To understand the impact of various physiological and anatomical factors on the prevalence of Chronic diseases in patients with history of sepsis or requiring mechanical ventilation on first day of ICU admission. It briefly highlights the overall impact on SAPS and SOFA score under such conditions.

SI 624 Final Project Plan

MIMIC dataset (Medical Information Mart for Intensive Care)
Exploration on Python Jupyter

Ashruti Tuteja

SI 624 Fall 2022:

Scoping the Research Project on MIMIC-II dataset

Team Members:

Abdul Haris Ibrahim

Ashruti Tuteja

Q.1. Give a brief summary of your research project.

We are planning to explore MIMIC data, to understand the trend and prevalence of chronic diseases in the presence of various physiological conditions amongst ICU admitted patients.

Objective of the project:

To understand the impact of various physiological and anatomical factors on the prevalence of Chronic diseases in patients with history of sepsis or requiring mechanical ventilation on first day of ICU admission. It briefly highlights the overall impact on SAPS and SOFA score under such conditions, where patient is on Indwelling Arterial Catheter when in ICU, and existence of such conditions help in predicting mortality based on the SAPS score or SOFA.

Q.2. What is being studied? Why? What purpose does this serve / What is the end objective or goal of this research project?

The study would help us to have an idea on how various physiological or anatomical factors impact the prevalence of chronic diseases to overall affect the SAPS score of the patient.
Our **population** are the *patients requiring mechanical ventilation who did not require vasopressors or have a diagnosis of sepsis were identified, and the primary outcome was 28-day mortality.*
The questions we foresee to answer from this dataset:

1. Chronic disease prevalence according to gender.
    1. Liver
    2. Kidney
    3. Heart

2. Impact of clinical indicators on the occurrence of chronic diseases in patients.
    a. Creatinine levels beyond range and is it causation of renal disease.
    creatinine_first and renal_flg ==1
    b. Hemoglobin count of patients having Congestive heart failure

3. Correlation between total number of chronic diseases a person has versus the number of days in hospitalization and also the number of days in ICU.

4. How does the number of patients with chronic diseases in each age group affects the SAPS score on ICU admission leading to ICU mortality.

5. Explore SOFA score:
    1. Understanding platelet count effect on SOFA score
    2. Understand SOFA score (categories) with hospital mortality

Q.3. What are the different steps involved in this research project?

1. Extracting the data
2. Understanding the Dataset
3. Cleaning the dataset:
    finding Null values,
    treating Null values,
    Extracting the metadata as required by the research questions.
4. Understanding the research question and manipulating data to align with this interest.
5. Data analysis using Pandas and other python libraries on Jupyter notebook, to find answers for healthcare related questions.
6. Understanding the output of these questions and brainstorming on best visualization method to depict the results.

Q.4. When is each step supposed to take place? What is the overall timeline of the project? (It doesn't need to be granular, you can give estimate timeframes for each milestone)

We plan to work weekly on each set of questions for the next couple of weeks, starting from October 13, 2022 to November 26, 2022. Our weekly meeting agenda would be to break down each research question and try to achieve the answer or at least the metadata of one question at a time eventually leading to answers in this journey of 1 month.

Q.5. Who are the beneficiaries and the stakeholders?

The primary use of this dataset is to carry out the case study in Chapter 16 of Secondary Analysis of Electronic Health Records. The case study data walks the reader through the process of examining the effect of indwelling arterial catheters (IAC) on 28-day mortality in the intensive care unit (ICU) in patients who were mechanically ventilated during the first day of ICU admission. Analysis of such data will benefit the CDC and WHO for the formulation of preventive measures like designing healthcare policies around having better access to healthcare for patients experiencing chronic conditions.
The study of the high prevalence of chronic diseases and deteriorating SAPS score will also

drive the attention of the state regulatory health bodies (like MDHHS for Michigan) to incorporate statewide chronic risk prevention-related measures.
Analysis of SAPS score and SOFA score will help clinical researchers to derive strategies, medications, or interventional protocols in contributing to reducing the number of mortality in such population.
The analysis study will also help academic researchers and students to further dig into the analysis and suggest or recommend ways in which we could understand the current trends and spread awareness with the aim of improved health outcomes.

Q.6. Who is your target research population?

