ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug1493

Numerical Computation and Epidemiological Risk Analysis of Sepsis Mortality among Adults with ICD-Coded Sepsis in MIMIC-IV

Innocent Obinna Okpara[1]; Roseline Toyin Abah[2]; Mary O. Durojaye[3]

[1,2,3]Department of Mathematics, Faculty of Physical Science, University of Abuja, Abuja, Nigeria

Publication Date: 2026/09/12

Abstract:

- ***Background:***

Transparent cohort characterization is a prerequisite for reliable sepsis prediction research, yet case definitions, admission units, outcome windows, and leakage controls vary considerably across studies.

- ***Methods:***

We conducted a retrospective descriptive analysis of adult hospital admissions in MIMIC-IV version 3.1. Admissions were identified using an explicit ICD-coded sepsis phenotype. The primary modelling population was defined as the chronologically first eligible coded-sepsis admission per patient; all eligible admissions were retained for descriptive analyses. Outcomes included in-hospital, 30-day, and 90-day mortality. We summarized demographic characteristics, ICU involvement, coded severity, first ICU care unit, and the availability of ten candidate physiological variables. Unadjusted mortality differences and risk ratios were calculated from the reported stratum counts.

- ***Results***:

Among 546,028 hospital admissions, 22,507 adult coded-sepsis admissions were identified in 18,041 patients. There were 4,395 in-hospital deaths (19.5%). The first-admission cohort contained 18,041 admissions and 3,716 deaths (20.6%). Thirty- and 90-day mortality were 22.5% and 31.4%, respectively. Median age was 68 years (IQR 56-79), 54.7% of admissions were male, and 63.0% involved ICU care. ICU-associated mortality was 28.2%, compared with 4.8% without ICU involvement, an unadjusted risk ratio of 5.84. Mortality increased across coded categories from 6.5% without a severe-sepsis or shock modifier, to 15.1% for severe sepsis without shock, and 35.2% for septic shock; septic shock had an unadjusted risk ratio of 5.39 relative to the no-modifier group. Vital-sign missingness ranged from 2.0% to 3.2%; laboratory missingness ranged from 4.1% to 16.2%.

- ***Conclusions:***

The cohort shows substantial heterogeneity in severity, care intensity, and mortality. These results support a patient-disjoint first-admission design for subsequent prediction studies, but they are descriptive and do not establish causality, predictive accuracy, or model fairness.

Keywords: *Sepsis; Mortality; MIMIC-IV; Retrospective Cohort; Intensive Care; Missing Data; Electronic Health Records.*

How to Cite: Innocent Obinna Okpara; Roseline Toyin Abah; Mary O. Durojaye (2026) Numerical Computation and Epidemiological Risk Analysis of Sepsis Mortality among Adults with ICD-Coded Sepsis in MIMIC-IV. *International Journal of Innovative Science and Research Technology*, 11(8), 3374-3383. https://doi.org/10.38124/ijisrt/26aug1493

I. INTRODUCTION

Sepsis is life-threatening organ dysfunction caused by a dysregulated host response to infection and remains a major challenge in acute and critical care [1, 2]. Its clinical course is heterogeneous: deterioration may be rapid, relevant measurements are irregularly collected, and mortality risk depends on the interaction of baseline vulnerability, physiologic disturbance, organ dysfunction, treatment, and care setting. Consequently, electronic-health-record studies have increasingly examined machine-learning approaches for early detection and outcome prediction [3-6].

The validity of such models depends first on the population to which they are applied. Reviews of sepsis prediction research identify substantial variation in phenotype definitions, prediction origins, observation windows, handling of repeated admissions, missing-data

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug1493

procedures, and validation strategies [3, 4, 7]. These differences can change apparent event prevalence and case mix before any algorithm is fitted. Diagnosis-code phenotypes are especially useful for reproducible retrospective cohort construction, but they should not be conflated with the consensus clinical Sepsis-3 definition or used as if they provide time-resolved physiological state labels [1].

MIMIC-IV provides deidentified hospital and ICU information suitable for transparent methodological research [8, 9]. Its linked admission, diagnosis, patient, and ICU tables allow hospital-wide cohorts to include both ICU and non-ICU episodes, avoiding the assumption that every sepsis admission is an independent ICU stay. This distinction matters because repeated admissions from one patient create dependence, ICU exposure is nested within the hospitalization, and variables such as discharge time and length of stay can introduce future-information leakage if used for early prediction.

This study describes an adult ICD-coded sepsis cohort in MIMIC-IV, quantifies mortality across clinically relevant strata, and audits the availability of candidate physiological measures. The objectives were to: (1) derive a reproducible hospital-wide cohort and a conservative first-admission population; (2) describe demographic and care-setting heterogeneity; (3) quantify mortality by coded severity and ICU setting; and (4) define the evidence and leakage boundaries for later longitudinal prediction research. In keeping with contemporary reporting guidance, this article separates descriptive patient-derived findings from predictive claims that require independent validation [10, 11].

II. METHODS

➢ *Design and Data Source*

This was a retrospective cohort description using MIMIC-IV version 3.1, a deidentified electronic health record database containing hospital and critical-care data [8, 9]. The analytical unit was a hospital admission. ICU stays were treated as care episodes nested within an admission rather than as independent patients or independent hospitalizations.

