Cost of Hospital Care

Analysing factors that drive cost of care for hospitalized patients

- Ng Geok Teng -

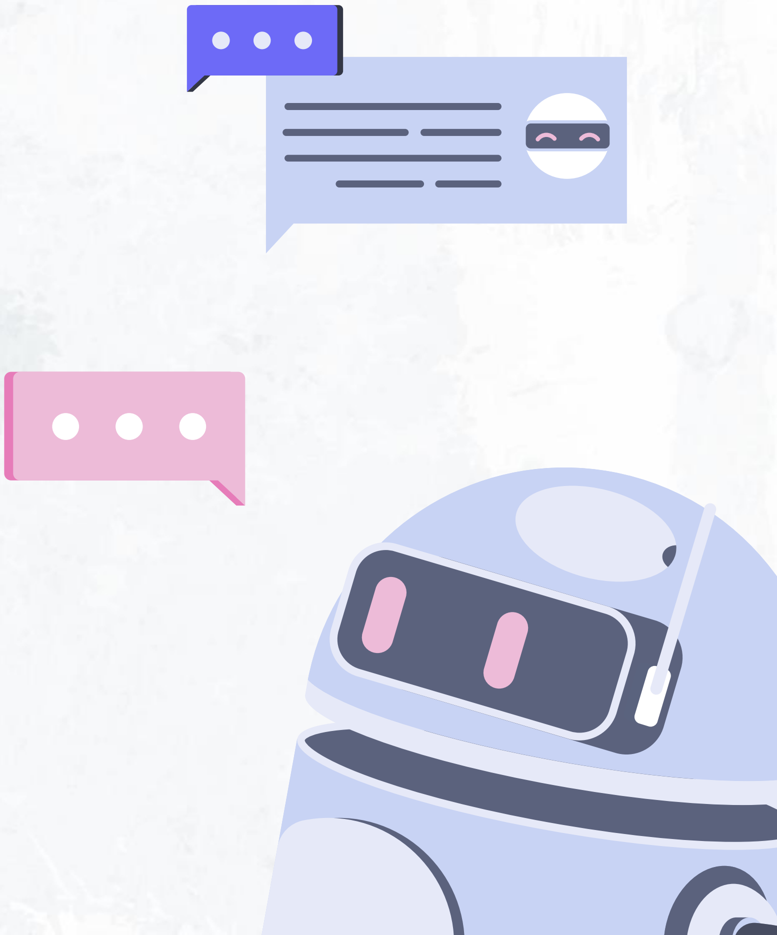


Table of contents

01

Task & Approach

01: What's the Task

Goals – To find out:

Qn. A **What factors drive cost of care?**

Plan: The driving factors will be identified through Data Analysis and evaluate post-modelling feature importance

Data Analysis

Post-Modelling Evaluation

Qn. B **What are the ways to estimate/predict cost of care?**

Plan: Different modelling techniques will be explored to produce one that can estimate cost accurately (Target: R2 score > 90%, MAPE score <10%)

Modelling

01: Some Background

Parkway Pantai hospitals launch AI-powered predictive hospital bill estimation system in Singapore

The new estimation system, which has been in use since November 2018, has made more than 10,000 predictions so far.

By Dean Koh | December 19, 2018 | 04:59 AM

Reference: https://www.healthcareitnews.com/news/asia/parkway-pantai-hospitals-launch-ai-powered-predictive-hospital-bill-estimation-system

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

Comparison of linear, penalized linear and machine learning models predicting hospital visit costs from chronic disease in Thailand

Wichayaporn Thongpeth, M.N.S. [a], Apiradee Lim, Ph.D. [a,*], Akemat Wongpairin, M.P.H. [a], Thaworn Thongpeth, M.D. [b], Santhana Chaimontree, Ph.D. [a]

[a] Department of Mathematics and Computer Science, Faculty of Science and Technology, Prince of Songkla University, Pattani Campus, Mueang, Pattani, 94000, Thailand

[b] Orthopedic Surgery and Preventive Medicine, Suratthani Hospital, Ministry of Public Health, Mueang, Surat Thani, 84000, Thailand

Reference: https://www.sciencedirect.com/science/article/pii/S2352914821002434

Use of AI in predicting hospital bill has been studied and even implemented in Singapore

01: What's the Approach

Data

What we have
Billing data
Clinical data
Demographic data

Data Processing
Merging data
Clean data
Feature engineering

Analysis

Type of analysis
Univariate Analysis
Multivariate Analysis

Data Visualisation:
Matplotlib
Seaborn

Modelling

Data Processing
RobustScaler
OneHotEncoder

Models
Statistical Models
(Linear Regression)
(Penalised Regressions)

Machine Learning Models
(RandomForestRegressor)
(XGBRegressor)

Post-Modelling Evaluation

Model Performance Metric
R2 (target>90%)
MAPE (target<10%)
RMSE

Feature Importance
Dependent on model type

02 →

Data Processing & Analysis

02: Our Data

Key Notes:

- 3000 unique patients
- 3400 unique hospitalisations from 2011-2015
- Some patients have multiple admissions
- Some patients have multiple bills per admission

- Each bill represented by ‘billing_id’
- Each hospitalisation represented by ‘patient_id’ and ‘date_of_admission’
- Each patient represented by ‘patient_id’

Legend:

Dataset obtained

Interim dataset created

Desired combined dataset created

billing_amt
contain bill_id, bill amount

billing_id
contain bill_id, patient_id date_of_admission

Join on ‘billing_id’

billing_df
- Each row represents a bill
- Each hospitalisation has 1 or more bills
- Data aggregated by sum of bill per hospitalisation (‘total_hosp_bill’)

clinical_df
Contain patient clinical features

Join on ‘patient_id’ ‘date_of_admission’

hospitalisation_df
Each row is a unique hospitalisation case of a patient

demographic_df
Contain patient demographic information

Join on ‘patient_id’

merged_df
Each row is a unique hospitalization case of a patient

02: Data Processing

merged_df → Data Processing →

3400 rows, 38 columns
Features include demographics, clinical features, hospitalisation information

Data Checking & Cleaning	
Duplicate Data	Each row is unique hospitalisation of a patient.
Missing Data	Found in ‘medical_history’ columns Assume patient do not have the medical history, as it is inappropriate to assume a patient had the medical history if is it not reported Impute `0`
Data Type	Ensure data in write data type for further processing and analysis: E.g. ‘`medical_history_3`’ is in *string* format when it should be *integer* format
Data Values	Check Categorical columns unique elements Check Quantitative columns value range

