An Extensive Data Processing Pipeline for MIMIC-IV Paper Reproduction Project Report

Nita Agarwal
nitaa2@illinois.edu

Group ID: 40
Paper ID: 62
Presentation link: https://mediaspace.illinois.edu/media/t/1_uwy57wzk
Code link: https://github.com/NitaAgarwal2022/DLH_project

1 Introduction

In this project I aim to replicate the main experiment of the paper "An Extensive Data Processing Pipeline for MIMIC-IV" by Mehak Gupta, Brennan Gallamoza, Nicolas Cutrona, Pranjal Dhakal, Raphael Poulain, Rahmatollah Beheshti (Gupta et al., 2022) .

The general problem the original paper aims to solves is to fill the gap of a standardized, flexible, customizable data pipeline to extract and pre-process the data available in MIMIC-IV dataset.

The number of clinical research to apply machine learning models/methods to EHR (electronic health record) data has increased in last few years.MIMIC-IV (Johnson et al., 2021; Goldberger et al., 2000 (June 13) is a public, free and widely used EHR dataset. However, the researchers have to extract and pre-process this dataset on their own before starting to utilize it in the machine learning programs. The existing cohort extraction and pre-processing pipelines are missing the flexibility of letting the user interact and decide as per their need of the type of cohort required or are processing only ICU data(and not the non-ICU data). The approach taken by the original paper to address this problem is by providing a configurable framework for data extraction, cleaning and pre-processing pipleine which is extendible, flexible, processes both ICU and non-ICU data.

I find this approach taken by the authors as new, interesting and challenging at the same time. Providing flexibility to the user to define their own cohort is challenging and also most rewarding as it will reduce the time needed to clean preprocess the dataset. It will thereby increase the usability of MIMIC-IV bymaking it more accessible to researchers.I am excited to be working on this project.

2 Scope of reproducibility

Data extraction and data pre-processing are the central contribution of the paper. The claims from the paper that I plan for the reproduction study are as below

- Pipleline can extract cohort features from MIMIC-IV based on user input of feature and on specific condition of the disease chosen by the user.
- Pipleline can extract cohort for prediction task of mortality in ICU patients.
- Pipleline can pre-process cohort from MIMIC-IV dataset for clinical grouping and conversion based on user input along with providing summary of the resultant cohort.
- Pipleline can pre-process cohort for outlier imputation and selection of feature based on user inputs.
- Pipleline can pre-process cohort by binning the sequential data into time intervals based on the timeseries length according to user inputs.

3 Methodology

3.1 Model descriptions

The original data processing pipeline consists of the Data Extraction unit and Data Preprocessing unit as shown in Figure 1. Data Extraction unit takes user input for version,type of data(ICU/non-ICU), specific cohort condition (ex. disease) and feature extraction. It then extracts and stores the generated cohort in form of csv files. Data Preprocessing unit clinically groups the data, selects the features, cleans the raw data by removing outliers, impute missing entries and also generate time-series dataset by binning the sequential data into

