Are elevated lactate levels associated with ICU mortality?

...

January, 2025

Contents

1 Project overview **3**

1.1 Data Description . . . 3

1.2 MIMIC-III v1.4 . . . 3

2 Data Extraction & Preparation **4**

2.1 Database setup, cohort extraction . . . 4

2.2 Data cleaning, variable creation . . . 8

2.3 Exploratory analysis . . . 13

2.3.1 Missing values and imputations . . . 13

3 Statistical Analysis **17**

3.1 Primary analysis (clustering, logistic regression, lineal model, etc., ROC curves) . . . 17

3.2 Subgroup analyses, sensitivity analyses . . . 18

3.3 Create all figures and tables . . . 18

4 Machine Learning modeling **18**

5 Next steps **18**

6 Conclusions **18**

7 References **18**

Authors' information

Nombre	Johanna Ursula Albers
Email	johanna.albers@estudiants.urv.cat
GitHub	https://github.com/
LinkedIn	https://www.linkedin.com/in/
Fecha	January 15, 2026

Nombre	Marco Russo
Email	mrussorb@uoc.edu
GitHub	https://github.com/marcusRB/
LinkedIn	https://www.linkedin.com/in/marcusrb/
Fecha	January 15, 2026

Nombre	Noa-Sibila Janer Oliver
Email	noa-sibila.janer@estudiants.urv.cat
GitHub	https://github.com/
LinkedIn	https://www.linkedin.com/in/
Fecha	January 15, 2026

1 Project overview

Research Question: Are elevated lactate levels associated with ICU mortality? Study Design: Retrospective cohort study using MIMIC-III database Primary Analysis: Association between first lactate measurement post-ICU admission and in-hospital mortality

1.1 Data Description

MIMIC-III is a relational database consisting of 26 tables. Tables are linked by identifiers which usually have the suffix 'ID'. For example, SUBJECT_ID refers to a unique patient, HADM_ID refers to a unique admission to the hospital, and ICUSTAY_ID refers to a unique admission to an intensive care unit.

Charted events such as notes, laboratory tests, and fluid balance are stored in a series of 'events' tables. For example the OUTPUTEVENTS table contains all measurements related to output for a given patient, while the LABEVENTS table contains laboratory test results for a patient.

Tables prefixed with 'D_' are dictionary tables and provide definitions for identifiers. For example, every row of CHARTEVENTS is associated with a single ITEMID which represents the concept measured, but it does not contain the actual name of the measurement. By joining CHARTEVENTS and D_ITEMS on ITEMID, it is possible to identify the concept represented by a given ITEMID.

Developing the MIMIC data model involved balancing simplicity of interpretation against closeness to ground truth. As such, the model is a reflection of underlying data sources, modified over iterations of the MIMIC database in response to user feedback. Care has been taken to avoid making assumptions about the underlying data when carrying out transformations, so MIMIC-III closely represents the raw hospital data.

Broadly speaking, five tables are used to define and track patient stays: ADMISSIONS; PATIENTS; ICUSTAYS; SERVICES; and TRANSFERS. Another five tables are dictionaries for cross-referencing codes against their respective definitions: D_CPT; D_ICD_DIAGNOSES; D_ICD_PROCEDURES; D_ITEMS; and D_LABITEMS. The remaining tables contain data associated with patient care, such as physiological measurements, caregiver observations, and billing information.

In some cases it would be possible to merge tables—for example, the D_ICD_PROCEDURES and CPTEVENTS tables both contain detail relating to procedures and could be combined—but our approach is to keep the tables independent for clarity, since the data sources are significantly different. Rather than combining the tables within MIMIC data model, we suggest researchers develop database views and transforms as appropriate.

1.2 MIMIC-III v1.4

The current version of the database is v1.4. When referencing this version, we recommend using the full title: MIMIC-III v1.4[1].

MIMIC-III v1.4 was released on 2 September 2016. It was a major release enhancing data quality and providing a large amount of additional data for Metavision patients.

[1] Johnson, A., Pollard, T., & Mark, R. (2016). MIMIC-III Clinical Database (version 1.4). PhysioNet. RRID:SCR_007345. https://doi.org/10.13026/C2XW26

2 Data Extraction & Preparation

2.1 Database setup, cohort extraction

```r
library(dplyr)
library(tidyr)
library(tibble)
library(lubridate)
library(readr)
library(stringr)
library(ggplot2)
library(data.table)
library(odbc)
library(RMariaDB)
```

```r
# Get the credentials file
# It looks like:
################
#username=your.user.name
#password=your.password.assigned
################
creds_file = "creds.txt"
```

```r
# Set the username and password
creds <- readLines(creds_file)
cred_list <- setNames(
  sub(".*=", "", creds),
  sub("=.*", "", creds)
)
```

```r
con <- dbConnect(
  drv = RMariaDB::MariaDB(),
  username = cred_list[["username"]],
  password = cred_list[["password"]],
  host = "ehr3.deim.urv.cat",
  dbname = "mimiciiiv14",
  port = 3306
)
```

We constructed a cohort at the level of individual lactate measurements obtained during ICU stays. Each row in the dataset corresponds to a single lactate measurement, and all lactate values recorded during any ICU stay were included.

Lactate measurements were assigned to ICU stays based on their timestamp: a measurement was linked to an ICU stay only if it occurred between the ICU admission (intime) and discharge (outtime) times.