➢ *Cohort Definition*

Eligible records were adult admissions with at least one explicit sepsis-related ICD-9 or ICD-10 diagnosis. The phenotype included sepsis, septicemia, severe sepsis, and septic shock terminology and excluded neonatal sepsis. It identified 24 ICD-9 and 49 ICD-10 codes in the source analysis. Diagnosis codes were used to identify the retrospective cohort, not as candidate predictors.

Two related populations were defined. The all-admission descriptive cohort retained every eligible hospitalization. The recommended primary modelling cohort retained only the chronologically first eligible admission for each patient. This design prevents one patient from contributing observations to both development and evaluation samples in subsequent studies. Any secondary analysis retaining repeated admissions should partition and resample by patient.

Let $C_{ij} = 1$ indicate that admission j of patient i satisfied the eligibility criteria. The alladmission coded-sepsis cohort was

$$\mathcal{C}_{all} = \{(i,j): C_{ij} = 1\}, N_{adm} = \sum_i \sum_j C_{ij} = 22{,}507. \quad (1)$$

The number of unique eligible patients was

$$N_{\text{patient}} = \sum_i \mathbb{1}\left(\sum_j C_{ij} \geq 1\right) = 18{,}041. \quad (2)$$

For each eligible patient, the first-admission index was

$$j_i^* = \underset{j:C_{ij}=1}{\arg\min} t_{ij}^{adm} \quad (3)$$

And the patient-unique primary cohort was

$$\mathcal{C}_{\text{first}} = \left\{(i, j_i^*): \sum_j C_{ij} \geq 1\right\}, |\mathcal{C}_{\text{first}}| = 18{,}041. \quad (4)$$

The number of eligible admissions beyond one first admission per patient was therefore

$$N_{\text{repeat}} = N_{\text{adm}} - N_{\text{patient}} = 22{,}507 - 18{,}041 = 4{,}466, \quad (5)$$

While the mean number of eligible admissions represented per patient was

$$\frac{N_{\text{adm}}}{N_{\text{patient}}} = \frac{22{,}507}{18{,}041} = 1.248. \quad (6)$$

The conservative systemic-sepsis sensitivity cohort retained 22,304 admissions, corresponding to

$$\frac{22{,}304}{22{,}507} \times 100\% = 99.10\% \quad (7)$$

Of the inclusive coded-sepsis cohort. Sepsis appeared as the principal diagnosis in

$$\frac{14{,}617}{22{,}507} \times 100\% = 64.94\% \quad (8)$$

Of admissions and as a secondary diagnosis in

$$\frac{7{,}890}{22{,}507} \times 100\% = 35.06\%. \quad (9)$$

These calculations define the relationship between the inclusive descriptive cohort, the conservative sensitivity cohort, and the patient-unique primary cohort.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug1493

➢ *Outcomes and Covariates*

The primary descriptive outcome was in-hospital death during the indexed admission. Secondary outcomes were death within 30 and 90 days of admission. Characteristics included age, sex, grouped race or ethnicity as recorded in the database, hospital length of stay, diagnosis position, ICU involvement, and first ICU care unit. Race and ethnicity categories were treated as social and administrative data fields rather than biological attributes [12, 13].

Coded severity was classified as: no explicit severe-sepsis or shock modifier; severe sepsis without shock; or septic shock. Because severity modifiers may be assigned after clinical deterioration, they are suitable for retrospective stratification but require verified timestamps before consideration in an early prediction model.

Ten candidate physiological variables were summarized: heart rate, respiratory rate, mean arterial pressure, temperature, oxygen saturation, lactate, creatinine, white-cell count, platelets, and total bilirubin. Only aggregate summaries were available for the current analysis. They describe availability and marginal distributions, not within-patient trajectories or survivor-nonsurvivor differences over time.

For admission j of patient i, in-hospital mortality was represented by

$$Y_{ij}^{H} = \mathbb{1}\left(t_i^{\text{death}} \geq t_{ij}^{adm}\right) \mathbb{1}\left(t_i^{\text{death}} \leq t_{ij}^{\text{disch}}\right), \quad (10)$$

Where t_{ij}^{adm} and t_{ij}^{disch} denote admission and discharge times. The secondary fixed-horizon outcome was

$$Y_{ij}^{(h)} = \mathbb{1}\left(0 \leq t_i^{\text{death}} - t_{ij}^{\text{adm}} \leq h\right), h \in \{30,90\} \text{ days.} \quad (11)$$

The coded-severity variable was defined as

$$G_{ij} = \begin{cases} 0, & \text{no explicit severe-sepsis or shock modifier,} \\ 1, & \text{severe sepsis without septic shock,} \\ 2, & \text{septic shock.} \end{cases} \quad (12)$$

The ten-dimensional physiological vector associated with the admission was

$$\mathbf{x}_{ij} = (HR, RR, MAP, \text{Temp}, SpO_2, \text{Lactate}, \text{Creatinine}, WBC, \text{Platelets}, \text{Bilirubin})^{\top}. \quad (13)$$

For a later longitudinal analysis, its standardized component would be calculated using trainingsample quantities only:

$$z_{ijk}(t) = \frac{x_{ijk}(t) - \bar{x}_{k,train}}{s_{k,\text{train}}}. \quad (14)$$

➢ *Leakage and Data-Quality Controls*

Hospital discharge time, discharge location, death time, total length of stay, and total ICU duration were designated descriptive outcomes and excluded from admission-time prediction. Twenty admissions had discharge times preceding admission times; their derived durations were treated as missing. Absence of ICU fields among admissions without a linked ICU stay was classified as structural non-applicability rather than ordinary missingness. A future longitudinal analysis should preserve event and store times, item identifiers, original and canonical units, duplicate-resolution rules, and the number of measurements excluded after a fixed prediction landmark.

