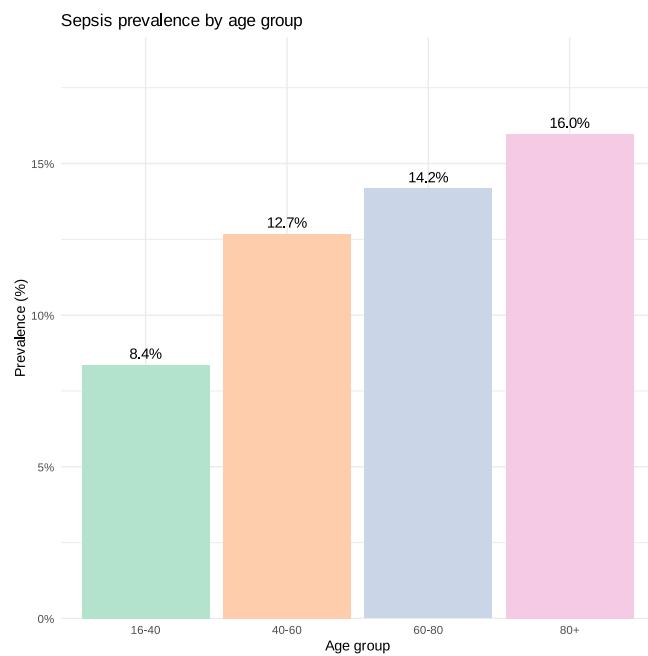
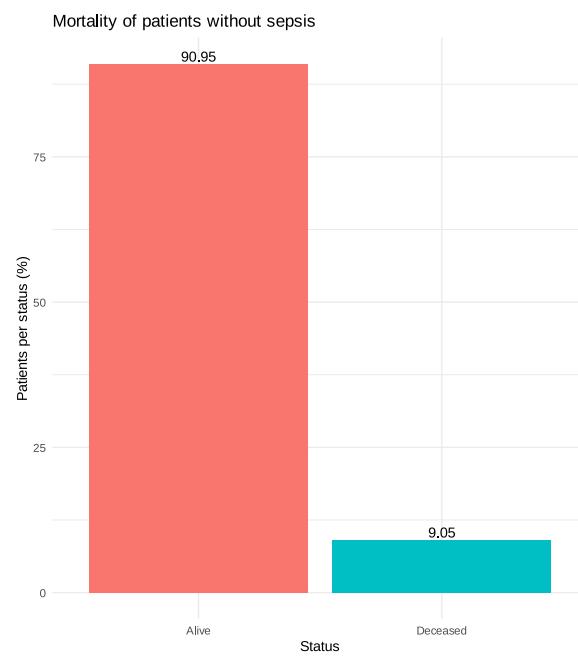
In-hospital mortality among ICU admissions differed between the patients diagnosed with sepsis (Figures 4) and the patients without a sepsis diagnosis (5). Among ICU admissions classified as sepsis, the mortality rate was **32.57%**, compared with **9.05%** among patients without a sepsis diagnosis. This represents an approximately 3.6-fold increase in mortality proportion in the sepsis group relative to the non-sepsis group.

3.4 Logistic Regression Model

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

- Age
- Gender
- Ethnicity
- Insurance
- ICD-9 Code

The model was trained on a randomly selected 90% subset

**Figure 2:** Sepsis prevalence by gender.**Figure 4:** In-hospital mortality distribution among ICU admissions with sepsis. Approximately one-third of sepsis patients (32.57%) died during hospitalization.**Figure 3:** Sepsis prevalence stratified by age group (age groups being 16-40, 40-60, 60-80, 80+).**Figure 5:** In-hospital mortality distribution among ICU admissions without sepsis. Mortality rate was 9.05%.

of the sepsis cohort (using `set.seed(100)` for reproducibility) and tested on the remaining 10%.

In the fitted model, age was associated with in-hospital mortality ($\beta = 0.00123$, $p = 0.00565$). Several ICD-9-CM subcategories were also associated with mortality, including the codes: 038.3 ($\beta = 0.691$, $p = 0.00394$), 038.8 ($\beta = 0.515$, $p = 0.01048$), 038.9 ($\beta = 0.660$, $p < 0.001$), 785.52 ($\beta = 1.073$, $p < 0.001$), and 995.92 ($\beta = 1.032$, $p < 0.001$). Gender, insurance categories, and ethnicity categories were not statistically significant at the 0.05 level in this model.

