**Data Wrangling Documentation**

**Capstone Project 1: Predicting Patient Clinical Deterioration at ICU Admission**

Abebual Demilew

**Objective**

We have all seen it. A patient’s health drastically deteriorates at Intensive Care Unit (ICU) and the rapid response team scrambles to save the patient. Currently, most hospitals in the United States use an Early Warning Score (ESW) assessment system to detect clinical change in patient’s clinical instability. The earliest, most ESW are predictors of clinical deterioration 2-day after ICU admission. We aim to develop a machine learning system that can predict patient’s clinical deterioration more accurately, earlier (within 24 hours) and based on fewer data points. This can lead to a significant life-saving difference by helping to prevent critical events before they happen. Furthermore, this will lead to cost-saving to the hospital from improved productivity and a reduction in human error.

**Data**

According to the Centers for Disease Control and Prevention (CDC) nearly all hospitals and 86% of office-based physicians in the United States use Electronic Medical Records/Electronic Health Records (EMRs/EHRs)[[1]](#footnote-1)[[2]](#footnote-2). This healthcare organizations are our potential clients and have the data required to implement our machine learning model.

For this project, I used publicly available EHRs datasets. The MIT Media Lab for Computational Physiology has developed MIMIC III dataset based on over forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center of Boston between 2001 and 2012. MIMIC-III dataset is freely available to researchers across the world[[3]](#footnote-3). Prior to requesting access to MIMIC-III dataset, I completed a required course on human research ‘Data or Specimens Only Research’[[4]](#footnote-4). I have full access to the MIMIC III datasets.

The dataset has 26 relational tables including patient’s hospital admission, callout information when patient was ready for discharge, caregiver information, electronic charted events including vital signs and any additional information relevant to patient care, patient demographic data, list of services the patient was admitted or transferred under, ICU stay types, diagnoses types, laboratory measurments, microbiology tests and sensitivity, prescription data and billing information.

**Data Wrangling**

The first challenge was working with large datasets. There are 26 relational tables and some of the tables have over 330 million rows (The chartevents table) and the lab events table has over 27 million rows. These are charts events and lab events of 46,520 patients admitted at the Beth Israel Deaconess Medical Center’s ICU between 2001 and 2012. I used Google Colab[[5]](#footnote-5) to use free GPU to load the massive dataset and take a random sample of 5000 patients (SUBJECT\_ID uniquely identifies patients). All admissions, ICU stays, and events (lab, microbiology, charts and other events) corresponding to the 5000 sampled patients were selected for the project. As a result, we have a sample of around 12 million rows of events and 106 thousand rows of ICU stays

Second, the various relational sampled data tables were merged using linking variables. For this purpose, various merging functions were defined to merge tables linked on one or more of the following variables SUBJECT\_ID (uniquely identifies individual patient), HADM\_ID (uniquely identifies individual admission to the hospital), ICUSTAY\_ID (uniquely identifies individual ICU admission), ITEMID (uniquely identifies individual event ), ICD9\_CODE (international classification of diseases version 9 codes for procedures), and CGID (uniquely identifies caregivers).

Third, the following columns were transformed or created or dropped.

* Datetime columns were stored as objects data type, all such arguments are converted to datetime.
* Added age at the time of ICU admission column based on Date of Birth (DOB) and ICU admission (INTIME).
* Added age group column with value labels ‘0’ for patients age below 18 and ‘1’ for age 18 and above. Our analysis will focus mainly on adults age 18 and above.
* Added in ICU unit mortality to accurately measure mortality during ICU stay and separate it from overall in hospital mortality or mortality outside of the hospital (outside mortality was captured form Social Security).
* Transformed gender, ethnicity, and diagnosis columns. ICD9\_CODE was used to transform diagnosis column.
* Transformed ‘ERROR’ in events dataset to NaN. Sometimes Glucose PH have ERROR as value.
* Transformed temperature values mapped in Fahrenheit to Celsius.
* Transformed weight values mapped in ounces to pounds. Sometimes ambiguous values, for example light adults (< 40 lb) or heavy adults (>450 lb) were cleaned.
* Transformed events datasets to timeseries.
* Removed ICU stays that are transferred to another hospital.
* Removed multiple ICU stays per individual hospital admission.
* Removed outliers for events columns using valid low and valid high ranges. Values outside of the ranges were removed.

1. https://www.cdc.gov/nchs/fastats/electronic-medical-records.htm [↑](#footnote-ref-1)
2. https://academic.oup.com/ajhp/article-abstract/74/17/1336/5102661?redirectedFrom=fulltext [↑](#footnote-ref-2)
3. <https://mimic.physionet.org/gettingstarted/access/> [↑](#footnote-ref-3)
4. Verify at: www.citiprogram.org/verify/?kb6607b78-5821-4de5-8cad-daf929f7fbbf-33486907 [↑](#footnote-ref-4)
5. <https://colab.research.google.com/notebooks/welcome.ipynb> [↑](#footnote-ref-5)