Let $t_{ij}^{*} = t_{ij}^{adm} + 12$ hours denote the proposed prediction landmark. A measurement x_{ijkm} recorded at event time t_{ijkm} is eligible only when

$$E_{ijkm} = \mathbb{1}\left(t_{ij}^{adm} \leq t_{ijkm} \leq t_{ij}^{*}\right) = 1. \quad (15)$$

The leakage-free history available at the landmark is therefore

$$\mathcal{H}_{ij}(t_{ij}^{*}) = \{x_{ijkm}: t_{ijkm} \leq t_{ij}^{*}, E_{ijkm} = 1\}, \quad (16)$$

Whereas the prohibited future-information set is

$$\mathcal{F}_{ij}^{+} = \{t_{ij}^{\text{disch}}, \text{discharge location}, t_i^{\text{death}}, \text{total LOS}, \text{total ICU duration}, x_{ijkm}: t_{ijkm} > t_{ij}^{*}\}. \quad (17)$$

The valid hospital length of stay was calculated as

$$LOS_{ij} = \begin{cases} t_{ij}^{disch} - t_{ij}^{adm}, & t_{ij}^{disch} \geq t_{ij}^{adm}, \\ \text{NA}, & t_{ij}^{disch} < t_{ij}^{adm}. \end{cases} \quad (18)$$

Thus, the 20 admissions with negative calculated duration contributed missing rather than negative length-of-stay values. Finally, ICU participation was represented by

$$U_{ij} = \mathbb{1}\left(\text{at least one ICU stay is linked to admission } (i,j)\right). \quad (19)$$

When $U_{ij} = 0$, ICU-specific fields are structurally undefined rather than imputed as ordinary missing observations.

➢ *Statistical Analysis*

Continuous variables were summarized using medians and interquartile ranges (IQRs) or means and standard deviations (SDs), as appropriate. Categorical variables were summarized with counts and column percentages. To make the analysis of mortality heterogeneity explicit, selected unadjusted contrasts were calculated from the reported counts. For exposed or higher-risk stratum 1 and reference stratum 0, mortality risk, absolute risk difference, and risk ratio were calculated as

$$\hat{p}_g = \frac{d_g}{n_g}, RD = \hat{p}_1 - \hat{p}_0, RR = \frac{\hat{p}_1}{\hat{p}_0},$$

Where d_g and n_g denote deaths and admissions in stratum g. These admission-level contrasts were used to describe the magnitude and direction of observed differences. Confidence intervals and hypothesis tests were not calculated

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug1493

from the aggregate table because repeated admissions require patient-clustered inference. No multivariable adjustment, causal modelling, predictive modelling, or hypothesis-driven subgroup fairness analysis was performed.

➢ *Mathematical Formulation of Cohort and Outcome Summaries*

The cohort construction can be written as an indicator rule. For admission j belonging to patient i, let A_{ij} denote age at admission and let D_{ij} be the set of recorded ICD diagnosis codes. Eligibility was

$$C_{ij} = \mathbb{1}(A_{ij} \geq 18)\,\mathbb{1}(D_{ij} \cap \mathcal{D}_{\text{sepsis}} \neq \emptyset)\,\mathbb{1}(D_{ij} \cap \mathcal{D}_{\text{neonatal}} = \emptyset), \quad (20)$$

Where $\mathcal{D}_{\text{sepsis}}$ is the prespecified sepsis-code dictionary and $\mathcal{D}_{\text{neonatal}}$ contains excluded neonatalsepsis codes. The primary admission for patient i was the earliest eligible admission,

$$j_i^* = \arg\min_{j:C_{ij}=1} t_{ij}^{adm} \quad (21)$$

Where t_{ij}^{adm} is hospital-admission time. This operation reduces the primary analytical population to one admission per patient and prevents repeated admissions from being treated as independent individuals.

For outcome horizon h, the binary mortality indicator was

$$Y_{ij}^{(h)} = \mathbb{1}\left(0 \leq t_i^{\text{death}} - t_{ij}^{\text{adm}} \leq h\right), \quad (22)$$

With in-hospital death additionally requiring death before discharge from the indexed admission. For a stratum g containing n_g admissions, the estimated mortality risk was

$$\hat{p}_g^{(h)} = \frac{1}{n_g} \sum_{(i,j)\in g} Y_{ij}^{(h)} = \frac{d_g^{(h)}}{n_g}. \quad (23)$$

For physiological variable k, aggregate missingness was calculated as

$$M_k = 1 - \frac{n_{k,\text{obs}}}{N}, \quad (24)$$

Where $n_{k,\text{obs}}$ is the number of admissions with an observed value and $N = 22{,}507$ is the descriptive cohort size.