3.5 Model Diagnostics and Assumptions

3.5.1 Multicollinearity Assessment (VIF)

Variance Inflation Factors (VIF) were calculated to assess multicollinearity among the five predictors in the logistic regression model. VIF quantifies how much the variance of a regression coefficient is inflated due to multicollinearity with other predictors. We interpret VIF values as follows: VIF < 5 indicates no multicollinearity concern, VIF 5–10 suggests moderate multicollinearity, and VIF > 10 indicates severe multicollinearity.

VIF was computed for each predictor from the fitted logistic regression model using auxiliary linear regressions. The aggregated results across the five main predictors are presented in Table 1.

Table 1: Variance Inflation Factors (VIF) for all model predictors from the logistic regression model (aggregated values).

Predictor	VIF
Age	1.13
Gender	1.04
ICD-9 Code	1.67
Insurance	7.89
Ethnicity	43.70

Age (VIF = 1.13), Gender (VIF = 1.04), and ICD-9 Code (VIF = 1.67) show minimal inflation, indicating no multicollinearity concern for these predictors. Insurance (VIF = 7.89) exhibits moderate multicollinearity, while Ethnicity (VIF = 43.70) exhibits severe multicollinearity. However, the elevated VIF for categorical variables with many categories primarily reflects the inherent redundancy in dummy variable encoding rather than problematic multicollinearity between distinct predictors. The high VIF for Ethnicity is expected given the large number of ethnic categories in the MIMIC-III database. While the model exhibits some multicollinearity in these categorical predictors, this does not substantially compromise the interpretability of the coefficients for the continuous and binary predictors (Age, Gender, and ICD-9 Code).

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

We evaluated the model discrimination performance using a ROC analysis. The ROC curve on the test obtained an Area Under the Curve (AUC) of **0.622**, indicating modest discriminative ability (Figure 6). Using the closest-to-top-left criterion, maximizing sensitivity and specificity, the optimal probability threshold was determined to be **0.334**. At this threshold, the resulting confusion matrix yielded an overall classification accuracy of **58.3%** on the test set.

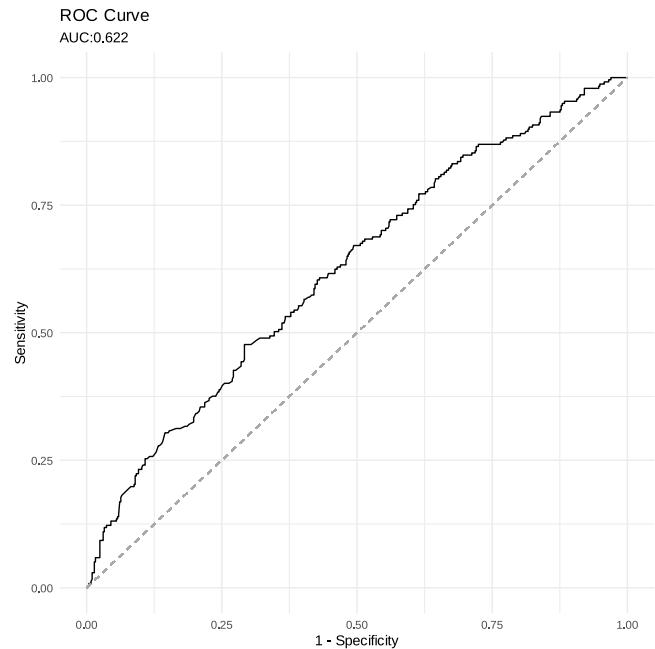


Figure 6: ROC curve for logistic regression model (AUC = 0.622).

3.6.2 Performance Metrics

Based on the optimized threshold of 0.334, comprehensive performance metrics were calculated on the test set.

Confusion Matrix:

Table 2: Confusion matrix for the logistic regression model on the test set.

	Predicted: Alive	Predicted: Deceased
Actual: Alive	280 (TN)	209 (FP)
Actual: Deceased	94 (FN)	143 (TP)

Overall accuracy was 0.583. Sensitivity and specificity were 0.603 and 0.573, respectively. The positive predictive value was 0.406, while the negative predictive value was 0.749. A summary of performance metrics is provided in Table 3.

Table 3: 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

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 indicate a substantial clinical burden of sepsis and meaningful associations with mortality outcomes.

4.1 Sepsis Prevalence and Epidemiology

The study identified sepsis in 13.59% of ICU admissions (7,261 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. Previous ICU-based epidemiological research has reported sepsis occurrence rates in the range of approximately 13% to 40% among critically ill patients, reflecting heterogeneity in populations and case definitions [12]. 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. A systematic review of ICU-treated sepsis studies has shown that prevalence measures vary widely but can include proportions comparable to our findings [13].