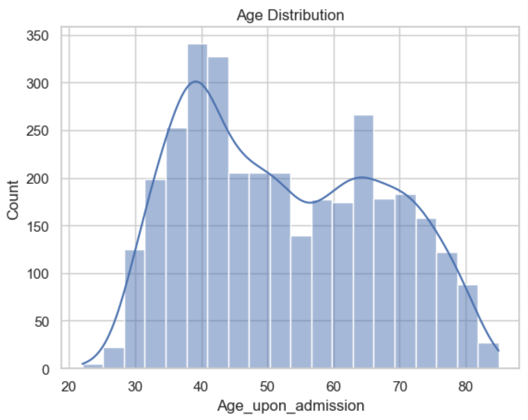
7 New Columns	
`` `total_hosp_bill` ``	Created earlier on before creating `merged_df` when aggregating the billing data
`` `had_prev_admission` ``	Boolean if patient had another admission prior
`` `Age_upon_admission` ``	Using patient `date_of_birth` and `date_of_admission`
`` `BMI` ``	Using patient `weight` and `height` To reflect if a patient has healthy body weight

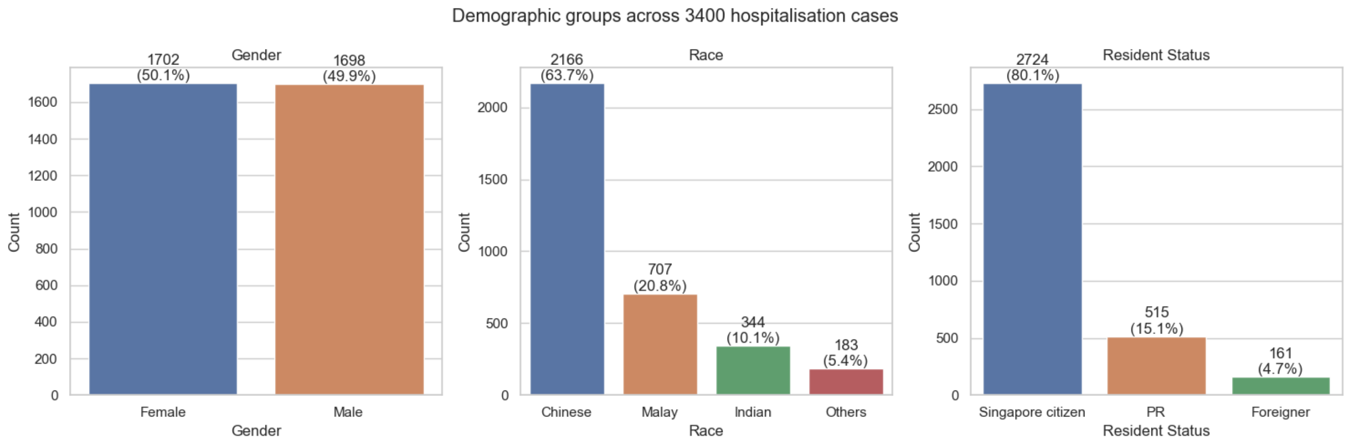
`` `sum_medical_history` ``
`` `sum_symptoms` ``
`` `sum_medications` ``
	Reflects total number of medical histories, symptoms and pre-op medications patient have respectively

Demographic → Clinical → Hospitalisation

02 Data Analysis

Univariate Demographic Analysis

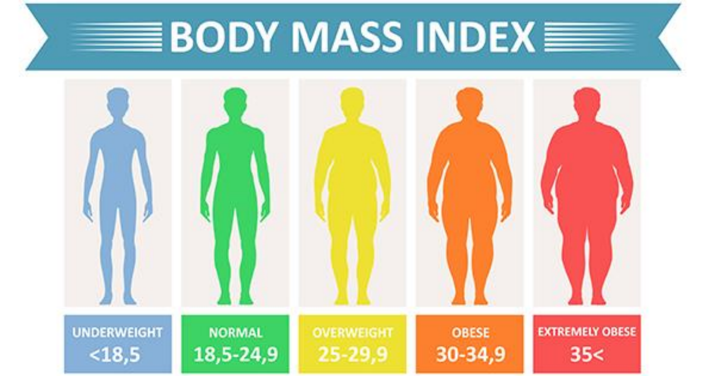




Age	Two bimodal distribution Younger group with median about 40yo Older group with median about 65yo
Gender	Balanced distribution between male and female
Race	Chinese-dominant, with smallest group being \`Other\` at 5.4%
Resident Status	80% Singaporeans, <5% Foreigners

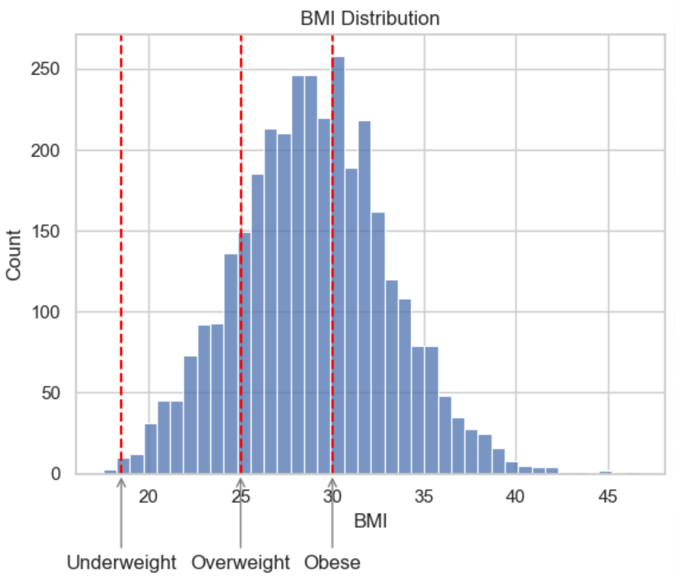
02 Data Analysis

Univariate Clinical Analysis



https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html

$$\text{BMI} = \frac{\text{Weight (in kilograms)}}{\text{Height}^2 \text{ (in meters)}}$$