ICU mortality was derived from hospital admission outcomes. For patients with multiple ICU stays within the same hospital admission, ICU mortality was assigned exclusively to the final ICU stay if the patient died during that admission; all preceding ICU stays were labeled as non-fatal. This approach allows ICU-level mortality attribution in the absence of a direct ICU death indicator. The final cohort comprised 129,966 lactate measurements, representing multiple ICU stays and hospital admissions across patients.

```
library(DBI)
dbExecute(con, "SET SQL_BIG_SELECTS=1;")
```

```
## [1] 0
```

```
cohort_query <- "
        WITH icu_ordered AS (
        SELECT
            i.subject_id,
            i.hadm_id,
            i.icustay_id,
            i.intime,
            i.outtime,
            a.hospital_expire_flag AS hospital_dead,
            ROW_NUMBER() OVER (
                PARTITION BY i.subject_id, i.hadm_id
                ORDER BY i.outtime DESC
            ) AS rn_icu
        FROM ICUSTAYS i
        JOIN ADMISSIONS a
            ON i.hadm_id = a.hadm_id
    ),
    lactate_first AS (
        SELECT
            io.subject_id,
            io.hadm_id,
            io.icustay_id,
            l.charttime as lactate_time,
            l.value as lactate_value,
            l.valuenum as lactate_value_num,
            l.valueuom as lactate_units,
            l.flag as lactate_flag,
            ROW_NUMBER() OVER (
                PARTITION BY io.icustay_id
                ORDER BY l.charttime ASC
            ) AS rn
        FROM LABEVENTS l
        INNER JOIN icu_ordered io USING(subject_id, hadm_id)
        WHERE l.itemid = 50813
          AND l.value IS NOT NULL
          AND l.charttime BETWEEN io.intime AND (io.intime + INTERVAL '24' HOUR)
    ),
    ck_first AS (
        SELECT
            io.subject_id,
            io.hadm_id,
            io.icustay_id,
            l.charttime as ck_time,
            l.value as ck_value,
            l.valuenum as ck_value_num,
            l.valueuom as ck_units,
            l.flag as ck_flag,
            ROW_NUMBER() OVER (
                PARTITION BY io.icustay_id
                ORDER BY l.charttime ASC
            ) AS rn
        FROM LABEVENTS l
        INNER JOIN icu_ordered io USING(subject_id, hadm_id)
        WHERE l.ITEMID = 50910
          AND l.charttime BETWEEN io.intime AND (io.intime + INTERVAL '24' HOUR)
    ),
    bilirubin_first AS (
        SELECT
            io.subject_id,
            io.hadm_id,
            io.icustay_id,
            l.charttime as bilirubin_time,
            l.value as bilirubin_value,
            l.valuenum as bilirubin_value_num,
            l.valueuom as bilirubin_units,
            l.flag as bilirubin_flag,
            ROW_NUMBER() OVER (
                PARTITION BY io.icustay_id
                ORDER BY l.charttime ASC
            ) AS rn
        FROM LABEVENTS l
        INNER JOIN icu_ordered io USING(subject_id, hadm_id)
        WHERE l.ITEMID = 50885
          AND l.charttime BETWEEN io.intime AND (io.intime + INTERVAL '24' HOUR)
    ),
    mabp_first AS (
        SELECT
            io.subject_id,
            io.hadm_id,
            io.icustay_id,
            c.CHARTTIME as mabp_time,
            c.value as mabp_value,
            c.valuenum as mabp_value_num,
            c.valueuom as mabp_units,
            ROW_NUMBER() OVER (
                PARTITION BY io.icustay_id
                ORDER BY c.charttime ASC
            ) AS rn
        FROM CHARTEVENTS c
        INNER JOIN icu_ordered io USING(subject_id, hadm_id, icustay_id)
        WHERE c.itemid = 220052
          AND c.valuenum IS NOT NULL
          AND c.charttime BETWEEN io.intime AND (io.intime + INTERVAL '24' HOUR)
    ),
    icu_with_lactate AS (
        SELECT DISTINCT subject_id, hadm_id, icustay_id
        FROM lactate_first
        WHERE rn = 1
    )
    SELECT
        ROW_NUMBER() OVER (ORDER BY iu.subject_id, iu.hadm_id, iu.icustay_id) AS row_id,
        iu.subject_id,
        p.dob,
        YEAR(iu.intime) - YEAR(p.DOB) AS age_icu_entry,
        YEAR(iu.outtime) - YEAR(p.DOB) AS age_icu_exit,
        p.gender,
        iu.hadm_id,
        iu.icustay_id,
        iu.intime,
        iu.outtime,
        iu.hospital_dead,
        CASE
            WHEN iu.hospital_dead = 1 AND iu.rn_icu = 1 THEN 1
            ELSE 0
        END AS icu_dead,
        lf.lactate_time,
        lf.lactate_value,
        lf.lactate_value_num,
        lf.lactate_units,
        lf.lactate_flag,
        ck.ck_time,
        ck.ck_value,
        ck.ck_value_num,
        ck.ck_units,
        ck.ck_flag,
        bil.bilirubin_time,
        bil.bilirubin_value,
        bil.bilirubin_value_num,
        bil.bilirubin_units,
        bil.bilirubin_flag,
        mb.mabp_time,
        mb.mabp_value,
        mb.mabp_value_num,
        mb.mabp_units
    FROM icu_ordered iu
    INNER JOIN icu_with_lactate iwl ON iwl.subject_id = iu.subject_id
        AND iwl.hadm_id = iu.hadm_id
        AND iwl.icustay_id = iu.icustay_id
    LEFT JOIN lactate_first lf ON lf.subject_id = iu.subject_id
        AND lf.hadm_id = iu.hadm_id
        AND lf.icustay_id = iu.icustay_id
        AND lf.rn = 1
    LEFT JOIN ck_first ck ON ck.subject_id = iu.subject_id
        AND ck.hadm_id = iu.hadm_id
        AND ck.icustay_id = iu.icustay_id
        AND ck.rn = 1
    LEFT JOIN bilirubin_first bil ON bil.subject_id = iu.subject_id
        AND bil.hadm_id = iu.hadm_id
        AND bil.icustay_id = iu.icustay_id
        AND bil.rn = 1
    LEFT JOIN mabp_first mb ON mb.subject_id = iu.subject_id
        AND mb.hadm_id = iu.hadm_id
        AND mb.icustay_id = iu.icustay_id
        AND mb.rn = 1
    LEFT JOIN PATIENTS p ON p.subject_id = iu.subject_id
    ORDER BY iu.subject_id, iu.hadm_id, iu.icustay_id;
```

```
"

initial_df <- dbGetQuery(con, cohort_query)
initial_df %>% glimpse()
```

```
## Rows: 22,307
## Columns: 31
## $ row_id              <int64> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 1~
## $ subject_id          <int> 3, 9, 12, 21, 25, 31, 33, 36, 36, 38, 41, 41, 45, ~
## $ dob                 <dttm> 2025-04-11, 2108-01-26, 2032-03-24, 2047-04-04, 2~
## $ age_icu_entry       <int> 76, 41, 72, 87, 59, 72, 82, 70, 73, 76, 57, 57, 42~
## $ age_icu_exit        <int> 76, 41, 72, 87, 59, 72, 82, 70, 73, 76, 57, 57, 42~
## $ gender              <chr> "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", ~
## $ hadm_id             <int> 145834, 150750, 112213, 109451, 129635, 128652, 17~
## $ icustay_id          <int> 211552, 220597, 232669, 217847, 203487, 254478, 29~
## $ intime              <dttm> 2101-10-20 19:10:11, 2149-11-09 13:07:02, 2104-08~
## $ outtime             <dttm> 2101-10-26 20:43:09, 2149-11-14 20:52:14, 2104-08~
## $ hospital_dead       <int> 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ icu_dead            <int> 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ lactate_time        <dttm> 2101-10-20 19:12:00, 2149-11-09 17:47:00, 2104-08~
## $ lactate_value       <chr> "4.3", "1.9", "13.8", "1.9", "1.6", "1.4", "6.0", ~
## $ lactate_value_num   <dbl> 4.3, 1.9, 13.8, 1.9, 1.6, 1.4, 6.0, 6.1, 1.0, 2.8,~
## $ lactate_units       <chr> "mmol/L", "mmol/L", "mmol/L", "mmol/L", "mmol/L", ~
## $ lactate_flag        <chr> "abnormal", NA, "abnormal", NA, NA, NA, "abnormal"~
## $ ck_time             <dttm> 2101-10-20 19:59:00, NA, 2104-08-08 07:30:00, 213~
## $ ck_value            <chr> "82", NA, "1554", "593", "296", NA, NA, NA, "347",~
## $ ck_value_num        <dbl> 82, NA, 1554, 593, 296, NA, NA, NA, 347, NA, NA, N~
## $ ck_units            <chr> "IU/L", NA, "IU/L", "IU/L", "IU/L", NA, NA, NA, "I~
## $ ck_flag             <chr> NA, NA, "abnormal", "abnormal", "abnormal", NA, NA~
## $ bilirubin_time      <dttm> 2101-10-20 19:59:00, 2149-11-10 09:40:00, NA, 213~
## $ bilirubin_value     <chr> "0.8", "0.4", NA, "0.4", "0.4", NA, NA, "0.4", NA,~
## $ bilirubin_value_num <dbl> 0.8, 0.4, NA, 0.4, 0.4, NA, NA, 0.4, NA, NA, NA, N~
## $ bilirubin_units     <chr> "mg/dL", "mg/dL", NA, "mg/dL", "mg/dL", NA, NA, "m~
## $ bilirubin_flag      <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ mabp_time           <dttm> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ mabp_value          <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ mabp_value_num      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ mabp_units          <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
```