4.2 Impact of Sepsis on In-Hospital Mortality

A striking finding was the substantial difference in in-hospital mortality between sepsis and non-sepsis patients (32.57% vs. 9.05%). This 3.6-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 [12]. The absolute risk difference of approximately 23 percentage points indicates that sepsis contributes substantially to in-hospital mortality in ICU patients.

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 60.3%, capturing a substantial proportion of patients who will die in-hospital. However, the specificity of 57.3% indicates moderate ability to cor-

rectly 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 that the model has limited clinical utility for individual risk stratification but confirms that demographic factors alone are insufficient for accurate mortality prediction.

4.4 Interpretation of model performance

The modest discriminative performance of the logistic regression model (AUC of 0.622) highlights a fundamental limitation of the approach: 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 [14].

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 [3]. 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.

Clinical interpretation:

- 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. Effective risk assessment requires integration of clinical parameters beyond demographic factors for mortality risk stratification in sepsis patients.

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, with sepsis potentially underdiagnosed or overdiagnosed in some

cases. Second, the model includes only demographic and administrative variables; clinical measurements such as vital signs, laboratory values, and organ dysfunction markers were not included and could improve discriminatory performance. Third, MIMIC-III represents a single tertiary academic medical center in the United States, limiting generalizability to other healthcare settings or populations. Fourth, the observational nature of the study precludes causal inference regarding risk factors and outcomes. Additionally, sepsis classification was performed at the admission level and may not fully capture temporal changes in sepsis status during the ICU stay. Finally, age values were used as provided in the MIMIC-III database, where ages above 89 years are intentionally obfuscated for privacy reasons. This may have introduced imprecision in age-related analyses, particularly among the oldest patients.

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, including other ICU populations or more recent data, would strengthen confidence in model generalizability. Investigation of temporal trends in sepsis epidemiology and outcomes could provide insight into changes in awareness and management strategies over time. Finally, comparative effectiveness research evaluating specific clinical interventions in sepsis management could help inform practice guidelines and improve patient 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 7,261 sepsis cases identified in a cohort of ICU admissions.

Sepsis prevalence of 13.59% among ICU admissions and a 3.6-fold elevation in mortality among sepsis patients (32.57% vs. 9.04%) 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 but demonstrates that clinical variables are necessary for meaningful risk stratification.

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 indicates that effective risk stratification requires incorporation of clinical and physiologic variables including vital signs, lactate levels, and SOFA scores.

Future research incorporating physiologic and laboratory

variables, validation in external cohorts, and evaluation of clinical interventions will be necessary to develop clinically meaningful prediction tools and improve outcomes for critically ill patients with sepsis.

References

- [1] Andrew Rhodes, Laura E. Evans, Waleed Alhazzani, et al. Surviving sepsis campaign: International guidelines for management of sepsis and septic shock: 2016. *Intensive Care Medicine*, 43(3):304–377, 2017. doi: 10.1007/s00134-017-4683-6.
- [2] Mitchell M. Levy, Mitchell P. Fink, John C. Marshall, Edward Abraham, Derek Angus, Deborah Cook, Jonathan Cohen, Steven M. Opal, Jean-Louis Vincent, and Graham Ramsay. 2001 sccm/esicm/accp/ats/sis international sepsis definitions conference. *Critical Care medicine*, 31(4):1250–1256, 2003. doi: 10.1097/01.ccm.0000050454.01978.3b.
- [3] Mervyn Singer, Clifford S Deutschman, Christopher W Seymour, et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA*, 315(8):801–810, 2016. doi: 10.1001/jama.2016.0287.
- [4] Shannon A. Novosad, Mathew R.P. Sapiano, Cheri Grigg, et al. Vital signs: Epidemiology of sepsis: Prevalence of health care factors and opportunities for prevention. *MMWR. Morbidity and Mortality Weekly Report*, 65(33):864–869, 2016. doi: 10.15585/mmwr.mm6533e1.
- [5] C. Bouza, T. Lopez-Cuadrado, and J. M. Amate-Blanco. Use of explicit icd9-cm codes to identify adult severe sepsis: impacts on epidemiological estimates. *Critical Care*, 20(1), 2016. doi: 10.1186/s13054-016-1497-9.
- [6] J. L. Vincent, R. Moreno, J. Takala, S. Willatts, A. De Mendonça, H. Bruining, C. K. Reinhart, P. M. Suter, and L. G. Thijs. The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure: On behalf of the working group on sepsis-related problems of the european society of intensive care medicine. *Intensive Care Medicine*, 22(7):707–710, 1996. doi: 10.1007/bf01709751.
- [7] Christopher W. Seymour, Vincent X. Liu, et al. Assessment of clinical criteria for sepsis: For the third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA*, 315 (8):762, 2016. doi: 10.1001/jama.2016.0288.
- [8] Alistair E W Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3(1):160035, 2016. doi: 10.1038/sdata.2016.35.
- [9] Derek C. Angus, Walter T. Linde-Zwirble, Jeffrey Lidicker, Gilles Clermont, Joseph Carcillo, and Michael R. Pinsky. Epidemiology of severe sepsis in the united states: Analysis of incidence, outcome, and associated costs of care. *Critical Care Medicine*, 29(7):1303–1310, 2001. doi: 10.1097/00003246-200107000-00002.
- [10] Rachel J Jolley, Keri Jo Sawka, Dean W Yergens, Hude Quan, Nathalie Jetté, and Christopher J Doig. Validity of administrative data in recording sepsis: a systematic review. *Critical Care*, 19(1), 2015. doi: 10.1186/s13054-015-0847-3.
- [11] National Center for Health Statistics. International classification of diseases, ninth revision, clinical modification (icd-9-cm). https://archive.cdc.gov/www_cdc_gov/nchs/icd/icd9cm.htm, 2021.
- [12] Yasser Sakr, Ulrich Jaschinski, Xavier Wittebole, Tamas Sza-