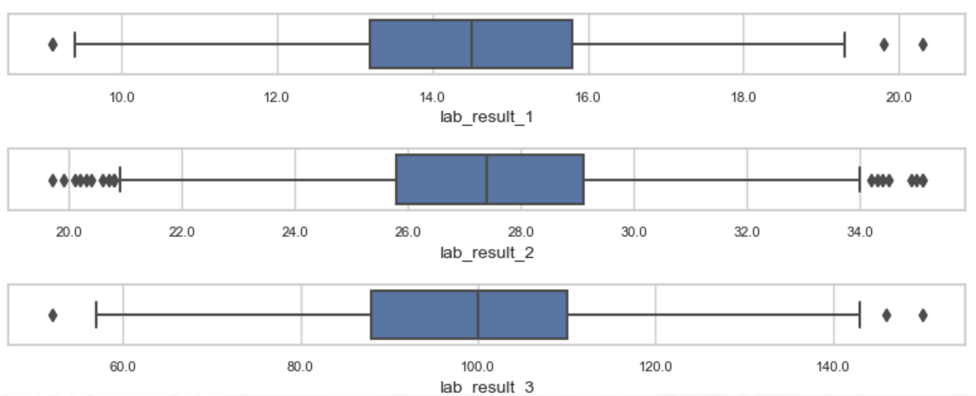


- Body mass index (BMI) allows easy screening if a person body weight is healthy
- BMI appears to be as strongly correlated with various metabolic and disease outcomes
- In our patient group, most of our patients are considered to not have healthy BMI levels (IQR 26.2-31.7). They are mostly overweight and close to half of them obese (median = 28.9)

02 Data Analysis

Univariate Clinical Analysis

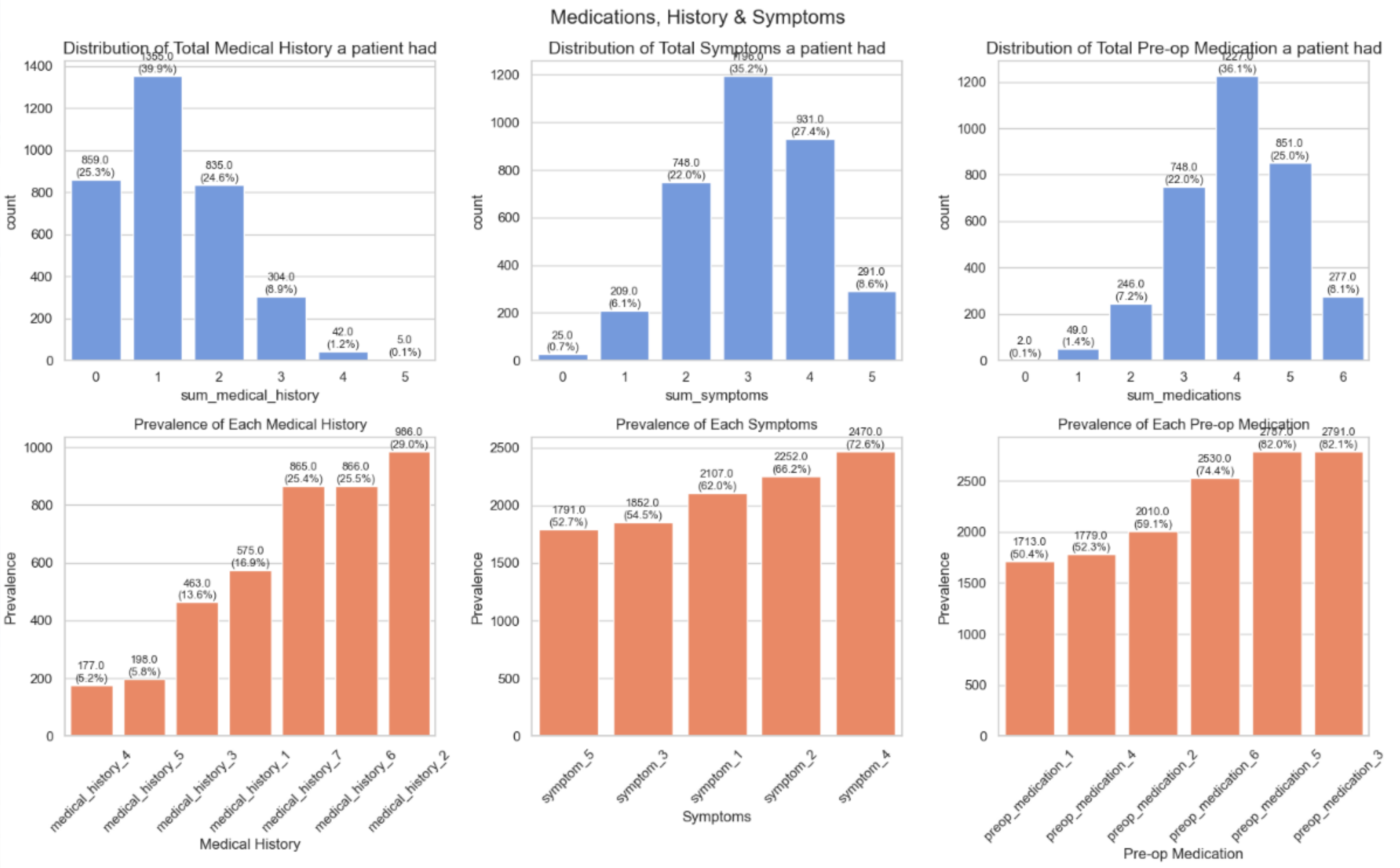
Boxplot for Lab Results



- Some outliers observed in the lab results.
- However, unable to comment about the lab result distribution as it is not to our knowledge what lab results are these. It is not clear how the lab results reflect a patient's health.

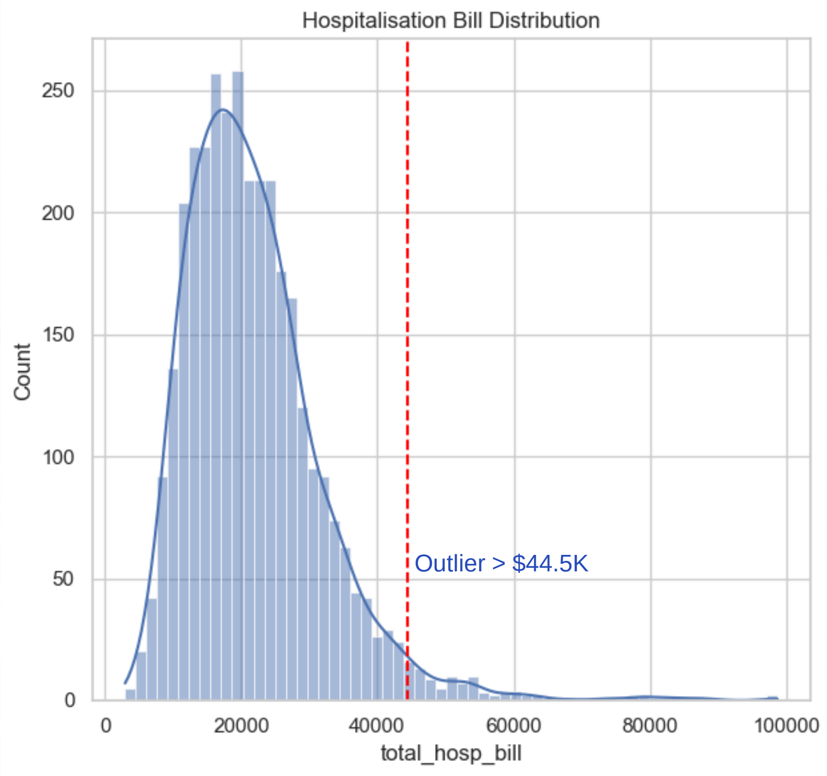
Patients tend to have 1-2 medical histories, 2-4 symptoms,
and 3 or more pre-op medications

