# CS598 DLH Final Project Spring 2022

Reproduce Paper:
INPREM: An Interpretable and Trustworthy Predictive Model for Healthcare

Nadia Wood (nadiaw2) & Diana Gonzalez Santillan (dianag4)

Github: https://github.com/nadiawoodninja/CS598DLHFinalProject

# General Problem

The focus of this paper was to develop a model called INPREM. This model addresses the problem in healthcare ML models of interpretability and trustworthiness.

The authors claim that the model proposed will enable physicians to determine which input variables impact the predictions of the model, hence making it more trustworthy and interpretable.

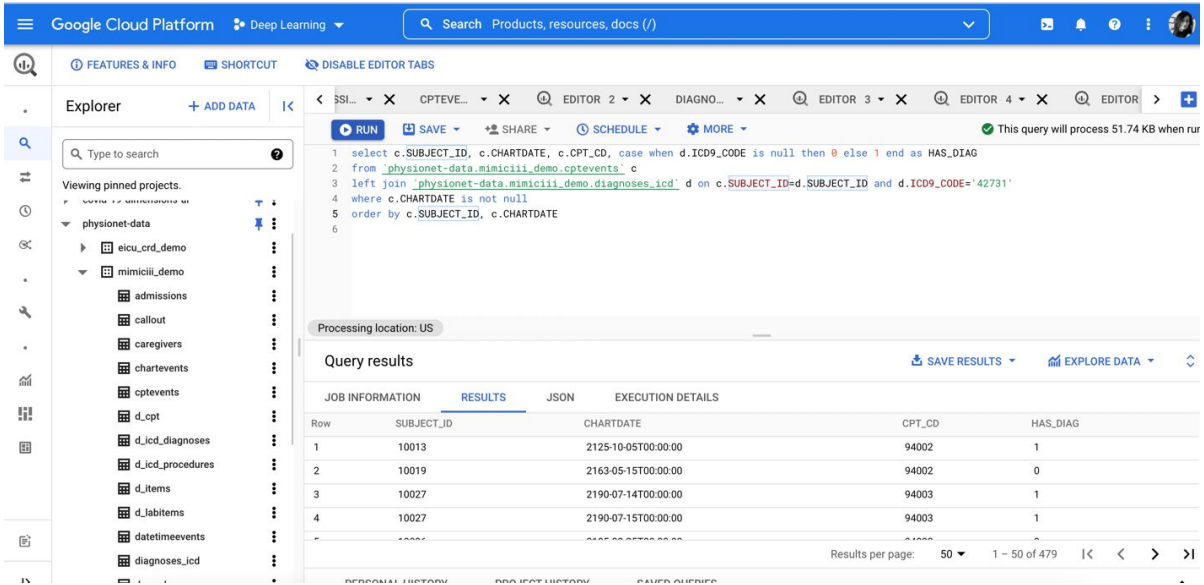
# Approach

Our approach to reproduce the paper included following steps:

- Understanding the code
  - “Mini” model code
- Data Wrangling
- Data Curation
- Data Cleansing
- Data formatting
- Data Profiling
- Changing the code to add missing logic
- Running the code

# Data Wrangling

- The paper required mimic iii and EHR longitudinal data
- Majority of time spent in determining how to gather data
- We exhausted options such as looking at other open data sources for EHR data
- We asked the authors for a data dictionary for the EHR data which they used, but were not provided with one.
- This limited us to just use mimic iii data
- In order to save time, we decided to use the BigQuery and the mimiciii demo data available in BigQuery




*Screen shot of bigquery and set up in google.*

# Data Curation

Our dataset included three files:

- **mimiciiiDemoData.csv**
  - This dataset contained data about the admission events
- **diagnosisCode.csv**
  - This dataset contained data about the ICD 9 codes related to the events
- **HeartFinalDataset.csv**
  - This csv file contained all patients with heart related events recorded in the mimic iii dataset.

To generate mimiciiiDemoData.csv we used the query below

```
SELECT * FROM 'physionet-data.mimiciii_demo.admissions
```

To generate diagnosisCode.csv we used the query below

```
SELECT * FROM 'physionet-data.mimiciii_demo.diagnoses_icd
```