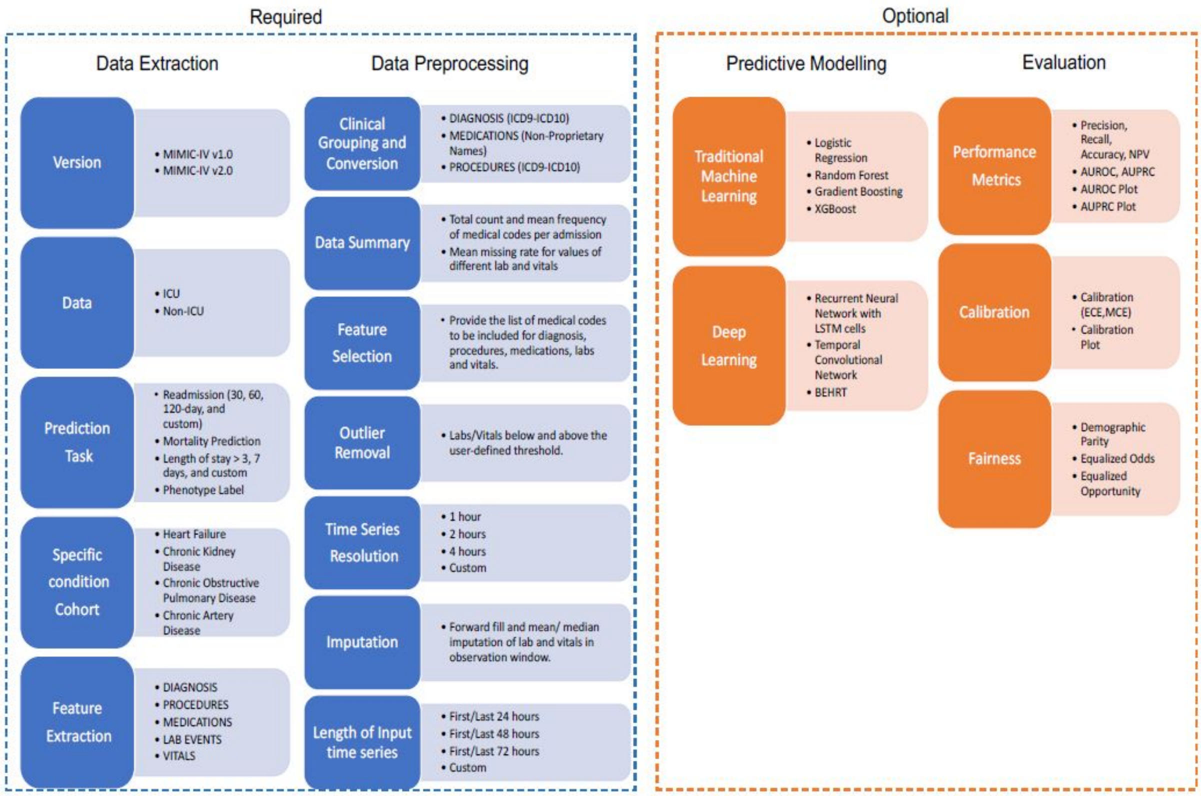


Figure 1: Pipeline overview.The left two parts(in blue) are required to produce the processed data

time intervals as per user's preference. The pipeline also includes two optional parts for modeling and evaluation.Users can construct a model to train for a prediction task and use performance metrics for study comparisons.

3.2 Data descriptions

The original paper uses MIMIC-IV dataset present at `https://physionet.org/content/mimiciv/1.0/`.This is a restricted-access resource. To access the files, I completed the the below requirements:

- Became a credentialed user by filling up the application for Credentiated Access using illinois email id and getting the request approved
- Completed required training:CITI Data or Specimens Only Research and submitting the result at `https://physionet.org/settings/training`
- Signed the data use agreement for the project

After receiving the access to data, the files can be downloaded from the terminal using below cmd: wget -r -N -c -np –user nitaa2 –ask-password https://physionet.org/files/mimiciv/1.0/

Distinct patients: 257,000
Admission records: 524,000
Data collection timeline: 2008 to 2019
Overall size: 6.9 GB of compressed files
Hosp(data derived from the hospital wide EHR): 4.4 GB (approx) compressed files
Core(patient tracking information necessary for any data analysis using MIMIC-IV data): 70 MB compressed files
ICU(data sourced from the clinical information system at the BIDMC): 2.5 GB (approx) compressed files

3.3 Hyperparameters

As data extraction and data pre-processing are the central contribution of the paper, I did not use any hyperparameter in the data processing pipeline.

3.4 Implementation

For this paper reproduction, I am re-using most of the author's code (`https://github.com/healthylaife/MIMIC-IV-Data-Pipeline`). I have added the code which enabled me to run the data pipleline in my environment, such as mounting the code and the MIMIC-IV data on google drive, installing the python libraries required for

