# Explainable artificial intelligence model to predict mortality in patients presenting with acute coronary syndromes

Edwin Kagereki - B00867154

17 October 2021

## Contents



## Business Understanding

### Introduction

Cardiovascular diseases (CVDs), principally ischemic heart disease (IHD) and cardiac stroke, are the leading cause of global mortality. In addition to their increasing global prevalence, they are often associated with poor survival(Roth 2020).

Acute Coronary Syndrome (ACS) is a term given to a continuum of CVDs ranging from ST-segment elevation myocardial infarction (STEMI) to non–ST-segment elevation myocardial infarction (NSTEMI) and unstable angina. Accurate estimation of risk for untoward outcomes after a suspected onset of an ACS may help clinicians chose the type and intensity of therapy. Such predictions may be helpful as patients predicted to be at higher risk may receive more aggressive surveillance and/or earlier treatment, while patients predicted to be at lower risk may be managed less aggressively.

**Problem statement** The establishment of prognosis model for patients with suspected ACS is important in critical care medicine.Numerous risk-prediction models for differing outcomes exist for the different types of ACS.

These models however have some limitations. First, most models have been developed from large randomized clinical trial populations in which the generalizability to risk prediction in the average clinician's experience is questionable(Eagle KA Lim MJ 2004).

Second, a model given the dynamic nature of the treatment environment, predicting future behavior while the treatment is underway may help the clinicians make more. Predictive models often perform calculations during live transactions—for example, to evaluate the risk or opportunity of a given customer or transaction to guide a decision.

This project aims at developing a risk-prediction tool for ACS, focusing on clinical end point of all-cause mortality. The motivation is to determine the utility of machine learning at the population level using

multiple linked ICU patients' datasets.

The learned representations of this comprehensible surrogate model, namely the decision trees and consequently extracted rules are then provided to the domain experts who aim to justify the decision for individual instances.

### Objectives

This project will develop a tool for application in the decision-making environment. This is done in two steps:

1. A binary classification model developed. This model will use:

- patient demographics
- Interventions within the golden hour (laboratory, medication, procedures and microbiology studies).

The concept of the golden hour refers to the vital period by which a patient with a suspected cardiovascular event should be receiving definitive treatment to prevent death or irreparable damage to the heart.Although not set in stone, the chances to save a patient are usually high if substantive medical attention is given within an hour of the cardiac event(Johnson 2016). In this study the golden hour cut-off was 60 minutes after initial contact with the hospital. This included all interventions that were done prior to admission.

2. Process mining will be used to explain the outcome of the patient based on the care pathway followed. Three aspects will be checked:

- Process discovery - Processes followed by the two classes will be described.
- Conformance checking - The care pathway followed by both patients will be checked for conformance with the American Heart Association(AHA) guidelines for CPR and ECC (ACLS 2020)

To run the process mining the timestamped intervention in the AHA guidelines are (laboratory, medication, procedures) are used.

### Success Criteria

The success of this project is to generate forward-looking, predictive insights to improve the management of the care pathway in patients suspected to have ACS by:

- Successfully predicting the ICU mortality outcome of the patient with suspected ACS based on the patient demographics and the interventions given within the first one hour.
- Identify patients in which undesirable events will likely be observed in the based on the interventions.

## Situation Assessment

This task involves more detailed fact-finding about all of the resources, constraints, assumptions, and other factors that should be considered in determining the data analysis goal and in developing the project plan.

### Inventory of resources

**Software** For this project the following software will be used:

1. SQL database server.
2. R
3. Python

The analysis will be done on a Windows desktop and a Linux server.

### Sources of Data and Knowledge

**MIMIC III Database** MIMIC-III is a large, freely-available database comprising deidentified health-related data associated with 46,520 patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. The database includes information such as demographics, vital sign measurements made at the bedside , laboratory test results, procedures, medications, caregiver notes, imaging reports, and mortality (including post-hospital discharge).

Although de-identified, the datasets described herein contain detailed information regarding the clinical care of patients, and as such it must be treated with appropriate care and respect.

Researchers seeking to use the database must:

1. Become a credentialed user on PhysioNet. This involves completion of a training course in human subjects research.
2. Sign the data use agreement. Adherence to the terms of the DUA is paramount.

**LOIC tables** The Loin(Institute 2021) data tables will be used to enrich the laboratory dataset.

**American Heart Association guidelines for CPR and ECC** This AHA guideline(ACLS 2020) will be used as the gold standard for ACS workflow. The workflows for the patients in this project will be assessed for cornformity with this workflow.