2.2 Data cleaning, variable creation

```
# Are there duplicates?
sum(duplicated(initial_df))
```

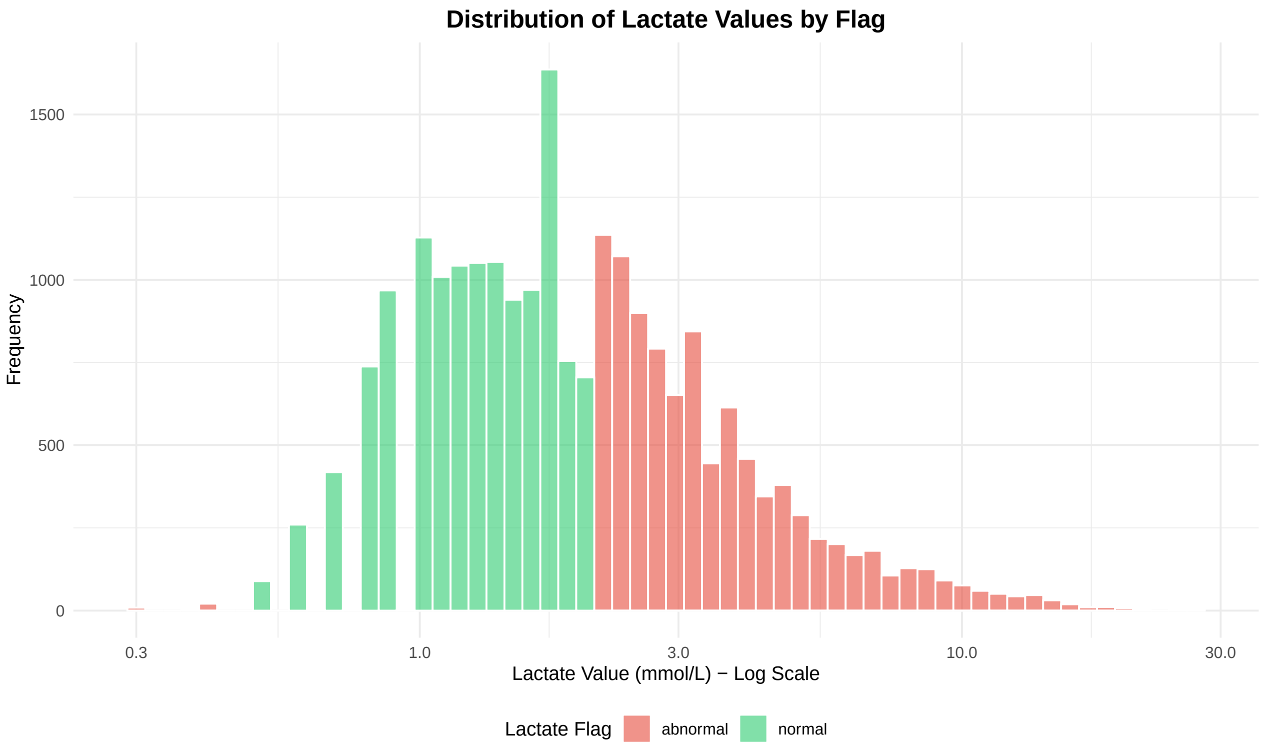
```
## [1] 0
```

```
clean_values_df <- initial_df %>%
  mutate(
    lactate_value_num = case_when(
      # For samples containing ">" or "GREATER THAN 30" assign value 31

      str_detect(lactate_value, fixed(">")) |
      str_detect(toupper(lactate_value), "GREATER THAN") ~ 31,

      TRUE ~ as.numeric(lactate_value) # Convert to numeric value,
      # samples with numbers and letters will be converted to NA
    )
  ) %>%
  filter(!is.na(lactate_value_num)) # Delete whole row if lactate value is null


# Check for suspicious values
clean_values_df %>%
  filter(lactate_value_num <= 0 | lactate_value_num > 31)
```

```
##  [1] row_id              subject_id          dob
##  [4] age_icu_entry       age_icu_exit        gender
##  [7] hadm_id             icustay_id          intime
## [10] outtime             hospital_dead       icu_dead
## [13] lactate_time        lactate_value       lactate_value_num
## [16] lactate_units       lactate_flag        ck_time
## [19] ck_value            ck_value_num        ck_units
## [22] ck_flag             bilirubin_time      bilirubin_value
## [25] bilirubin_value_num bilirubin_units     bilirubin_flag
## [28] mabp_time           mabp_value          mabp_value_num
## [31] mabp_units
## <0 rows> (or 0-length row.names)
```

Imputed ‘normal’ for all missing records in the lactate_flag variable.

```
clean_values_df <- clean_values_df %>%
  mutate(
    lactate_flag = if_else(is.na(lactate_flag), "normal", lactate_flag)
  )


ggplot(clean_values_df, aes(x = lactate_value_num, fill = lactate_flag)) +
  geom_histogram(bins = 60, alpha = 0.6, position = "identity", color = "white") +

  scale_x_log10() +

  scale_fill_manual(values = c("normal" = "#2ecc71", "abnormal" = "#e74c3c")) +

  labs(
    title = "Distribution of Lactate Values by Flag",
    x = "Lactate Value (mmol/L) - Log Scale",
    y = "Frequency",
    fill = "Lactate Flag"
  ) +

  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", size = 14),
    plot.subtitle = element_text(hjust = 0.5),
    legend.position = "bottom"
  )
```



Validate mortality coherence 1: icu_dead cannot be true if hospital_dead is false.

```
clean_values_df %>%
  filter(icu_dead == 1 & hospital_dead == 0) %>%
  nrow()
```

```
## [1] 0
```

Validate mortality coherence 2: each subject should have at most one recorded death across ICU and hospital stays.