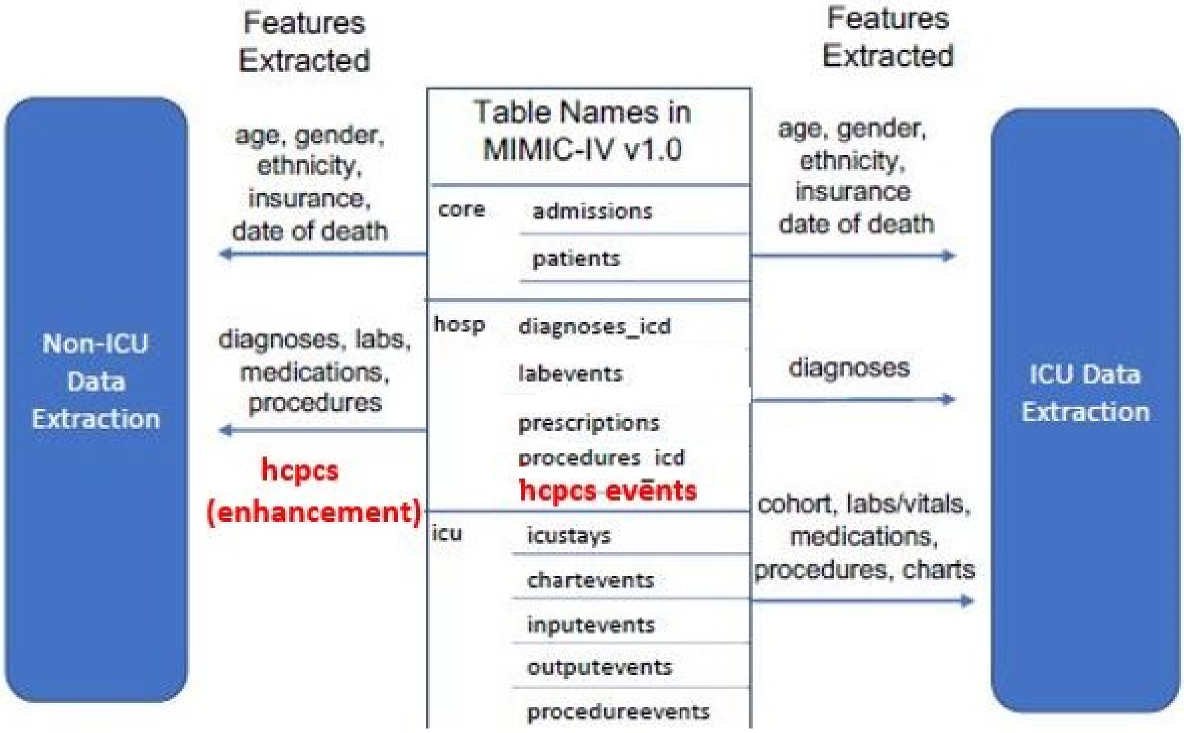


Figure 2: Extracted features

executing the code.

I have **implemented the code to enhance the original data processing pipeline** (https://github.com/NitaAgarwal2022/DLH_project).

- I have added feature extraction logic for hcpcs events (https://physionet.org/content/mimiciv/1.0/hosp/hcpcsevents.csv.gz). The code works for both 1.0 and 2.0 version of MIMIC-IV data. Feature extraction enhancement is shown in Figure 2
- I have added code for extracting data for additional disease such as Cancer.
- I have added the code to provide user with a flexibility to choose to start the processing from the already extracted intermediate cohort. This will skip few of the steps in the processing pipeline thereby taking less processing time to reach to later steps.

3.5 Computational requirements

The volume of the data is very high as it is processing the complete dataset of MIMIC IV based on the cohort selection. Replicating the work in the paper is computationally feasible. I have taken the paid version Google Colab pro+ as it gives 500 compute units with faster GPUs , enabled background processing and more memory is helpful.

Hardware:

Operating system: Windows 10, RAM: 12 GB, Disk space: 225 GB, Faster GPUs, 500 compute units **Software:** Pyhton: 3.7 import_ipynb:0.1.3 ipywidgets:7.5.1 numpy:1.18.5 pandas:1.0.5 tqdm:4.47.0

4 Results

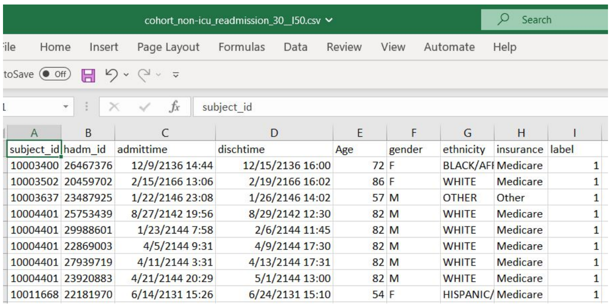
I am able to successfully run the original data processing pipeline and reproduce all the claims in section 2. Please find the details of the results below

4.1 Result 1

I am able to successfully extract cohort features from MIMIC-IV based on user input of feature(non-icu, readmission of 30 days) and on specific condition of the disease(Heart Failure) chosen by the user. The records are stored in compressed format csv file as shown in Figure 3 and **analysis** is in Figure 4

