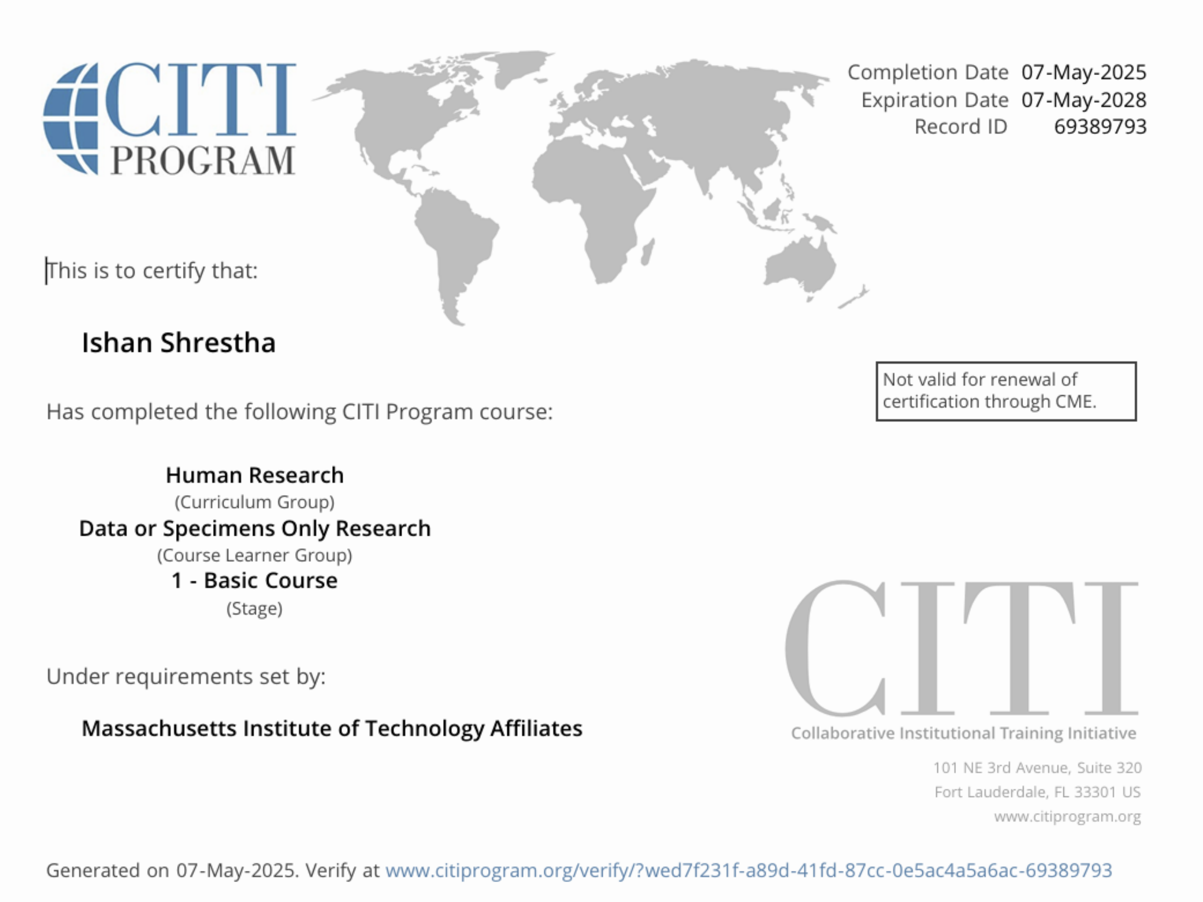
MIMIC-IV Data Extraction

Replication of Liu et al. (2022): Weaning Model for Sepsis Patients

1. Part1: Prerequisites

1.1. CITI Data Certification



CITI PROGRAM

Completion Date 07-May-2025
Expiration Date 07-May-2028
Record ID 69389793

This is to certify that:

Ishan Shrestha

Not valid for renewal of certification through CME.

Has completed the following CITI Program course:

Human Research
(Curriculum Group)
Data or Specimens Only Research
(Course Learner Group)
1 - Basic Course
(Stage)

Under requirements set by:

Massachusetts Institute of Technology Affiliates

CITI
Collaborative Institutional Training Initiative

101 NE 3rd Avenue, Suite 320
Fort Lauderdale, FL 33301 US
www.citiprogram.org

Generated on 07-May-2025. Verify at www.citiprogram.org/verify/?wed7f231f-a89d-41fd-87cc-0e5ac4a5a6ac-69389793

Prior to using the MIMIC-IV database, I completed the CITI Program's "Data or Specimens Only Research" course, which provided essential training on ethical principles, data privacy, and responsible handling of de-identified human subject data.

1.2. Database format

The dataset was queried directly within the Google BigQuery environment using standard SQL. All joins, filters, and transformations were modularized using Common Table Expressions (CTEs), which enabled structured and efficient cohort construction and feature extraction.

2. Part2: Data extraction and Conceptualization

2.1. MIMIC version

The dataset used for this assignment was MIMIC-IV version 1.0.

2.2. Brief summary of the dataset

MIMIC-IV version 1.0 is a freely available clinical database developed by the MIT Lab for Computational Physiology. It contains de-identified health data from ICU patients at Beth Israel Deaconess Medical Center between 2008 and 2019.