Most prevailing observations are medical_history 2, symptom 4 and pre-op medication 3 and 5



02 Data Analysis

Univariate Hospitalisation Analysis

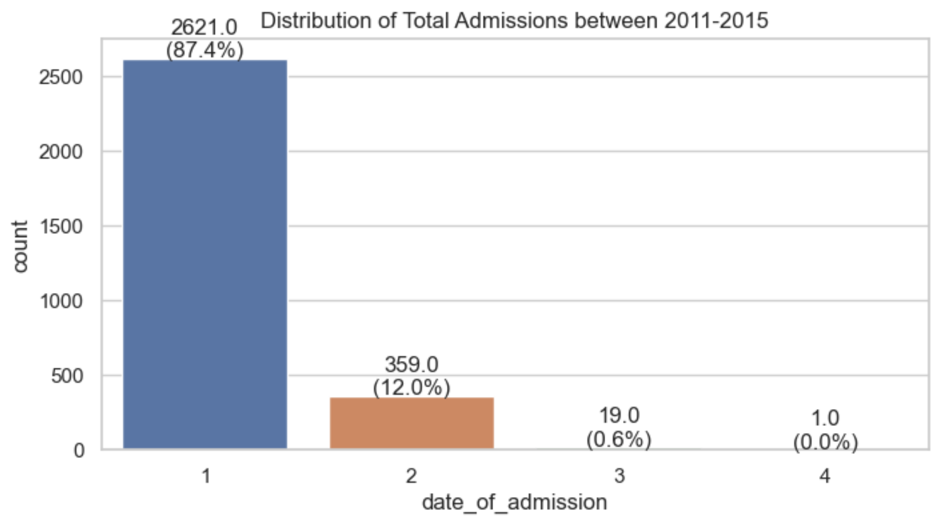


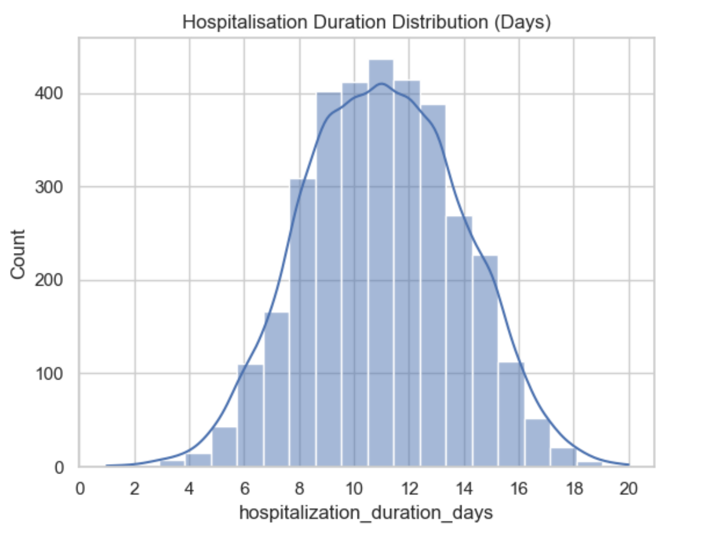
- The hospitalization bill distribution is positively skewed as the histogram shows it tails off on the right.
- As the Shapiro-Wilk test shows p<0.05, we reject the null hypothesis that the distribution is normal → it **does not** follow a normal distribution.
- The bill ranges from ~$2.9K to ~$99K.
- Average bill is about $20K.
- 2.8% of the bill is very large and are outliers (>$44.5K)

Demographic → Clinical → **Hospitalisation**

02 Data Analysis

Univariate Hospitalisation Analysis





Most patients were admitted once between 2011-2015, while some are admitted more than 3x.

Hospitalization days range 1-20 days, most 1-2weeks durations

02 Summary about the patients

Who are they?
(Demographic)

- Patients hospitalised between 2011-2015
- Main two age groups
  (20+ to mid-50s & mid-50s to 80s)
- Mostly Chinese
- Mostly Singaporeans
- Balanced distribution of male and female

What are their medical conditions?
(Clinical)

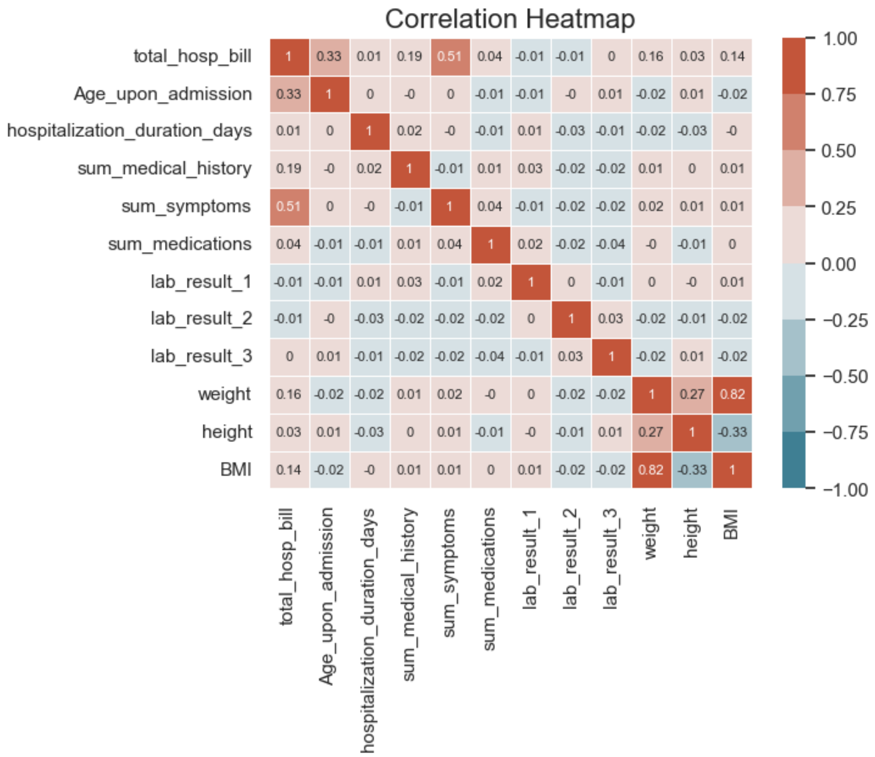
- Mostly overweight or obese
- Mostly have
  1-2 medical history,
  2-4 symptoms,
  3-5 pre-op medications
- Highest prevalence observed for medical history 2,
  symptoms 4 and
  pre-op medicine 3 & 5