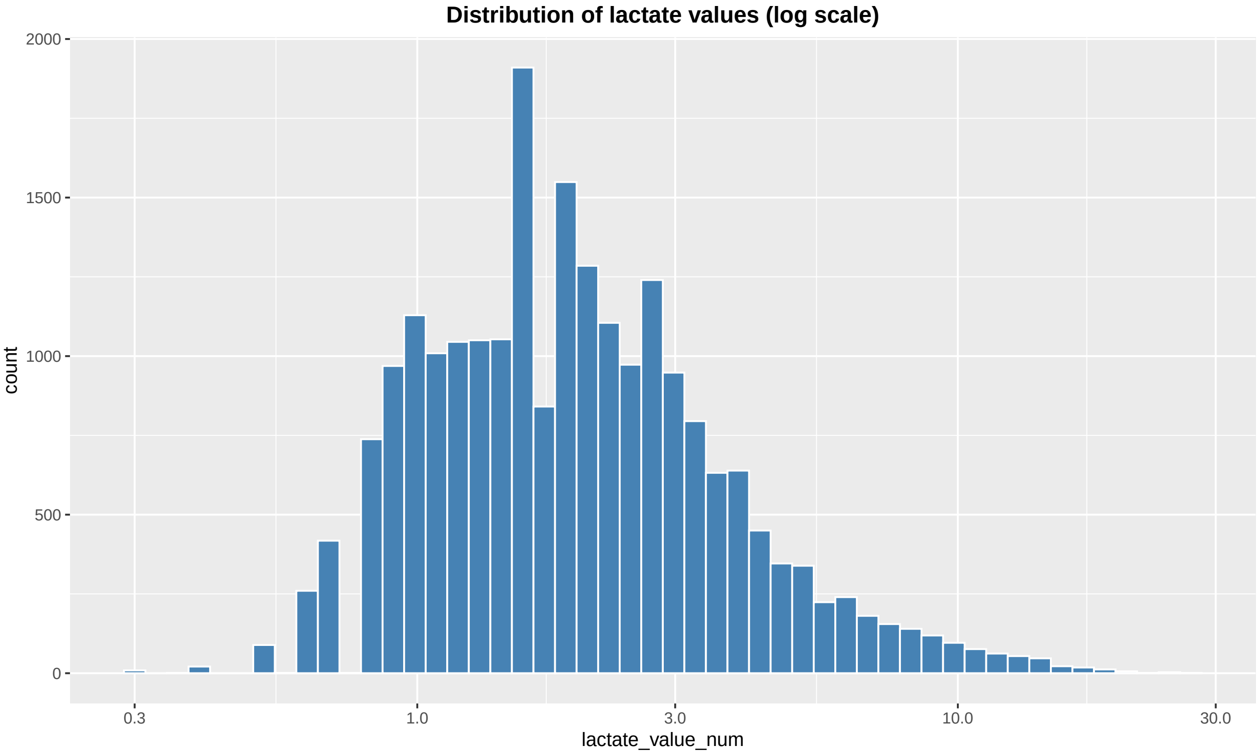
```
death_counts <- clean_values_df %>%
  group_by(subject_id) %>%
  summarise(
    hosp_dead = n_distinct(hadm_id[hospital_dead == 1], na.rm = TRUE),
    icu_dead  = n_distinct(icustay_id[icu_dead == 1], na.rm = TRUE)
  ) %>%
  filter(icu_dead > 1)
death_counts
```

```
## # A tibble: 1 x 3
##   subject_id hosp_dead icu_dead
##        <int>     <int>    <int>
## 1      17796         2        2
```

A data inconsistency was identified involving a single subject with two incorrectly assigned mortality flags; these records were manually rectified to ensure data integrity.

```
clean_values_df <- clean_values_df %>%
  mutate(
    # Reassign hospital_dead from first admission
    hospital_dead = if_else(subject_id == 17796 & hadm_id == 119823, 0, hospital_dead),

    # Reassign icu_dead from first ICU stay
    icu_dead = if_else(subject_id == 17796 & hadm_id == 119823, 0, icu_dead)
  )
```

```
# Show the distribution of lactate values (log scale)
ggplot(clean_values_df, aes(lactate_value_num)) +
  geom_histogram(bins = 50, fill = "steelblue", color = "white") +
  scale_x_log10() +
  labs(title = "Distribution of lactate values (log scale)") +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
```



The final cohort for exploratory analysis comprises 21,016 unique subjects, accounting for 24,622 hospital admissions and 25,743 ICU stays. The dataset includes a total of 128,715 lactate observations, with an average of 5 measurements per ICU stay, providing a robust longitudinal perspective on patient biomarkers.

```
summary(clean_values_df)
```

```
##      row_id        subject_id         dob                          age_icu_entry
##  Min.   :    1   Min.   :    3   Min.   :1800-07-02 00:00:00   Min.   :  0.00
##  1st Qu.: 5577   1st Qu.:15378   1st Qu.:2057-09-06 12:00:00   1st Qu.: 54.00
##  Median :11155   Median :29470   Median :2085-12-29 12:00:00   Median : 66.00
##  Mean   :11154   Mean   :39821   Mean   :2077-03-11 01:43:53   Mean   : 74.07
##  3rd Qu.:16732   3rd Qu.:64588   3rd Qu.:2112-01-26 06:00:00   3rd Qu.: 78.00
##  Max.   :22307   Max.   :99995   Max.   :2185-04-04 00:00:00   Max.   :311.00
##
##   age_icu_exit       gender             hadm_id          icustay_id
##  Min.   :  0.00   Length:22302       Min.   :100003   Min.   :200001
##  1st Qu.: 54.00   Class :character   1st Qu.:124882   1st Qu.:224940
##  Median : 66.00   Mode  :character   Median :150028   Median :249918
##  Mean   : 74.09                      Mean   :150036   Mean   :249906
##  3rd Qu.: 78.00                      3rd Qu.:175221   3rd Qu.:275104
##  Max.   :311.00                      Max.   :199998   Max.   :299993
##
##      intime                         outtime                        hospital_dead
##  Min.   :2100-06-22 06:34:52   Min.   :2100-06-24 13:35:56   Min.   :0.0000
##  1st Qu.:2126-02-10 09:37:14   1st Qu.:2126-02-16 18:06:04   1st Qu.:0.0000
##  Median :2150-12-11 09:24:17   Median :2150-12-18 01:27:12   Median :0.0000
##  Mean   :2151-04-07 12:30:37   Mean   :2151-04-13 02:22:53   Mean   :0.1766
##  3rd Qu.:2176-08-30 02:38:42   3rd Qu.:2176-09-02 15:34:24   3rd Qu.:0.0000
##  Max.   :2210-08-18 12:34:24   Max.   :2210-08-20 18:35:13   Max.   :1.0000
##
##     icu_dead       lactate_time                  lactate_value
##  Min.   :0.0000   Min.   :2100-06-22 13:00:00   Length:22302
##  1st Qu.:0.0000   1st Qu.:2126-02-10 12:41:45   Class :character
##  Median :0.0000   Median :2150-12-11 13:47:30   Mode  :character
##  Mean   :0.1598   Mean   :2151-04-07 17:14:25
##  3rd Qu.:0.0000   3rd Qu.:2176-08-30 08:00:45
##  Max.   :1.0000   Max.   :2210-08-18 14:52:00
##
##  lactate_value_num lactate_units      lactate_flag
##  Min.   : 0.300    Length:22302       Length:22302
##  1st Qu.: 1.200    Class :character   Class :character
##  Median : 1.800    Mode  :character   Mode  :character
##  Mean   : 2.436
##  3rd Qu.: 2.900
##  Max.   :27.000
##
##     ck_time                           ck_value          ck_value_num
##  Min.   :2100-07-22 04:18:00   Length:22302       Min.   :     3
##  1st Qu.:2126-01-19 14:23:30   Class :character   1st Qu.:    60
##  Median :2150-02-18 11:25:00   Mode  :character   Median :   140
##  Mean   :2151-02-14 14:43:53                      Mean   :  1059
##  3rd Qu.:2176-10-15 19:39:00                      3rd Qu.:   433
##  Max.   :2209-07-31 17:50:00                      Max.   :266720
##  NA's   :13111                                    NA's   :13112
##    ck_units           ck_flag          bilirubin_time
##  Length:22302       Length:22302       Min.   :2100-07-10 04:16:00
##  Class :character   Class :character   1st Qu.:2126-01-06 10:45:00
##  Mode  :character   Mode  :character   Median :2150-12-26 21:00:00
##                                        Mean   :2151-03-21 07:00:32
##                                        3rd Qu.:2176-06-28 19:23:00
##                                        Max.   :2210-08-18 17:07:00
##                                        NA's   :12613
##  bilirubin_value    bilirubin_value_num bilirubin_units    bilirubin_flag
##  Length:22302       Min.   : 0.000      Length:22302       Length:22302
##  Class :character   1st Qu.: 0.400      Class :character   Class :character
##  Mode  :character   Median : 0.800      Mode  :character   Mode  :character
##                     Mean   : 2.166
##                     3rd Qu.: 1.800
##                     Max.   :79.000
##                     NA's   :12614
##    mabp_time                      mabp_value        mabp_value_num
##  Min.   :2100-07-20 11:23:00   Length:22302       Min.   :-38.00
##  1st Qu.:2126-07-31 07:58:30   Class :character   1st Qu.: 66.00
##  Median :2151-11-24 05:50:00   Mode  :character   Median : 78.00
##  Mean   :2151-09-09 00:16:00                      Mean   : 77.86
##  3rd Qu.:2176-12-13 21:15:30                      3rd Qu.: 90.00
##  Max.   :2208-08-19 14:08:00                      Max.   :361.00
##  NA's   :15767                                    NA's   :15767
##   mabp_units
##  Length:22302
##  Class :character
##  Mode  :character
##
##
##
##
```