To generate HeartFinalDataset.csv we used the query below

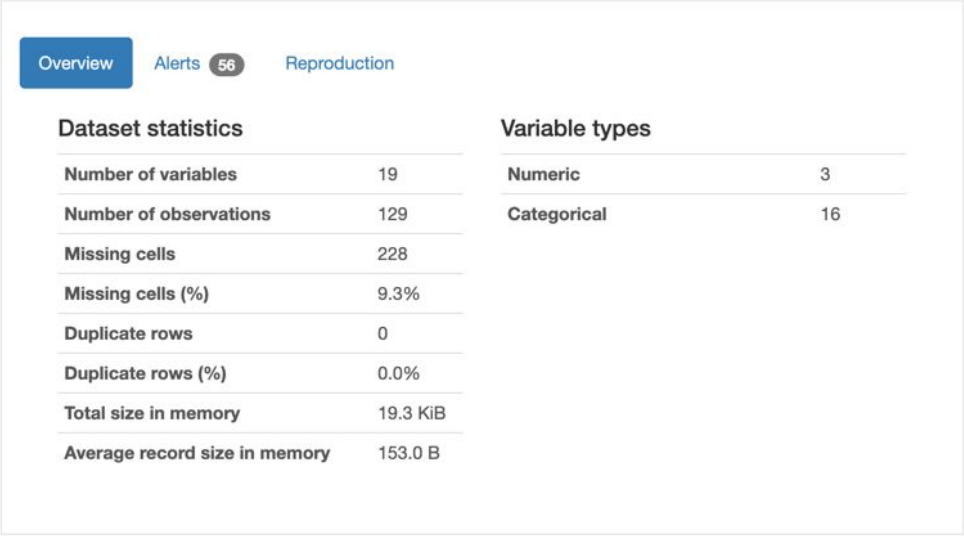
```
SELECT c.SUBJECT_ID, c.CHARTDATE, c.CPT_CD,
CASE WHEN d.ICD9_CODE IS null THEN 0 ELSE 1 END AS HAS_DIAG
FROM 'physionet-data.mimiciii_demo.cptevents' c
LEFT JOIN 'physionet-data.mimiciii_demo.diagnoses_icd' d
ON c.SUBJECT_ID=d.SUBJECT_ID and d.ICD9_CODE='42731'
WHERE c.CHARTDATE IS NOT null
ORDER BY c.SUBJECT_ID, c.CHARTDATE
```

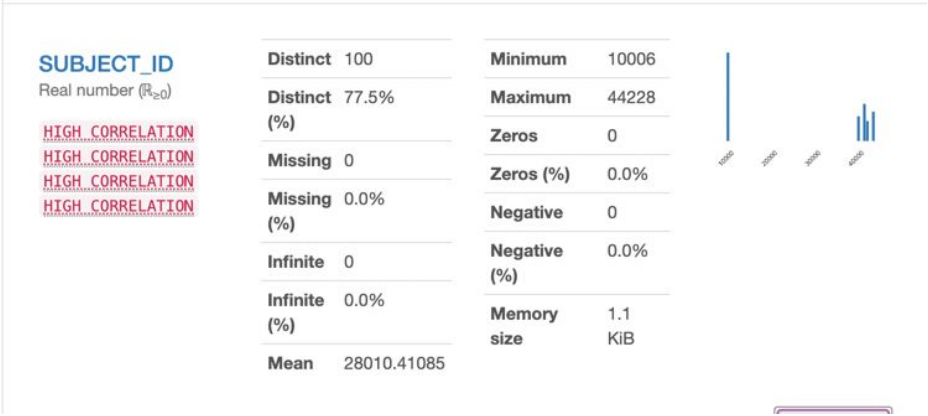
*Queries used to generate the dataset from mimic iii dataset in bigquery*

# Data Profiling

## Mimiciii Demo Admissions Data Stats

- Using panda-profiling we were able to get a sense of the data and statistics about this dataset.
- This was a dataset for us to get an understanding of what the admissions data looks like.
- We did not use this dataset in the model but it gave us insights into what columns made sense to include in the final two datasets
- Important insight was sampling size of patients
- Distinct sampling of 100