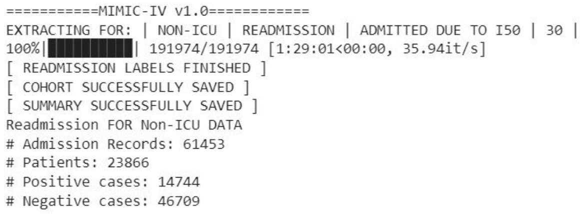
4.2 Result 2

I am able to successfully extract cohort for prediction task of mortality in ICU patients. The records



subject_id	hadm_id	admittime	dischtime	Age	gender	ethnicity	insurance	label
10003400	26467376	12/9/2136 14:44	12/15/2136 16:00	72	F	BLACK/AFI	Medicare	1
10003502	20459702	2/15/2166 13:06	2/19/2166 16:02	86	F	WHITE	Medicare	1
10003637	23487925	1/22/2146 23:08	1/26/2146 14:02	57	M	OTHER	Other	1
10004401	25753439	8/27/2142 19:56	8/29/2142 12:30	82	M	WHITE	Medicare	1
10004401	29988601	1/23/2144 7:58	2/6/2144 11:45	82	M	WHITE	Medicare	1
10004401	22869003	4/5/2144 9:31	4/9/2144 17:30	82	M	WHITE	Medicare	1
10004401	27939719	4/11/2144 3:31	4/13/2144 17:31	82	M	WHITE	Medicare	1
10004401	23920883	4/21/2144 20:29	5/1/2144 13:00	82	M	WHITE	Medicare	1
10011668	22181970	6/14/2131 15:26	6/24/2131 15:10	54	F	HISPANIC/	Medicare	1

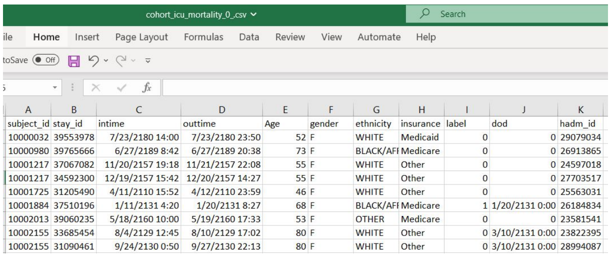
Figure 3: Extracted Cohort: Readmission Non-ICU Heart Failure patients



```
===========MIMIC-IV v1.0============
EXTRACTING FOR: | NON-ICU | READMISSION | ADMITTED DUE TO I50 | 30 |
100%|██████████| 191974/191974 [1:29:01<00:00, 35.94it/s]
[ READMISSION LABELS FINISHED ]
[ COHORT SUCCESSFULLY SAVED ]
[ SUMMARY SUCCESSFULLY SAVED ]
Readmission FOR Non-ICU DATA
# Admission Records: 61453
# Patients: 23866
# Positive cases: 14744
# Negative cases: 46709
```

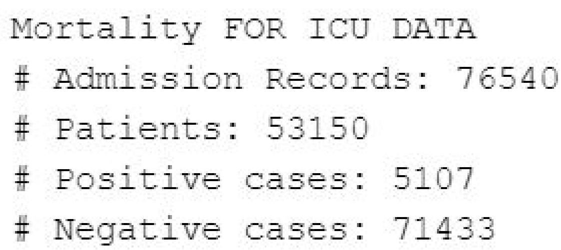
Figure 4: Extracted Cohort Analysis Summary: Readmission Non-ICU Heart Failure patients

are stored in compressed format csv file as shown in Figure 5. The **analysis** summary of the extracted cohort is as shown in Figure 6



subject_id	stay_id	intime	outtime	Age	gender	ethnicity	insurance	label	dod	hadm_id
10000032	39553978	7/23/2180 14:00	7/23/2180 23:50	52	F	WHITE	Medicaid	0	0	29079034
10000980	39765666	6/27/2189 8:42	6/27/2189 20:38	73	F	BLACK/AFI	Medicare	0	0	26913865
10001217	37067082	11/20/2157 19:18	11/21/2157 22:08	55	F	WHITE	Other	0	0	24597018
10001217	34592300	12/19/2157 15:42	12/20/2157 14:27	55	F	WHITE	Other	0	0	27703517
10001725	31205490	4/11/2110 15:52	4/12/2110 23:59	46	F	WHITE	Other	0	0	25563031
10001884	37510196	1/11/2131 4:20	1/20/2131 8:27	68	F	BLACK/AFI	Medicare	1	1/20/2131 0:00	26184834
10002013	39060235	5/18/2160 10:00	5/19/2160 17:33	53	F	OTHER	Medicare	0	0	23581541
10002155	33685454	8/4/2129 12:45	8/10/2129 17:02	80	F	WHITE	Other	0	3/10/2131 0:00	23822395
10002155	31090461	9/24/2130 0:50	9/27/2130 22:13	80	F	WHITE	Other	0	3/10/2131 0:00	28994087