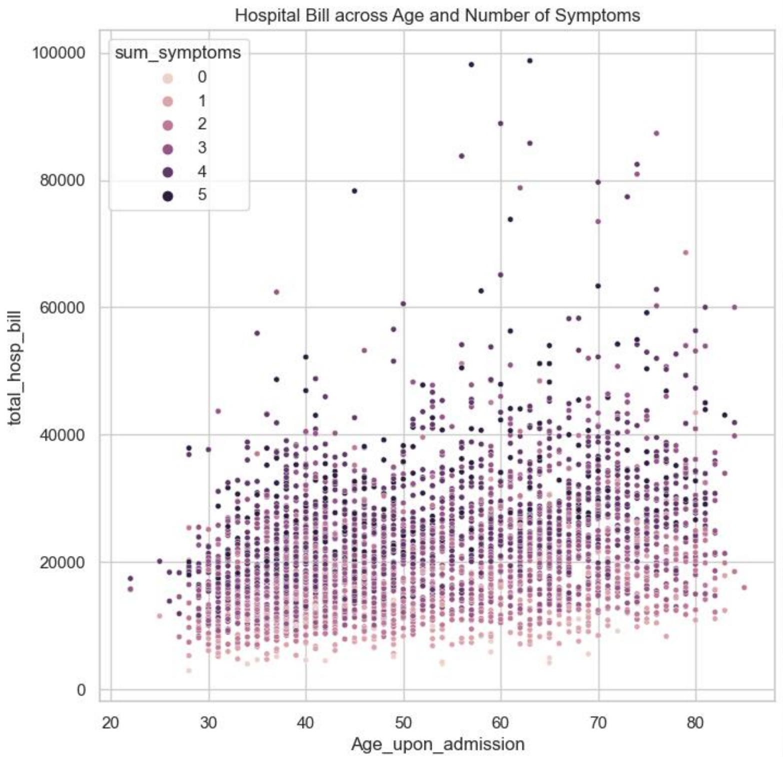
How were their hospitalisation like?
(Hospitalisation)

- Hospitalisation bills range from $2.9-99 K,
  and centred around $20K
- Hospitalisation range from 1 to 20 days,
  most patients stayed
  1 to 2 weeks
- Mostly had 1 hospitalisation between 2011-2015

02 Data Analysis

Factors correlating with Hospitalization bill





FACTORS DRIVING COST OF CARE:

- It appears that `sum_symptoms` has the greatest correlation, with moderate positive correlation with hospital bill (r=0.51)
- It is followed by `Age_upon_admission` which has weak positive correlation with bill at r = 0.33.
- We can infer that patients older in age and patients with more symptoms tend to have higher hospital bill, likely due to greater medical care needed.

02 Data Analysis

Factors correlating with Hospitalization bill

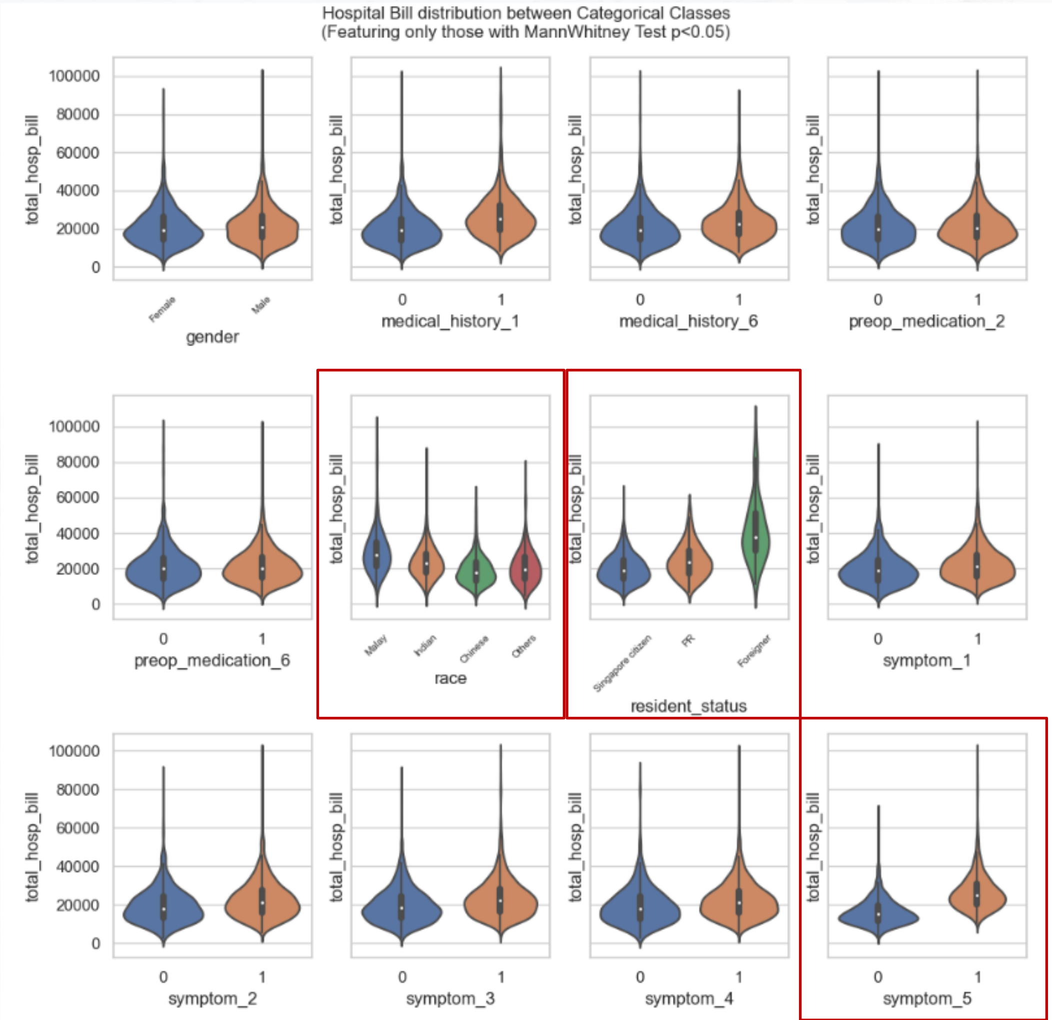
FACTORS DRIVING COST OF CARE:

Mann-Whitney U Test is used to compare difference in hospital bill between groups.