- kmany, Jeffrey Lipman, Silvio A Namendys Silva, Ignacio Martin-Loeches, Marc Leone, Mary-Nicoleta Lupu, Jean-Louis Vincent, and ICON Investigators. Sepsis in intensive care unit patients: Worldwide data from the intensive care over nations audit. *Open Forum Infectious Diseases*, 5(12):ofy313, 2018. doi: 10.1093/ofid/ofy313.
- [13] Robby Markwart, Hiroki Saito, Thomas Harder, Sara Tomczyk, Alessandro Cassini, Carolin Fleischmann-Struzek, Felix Reichert, Tim Eckmanns, and Benedetta Allegranzi. Epidemiology and burden of sepsis acquired in hospitals and intensive care units: a systematic review and meta-analysis. *Intensive Care Medicine*, 46(8):1536–1551, 2020. doi: 10.1007/s00134-020-06106-2.
- [14] Christopher W. Seymour, Vincent X. Liu, Theodore J. Iwashyna, Frank M. Brunkhorst, Thomas D. Rea, André Scherag, Gordon Rubenfeld, Jeremy M. Kahn, Manu Shankar-Hari, Mervyn Singer, Clifford S. Deutschman, Gabriel J. Escobar, and Derek C. Angus. Assessment of clinical criteria for sepsis: For the third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA*, 315(8):762, 2016. doi: 10.1001/jama.2016.0288.

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. This is the complete R code in Github: <https://github.com/Blanqui04/EHR-A3-Group-B.git>

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)

```

A.5 Variance Inflation Factor (VIF) Calculation

VIF was calculated using manual auxiliary linear regressions, as the `car` package was unavailable in the computational environment. The VIF for each predictor is calculated as $VIF = 1/(1 - R^2)$, where R^2 is obtained from regressing each predictor on all other predictors.

```
# Manual VIF calculation function
vif_manual <- function(model_object) {
  X <- model.matrix(model_object)
  X <- X[, -1] # Remove intercept
  col_names <- colnames(X)
  vif_values <- numeric(ncol(X))

  for (i in 1:ncol(X)) {
    # Escape column names with backticks
    response_var <- paste0(``, col_names[i], ``)
    predictor_vars <- paste0(``, col_names[-i], ``,
                             collapse=" + ")

    # Fit auxiliary regression
    aux_formula <- as.formula(paste(response_var, "~",
                                      predictor_vars))
    aux_model <- lm(aux_formula, data=as.data.frame(X))

    # Calculate VIF
    r_squared <- summary(aux_model)$r.squared
    vif_values[i] <- 1 / (1 - r_squared)
  }

  return(data.frame(Predictor = col_names, VIF = vif_values))
}

# Calculate VIF for logistic model
vif_results <- vif_manual(logistic)

# Aggregate by main predictor
vif_results_agg <- vif_results %>%
  mutate(
    main_predictor = case_when(
      Predictor == "age" ~ "age",
      Predictor == "genderM" ~ "gender",
      str_detect(Predictor, "^ethnicity") ~ "ethnicity",
      str_detect(Predictor, "^insurance") ~ "insurance",
      str_detect(Predictor, "^icd9_code") ~ "icd9_code",
      TRUE ~ Predictor
    )
  ) %>%
  group_by(main_predictor) %>%
  summarise(VIF_avg = mean(VIF), .groups = "drop") %>%
  rename(Predictor = main_predictor, VIF = VIF_avg)

# Display results
print(vif_results_agg)
```

A.6 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