2.3 Exploratory analysis

2.3.1 Missing values and imputations

After the imputation task the first step is select only the important features to doing the exploratory analysis

```
names(clean_values_df)
```

```
##  [1] "row_id"              "subject_id"          "dob"
##  [4] "age_icu_entry"       "age_icu_exit"        "gender"
##  [7] "hadm_id"             "icustay_id"          "intime"
## [10] "outtime"             "hospital_dead"       "icu_dead"
## [13] "lactate_time"        "lactate_value"       "lactate_value_num"
## [16] "lactate_units"       "lactate_flag"        "ck_time"
## [19] "ck_value"            "ck_value_num"        "ck_units"
## [22] "ck_flag"             "bilirubin_time"      "bilirubin_value"
## [25] "bilirubin_value_num" "bilirubin_units"     "bilirubin_flag"
## [28] "mabp_time"           "mabp_value"          "mabp_value_num"
## [31] "mabp_units"
```

```
eda_columns <- c("subject_id", "dob", "age_icu_entry", "age_icu_exit",
                 "gender", "hadm_id", "icustay_id", "intime", "outtime",
                 "hospital_dead", "icu_dead", "lactate_time", "lactate_value_num",
                 "lactate_units", "lactate_flag", "ck_time", "ck_value_num", "ck_units",
                 "ck_flag", "bilirubin_time", "bilirubin_value_num", "bilirubin_units",
                 "bilirubin_flag", "mabp_time", "mabp_value_num", "mabp_units")

eda_df <- clean_values_df %>%
  select(eda_columns)
head(eda_df)
```