Our target research population is healthcare regulatory bodies and the population suffering from chronic diseases with history of sepsis or with the need of mechanical ventilation on ICU admission day zero. The study results will give a broader understanding of effect of various factors on SAPS score and SOFA score which would help the regulatory bodies to further predict mortalities and suggest ways to improve prevalence of chronic diseases amongst vulnerable populations.

About the Dataset

Indwelling arterial catheters (IACs) are used extensively in the ICU for hemodynamic monitoring and for blood gas analysis. IAC use also poses potentially serious risks, including bloodstream infections and vascular complications. In 2015, Hsu et al published a study to assess whether IAC use was associated with mortality in patients who are mechanically ventilated and do not require vasopressor support. This dataset was created for the purpose of a case study in the book: Secondary Analysis of Electronic Health Records, published by Springer in 2016. The dataset in question was used throughout Chapter 16 (Data Analysis) by Raffa J. et al. to investigate the effectiveness of indwelling arterial catheters in hemodynamically stable patients with respiratory failure for mortality outcomes. The dataset is derived from MIMIC-II, the publicly-accessible critical care database. It contains a summary of clinical data and outcomes for 1,776 patients. The dataset (full_cohort_data.csv) is a comma-separated value file that includes a header with descriptive variable names.

To Access the dataset
Clinical data from the MIMIC-II database for a case study on indwelling arterial catheters. Accessed on: 1st February 2022. https://physionet.org/content/mimic2-iaccd/1.0/

Data Dictionary

S.NO	Parameter Name	Meaning
1	aline_flg	IAC used (binary, 1 = year, 0 = no)

2	icu_los_day	length of stay in ICU (days, numeric)
3	hospital_los_day	length of stay in hospital (days, numeric)
4	age	age at baseline (years, numeric)
5	gender_num	patient gender (1 = male; 0=female)
6	weight_first	first weight, (kg, numeric)
7	bmi	patient BMI, (numeric)
8	sapsi_first	first SAPS I score (numeric)
9	sofa_first	first SOFA score (numeric)
10	service_unit	type of service unit (character: FICU, MICU, SICU)
11	service_num	service as a numeric (binary: 0 = MICU or FICU, 1 = SICU)
12	day_icu_intime	day of week of ICU admission (character)
13	day_icu_intime_num	day of week of ICU admission (numeric, corresponds with day_icu_intime)
14	hour_icu_intime	hour of ICU admission (numeric, hour of admission using 24hr clock)
15	hosp_exp_flg	death in hospital (binary: 1 = yes, 0 = no)
16	icu_exp_flg	death in ICU (binary: 1 = yes, 0 = no)
17	day_28_flg	death within 28 days (binary: 1 = yes, 0 = no)
18	mort_day_censored	day post ICU admission of censoring or death (days, numeric)
19	censor_flg	censored or death (binary: 0 = death, 1 = censored)
20	sepsis_flg	sepsis present (binary: 0 = no, 1 = yes -- absent (0) for all)
21	chf_flg	Congestive heart failure (binary: 0 = no, 1 = yes)
22	afib_flg	Atrial fibrillation (binary: 0 = no, 1 = yes)

23	renal_flg	Chronic renal disease (binary: 0 = no, 1 = yes)
24	liver_flg	Liver Disease (binary: 0 = no, 1 = yes)
25	copd flg	Chronic obstructive pulmonary disease (binary: 0 = no, 1 = yes)
26	cad_flg	Coronary artery disease (binary: 0 = no, 1 = yes)
27	stroke_flg	Stroke (binary: 0 = no, 1 = yes)
28	mal flg	Malignancy (binary: 0 = no, 1 = yes)
29	resp_flg	Respiratory disease (non-COPD) (binary: 0 = no, 1 = yes)
30	map_1st	Mean arterial pressure (mmHg, numeric)
31	hr_1st	Heart Rate (numeric)
32	temp_1st	Temperature (F, numeric)
33	spo2_1st	S_pO_2 (%, numeric)
34	abg_count	arterial blood gas count (number of tests, numeric)
35	wbc first	first White blood cell count (K/uL, numeric)
36	hgb_first	first Hemoglobin (g/dL, numeric)
37	platelet_first	first Platelets (K/u, numericL)
38	sodium_first	first Sodium (mEq/L, numeric)
39	potassium_first	first Potassium (mEq/L, numeric)
40	tco2 first	first Bicarbonate (mEq/L, numeric)
41	chloride_first	first Chloride (mEq/L, numeric)
42	bun_first	first Blood urea nitrogen (mg/dL, numeric)
43	creatinine_first	first Creatinine (mg/dL, numeric)
44	po2_first	first PaO_2 (mmHg, numeric)
45	pco2_first	first PaCO_2 (mmHg, numeric)