The source research also embeds mortality within an absorbing probability-flow model. If $I(t)$, $D(t)$, $O(t)$, and $K(t)$ represent infection, dysregulated response, organ dysfunction, and shock probabilities, cumulative mortality obeys

$$\frac{dC_M(t)}{dt} = \mu_I I(t) + \mu_D D(t) + \mu_O O(t) + \mu_K K(t) \geq 0, \quad (25)$$

And the augmented probability mass satisfies

$$S + I + D + O + K + R + Q + C_M = 1. \quad (26)$$

Equations (25)-(26) imply that cumulative discharge and mortality cannot decrease and that any later numerical implementation must preserve nonnegative probability mass. The aggregate data in this paper do not estimate the transition or mortality-rate parameters μ_I, μ_D, μ_O, and μ_K; the equations are included to show how the descriptive outcome analysis connects to the wider mathematical framework.

III. RESULTS

➢ *Cohort Derivation*

Of 546,028 hospital admissions, 545,497 had at least one diagnosis record. The explicit phenotype identified 22,507 adult coded-sepsis admissions (4.1% of all admissions) among 18,041 unique patients (Table 1). Sepsis was the principal diagnosis in 14,617 admissions (64.9%) and a secondary diagnosis in 7,890 (35.1%). The maximum number of eligible admissions contributed by one patient was 19, supporting use of one first admission per patient for primary modelling.

Table 1 Derivation of the Adult Coded-Sepsis Cohort.

Population	Number
All hospital admissions	546,028
Admissions with at least one diagnosis record	545,497
Adult admissions with an explicit sepsis-related ICD code	22,507
Conservative systemic-sepsis-code sensitivity cohort	22,304
Unique patients with at least one coded-sepsis admission	18,041
First coded-sepsis admission per patient	18,041
Sepsis admissions involving ICU care	14,188
Individual ICU stays linked to sepsis admissions	17,240

➢ *Worked Calculations from the Observed Cohort*

The overall in-hospital mortality calculation was

$$\hat{p}_{\text{hospital}} = \frac{4{,}395}{22{,}507} = 0.1953 \approx 19.5\%. \quad (27)$$

For the recommended first-admission cohort, the corresponding estimate was

$$\hat{p}_{\text{first}} = \frac{3{,}716}{18{,}041} = 0.2060 \approx 20.6\%. \quad (28)$$

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug1493

The difference between these estimates is $0.2060 - 0.1953 = 0.0107$, or 1.07 percentage points. This difference reflects the change in analytical unit and case mix after retaining one admission per patient; it is not evidence that selecting the first admission changes an individual's mortality.

The ICU contrast was obtained directly from the stratum counts:

$$\hat{p}_{ICU} = \frac{3,994}{14,188} = 0.2815, \quad \hat{p}_{\text{nonICU}} = \frac{401}{8,319} = 0.0482,$$
$$RD_{ICU} = 0.2815 - 0.0482 = 0.2333, \quad RR_{ICU} = \frac{0.2815}{0.0482} = 5.84. \tag{29}$$

Thus, ICU-associated mortality was 23.33 percentage points higher and 5.84 times the non-ICU mortality risk at the unadjusted admission level.

For coded severity, the observed risks were

$$\hat{p}_0 = \frac{628}{9,619} = 0.0653, \hat{p}_{\text{severe}} = \frac{576}{3,816} = 0.1509, \hat{p}_{\text{shock}} = \frac{3,191}{9,072} = 0.3517. \tag{30}$$

Consequently,

$$\hat{p}_0 < \hat{p}_{\text{severe}} < \hat{p}_{\text{shock}}, RR_{\text{shock :0}} = \frac{0.3517}{0.0653} = 5.39. \tag{31}$$

This ordered gradient is qualitatively consistent with a progression framework containing distinct organ-dysfunction and shock states, although diagnosis groups are not timed transitions and cannot estimate the differential-equation parameters.

The age contrast was similarly calculated as

$$RR_{\geq 80:18-39} = \frac{1,366/5,290}{159/1,783} = \frac{0.2582}{0.0892} = 2.90, \tag{32}$$

With an absolute difference of $0.2582 - 0.0892 = 0.1690$, or 16.90 percentage points. Finally, the missingness calculation for lactate illustrates Equation (24):

$$M_{\text{lactate}} = 1 - \frac{19,774}{22,507} = 0.1214 \approx 12.1\%. \tag{33}$$

For bilirubin, $1 - 18,872/22,507 = 0.1615 \approx 16.2\%$, the highest missingness among the ten candidate physiological variables.

➢ *Patient Characteristics and in-Hospital Outcome*

There were 4,395 in-hospital deaths among 22,507 admissions (19.5%). In the patient-unique first-admission cohort, 3,716 of 18,041 patients died in hospital (20.6%). Median age was 68 years (IQR 56-79), 12,309 admissions (54.7%) were male, and median valid hospital stay was 8.3 days (IQR 4.7-16.2). Patients who died were older and substantially more likely to have ICU involvement (Table 2). These unadjusted patterns do not establish causal effects.

Table 2 Selected Characteristics by in-Hospital Outcome.