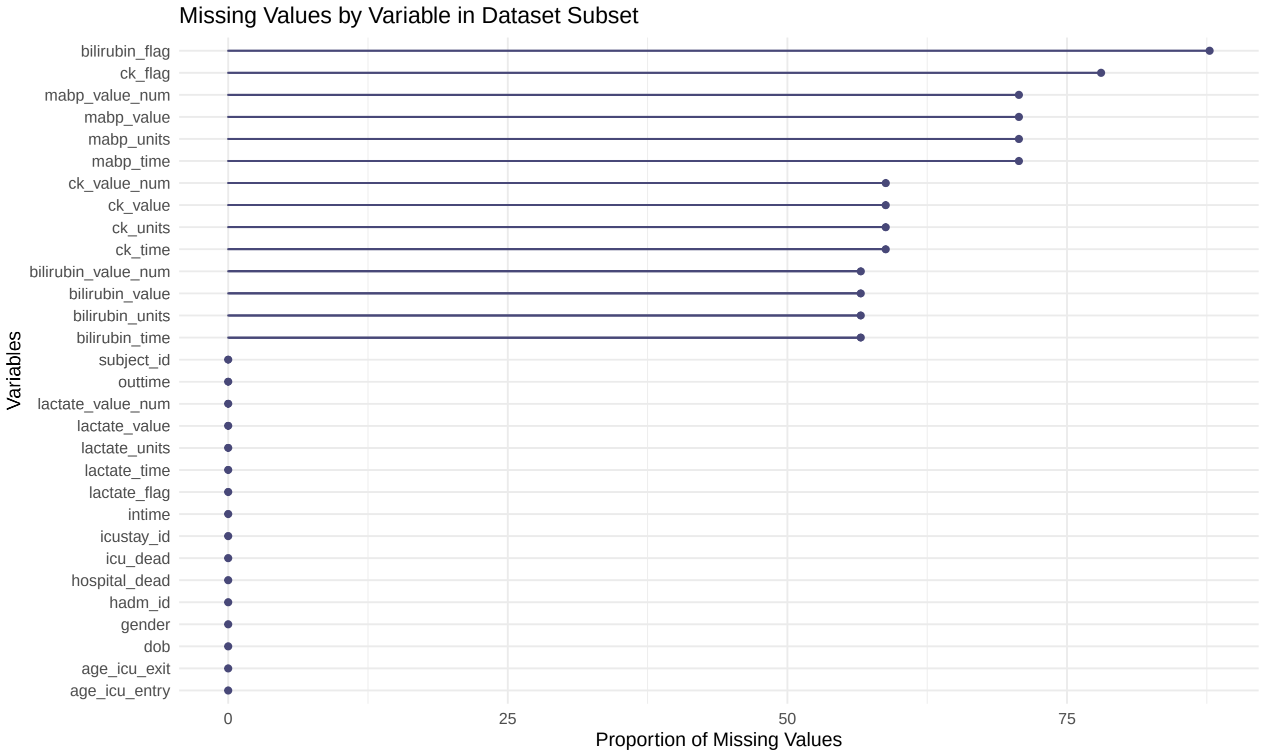
```
##   subject_id        dob age_icu_entry age_icu_exit gender hadm_id icustay_id
## 1          3 2025-04-11            76           76      M  145834     211552
## 2          9 2108-01-26            41           41      M  150750     220597
## 3         12 2032-03-24            72           72      M  112213     232669
## 4         21 2047-04-04            87           87      M  109451     217847
## 5         25 2101-11-21            59           59      M  129635     203487
## 6         31 2036-05-17            72           72      M  128652     254478
##                intime             outtime hospital_dead icu_dead
## 1 2101-10-20 19:10:11 2101-10-26 20:43:09             0        0
## 2 2149-11-09 13:07:02 2149-11-14 20:52:14             1        1
## 3 2104-08-08 02:08:17 2104-08-15 17:22:25             1        1
## 4 2134-09-11 20:50:04 2134-09-17 18:28:32             0        0
## 5 2160-11-02 03:16:23 2160-11-05 16:23:27             0        0
## 6 2108-08-22 23:28:42 2108-08-30 21:59:20             1        1
##          lactate_time lactate_value_num lactate_units lactate_flag
## 1 2101-10-20 19:12:00               4.3        mmol/L     abnormal
## 2 2149-11-09 17:47:00               1.9        mmol/L       normal
## 3 2104-08-08 02:15:00              13.8        mmol/L     abnormal
## 4 2134-09-12 09:21:00               1.9        mmol/L       normal
## 5 2160-11-02 06:05:00               1.6        mmol/L       normal
## 6 2108-08-23 00:21:00               1.4        mmol/L       normal
##               ck_time ck_value_num ck_units  ck_flag      bilirubin_time
## 1 2101-10-20 19:59:00           82     IU/L     <NA> 2101-10-20 19:59:00
## 2                <NA>           NA     <NA>     <NA> 2149-11-10 09:40:00
## 3 2104-08-08 07:30:00         1554     IU/L abnormal                <NA>
## 4 2134-09-12 04:30:00          593     IU/L abnormal 2134-09-12 04:30:00
## 5 2160-11-02 08:55:00          296     IU/L abnormal 2160-11-02 23:18:00
## 6                <NA>           NA     <NA>     <NA>                <NA>
##   bilirubin_value_num bilirubin_units bilirubin_flag mabp_time mabp_value_num
## 1                 0.8           mg/dL           <NA>      <NA>             NA
## 2                 0.4           mg/dL           <NA>      <NA>             NA
## 3                  NA            <NA>           <NA>      <NA>             NA
## 4                 0.4           mg/dL           <NA>      <NA>             NA
## 5                 0.4           mg/dL           <NA>      <NA>             NA
## 6                  NA            <NA>           <NA>      <NA>             NA
##   mabp_units
## 1       <NA>
## 2       <NA>
## 3       <NA>
## 4       <NA>
## 5       <NA>
## 6       <NA>
```