### Requirements, Assumptions, and Constraints

For this analysis, the all the medical records were not analysed.

It is also assumed that:

- The patients in this population were only treated in this hospital, therefore mortalities are only captured in this hospital.
- All the pre-hospitalization interventions were captured.
- The unit of analysis is the admission.

### Risks and Contingencies

- Although the process mining will give a better predictive description of the patient outcomes, alternative surrogate modeling methods like decision tree maybe used.

### Terminology

1. **Acute Cardiac syndrome:** Candidates of acute cardiac syndrome were identified using the *DIAGNOSIS* in the *ADMISSIONS* table which provides a preliminary. This column was is a free text diagnosis for the patient on hospital admission.

*"stemi,""acute coronary syndrome,""angina,""tachycardia,""aortic aneurysm,""pericardi,""ortic dissection,""coronary artery dissection,""cardiomyopathy,""heart failure,""mitral valve disease,""mitral stenosis,""coronary artery disease,""chf,""congestive heart failure,""heart failure,""telemetry,""myocardial infaction,""cardiac arrest,""myocardial infarction,""aortic stenosis,""st elevated,""pericardial effusion," "cardiomyopathy,""cath lab,""tamponade,""tamponede"*

2. Angiotensin-converting enzyme (ACE). The following terms will be used to identify ACE's from the list of *DRUG* column of the *PRESCRIPTIONS TABLE* table:

*"benazepril," "captopril," "enalapril,""enalaprilat,""fosinopril," "lisinopril," "moexipril," "perindopril," "quinapril," "ramipril,""trandolapril"*

3. Beta blockers (beta-adrenergic blocking agents) The following terms will be used to identify Beta blocker from the list of *DRUG* column of the *PRESCRIPTIONS TABLE* table:

*"acebutolol,""atenolol,""betaxolol,""bisoprolol,""carteolol,""carvedilol,""labetalol,""metoprolol,""nadolol,""nebivolol,""penbutolol,""pindolol,""propanolol,""sotalol,""timolol"*

4. Glycoprotein IIb/IIIa inhibitors. The following words will be used to identify Glycoprotein IIb/IIIa inhibitors's from the list of *DRUG* column of the *PRESCRIPTIONS TABLE* table:

*"abciximab,""eptifibatide,""tirofiban,""roxifiban,""orbofiban"*

5. P2Y12 inhibitors. The followeing words were used to identify the P2Y12 inhibitors from the list of *DRUG* column of the *PRESCRIPTIONS TABLE* table:

*"clopidogrel,""prasugrel,""ticlopidine,""ticagrelor"*

6. HMGCoA The following words were used to identify HMGCoA from the list of *DRUG* column of the *PRESCRIPTIONS TABLE* table: *"altoprev,""amlodipine,""atorvastatin,""caduet,""crestor,""ezallor,""fluvastatin,""lescol,""l*

## Data Mining Goals

1. Build a binary classification machine learning model to predict all-cause mortality of patients based on demographics and interventions within first hour of suspected ACS event.
2. Compare the conformity of care path in patients who died and patients who survived with the with the ACLS care path.
3. Assess and report any work flow variants, and differences between patients who died and those who survived.

### Data Mining Success Criteria

After training the binary classifier, evaluation measures will be used to assess the performance of the model. The predictive performance of the classifier will be assessed by calculating the number of correctly identified class patients (true positives), the number of correctly recognized patients that are not member of the class (true negatives), the number of the patients that are wrongly recognized (false positives) and the number of the examples that were not identified (false negatives). By using these measures, a confusion matrix will be constituted.

| | | Ground Truth Values | |
|---|---|---|---|
| | | Positive | Negative |
| Predicted Values | Positive | true positive (tp) | false positive (fp) |
| | Negative | false negative (fn) | true negative (tn) |

The following measures will be calculated from this table:

- Accuracy.
- Precision.
- Recall.
- Specificity.

# Data Understanding

## Data collection

When allowed access to the MIMIC-III database, I transferred all of this information into a RDMS (relational database management system). This was done with the help of open source scripts(Ganas 2018). After connecting to the PostgreSQL database, I was able to make SQL queries and connect my database and access the data that I was interested in. Although the database includes 26 tables, only the following tables will be included in the analysis:

- **ADMISSIONS:** Contains information regarding a patients admission to the hospital. Information available includes timing information for admission and discharge, demographic information, the source of the admission, and so on. Record of 58,976 unique admissions.
- **PATIENTS:**Defines each patient in the database, i.e. defines a single patient. There are 46,520 patients recorded.
- **SERVICES:**Lists services that a patient was admitted/transferred under.This table contains 73,343 entries.
- **DIAGNOSIS_ICD**Identify type of data sources (online sources, experts, written documentation, etc.)
- **MICROBIOLOGYEVENTS:**Contains microbiology information, including cultures acquired and associated sensitivities.There are 631,726 rows in this table.
- **PRESCRIPTIONS:**Contains medication related order entries, i.e. prescriptions.This table contains 4,156,450 rows.
- **PROCEDUREEVENTS_MV:**Contains procedures for patients. This table has 258,066 rows.
- **D_ITEMS:** Definition table for all 12,487 items in the ICU databases.
- **D_LABITEMS:** Definition table for 753 laboratory measurements.