Characteristic	Overall	Survived	Died
Admissions, n	22,507	18,112	4,395
Age, median (IQR), years	68 (56-79)	67 (55-78)	72 (61-82)
Hospital stay, median (IQR), days	8.3 (4.7-16.2)	8.3 (4.9-16.0)	8.1 (2.8-17.3)
Female, n (column %)	10,198 (45.3)	8,252 (45.6)	1,946 (44.3)
Male, n (column %)	12,309 (54.7)	9,860 (54.4)	2,449 (55.7)
Age 18-39 years, n (column %)	1,783 (7.9)	1,624 (9.0)	159 (3.6)
Age 40-64 years, n (column %)	7,658 (34.0)	6,386 (35.3)	1,272 (28.9)
Age 65-79 years, n (column %)	7,776 (34.5)	6,178 (34.1)	1,598 (36.4)
Age ≥ 80 years, n (column %)	5,290 (23.5)	3,924 (21.7)	1,366 (31.1)
Unknown/not stated race or ethnicity, n (column %)	1,664 (7.4)	993 (5.5)	671 (15.3)
ICU involvement, n (column %)	14,188 (63.0)	10,194 (56.3)	3,994 (90.9)

Thirty-day mortality was 22.5% (5,072 deaths) and 90-day mortality was 31.4% (7,060 deaths). Mortality was 16.4% when sepsis was the principal diagnosis and 25.4% when it was secondary. Diagnosis sequence may reflect the complexity and timing of hospitalization as well as documentation practices, so this contrast should not be interpreted causally.

Age-specific mortality increased across the four prespecified categories. The calculations were

$$\hat{p}_{18-39} = \frac{159}{1,783} = 0.0892, \quad \hat{p}_{40-64} = \frac{1,272}{7,658} = 0.1661,$$
$$\hat{p}_{65-79} = \frac{1,598}{7,776} = 0.2055, \quad \hat{p}_{\geq 80} = \frac{1,366}{5,290} = 0.2582. \tag{34}$$

Thus,

$$0.0892 < 0.1661 < 0.2055 < 0.2582, \tag{35}$$

Showing a monotone unadjusted age gradient in the observed cohort. Sex-specific mortality was

$$\hat{p}_{\text{male}} = \frac{2,449}{12,309} = 0.1990, \hat{p}_{\text{female}} = \frac{1,946}{10,198} = 0.1908. \tag{36}$$

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug1493

The absolute difference was $0.1990 - 0.1908 = 0.0081$, or 0.81 percentage points, and the unadjusted ratio was $0.1990/0.1908 = 1.04$. These marginal calculations do not adjust for age, severity, ICU involvement, comorbidity, or repeated admissions.

➢ *Comparative Descriptive Analysis of Mortality*

The derived contrasts quantify the magnitude of mortality heterogeneity already visible in the cohort counts (Table 3). ICU-associated admissions had a mortality risk 23.33 percentage points higher than non-ICU admissions and an unadjusted risk ratio of 5.84. Relative to admissions without an explicit severe-sepsis or shock modifier, the mortality risk ratio was 2.31 for severe sepsis without shock and 5.39 for septic shock. Mortality among patients aged 80 years or older was 25.82%, compared with 8.92% among those aged 18-39 years, giving an absolute difference of 16.90 percentage points and an unadjusted risk ratio of 2.90. The male-female difference was comparatively small: 19.90% versus 19.08%, respectively, corresponding to a risk ratio of 1.04 .

Table 3 Selected Unadjusted Descriptive Contrasts in in-Hospital Mortality.

Contrast (stratum 1 vs reference)	Risk 1 (%)	Risk 0 (%)	Difference (percentage points)	Risk ratio
ICU involvement vs no ICU involvement	28.15	4.82	23.33	5.84
Severe sepsis without shock vs no severe/shock modifier	15.09	6.53	8.57	2.31
Septic shock vs no severe/shock modifier	35.17	6.53	28.65	5.39
Age $\geq$ 80 years vs age 18-39 years	25.82	8.92	16.90	2.90
Male vs female	19.90	19.08	0.81	1.04

For each contrast, the proportional increase relative to the reference risk was also calculated as

$$RI = \frac{\hat{p}_1 - \hat{p}_0}{\hat{p}_0} \times 100\% = (RR - 1) \times 100\%. \quad (37)$$

Using the unrounded stratum risks, the ICU contrast gives

$$RI_{ICU} = \frac{0.2815 - 0.0482}{0.0482} \times 100\% \approx 484.1\%, \quad (38)$$

While severe sepsis and septic shock give approximately 131.2% and 438.9% increases relative to the no-modifier group. The oldest-versus-youngest age contrast gives approximately 189.6%, whereas the male-versus-female contrast gives approximately 4.3%. These percentages restate the unadjusted risk ratios on a relative-change scale and are not estimates of causal effects.

➢ *Coded Severity and ICU Involvement*

Mortality rose monotonically with coded severity (Table 4). Among admissions without an explicit severe-sepsis or shock modifier, mortality was 6.5%. It was 15.1% for severe sepsis without shock and 35.2% for septic shock. The gradient supports retention of clinically distinct severity strata in descriptive and sensitivity analyses, but diagnosis codes do not provide timed transitions between states.

Table 4 In-Hospital Outcomes by ICD-Coded Sepsis Severity.