46	iv_day_1	input fluids by IV on day 1 (mL, numeric)

Visualizations we plan to create:
Scatterplot, tables, histogram, correlation plot

Statistics we plan to perform:
Chi-square test, f-statistic, correlation, regression

Strengths of the Dataset:

- Reliable dataset – MIMIC is reputed and open data source for medical data
- Extensive – incorporating many attributes (~46 columns)
- Meaningfulness - Data dictionary is self-explanatory
- Completeness – Less missing values

Weaknesses of the Dataset:

- Less instances – 1776 rows depicting 1776 patients
    - If there were more instances or patients recorded as part of the dataset, the subsequent study and its finding would be more inclusive and meaningful which can stand true in numerous cases.
- Validity – The data was collected using IAC and it is not known whether the readings obtained were cross-verified against other standard tests for each parameter.

Glossary:

SAPS score: Estimates the probability of mortality for ICU patients on admission.

SOFA score: The Sequential Organ Failure Assessment (SOFA) score is a scoring system that assesses the performance of several organ systems in the body (neurologic, blood, liver, kidney, and blood pressure/hemodynamics)

OVERVIEW OF THE DATASET

1. Dataset Size

▾ Data Shape

```
[ ]  print("The shape of the original dataset is: ", data.shape)
```

```
The shape of the original dataset is:  (1776, 46)
```

There are 1776 rows with 46 different columns in this dataset

2. Sample of Dataset

▾ Data Sample

```
data.head(5)
```

	aline_flg	icu_los_day	hospital_los_day	age	gender_num	weight_first	bmi	sapsi_first	sofa_first	service_unit	...	platelet_first	sodium_firs
0	1	7.63	13	72.36841	1.0	75.0	29.912791	15.0	9.0	SICU	...	354.0	138
1	0	1.14	1	64.92076	0.0	55.0	20.121312	NaN	5.0	MICU	...	NaN	Na
2	0	2.86	5	36.50000	0.0	70.0	27.118272	16.0	5.0	MICU	...	295.0	144
3	1	0.58	3	44.49191	0.0	NaN	NaN	21.0	7.0	SICU	...	262.0	139
4	1	1.75	5	23.74217	1.0	95.2	28.464563	18.0	7.0	SICU	...	22.0	146

5 rows × 46 columns

3. Basic Statistics in Dataset Before doing manipulation

▾ Basic Statistics

```
data.describe()
```

	aline_flg	icu_los_day	hospital_los_day	age	gender_num	weight_first	bmi	sapsi_first	sofa_first	service_num	...	platelet_first	soc
count	1776.000000	1776.000000	1776.000000	1776.000000	1775.000000	1666.000000	1310.000000	1691.000000	1770.000000	1776.000000	...	1768.000000	
mean	0.554054	3.346498	8.110923	54.379660	0.577465	80.075948	27.827316	14.136606	5.820904	0.552928	...	246.083145	
std	0.497210	3.356261	8.157159	21.062854	0.494102	22.490516	8.210074	4.114302	2.334666	0.497331	...	99.865469	
min	0.000000	0.500000	1.000000	15.180230	0.000000	30.000000	12.784877	3.000000	0.000000	0.000000	...	7.000000	
25%	0.000000	1.370000	3.000000	38.247318	0.000000	65.400000	22.617307	11.000000	4.000000	0.000000	...	182.000000	
50%	1.000000	2.185000	6.000000	53.678585	1.000000	77.000000	26.324846	14.000000	6.000000	1.000000	...	239.000000	
75%	1.000000	4.002500	10.000000	72.762992	1.000000	90.000000	30.796551	17.000000	7.000000	1.000000	...	297.000000	
max	1.000000	28.240000	112.000000	99.110950	1.000000	257.600000	98.797134	32.000000	17.000000	1.000000	...	988.000000	