- Core: Demographics and hospital admissions (patients, admissions)
- Hosp: In-hospital events (diagnoses_icd, prescriptions)
- ICU: ICU-specific data (icustays, chartevents)
- Derived: Pre-calculated clinical scores and standardized variables (sofa, gcs, ventilation, urine_output)

2.3. Conceptualizing Liu et al. (2022) paper

The goal was to replicate the cohort and features described in the study “A Simple Weaning Model Based on Interpretable Machine Learning Algorithm for Patients With Sepsis” using SQL.

2.3.1. Inclusion criteria

The following criteria were used to select eligible patients for the study:

- Patients must be 18 years of age or older at the time of ICU admission.
- Patients with a SOFA score greater or equal to 2 within the first 24 hours of ICU admission were included.
- Only patients who received invasive mechanical ventilation were selected.
- Patients must have a documented extubation time, derived as the last recorded endtime for invasive ventilation.
- 24-hour window prior to extubation was required to extract clinical features for the model.

2.3.2. Exclusion criteria

The following conditions led to exclusion from the cohort:

- Only the last stay per subject was retained in cases where multiple ICU stays existed.
- Patients without SOFA score, ventilation records, or extubation timestamps were excluded.

2.4. Data Extraction

The sql query builds a cohort and extracts clinical features from the MIMIC-IV database to replicate the dataset construction described in the paper.

2.4.1. Cohort construction

- Latest extubation time for ICU stays with InvasiveVent status
- Most recent invasive ventilation entry per ICU stay
- Adult patients with age greater than or equal to 18
- Diagnosed with sepsis approximated using SOFA score greater or equal to 2
- Patients receiving IMV
- Only the last ICU stay per patient is included

2.4.2. Feature extraction

- Vitals: max HR, max RR, min MAP, max Temp, min SpO2
- Lab Tests: WBC, Hb, Platelets, Creatinine, Anion Gap
- Blood Gas: pH, PaO_2, $PaCO_2$, Base Excess
- GCS: min and max
- Urine Output: average per hour
- Ventilation Settings: FiO_2, PEEP, tidal volume
- Demographics: sex, BMI
- Comorbidities & Charlson Index: Based on ICD codes
- Oxygenation Index (OI)
- Antibiotic and CRRT durations
- Ventilation duration
- Vasopressor use in last 24 hours
- Death Time

2.4.3. Summary

- Columns: 44
- Rows: 20421 (16,765 weaning success and 3,656 weaning failure)

2.4.4. Extracted columns

- Data joining/Identification features: subject_id, hadm_id, stay_id, extubation_time, deathtime
- Clinical Features

Category	**Column Name**
Demographics	age, male, bmi
Vitals	highest_heart_rate, highest_respiratory_rate, lowest_map, highest_temperature, lowest_spo2

Labs	highest_wbc, lowest_hemoglobin, lowest_platelets, highest_creatinine, highest_anion_gap
Blood Gas	lowest_ph_level, lowest_PaO2, highest_PaCO2, lowest_base_excess
GCS	min_gcs, max_gcs
Output	urine_output
Ventilation	highest_FiO2, highest_peep, lowest_tidal_volume
Comorbidities	chronic_pulomary_disease, congestive_heart_failure, dementia, severe_liver_disease, renal_disease, diabetes
Derived	sofa, Charlson_comorbidity_index, lowest_OI, antibiotic_duration, crrt_duration, imv_duration, vasopressor_used_1__day_before_weaning

- Target variable: weaning
  - Failure – either reintubation or death within 72 hours after extubation
  - Success – no reintubation or death with 72 hours after extubation

2.5. Comparison with Liu et al. (2022)

Aspect	**Query**	**Paper**
Cohort	Adult, sepsis (SOFA≥2), invasive vent, both alive and dead after 72h post-extubation	Adult, sepsis (SOFA≥2), invasive vent, alive ≥72h post-extubation
Single ICU Stay	Only latest ICU stay per subject	Matches paper (only one ICU stay per subject)
Feature Extraction Window	Last 24h before extubation	Matches the paper exactly
Variables Extracted	Matches 35 clinical variables mentioned in the paper	Fully aligned
Model Target	Weaning success and failure as binary target variable for classification matching the paper	Paper builds binary classifier for weaning success

The SQL script closely mirrors the methodology and dataset preparation outlined in Liu et al. (2022), allowing for reproducible and explainable downstream ML modeling.

2.6. Reflection

Working with MIMIC-IV, even though it is a de-identified dataset, carries ethical responsibilities. My CITI certification on “Data or Specimens Only Research” was not merely a formal requirement, but a guiding framework that influenced the entire data pipeline.

I approached this project with attention to:

- Fairness: Avoiding overrepresentation by including only the final ICU stay per patient
- Data minimization: Extracting only clinically relevant variables
- Respect: Never attempt to identify individuals or link back to identities
- Scientific Integrity: Documenting my work for reproducibility and transparency

The inclusion/exclusion criteria ensured the population was ethically and clinically appropriate. I was mindful that de-identification doesn't eliminate risk, and I avoided unnecessary data merge or any manipulations that could compromise privacy.

In line with CITI training, I maintained:

- Transparency: All CTEs in SQL were structured and labeled
- Accountability: Consistent clinical logic across feature extraction
- Reproducibility: All filters, joins, and assumptions clearly stated

This certification reinforced the importance of respecting the dignity and rights of individuals, even when their identities are masked and informed my every decision from inclusion logic to the final CSV export.