Genders, race, resident status and specific symptom, medical history and medications (12 categorical features) showed significant difference using the statistical test ($p<0.05$)

However, visual inspection on the violin plots, the difference in bill is more distinctly, observed among race, resident status, and between patients with and without symptom 5.

	Column_name	Category 1	Category 2	p-value
0	symptom_5	1	0	0.00000
1	race	Malay	Indian	0.00000
2	race	Malay	Chinese	0.00000
3	race	Malay	Others	0.00000
4	race	Indian	Chinese	0.00000
5	symptom_3	1	0	0.00000
6	symptom_2	0	1	0.00000
7	resident_status	Singapore citizen	PR	0.00000
8	resident_status	Singapore citizen	Foreigner	0.00000
9	resident_status	PR	Foreigner	0.00000
10	symptom_1	0	1	0.00000
11	medical_history_1	0	1	0.00000
12	medical_history_6	0	1	0.00000
13	symptom_4	1	0	0.00000
14	gender	Female	Male	0.00021
15	race	Indian	Others	0.00043
16	race	Chinese	Others	0.00519
17	preop_medication_2	1	0	0.02228
18	preop_medication_6	0	1	0.04031



02 Data Feature Selections

Column Set	Description
`allcol`	All original features included.
`sub1col`	Exclude categorical features where there is no significant hospital bills between the classes 18 features (12 categorical + 6 quantitative)
`sub2col`	Exclude categorical features where there is no significant hospital bills between the classes AND quantitative features showed poor correlation with hospital bill (|r|<0.3). 14 features (12 categorical + 2 quantitative)

We will be using insights gathered from the analysis to create three sets of feature columns to compare model's performance when different set of features are used.

12 Categorical features	6 Quantitative features
'gender', 'medical_history_1', 'medical_history_6', 'preop_medication_2', 'preop_medication_6', 'race', 'resident_status', 'symptom_1', 'symptom_2', 'symptom_3', 'symptom_4', 'symptom_5',	'Age_upon_admission', 'hospitalization_duration_days', 'sum_medical_history', 'sum_symptoms', 'sum_medications', 'BMI' *Excluded:* - *Lab results* - *Weight & Height*

03 →

Modelling & Evaluation

03: Approach

merged_df

Train-test-split (75:25)

X_train, X_test, y_train, y_test

Bootstrapping to create data with 3 different sizes (original/ 2fold/ 4fold)
Created data with 3 different sets of X features (allcol, sub1col, sub2col)

9 X_train of different size and columns
9 X_test of different size and columns
3 y_train of different data size
1 y_test

RobustScaler quantitative columns
OneHotEncode categorical columns

9 transformed X_train of different size and columns
9 transformed X_test of different size and columns
3 y_train of different data size
1 y_test

Statistical Models

(Linear Regression)
(Penalised Regressions **(L1,L2,ElasticNet)**)

9 Datasets, 4 model types = 36 trained models

Created new X datasets

Using selected features using coefficient values from statistical model

Different X features: *allcol, sub1col, sub2col, sub3col*
Creating to total 12 sets of X data

Selected 4 data sets

Size: 4fold
Column sets: allcol, sub1col, sub2col, sub3col

Machine Learning Models

(RandomForestRegressor)
(XGBRegressor)
4 Datasets, 2 model types = 8 trained models

Metric
R2
RMSE
MAPE

Hyperparameter Tuning
RandomSearchCH
GridSearchCV

03 Statistical Model

- 4 Model types: Linear Regression, LassoCV, RidgeCV, ElasticNetCV
- 9 X data sets used
- Total 36 trained statistical models

Top 5 performing models based on R2 score on test set

Model	Model_type	Data_size	Column_set	R2_train	R2_test	MAPE_train	MAPE_test	RMSE_train	RMSE_test
ElasticNet_4fold_sub1col	ElasticNet	4fold	sub1col	91.988	93.577	9.952	9.734	2990.76526	2251.15433
ElasticNet_4fold_allcol	ElasticNet	4fold	allcol	92.091	93.700	9.877	9.755	2971.48339	2229.45249
ElasticNet_original_allcol	ElasticNet	original	allcol	92.316	93.686	9.648	9.806	2922.14372	2231.91596
ElasticNet_original_sub1col	ElasticNet	original	sub1col	92.221	93.543	9.765	9.819	2940.12220	2257.09867
Lasso_4fold_allcol	Lasso	4fold	allcol	92.118	93.406	9.984	9.883	2966.32151	2280.88374

03 Statistical Model Performance

Generally, all the model performance are relatively similar.

There are not overfitting: Differenced in train and test R2 score **<5%**
There are not underfitting: The models have **>90% test R2 score & >12% test MAPE score** with very similar performance (**R2: 90-94%, MAPE: 9.8-11.9%**)