Figure 5: Extracted Cohort: Mortality ICU patients

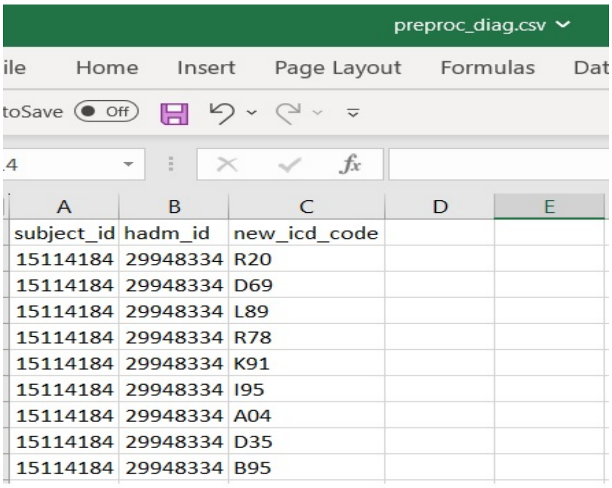


```
Mortality FOR ICU DATA
# Admission Records: 76540
# Patients: 53150
# Positive cases: 5107
# Negative cases: 71433
```

Figure 6: Extracted Cohort Analysis Summary: Mortality ICU patients

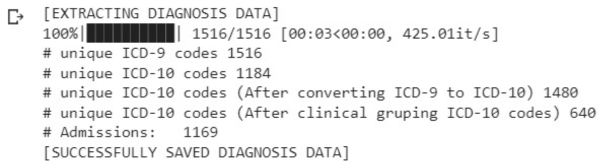
4.3 Result 3

I am able to successfully pre-process cohort from MIMIC-IV dataset for clinical grouping and conversion based on user input along with providing summary of the resultant cohort as shown in Figure 7 and Figure 8



subject_id	hadm_id	new_icd_code
15114184	29948334	R20
15114184	29948334	D69
15114184	29948334	L89
15114184	29948334	R78
15114184	29948334	K91
15114184	29948334	I95
15114184	29948334	A04
15114184	29948334	D35
15114184	29948334	B95

Figure 7: Pre-process Cohort for clinical grouping: Convert ICD-9 to ICD-10 and group ICD-10 codes

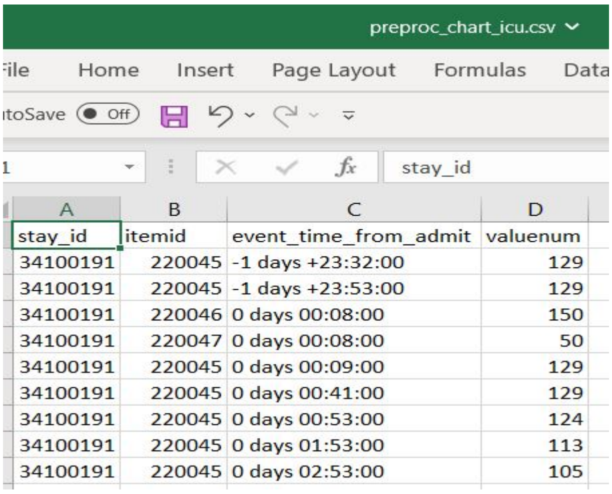


```
[EXTRACTING DIAGNOSIS DATA]
100%|██████████| 1516/1516 [00:03<00:00, 425.01it/s]
# unique ICD-9 codes 1516
# unique ICD-10 codes 1184
# unique ICD-10 codes (After converting ICD-9 to ICD-10) 1480
# unique ICD-10 codes (After clinical gruping ICD-10 codes) 640
# Admissions:   1169
[SUCCESSFULLY SAVED DIAGNOSIS DATA]
```

Figure 8: Pre-process Cohort Summary for clinical grouping

4.4 Result 4

I am able to successfully pre-process cohort for outlier imputation,and selection of feature based on user inputs as shown in Figure 9



stay_id	itemid	event_time_from_admit	valuenum
34100191	220045	-1 days +23:32:00	129
34100191	220045	-1 days +23:53:00	129
34100191	220046	0 days 00:08:00	150
34100191	220047	0 days 00:08:00	50
34100191	220045	0 days 00:09:00	129
34100191	220045	0 days 00:41:00	129
34100191	220045	0 days 00:53:00	124
34100191	220045	0 days 01:53:00	113
34100191	220045	0 days 02:53:00	105