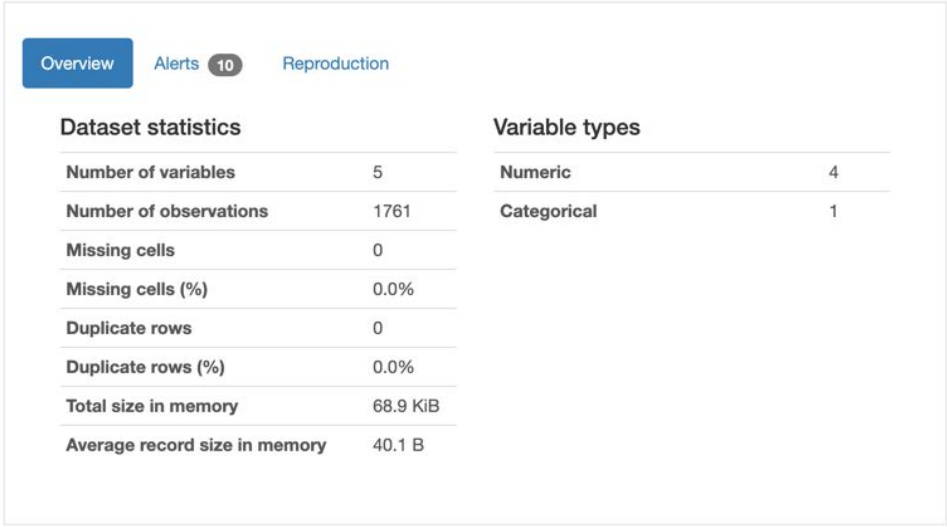


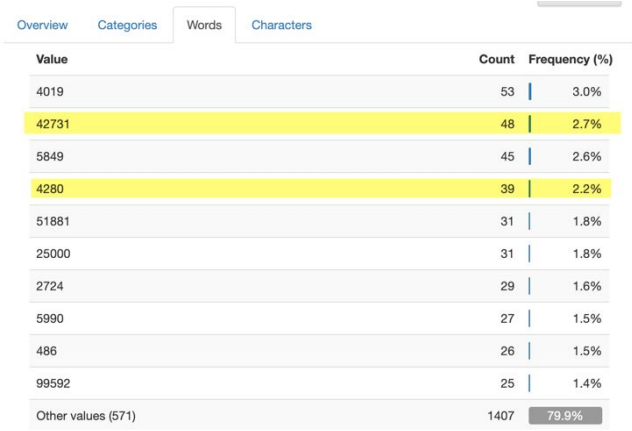



# Data Profiling

## Diagnosis Code Data Stats

- Using panda-profiling we were able to get a sense of the data and statistics about this dataset.
- This dataset was the dataset used in the model and contained information about the diagnosis events.
- There were 1761 observations in the dataset with no duplicate rows
- There were 581 distinct ICD9 codes in the dataset
- ICD9 42731, 4280  are heart disease, related code which gave us about 87 records



# Reproduction Attempts

Reproduction was tough given that we did not have access to EHR data that the authors used.

# Reproduction attempts

- There were two major roadblocks for reproducibility
  - Having “MINI” model of INPREM
  - Not having data dictionary of EHR data used in the model
  - Not having EHR longitudinal data
- There were several attempts to ask the authors for data dictionary so that we can reproduce the datasets but to no avail
- We ended up using mimic iii demo data to produce accuracy and F1 results of the model and compared it to results presented in the model

# Results

- The accuracy results, as you can see given the small dataset, is not conclusive to prove or disprove the authors claim of accuracy
- In case of 10 epochs, our experiment scored higher accuracy than the author's baseline and INPREM best case
- Table 2 shows the hyperparameters used and the accuracy and F1 scores
- We decided to use F1 scores to take a weighted average of precision and recall.
- As illustrated, given the hyperparameters, epoch 10 experiment out performed the best compared to baseline and INPREM best cases
- We hypothesis that this is due to small dataset

| Number of epochs: | 5 | 10 | 15 | 20 | 25 | 30 |
|---|---|---|---|---|---|---|
| Baselines Best Case: | 0.6399 | 0.5840 | 0.6318 | 0.7018 | 0.7687 | 0.8212 |
| INPREM Best Case: | 0.6902 | 0.6253 | 0.6626 | 0.7306 | 0.7890 | 0.8314 |
| Our Model: | 0.4286 | 0.8900 | 0.7143 | 0.5714 | 0.5714 | 0.5714 |

Table 1: Comparison of ACCURACY results for section 4.2

| Epocs | Batch Size | Drop Rate | Learning Rate | Weight Decay | Accuracy | F1 Score |
|---|---|---|---|---|---|---|
| 5 | 32 | 0.5 | 0.0005 | 0.0001 | 43% | 0.33 |
| 10 | 32 | 0.5 | 0.0005 | 0.0001 | 86% | 0.89 |
| 15 | 32 | 0.5 | 0.0005 | 0.0001 | 71% | 0.75 |
| 20 | 32 | 0.5 | 0.0005 | 0.0001 | 57% | 0.73 |
| 25 | 32 | 0.5 | 0.0005 | 0.0001 | 57% | 0.67 |
| 30 | 32 | 0.5 | 0.0005 | 0.0001 | 57% | 0.40 |

Table 2: Hyperparameters with accuracy and F1 score

# Potential Next Steps

1. If we had more time and the appropriate dataset, we would have loved to compare the results with Keras using tensorflow
2. If given access to full model code and the longitudinal EHR dataset, we will be able to uncover more insights
3. Other potential steps would be to add clinical notes and some NLP to contextualize the model with text data