8 rows × 44 columns

4. Missing value for the columns

▾ Check Missing Values

```
data.isna().sum()
```

```
aline_flg               0
icu_los_day             0
hospital_los_day        0
age                     0
gender_num              1
weight_first          110
bmi                   466
sapsi_first            85
sofa_first              6
service_unit            0
service_num             0
day_icu_intime          0
day_icu_intime_num      0
hour_icu_intime         0
hosp_exp_flg            0
icu_exp_flg             0
day_28_flg              0
mort_day_censored       0
censor_flg              0
sepsis_flg              0
chf_flg                 0
afib_flg                0
renal_flg               0
liver_flg               0
copd_flg                0
cad_flg                 0
stroke_flg              0
mal_flg                 0
```

5. Dataset Column Name

▾ Remove Missing Values

```
[ ] # Finding Index of Missing Values
    gender_index = data[data['gender_num'].isna()].index.tolist()
    sapsi_index = data[data['sapsi_first'].isna()].index.tolist()

    # Accumulate all the indexes in one list
    missing_data = [y for x in [sapsi_index, gender_index ] for y in x]

    # Remove missing values in each row of every column which will be used for further analysis
    new_data = data.drop(labels=missing_data, axis=0)
    new_data.shape
```

```
(1690, 46)
```

Websites/References we would look forward to take help from:

https://pubmed.ncbi.nlm.nih.gov/15593047/#:~:text=Long%2Dstanding%20anemia%20of%20any,anemia%20and%20the%20CHF%20further.

https://www.mdcalc.com/calc/4044/simplified-acute-physiology-score-saps-ii

https://clincalc.com/IcuMortality/SOFA.aspx

[1] https://link.springer.com/book/10.1007/978-3-319-43742-2
[2] https://link.springer.com/chapter/10.1007/978-3-319-43742-2_16
[3] https://physionet.org/content/mimic2-iaccd/1.0/
[4] https://archive.physionet.org/mimic2/#:~:text=The%20MIMIC%20II%20(Multiparameter%20Intelligent,Unit%20(ICU)%20patients*.

[5] https://www.mdcalc.com/calc/10403/simplified-acute-physiology-score-saps-3#:~:text=Estimates%20the%20probability%20of%20mortality%20for%20ICU%20patients%20on%20admission.&text=The%20SAPS%203%20Score%20predicts,physiologic%20derangement%20upon%20ICU%20admission.

[6] https://files.asprtracie.hhs.gov/documents/aspr-tracie-sofa-score-fact-sheet.pdf

[7] *https://www.kidney.org/atoz/content/kidneytests#:~:text=A%20creatinine%20level%20of%20greater,creatinine%20in%20the%20blood%20rises*

[8] https://www.cdc.gov/kidneydisease/publications-resources/annual-report/ckd-risk-prevention.html#:~:text=Diabetes%20and%20high%20blood%20pressure,can%20help%20keep%20kidneys%20healthy.

[9] *https://www.mayoclinic.org/symptoms/low-hemoglobin/basics/definition/sym-20050760#:~:text=A%20low%20hemoglobin%20count%20is%20generally%20defined%20as%20less%20than,varies%20with%20age%20and%20sex*

[10] *https://www.nhlbi.nih.gov/health/heart-failure/causes*

[11] https://www.merckmanuals.com/medical-calculators/SOFA.htm

[12] *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9290429/#:~:text=Patients%27%20demographi*

cs%20and%20history%20of%20co%2Dmorbidities.&text=A%20SOFA%20score%20from%200,mortality%20(Table%20%E2%80%8B2).

[13] https://www.omnicalculator.com/health/saps-ii

[14] *https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0164828*

[15] https://www.nature.com/articles/s41598-021-03397-3.pdf?proof=t+target%3D
[16] https://www.frontiersin.org/articles/10.3389/fcvm.2021.774935/full
[17] https://www.ijidonline.com/article/S1201-9712(21)00863-8/fulltext
[18] https://clincalc.com/IcuMortality/SOFA.aspx

Note:

1. Inferences to analysis and visualizations are in green color.
2. When a particular kind of Method is repeated then the method definition/rationale is not mentioned again, as it would be repetitive.
3. Because some snippets are part of one whole image (they are throwing an error by showing same snippet each time), so it would look as a duplicate, but it is not.