Figure 9: Pre-process Cohort: Outlier imputation

4.5 Result 5

I am able to successfully pre-process cohort by binning the sequential data into time intervals based on the timeseries length according to user inputs as shown in Figure 10 and 11

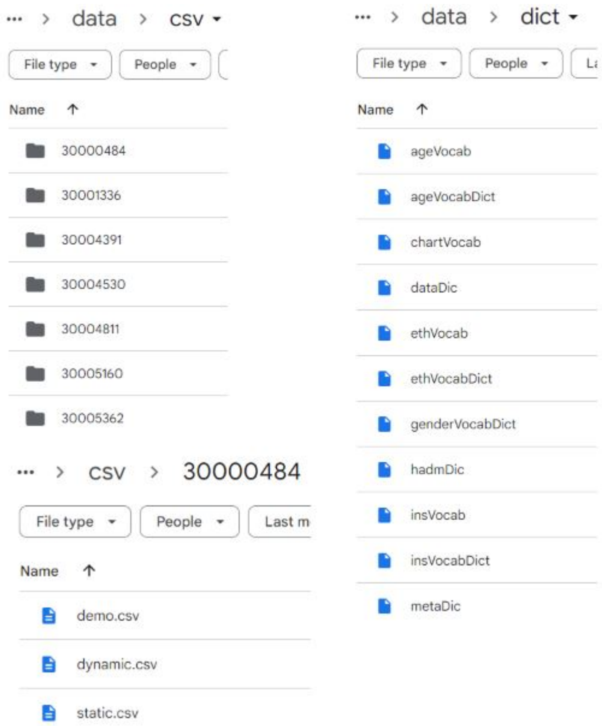
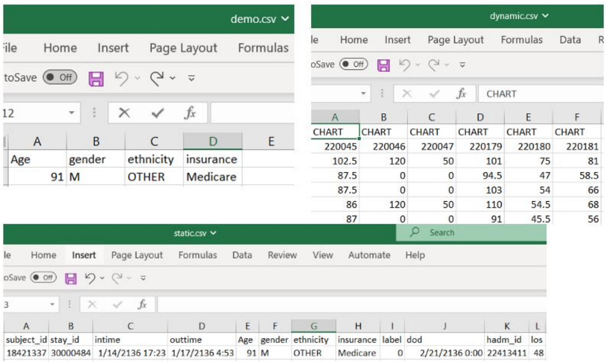


Figure 10: Pre-process Cohort: Timeseries and dictionary folder structure



demo.csv

Age	gender	ethnicity	insurance
91	M	OTHER	Medicare

dynamic.csv

CHART	CHART	CHART	CHART	CHART	CHART
220045	220046	220047	220179	220180	220181
102.5	120	50	101	75	81
87.5	0	0	94.5	47	58.5
87.5	0	0	103	54	66
86	120	50	110	54.5	68
87	0	0	91	45.5	56

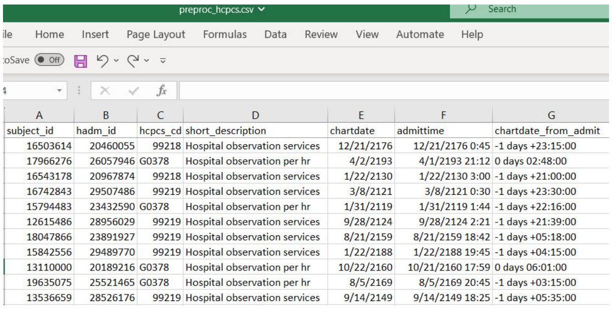
static.csv

subject_id	stay_id	intime	outtime	Age	gender	ethnicity	insurance	label	dod	hadm_id	los
18421337	30000484	1/14/2136 17:23	1/17/2136 4:53	91	M	OTHER	Medicare	0	2/21/2136 0:00	22413411	48