Severity	Admissions	Deaths	Mortality (%)	ICU admissions
No explicit severe/shock modifier	9,619	628	6.5	3,493
Severe sepsis without shock	3,816	576	15.1	2,093
Septic shock	9,072	3,191	35.2	8,602
All coded sepsis	22,507	4,395	19.5	14,188

The severity categories were also compared by their shares of all admissions and all in-hospital deaths. Septic shock accounted for

$$\frac{9{,}072}{22{,}507} \times 100\% = 40.31\% \quad (39)$$

Of coded-sepsis admissions but

$$\frac{3{,}191}{4{,}395} \times 100\% = 72.61\% \quad (40)$$

Of in-hospital deaths. The ratio of these shares was $72.61/40.31 = 1.80$, indicating that shock admissions were disproportionately represented among deaths. By comparison, the no-modifier group represented 42.74% of admissions but 14.29% of deaths.

The 14,188 ICU-associated admissions had 3,994 deaths (28.2%), compared with 401 deaths among 8,319 admissions without ICU involvement (4.8%). ICU-associated admissions contained 17,240 individual ICU stays. Median total ICU time per hospital admission was 3.43 days (IQR 1.66-8.64), but ICU duration is a downstream care outcome and is unsuitable as an admission-time predictor. Mortality also differed across the first recorded ICU care units (Table 5).

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug1493

Table 5 Mortality by First Care Unit Among ICU-Associated Coded-Sepsis Admissions.

First care unit	Admissions	Deaths	Mortality (%)
Medical ICU	5,149	1,518	29.5
Medical/Surgical ICU	4,341	1,078	24.8
Surgical ICU	1,463	380	26.0
Trauma Surgical ICU	1,228	330	26.9
Coronary Care Unit	1,036	406	39.2
Cardiac Vascular ICU	537	154	28.7
Neuro Surgical ICU	155	58	37.4
Other units	279	70	25.1
All ICU coded-sepsis admissions	14,188	3,994	28.2

For example, mortality in the coronary care unit was

$$\hat{p}_{CCU} = \frac{406}{1{,}036} = 0.3919, \quad (41)$$

Whereas mortality in the medical/surgical ICU was

$$\hat{p}_{MSICU} = \frac{1{,}078}{4{,}341} = 0.2483. \quad (42)$$

The unadjusted absolute difference was $0.3919 - 0.2483 = 0.1436$, or 14.36 percentage points, and the ratio was $0.3919/0.2483 = 1.58$. This contrast describes different case mixes and must not be interpreted as a comparative effect of the care units.

➢ *Physiological Variables and Missingness*

Five vital signs and five laboratory variables formed the candidate state vector for later longitudinal work (Table 6). Vital-sign missingness was low at the aggregate admission level, ranging from 2.0% for heart rate to 3.2% for temperature. Laboratory missingness was higher, particularly for lactate (12.1%) and bilirubin (16.2%). These values do not reveal whether measurements occurred before a future prediction landmark or how observation frequency differed by outcome.

Table 6 Aggregate Availability and Distribution of Candidate Physiological Variables.

Variable	Unit	Observed n	Missing (%)	Mean (SD)	Median (IQR)
Heart rate	bpm	22,060	2.0	102.95 (19.92)	102.96 (89.53-116.44)
Respiratory rate	breaths/min	21,964	2.4	23.05 (5.96)	23.03 (19.01-27.10)
Mean arterial pressure	mmHg	21,839	3.0	70.09 (11.88)	70.01 (62.08-78.10)
Temperature	˚C	21,778	3.2	38.00 (0.90)	37.99 (37.39-38.61)
Oxygen saturation	%	22,022	2.2	93.91 (3.75)	94.04 (91.32-96.71)
Lactate	mmol/L	19,774	12.1	2.43 (1.15)	2.19 (1.62-2.98)
Creatinine	mg/dL	21,151	6.0	1.53 (0.81)	1.35 (0.96-1.89)
White-cell count	$\times 10^9$/L	21,586	4.1	13.56 (5.87)	13.52 (9.49-17.59)
Platelets	$\times 10^9$/L	21,555	4.2	185.04 (74.18)	184.78 (133.42-235.05)
Total bilirubin	mg/dL	18,872	16.2	1.62 (0.97)	1.40 (0.96-2.02)

Across the five vital signs, the total number of missing variable-admission cells was

$$447 + 543 + 668 + 729 + 485 = 2{,}872. \quad (43)$$

Because five variables were assessed across 22,507 admissions, the mean vital-sign missingness was

$$\bar{M}_{vital} = \frac{2{,}872}{5(22{,}507)} \times 100\% = 2.55\%. \quad (44)$$

For the five laboratory variables, the corresponding calculation was

$$\bar{M}_{\text{laboratory}} = \frac{2{,}733 + 1{,}356 + 921 + 952 + 3{,}635}{5(22{,}507)} \times 100\% = 8.53\%. \quad (45)$$

Therefore, the average laboratory missingness was $8.53/2.55 = 3.34$ times the average vital-sign missingness.

Relative physiological dispersion was summarized for positive ratio-scale variables using the coefficient of variation,

$$CV_k = \frac{SD_k}{\bar{x}_k} \times 100\%. \quad (46)$$

For heart rate, lactate, platelets, and bilirubin, respectively,

$$CV_{HR} = \frac{19.92}{102.95} = 19.35\%, CV_{\text{lactate}} = \frac{1.15}{2.43} = 47.33\% \quad (47)$$

$$CV_{\text{platelets}} = \frac{74.18}{185.04} = 40.09\%, CV_{\text{bilirubin}} = \frac{0.97}{1.62} = 59.88\% \quad (48)$$

The larger relative dispersion of lactate and bilirubin reinforces the physiological heterogeneity of the cohort, although these aggregate calculations cannot describe within-patient temporal variation.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug1493

Demographic and admission-outcome fields were complete for age, sex, grouped race or ethnicity, admission type, admission location, and in-hospital mortality. Insurance was missing in 187 admissions (0.8%), language in 93 (0.4%), and marital status in 1,284 (5.7%). Missing ICU fields in 8,319 non-ICU admissions reflected the study's hospital-wide design rather than failed measurement.