#### Initial Data Collection Report

Due to the complexity of the MIMIC-III database and the massive size of the available source data (over 100 gigabytes), an SQL database was created. This was created using SQL scripts to load the MIMIC-III data into a SQL Server 2016 database.(Ganas 2018)

Respective tables were queried, tables joined and the some columns dropped. Flat files(csv) were saved for further analysis.

#### Data quality in flat files

- 6 csv files were saved for further analysis. These will be joined for the modeling.
- Only number of fields to be used were retained.

#### Rationale for Inclusion/Exclusion

- All the clinical interventions will be included in the analysis.
- Although there were some tables with some data on the interventions, they were excluded because:

1. TEXT is often large and contains many newline characters were excluded because of the complexity of analyzing the data.
2. Echo reports, ECG reports, and radiology reports were excluded because of the complexity of the data.

## Data overview

Once the data was accessed and summary analysis was done. The patients details are summarized in table 2. The other variables were:

1. Procedures:
2. Laboratory:
3. Medicine:
4. Microbiology:
5. Services:

Each of these table was merged with the demographics dataset to form a final csv that has

Table 2: **Patient Characteristics**

| **Characteristic** | **Deceased**, N = 3,237 | **Survived**, N = 7,982 |
|---|---|---|
| **GENDER, n (%)** | | |
| F | 1,423 (44%) | 2,843 (36%) |
| M | 1,814 (56%) | 5,139 (64%) |
| **AGE, Mean (SD)** | 74.90 (12.16) | 66.26 (13.53) |
| **INSURANCE, n (%)** | | |
| Private | 465 (14%) | 2,835 (36%) |
| Public | 2,762 (85%) | 5,107 (64%) |
| Self Pay | 10 (0.3%) | 40 (0.5%) |
| **MARITAL_STATUS, n (%)** | | |
| | 211 (6.5%) | 225 (2.8%) |
| Living alone | 1,455 (45%) | 3,099 (39%) |
| Living with Partner | 1,533 (47%) | 4,624 (58%) |
| UNKNOWN (DEFAULT) | 38 (1.2%) | 34 (0.4%) |
| **Ethinicity, n (%)** | | |
| ASIAN | 42 (1.3%) | 147 (1.8%) |
| BLACK | 186 (5.7%) | 593 (7.4%) |
| CAUCASIAN | 2,224 (69%) | 5,736 (72%) |
| HISPANIC | 53 (1.6%) | 249 (3.1%) |
| OTHER | 47 (1.5%) | 187 (2.3%) |
| UNKNOWN | 685 (21%) | 1,070 (13%) |

# References


ACLS. 2020. "Acute Coronary Syndromes Algorithm - ACLS Version Control: This Document Follows 2020 American Heart Association Guidelines for CPR and ECC. American Heart Association Guidelines Are Updated Every Five Years." https://www.acls.net/images/algo-acs.pdf.

Eagle KA Lim MJ, et al., Dabbous OH. 2004. "A Validated Prediction Model for All Forms of Acute Coronary Syndrome: Estimating the Risk of 6-Month Postdischarge Death in an International Registry." *Jama*. https://doi.org/doi:10.1001/jama.291.22.2727.

Ganas, Spiro. 2018. "MIMIC-on-SQL-Server." *GitHub Repository*. GitHub. https://github.com/SpiroGanas/MIMIC-on-SQL-Server.

Institute, Regenstrief. 2021. "LOINC." *LOINC*. https://loinc.org/downloads/.

Johnson, et al, A. 2016. "MIMIC-III Clinical Database." https://doi.org/https://doi.org/10.13026/C2XW26.

Roth, et al, Gregory A. 2020. "Global Burden of Cardiovascular Diseases and Risk Factors, 1990-2019: Update from the GBD 2019 Study." *Journal of the American College of Cardiology* 76 (25): 2982–3021. https://doi.org/10.1016/j.jacc.2020.11.010.