Figure 11: Pre-process Cohort: Timeseries binning files

4.6 Additional results not present in the original paper

- I have successfully added feature extraction logic for hcpcs events. The data records from MIMIC-IV (https://physionet.org/content/mimiciv/1.0/hosp/hcpcsevents.csv.gz) are extracted and merged with the cohort based on user selection inputs. The final records are stored in compressed format csv file as shown in Figure 12 and **analysis** summary in Figure 13



preproc_hcpcs.csv

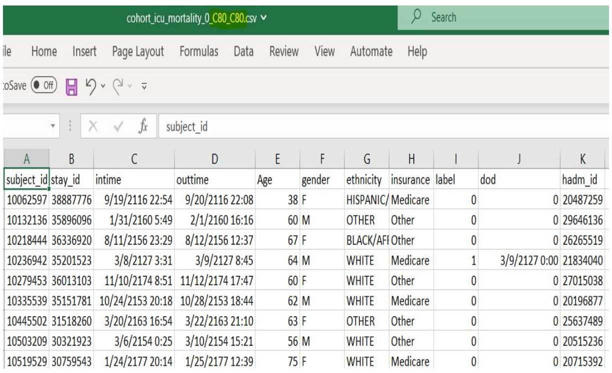
subject_id	hadm_id	hcpcs_cd	short_description	chartdate	admittime	chartdate_from_admit
16503614	20460055	99218	Hospital observation services	12/21/2176	12/21/2176 0:45	-1 days +23:15:00
17966276	26057946	G0378	Hospital observation per hr	4/2/2193	4/1/2193 21:12	0 days 02:48:00
16543178	20967874	99218	Hospital observation services	1/22/2130	1/22/2130 3:00	-1 days +21:00:00
16742843	29507486	99219	Hospital observation services	3/8/2121	3/8/2121 0:30	-1 days +23:30:00
15794483	23432590	G0378	Hospital observation per hr	1/31/2119	1/31/2119 1:44	-1 days +22:16:00
12615486	28956029	99219	Hospital observation services	9/28/2124	9/28/2124 2:21	-1 days +21:39:00
18047866	23891927	99219	Hospital observation services	8/21/2159	8/21/2159 18:42	-1 days +05:18:00
15842556	29489770	99219	Hospital observation services	1/22/2188	1/22/2188 19:45	-1 days +04:15:00
13110000	20189216	G0378	Hospital observation per hr	10/22/2160	10/21/2160 17:59	0 days 06:01:00
19635075	25521465	G0378	Hospital observation per hr	8/5/2169	8/5/2169 20:45	-1 days +03:15:00
13536659	28526176	99219	Hospital observation services	9/14/2149	9/14/2149 18:25	-1 days +05:35:00

Figure 12: Extracted Feature: hcpcs

```
[EXTRACTING HCPCS DATA  ]
cohort_non-icu_mortality_0_
mimiciv/1.0
Number of unique type of hcpcs code:  2239
Total number of rows:  160720
# Admissions:   126029
[SUCCESSFULLY SAVED HCPCS DATA]
```

Figure 13: Feature Analysis Summary: hcpcs

- I have successfully extracted data for additional disease such as Cancer. The ICD-10 Diagnosis code for Cancer is C80. The final records are stored in compressed format csv file as shown in Figure 14 and **analysis** in Figure 15



cohort_icu_mortality_0_C80_C80.csv

subject_id	stay_id	intime	outtime	Age	gender	ethnicity	insurance	label	dod	hadm_id
10062597	38887776	9/19/2116 22:54	9/20/2116 22:08	38	F	HISPANIC/	Medicare	0	0	20487259
10132136	35896096	1/31/2160 5:49	2/1/2160 16:16	60	M	OTHER	Other	0	0	29646136
10218444	36336920	8/11/2156 23:29	8/12/2156 12:37	67	F	BLACK/AFI	Other	0	0	26265519
10236942	35201523	3/8/2127 3:31	3/9/2127 8:45	64	M	WHITE	Medicare	1	3/9/2127 0:00	21834040
10279453	36013103	11/10/2174 8:51	11/12/2174 17:47	60	F	WHITE	Other	0	0	27015038
10335539	35151781	10/24/2153 20:18	10/28/2153 18:44	62	M	WHITE	Medicare	0	0	20196877
10445502	31518260	3/20/2163 16:54	3/22/2163 21:10	63	F	OTHER	Other	0	0	25637489
10503209	30321923	3/6/2154 0:25	3/10/2154 15:21	56	M	WHITE	Other	0	0	20515236
10519529	30759543	1/24/2177 20:14	1/25/2177 12:39	75	F	WHITE	Medicare	0	0	20715392