IV. DISCUSSIONS

➢ *Principal Findings*

This study establishes a hospital-wide coded-sepsis cohort of 22,507 admissions among 18,041 patients. One in five admissions ended in hospital death, and mortality increased further over the 30- and 90-day windows. The cohort was clinically heterogeneous: mortality was markedly higher with ICU involvement and increased more than fivefold from the lowest coded-severity stratum to septic shock. These differences show why a single pooled event rate is insufficient to describe the intended population for prediction research.

The comparative analysis adds scale to these observations. ICU involvement and septic shock were each associated with mortality risks more than five times their respective reference groups, while mortality among the oldest age category was almost three times that of the youngest adult category. By contrast, the unadjusted difference between male and female admissions was less than one percentage point. These contrasts are useful for describing case-mix heterogeneity and planning stratified validation, but they cannot separate baseline severity, treatment selection, comorbidity, measurement intensity, or other confounding mechanisms.

The number of admissions exceeded the number of patients by 4,466, and some patients had many recurrent eligible admissions. A random admission-level split could therefore place records from the same person in both model development and evaluation, allowing stable patient characteristics to inflate apparent generalization. The chronologically first eligible admission per patient is consequently the clearest primary design. All-admission analyses remain informative if partitioning, resampling, and uncertainty estimation are clustered by patient.

The results also distinguish a hospital-based cohort from an ICU-only cohort. More than one third of eligible admissions had no linked ICU stay. Restriction to ICU records would omit this lowermortality group and change both the clinical target population and outcome prevalence. Conversely, ICU duration and first care unit are important descriptive indicators but can encode downstream decisions and future information. This illustrates the broader requirement that cohort-description variables should not automatically become predictors.

➢ *Implications for Sepsis Prediction Studies*

Existing studies have reported promising performance using dynamic vital signs, laboratory values, and machine-learning models [14-17]. Nevertheless, systematic reviews emphasize that retrospective discrimination alone does not establish safe clinical benefit [3, 4, 18]. External evaluation has shown that deployed sepsis scores may perform differently from their original reports [19], while prospective evaluation requires attention to workflow, calibration, alerts, and patient outcomes [11, 20].

The ten candidate physiological variables in this cohort are clinically coherent for subsequent longitudinal modelling. However, aggregate availability cannot answer the most important temporal questions: whether a value occurred before the prediction origin, how long it had been since the last measurement, whether measurement intensity was informative, or whether missingness differed systematically by outcome. A defensible prediction analysis should therefore use timestamped events, fit imputation and scaling only within training data, retain time-since-measurement information where appropriate, and audit every post-landmark exclusion.

Recorded demographic fields also require cautious interpretation. The greater frequency of unknown or unstated race or ethnicity among deaths may reflect documentation and access processes rather than biology. Fairness assessment should not silently remove unknown values or collapse categories solely to obtain favorable metrics. It should report subgroup support, calibration, error rates, uncertainty, and threshold consequences, recognizing that formal metric parity alone is not equivalent to health equity [13, 21-23].

➢ *Strengths and Limitations*

Strengths include a large hospital-wide cohort, explicit separation of admissions from patients and ICU stays, transparent mortality windows, a first-admission primary design, and advance identification of leakage-prone variables. The analysis also distinguishes patient-derived descriptive evidence from model-validation evidence, consistent with TRIPOD+AI and early-stage clinical-AI reporting principles [10, 11].

Several limitations should temper interpretation. First, the phenotype is diagnosis-code based and is not a direct implementation of Sepsis-3. Coding and diagnosis sequencing may vary across time and clinical services. Second, MIMIC-IV is retrospective and single-centre, limiting transportability.

V. CONCLUSIONS

Adult coded-sepsis admissions in MIMIC-IV form a heterogeneous hospital-wide cohort with substantial mortality and strong descriptive gradients by coded severity and ICU involvement. A patient-unique first-admission population, explicit leakage controls, and timestamped physiological extraction are necessary foundations for subsequent prediction research. The current findings characterize the cohort and its data-quality boundaries; they should not be interpreted as evidence of causality, clinical prediction performance, or readiness for bedside use.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug1493

DECLARATIONS

Ethics and data access. MIMIC-IV is a controlled-access, deidentified research database. The lead author received approval from PhysioNet to access MIMIC-IV after completing the required credentialing and accepting the applicable data-use requirements. No patient-level source data are redistributed with this manuscript.

Funding. This research received no specific grant or financial support from any public, commercial, or not-for-profit funding agency.

Data availability. The lead author received approved access to MIMIC-IV through PhysioNet under its credentialing and data-use requirements [9]. The database remains available to qualified researchers who independently satisfy PhysioNet's access requirements.

REFERENCES

[1]. 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.

