SOFA Unpacked

Finding Which Parts of a Common ICU Score Treat Patient Groups Differently

Applicant	Suleyman Akkaya
Track	Computational Medicine
Submission Date	July 2026

1 · Problem

SOFA is a score doctors use in the ICU to measure how sick a patient's organs are. It is built from six parts: lungs, blood clotting, liver, blood pressure, kidneys, and brain function. Past studies have already shown that the overall SOFA score does not predict death equally well for every patient group, it tends to be less accurate for some racial and age groups. But no study has looked inside the score to ask which of the six individual parts is causing that gap. This project looks inside the score instead of just checking the final number.

2 · Approach

This project breaks the SOFA score into its six separate parts and checks whether each part predicts death the same way across different patient groups. A simple, easy to understand model, logistic regression, is trained using the six SOFA parts as inputs instead of the single combined score. The model is then compared across groups to see which of the six parts is weighted differently or performs differently for each group. For example, the kidney score might matter a lot more for predicting death in older patients than in younger ones, even if the total SOFA score looks the same.

The model is simple on purpose. The goal is not to build the most accurate model possible, it is to use a model that is easy to read so the individual parts driving any group differences can be clearly explained.

Why this is new: existing studies already show the total SOFA score is less fair for some groups. None of the ones found so far break the score apart to explain which specific organ system is responsible. That is the open question this project answers.

Is this really machine learning? Yes. Logistic regression is a standard machine learning model. It is used here instead of something bigger because it gives a clear, readable number for how much each of the six SOFA parts matters, called a coefficient. A bigger model like a neural network would hide that detail instead of showing it, which would work against the actual goal of this project.

3 · Dataset

Source: MIMIC-IV, a public ICU dataset from PhysioNet. Access is free but requires a short training course and application, which should be started right away since it can take a couple weeks to be approved.

Who is included: adult ICU patients, first ICU stay only, staying between 1 and 30 days. SOFA scores and their six individual parts are already calculated in MIMIC-IV, so this does not need to be built from scratch.

Size: MIMIC-IV covers roughly 50,000 to 90,000 ICU stays depending on the exact filters applied. After removing patients with missing group information or missing SOFA parts, a working cohort in the tens of thousands of patients is expected, which is more than enough for this kind of group comparison.

Groups compared: age group and race or ethnicity, since those are the groups where past studies already found the overall SOFA score works less well.

Cleanup needed: match each patient's SOFA parts to whether they survived the hospital stay, remove patients missing group information or missing SOFA parts, and organize everyone into their comparison groups.

4 · Metrics

- How well each of the six SOFA parts predicts death on its own, measured separately for each patient group
- Whether the simple model weighs each part differently depending on group
- A basic statistical check to make sure any difference found is real and not just random chance

Success means being able to clearly say which specific SOFA part behaves differently across groups, backed by a real statistical check, not just a guess based on the numbers looking different.

5 · Compute

- Training time: the logistic regression model trains in seconds, not hours, since it only uses six input numbers per patient
- Memory: the full dataset fits easily in a few gigabytes of RAM, well within a normal laptop or a free cloud notebook
- Dependencies: Python, with standard, free libraries only (pandas for data handling, scikit-learn for the model, matplotlib for charts)
- No GPU is needed at any point in this project

6 · Deliverables

- Model: the trained logistic regression model and the code to rebuild it, shared on GitHub
- Paper: a short written report explaining which SOFA part drives the biggest group differences and why that matters
- Demo: a simple notebook or set of charts showing, for each patient group, how much each SOFA part is weighted, so the result can be seen at a glance without reading code

7 · Risk

What could go wrong	What to do instead
PhysioNet approval takes a while	Apply for access right away, before anything else

One group has too few patients to compare fairly	Combine it with a similar group and be upfront about the smaller sample size
No clear difference shows up between groups	Still a useful, honest answer worth writing up, not a failed project
Trying to compare too many groups or add extra models	Stick to two group comparisons and one simple model as the full plan, anything more is optional later

---

This project checks an existing clinical tool using past, fully anonymous patient data. It does not change how any patient is treated and does not claim to fix anything on its own. The goal is a clear, honest answer about which part of a widely used score may need a closer look.