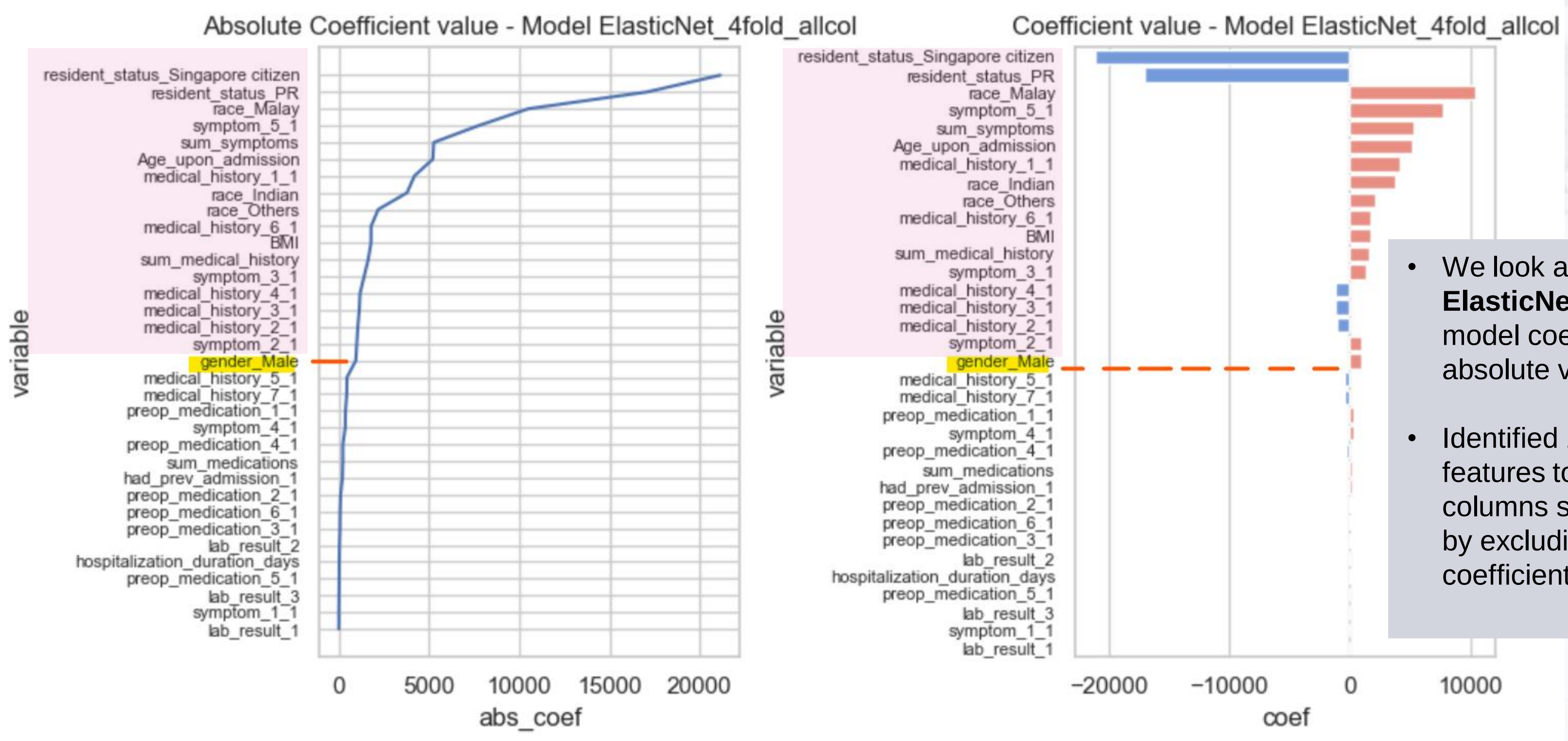
General observation:

- All columns appear to give better model performance.
- 4-fold data size appear to give better model performance.
- Elastic Net appear to perform better than other statistical models.

This make sense as:

- All columns allows more columns subjected to the regularization.
- 4-fold all model to train better with better representation of the population
- Elastic Net allow use of l1 and l2 regularization

03 Statistical Model Feature Importance



- We look at **ElasticNet_4fold_allcol** model coefficients and its absolute values
- Identified 18 transformed features to form a new columns set: `sub3col` by excluding features with coefficient values close to 0

Further discussion on feature importance include in later slide

03 Machine Learning Models

- 2 Model types: RandomForestRegressor, XGBoostRegressor
- 4 X data sets used (Datasize: 4fold, Column set: ‘allcol’, ‘sub1col’, ’sub2col’, ’sub3col’)
- Total 8 trained machine learning models

Model	R2_train	R2_test	MAPE_train	MAPE_test	RMSE_train	RMSE_test
RandomForestRegressor_4fold_allcol	96.978	87.576	4.380	12.034	1836.77105	3130.89750
RandomForestRegressor_4fold_sub1col	95.571	90.485	5.577	9.959	2223.63392	2739.98994
RandomForestRegressor_4fold_sub2col	94.954	90.079	6.551	9.803	2373.40365	2797.87660
RandomForestRegressor_4fold_sub3col	97.339	92.545	4.265	8.363	1723.51501	2425.28883
XGBRegressor_4fold_allcol	99.987	97.732	0.375	4.932	120.78346	1337.71927
XGBRegressor_4fold_sub1col	99.985	97.804	0.454	4.616	128.17734	1316.16081
XGBRegressor_4fold_sub2col	99.935	92.694	0.660	8.758	269.56463	2400.98342
XGBRegressor_4fold_sub3col	99.979	97.798	0.554	4.764	151.68395	1318.01607

Compared to statistical models
(with R2 test score 90-94%, MAPE test score 9.8-11.9%):

RandomForestRegressor appear to perform poorer:
- Best R2 score – 92.5%
- It also has overfitting model with greatest train test R2 difference of 9%

XGBRegressor appear to perform best:
- Best R2 score – 98%
- ¾ model not overfitting with train test R2 ~2% difference

03 Machine Learning Model

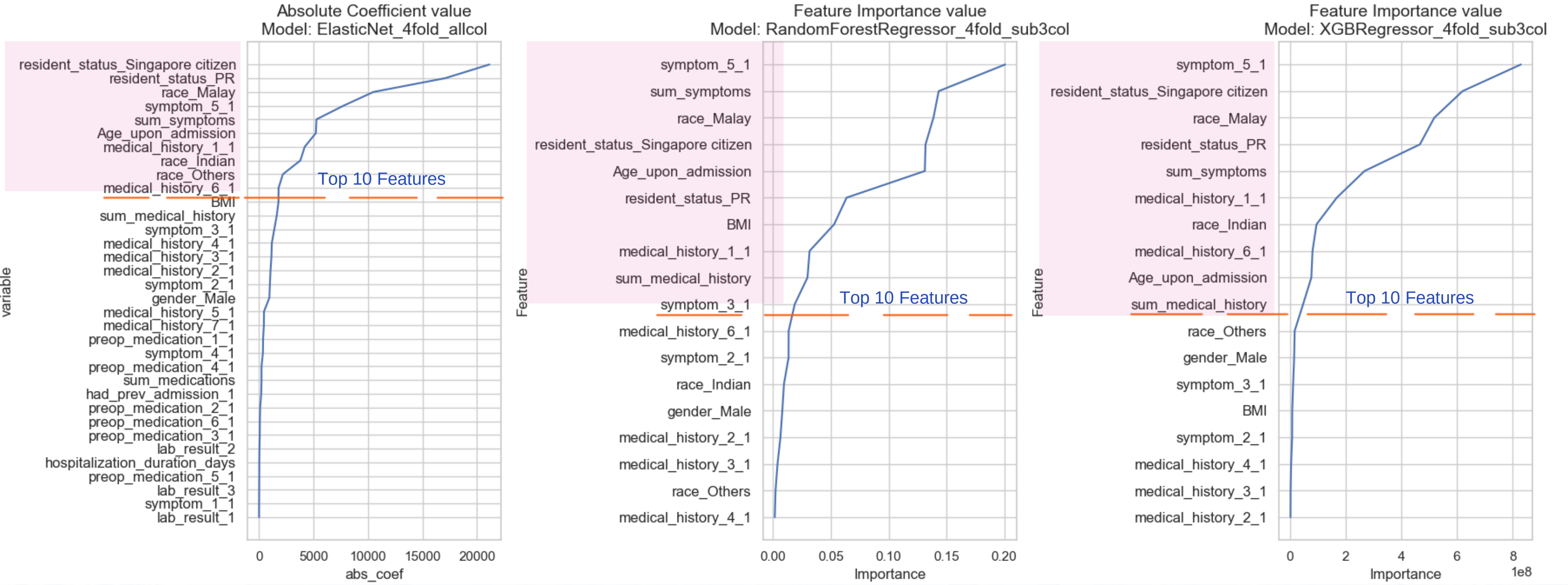
Other Observations:

- RandomForestRegressor performs better when trained on specific important features, (‘sub3col’ columns)
- While XGBRegressor showed better performance when more columns (it perform least well on sub2col which had least columns)

Remarks:

- We can infer XGBRegressor is better selecting important features, and this is make sense.
- XGBoost offers the ability to adjust regularization parameters, granting users precise control to fine-tune the complexity of the model. It provides choices like L1 and L2 regularization, which aid in mitigating overfitting and enhancing generalization.
- In contrast, Random Forest Regressor relies on the natural randomness inherent in its ensemble construction to regulate the model, lacking direct influence over regularization parameters.

03 Feature Importance Across Models



'Resident Status, Race (Malay), Sum of symptoms and Symptom_5 appear to be top few most common features across models

03 Potential Cost Driving Factors

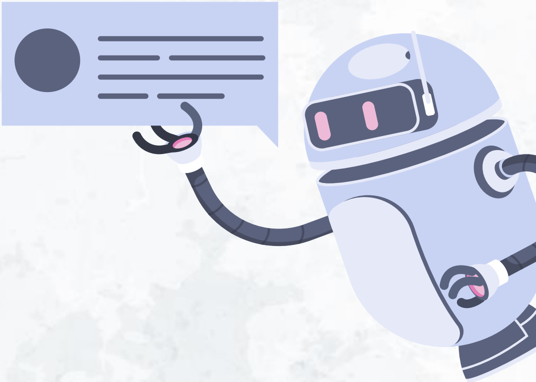
	Correlation Heat Map	Mann-Whiteney U Test	Machine Learning Models	
Analysis	*Using Pearson correlation to identify linear relationship between quantitative variable with hospital bill*	Identify categorical categories that have significant difference in hospital bill between the classes	**Models and feature importance:** ElasticNet – Absolute Coefficient values Random Forest Regressor – Impurity-based Feature Importance value XGBoost Regressor – 'gain' score Feature Importance value	
Identified Potential Cost Driving Factors	'sum_symptoms' 'Age_upon_admission'	'gender', 'medical_history_1', 'medical_history_6', 'preop_medication_2', 'preop_medication_6', 'race', 'resident_status', 'symptom_1', 'symptom_2', 'symptom_3', 'symptom_4', 'symptom_5',	**Common Top 10 Features:** 'medical_history_6_1', 'resident_status_PR', 'medical_history_1_1', 'race_Malay', 'Age_upon_admission', 'resident_status_Singapore citizen', 'sum_symptoms', 'symptom_5_1'	**Common Top 5 Features:** 'resident_status_PR', 'race_Malay', 'resident_status_Singapore citizen', 'sum_symptoms', 'symptom_5_1'

- Generally, top features identified post-modelling are also identified in the analysis conducted during data analysis.
- Most notable features are 'Resident Status, Race (Malay), Sum of symptoms and Symptom_5
- Medically, a patient tend to have higher bill if they have more symptoms, or particularly symptom 5, as they would require more treatments or more complex treatments.
- Resident Status evidently affects hospital bill as we are aware it affects one's eligibility to subsidies
- Malay patients may observe higher hospital bills, possibly due to higher prevalence of certain medical conditions and treatment needed that are more costly. However, more medical background information would be preferable to assess this possibility.

04 →

Discussion & Conclusion

(AI)



04 Recap: Task Goal

Qn. A

What factors drive cost of care?

Plan: The driving factors will be identified through Data Analysis and evaluate post-modelling feature importance

Data Analysis

Post-Modelling Evaluation

Qn. B

What are the ways to estimate/predict cost of care?

Plan: Different modelling techniques will be explored to produce one that can estimate cost accurately (Target: R2 score > 90%, MAPE score <10%)

Modelling

04 Our Findings:

Key factors influence cost

Cost tends to be higher when:

- Patient is foreigner
- Patient is Malay (likely due to higher prevalence of medical issues and need for costly medical attentions)
- Patient has more symptoms
- Patient has symptom 5

Model

- Selected model: XGBRegressor_4fold_sub3col
- R2 score: 97.8%
- MAPE score: 4.76%
- RMSE score: 1318

What this translates to is:

Given a patient information during hospitalization (found in sub3col), it is able to estimate the hospital bill with about ~4.8% and $1.3K off the true bill value.

04 Limitations & Future Work

1. Did not explore non-linear relationship between hospitalisation bill with other quantitative features:
    - Such non-linear relationships if identified will allow us to obtain other prominent quantitative cost driving factors, as we observed minimal of among the other important features assessed post-modelling.
    - We can achieve this by transforming the quantitative variables (log2, exponential, square root, polynomial transformations) and assess their correlation with the hospital bill
2. Patient's personal and family financial well-being not accounted for:
    - This affects the level of subsidy a patient is eligible for, which will significantly affect a patient's bill (as high as 80% subsidies[1])
3. Patient preference:
    - Patient's preference in choice of basic VS premium services affected the hospitalization cost
    - For example, in Singapore General Hospital[2], a class C ward which has 8 beds in a room cost SGD37 a day, compared that to class A ward which is a single room cost SGD 540 a day

[1] https://www.moh.gov.sg/healthcare-schemes-subsidies/subsidies-for-acute-inpatient-care-at-public-healthcare-institutions
[2] https://www.sgh.com.sg/patient-care/inpatient-day-surgery/type-of-wards-singapore-general-hospital

04 Concluding Statements

We have achieved what the task set out to do. We also identify there is limited information in the data set and further analysis that could have been explored.

The use of other advanced machine learning models could be further explored as well (such as deep learning models), which has also shown great results as seen in price estimator model by UCARE.AI.

AI has so much potential in not only in the advancement of medical treatments, but also in the delivery of these treatments through our healthcare systems. We can look forward to discovery new innovative ways to advance our healthcare systems as the world continues to see rising demands of medical needs.

~Thank you~