[2]. Laura Evans, Andrew Rhodes, Waleed Alhazzani, et al. Surviving sepsis campaign: International guidelines for management of sepsis and septic shock 2021. Intensive Care Medicine, 47:1181-1247, 2021. doi: 10.1007/s00134-021-06506-y.

[3]. Michael Moor, Bastian Rieck, Max Horn, Catherine R. Jutzeler, and Karsten Borgwardt. Early prediction of sepsis in the icu using machine learning: A systematic review. Frontiers in Medicine, 8:607952, 2021. doi: 10.3389/fmed.2021.607952.

[4]. Hai-Feng Deng, Ming-Wei Sun, Yan Wang, et al. Evaluating machine learning models for sepsis prediction: A systematic review of methodologies. iScience, 25:103651, 2022. doi: 10.1016/j.isci.2021.103651.

[5]. S. Bomrah, M. Uddin, and U. Upadhyay. A scoping review of machine learning for sepsis prediction: Feature engineering strategies and model performance. Critical Care, 28:180, 2024. doi: 10.1186/s13054-024-04948-6.

[6]. Florin Muşat. Machine learning models in sepsis outcome prediction: A systematic review. Biomedicines, 12:2892, 2024. doi: 10.3390/biomedicines12122892.

[7]. M. Y. Yadgarov, G. Landoni, L. Berikashvili, et al. Early detection of sepsis using machine learning algorithms: A systematic review and network meta-analysis. Frontiers in Medicine, 11:1491358, 2024. doi: 10.3389/fmed.2024.1491358.

[8]. Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, et al. Mimic-iv, a freely accessible electronic health record dataset. Scientific Data, 10:1, 2023. doi: 10.1038/s41597-022-01899-x.

[9]. Alistair Johnson, Lucas Bulgarelli, Tom Pollard, et al. Mimic-iv (version 3.1), 2024.

[10]. Gary S. Collins, Karel G. M. Moons, Paul Dhiman, et al. Tripod+ai statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ, 385:e078378, 2024. doi: 10.1136/bmj-2023-078378.

[11]. Baptiste Vasey, Myura Nagendran, Bruce Campbell, et al. Reporting guideline for the earlystage clinical evaluation of decision support systems driven by artificial intelligence: Decide-ai. Nature Medicine, 28:924-933, 2022. doi: 10.1038/s41591-022-01772-9.

[12]. Kadija Ferryman, Michael Mackintosh, and Marzyeh Ghassemi. Considering biased data as informative artifacts in ai-assisted health care. New England Journal of Medicine, 389(9): 833-838, 2023. doi: 10.1056/NEJMra2214964.

[13]. Richard J. Chen, Judy J. Wang, Drew F. K. Williamson, et al. Algorithmic fairness in artificial intelligence for medicine and healthcare. Nature Biomedical Engineering, 7:719-742, 2023. doi: 10.1038/s41551-023-01056-8.

[14]. Chih-Yang Cheng. Machine learning models for predicting in-hospital mortality in patients with sepsis based on dynamic vital signs. Frontiers in Medicine, 9:964667, 2022. doi: 10.3389/ fmed.2022.964667.

[15]. J. Gao. Prediction of sepsis mortality in icu patients using machine learning methods. BMC Medical Informatics and Decision Making, 24, 2024. doi: 10.1186/s12911-024-02630-z.

[16]. D. Wang. A machine learning model for accurate prediction of sepsis in icu patients. Frontiers in Public Health, 9:754348, 2021. doi: 10.3389/fpubh.2021.754348.

[17]. Z. Zeng. Development and validation of a novel blending machine learning model for hospital mortality prediction in icu patients with sepsis. BioData Mining, 14:40, 2021. doi: 10.1186/ s13040-021-00276-5.

[18]. Luisa Agnello, Matteo Vidali, and Andrea Padoan. Machine learning algorithms in sepsis. Clinica Chimica Acta, 553:117738, 2024. doi: 10.1016/j.cca.2023.117738.

[19]. Andrew Wong, Erkin Otles, John P. Donnelly, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Internal Medicine, 181: 1065-1070, 2021. doi: 10.1001/jamainternmed.2021.2626.

[20]. Roy Adams, Katherine E. Henry, Aditya Sridharan, et al. Prospective, multi-site study of patient outcomes after implementation of the trews machine learning-based early warning system for sepsis. Nature Medicine, 28:1455-1460, 2022. doi: 10.1038/s41591-022-01894-0.

[21]. Alvin Rajkomar, Michaela Hardt, Michael D. Howell, Greg Corrado, and Marshall H. Chin. Ensuring fairness in machine learning to advance health equity. Annals of Internal Medicine, 174:866-872, 2021. doi: 10.7326/M20-6261.

[22]. Y. Huang, J. Guo, W.-H. Chen, et al. A scoping review of fair machine learning techniques when using real-world data. Journal of Biomedical Informatics, 151:104622, 2024. doi: 10.1016/j.jbi.2024.104622.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug1493

[23]. Daiju Ueda, T. Kakinuma, S. Fujita, et al. Fairness of artificial intelligence in healthcare: Review and recommendations. Japanese Journal of Radiology, 42:3-15, 2024. doi: 10.1007/s11604-023-01474-3.

Note: Values are unadjusted admission-level descriptive measures calculated from the counts in the source analysis. They do not account for confounding or repeated admissions and should not be interpreted as causal effects.