Figure 14: Extracted Cancer Cohort: Mortality ICU patients

- **Ablation:** I have successfully added the code to provide user with a flexibility to choose to start the processing from the already extracted intermediate cohort. This will skip few of the

```
Mortality FOR ICU DATA
# Admission Records: 241
# Patients: 215
# Positive cases: 37
# Negative cases: 204
```

Figure 15: Extracted Cancer Cohort Analysis Summary: Mortality ICU patients

steps in the extraction pipeline thereby taking less processing time to reach to later steps as shown in Figure 16,17 and 18

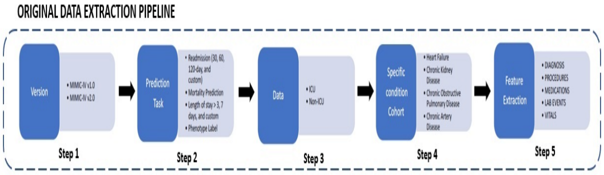


Figure 16: Original Data Extraction Pipeline

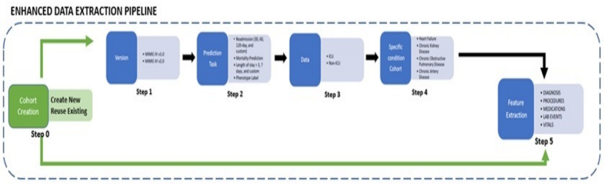


Figure 17: Enhanced Data Extraction Pipeline

```
Please select the metadata for the existing cohort
Cohort path
cohort_non-icu_mortality_0_
Version Path
mimiciv/1.0
ICU Flag
False
Mortality Flag
True
Length of Stay Flag
False
Readmission Flag
False

Using existing cohort for feature selection: cohort_non-icu_mortality_0_

[EXTRACTING HCPCS DATA  ]
cohort_non-icu_mortality_0_
mimiciv/1.0
Number of unique type of hcpcs code:  2239
Total number of rows:  160720
# Admissions:  126029
[SUCCESSFULLY SAVED HCPCS DATA]
```

Figure 18: Using existing cohort for feature selection

- **Bonus initiatives**
  1. **Notebook**: I have added Jupiter notebook https://github.com/NitaAgarwal2022/DLH_project/blob/main/projectPipeline.ipynb
  2. **PyHealth**: (Yang et al., 2022) I was able to successfully contribute by adding hcpcsevents table in MIMIC4 dataset. **Approved PR** link https://github.com/sunlabuiuc/PyHealth/pull/134

5 Discussion

The data extraction and pre-processing pipeline of the original paper was reproducible. I was able to reproduce all the claims mentioned in section 2

5.1 What was easy

The original data pipeline code is documented well. With few changes to the original code I was able to run it on my local setup. Also, the authors were prompt in their response to my query so that helped.

5.2 What was difficult

The difficulty I faced was during the environment restart. The original data pipeline seems to be executed in steps one after the other. If the environment is restarted, the data extracted by the pipeline in last cycle cannot be used this time for further processing . In this case the pipeline has to be again started from step 1 and extract the already extracted files again instead of reusing them.
This challenge in the existing pipeline inspired me for my ablation topic and I am able to successfully add the code to provide user with a flexibility to choose to start the processing from the already extracted intermediate cohort.

5.3 Recommendations for reproducibility

- I have added the preliminary steps cell in my Jupiter notebook(https://github.com/NitaAgarwal2022/DLH_project/blob/main/projectPipeline.ipynb) as well as readme for the local setup to work successfully.

- As the volume of MIMIC-IV dataset is high, using paid version of Google Colab pro+ will provide helpful hardware support for extracting and pre-processing the data

6 Communication with original authors

I communicated with the original authors regarding the system hardware configuration to run the data pipeline https://github.com/healthylaife/MIMIC-IV-Data-Pipeline/issues/34 and their response was helpful.

References

A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. 2000 (June 13). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220. Circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.

Mehak Gupta, Brennan Gallamoza, Nicolas Cutrona, Pranjal Dhakal, Raphael Poulain, and Rahmatollah Beheshti. 2022. An Extensive Data Processing Pipeline for MIMIC-IV. In *Proceedings of the 2nd Machine Learning for Health symposium*, volume 193 of *Proceedings of Machine Learning Research*, pages 311–325. PMLR.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2021. Mimic-iv.

Chaoqi Yang, Zhenbang Wu, Patrick Jiang, Zhen Lin, and Jimeng Sun. 2022. PyHealth: A deep learning toolkit for healthcare predictive modeling.