Checking the missing values

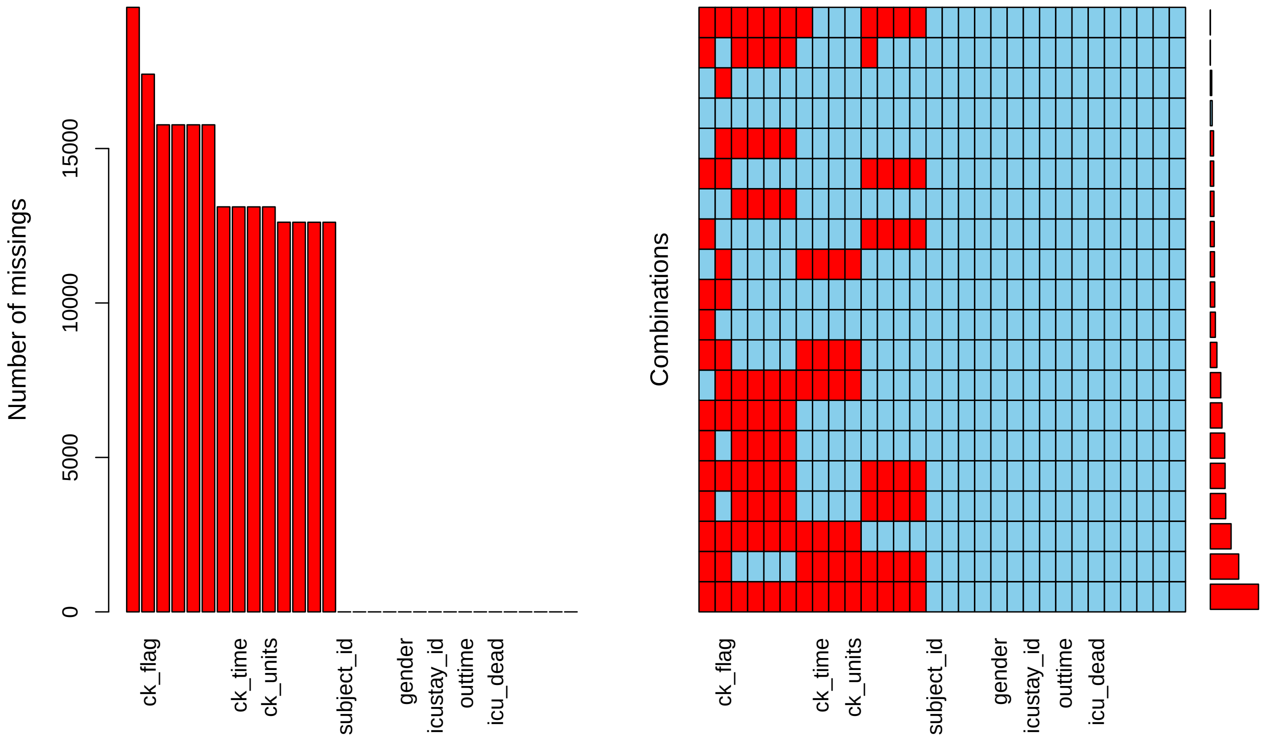
```
as.data.frame(eda_df) %>%
  summarise(across(everything(), ~ sum(is.na(.)))) %>%
  select(where(~ . > 0))
```

```
##   ck_time ck_value_num ck_units ck_flag bilirubin_time bilirubin_value_num
## 1   13111        13112    13111   17406          12613               12614
##   bilirubin_units bilirubin_flag mabp_time mabp_value_num mabp_units
## 1           12613          19570     15767          15767      15767
```

```
library(naniar)
library(ggplot2)
# Visualize missing values by variable

missing_df <- clean_values_df %>%
  select(-row_id)

gg_miss_var(missing_df, show_pct = TRUE) +
labs(title = "Missing Values by Variable in Dataset Subset",
x = "Variables",
y = "Proportion of Missing Values")
```



```
library(VIM)
aggr(missing_df, numbers = FALSE, prop = FALSE, sortVar = TRUE)
```



```
##
##  Variables sorted by number of missings:
##             Variable Count
##       bilirubin_flag 19570
##              ck_flag 17406
##            mabp_time 15767
##           mabp_value 15767
##       mabp_value_num 15767
##           mabp_units 15767
##         ck_value_num 13112
##              ck_time 13111
##             ck_value 13111
##             ck_units 13111
##  bilirubin_value_num 12614
##       bilirubin_time 12613
##      bilirubin_value 12613
##      bilirubin_units 12613
##           subject_id     0
##                  dob     0
##        age_icu_entry     0
##         age_icu_exit     0
##               gender     0
##              hadm_id     0
##           icustay_id     0
##               intime     0
##              outtime     0
##        hospital_dead     0
##             icu_dead     0
##         lactate_time     0
##        lactate_value     0
##    lactate_value_num     0
##        lactate_units     0
##         lactate_flag     0
```

```r
# Check missing values
missing_summary <- sapply(eda_df, function(x) sum(is.na(x)))
missing_df <- data.frame(
    Column = names(missing_summary),
    Missing_Count = missing_summary,
    Missing_Percent = round(missing_summary/nrow(eda_df)*100, 2)
  ) %>%
  arrange(desc(Missing_Count))

print(missing_df)
```

```
##                                  Column Missing_Count Missing_Percent
## bilirubin_flag           bilirubin_flag         19570           87.75
## ck_flag                         ck_flag         17406           78.05
## mabp_time                     mabp_time         15767           70.70
## mabp_value_num           mabp_value_num         15767           70.70
## mabp_units                   mabp_units         15767           70.70
## ck_value_num               ck_value_num         13112           58.79
## ck_time                         ck_time         13111           58.79
## ck_units                       ck_units         13111           58.79
## bilirubin_value_num bilirubin_value_num         12614           56.56
## bilirubin_time           bilirubin_time         12613           56.56
## bilirubin_units         bilirubin_units         12613           56.56
## subject_id                   subject_id             0            0.00
## dob                                 dob             0            0.00
## age_icu_entry             age_icu_entry             0            0.00
## age_icu_exit               age_icu_exit             0            0.00
## gender                           gender             0            0.00
## hadm_id                         hadm_id             0            0.00
## icustay_id                   icustay_id             0            0.00
## intime                           intime             0            0.00
## outtime                         outtime             0            0.00
## hospital_dead             hospital_dead             0            0.00
## icu_dead                       icu_dead             0            0.00
## lactate_time               lactate_time             0            0.00
## lactate_value_num     lactate_value_num             0            0.00
## lactate_units             lactate_units             0            0.00
## lactate_flag               lactate_flag             0            0.00
```

3 Statistical Analysis

3.1 Primary analysis (clustering, logistic regression, lineal model, etc., ROC curves)

3.2 Subgroup analyses, sensitivity analyses

3.3 Create all figures and tables

4 Machine Learning modeling

5 Next steps

6 Conclusions

7 References