

**ARTIFICIAL INTELLIGENCE PROGRAMMING PROJECT**

**Report 3 – Data Collection**

**Group 3: Apply attention-based model in Healthcare Representation Learning and Diagnosis Prediction**

– Hanoi, May 2021 –

**Table of Contents**

[III. Data Collection 3](#_Toc68809802)

[1. Methodology 3](#_Toc68809803)

[2. Data description 3](#_Toc68809804)

[3. Data Exploration 3](#_Toc68809805)

# III. Data Collection

## Methodology

MIMIC-III is a open dataset for international researchers with a credential access to use in their clinical studies. It integrates deidentifed, comprehensive clinical data of patients admitted to the Beth Israel Deaconess Medical Center in Boston, Massachusetts.

The MIMIC-III database was populated with data that had been acquired during routine hospital care, so there was no associated burden on caregivers and no interference with their workflow. Data was downloaded from several sources, including:

+ Archives from critical care information systems

+ Hospital electronic health record databases

+ Social Security Administration Death Master File

Two different critical care information systems were in place over the data collection period: Philips CareVue Clinical Information System, and iMDsoft MetaVision ICU. These systems were the source of clinical data such as:

+ Time-stamped nurse verified physiological measurements (ex: hourly documentation of heart rate, arterial blood pressure, or respiratory rate)

+ Documented progress notes by care providers

+ Continuous intravenous drip medications and fluid balances

**Out-of-hospital mortality dates** were obtained using the Social Security Administration Death Master File.

**Protected health information** was removed from free text fields, such cas diagnostic reports and physician notes, using a rigorously evaluated deidentification system based on extensive dictionary look-ups and pattern-matching with regular expressions. The components of this deidentification system are continually expanded as new data is acquired. [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.

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. [1]

## Data Exploration

I choose this data for 3 main reason: it is freely available to researchers worldwide; it encompasses a diverse and very large population of ICU patients; and it contains highly granular data, including vital sign, laboratory results, and medications.

In this project, we will make use of the following MIMIC III tables

+ ADMISSIONS – a table containing admission and discharge dates (has a unique identifier HADM\_ID for each admission)

+ NOTEEVENTS – contains all notes for each hospitalization (links with HADM\_ID)

To maintain anonymity, all dates have been shifted far into the future for each patient, but the time between two consecutive events for a patient is maintained in the database. This is important as it maintains the time between two hospitalization for a specific patient.

1. **References**

Charles, D., King, J., Patel, V. & Furukawa, M. Adoption of Electronic Health record Systems among U.S. Non-federal Acute Care Hospitals. ONC Data Brief No. 9, 1–9 (2013).

Collins, F. S. & Tabak, L. A. NIH plans to enhance reproducibility. Nature 505, 612–613 (2014).