Diabetes Care Prediction

Predict quality of diabetes care in a healthcare system

12th February 2022

Adelyn | Xin Jie | Shannon | Jun Li | Jiadong

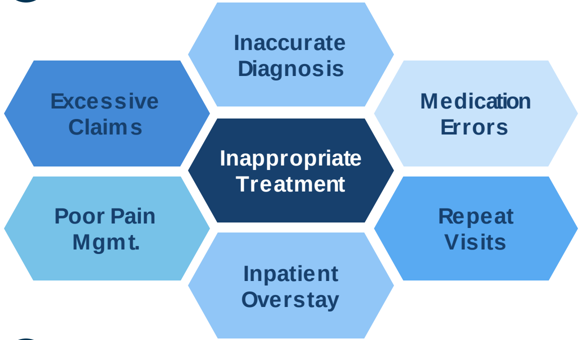
Executive Summary

Introduction	• Hospitals often struggle from the trade-off between operational efficiency & saving costs and maximising patients quality of care • Previously, doctors review patients' cases on a case-by-case basis to understand whether there is a need for medical intervention
Objective	• Management has asked to predict quality of diabetes care in healthcare system so as to provide appropriate medical interventi on to patients receiving poor care
Error Bias	• We prefer a low threshold (i.e., higher sensitivity) as the error cost for higher sensitivity is higher operating costs but the error cost for higher specificity is misidentification of poor care patients as having good care
Approach	• Exploratory data analysis coupled with objective assessment to filter list of likely variables to include in the model • Oversampling to make up for data imbalance in dependent variable • Data transformation, hyperparameter tuning and cross-validation
Model	• Focuses on three variables – ER + Office Visits, Narcotics and Started on Combination • This means that the likelihood of poor care is higher when the patient experiences a higher number of ER and Office visits, higher number of times prescribed and/or is given a drug combination.
Evaluation	• Out-of-sample accuracy of 0.825, Sensitivity of 0.778 • Model generalises well from our training data to unseen data, no overfitting problems
Limitation	• Imbalanced dataset • Dataset not fully representative of all patient experience - only those that are declared to the insurers
Recommendation	• Look into the following plausible factors - Patient throughput, Waiting Times, Consultation Times, Appointment Schedules • Model Lifecycle and Management – for progress tracking and patient & problem prioritisation

Improving quality of care is important, but currently inefficient and subjective

Improving care lessens economic and social costs…

1 **Low quality care increases burden of illness and health costs…**



15% of hospital expenditure is due to poor care

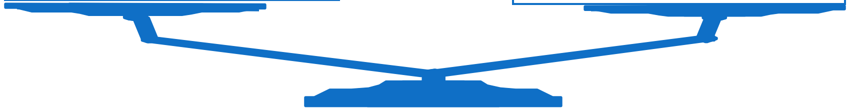
27% cost savings if hospitals implement quality management

2 **… but hospitals struggle with the tradeoff of operational efficiency and quality of care**

Operational Efficiency
Patient throughput
Bed utilisation
ED capacity
Profit motives

Resource Constraints

Quality of Care
Patient satisfaction
Readmission rates
Treatment modality
Unnecessary treatment



… and should be data-driven and objective

3 **Data-driven models should supplement expert physicians' evaluation…**

	Time Consuming	*Subjective*
Current: *Doctors' Evaluation*	Assess on a case-by-case basis, **impractical** to review for millions of patient	Different guidelines to define quality of healthcare
New: *Models*	*Able to analyse a large number of observations*	Should be **easy to understand, actionable** and eliminates human biases

4 **… and glean new insights through alternative data sources by partnering with players across the healthcare value chain**

Consumers	*Fiscal Intermediaries*	*Providers*
Individuals **Employers** **Government**	**Insurers** **HMOs** **Pharmacy Benefit Managers**	**Hospitals** **Physicians** **Pharmacies**

Purchases and provides care data

Feedback loop: our dataset

Offers healthcare services

Source: IQVIA, WHO, Lit Analysis

We use a logistic regression model that is rationally guided in our analysis

Our framework in approaching the problem

Logistic Regression	**Dependent Var. is binary** - Good Care ("0"), Poor Care ("1") Linear regression would predict a continuous outcome.
Objective	Model should effectively differentiate **poor care and good care** cases so as to provide **timely medical intervention** to poor care patients
Error Preference	<u>**Why we prefer a lower threshold model?**</u> **High sensitivity**, $\frac{TP}{TP+FN}$, is preferred to high specificity $\frac{TN}{TN+FP}$. • The model should prioritise accurately predicting patients that receive poor care for timely intervention than patients already receiving good care. • The <u>error cost of a highly sensitive model</u> is that the hospital pays more to **provide better quality care for those who are already receiving good quality care (FN)** • The <u>error cost of a highly specific model</u> is that the hospital **wrongly deduce** patients receiving poor care as having good care, **and takes no action to improve care quality**.
Model Priority	A **simple** model with few covariates not only **reduces probability of overfitting**, but **helps to prioritise management focus**.

	Predicted = Good Care (0)	*Predicted = Poor Care (1)*
Actual = Good Care (0)	True Negatives (TN)	False Positives (FP)
Actual = Poor Care (1)	*False Negatives (FN)*	True Positives (TP)

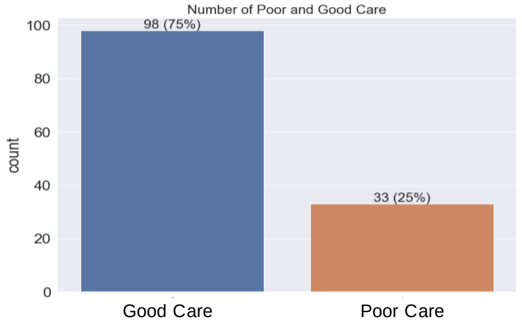
Our methodology in applying our framework

1	**Problem Analysis**	• Identify **objective** of model and any **error preference** embedded in the model
2	**Explor-atory Data Analysis**	• **Initial qualitative assessment** and **weighing of variable likelihood** in impacting quality of care • **Data visualisation** to prove or disprove hypothesis • **List possible variables** to include in the model
3	**Model Dev.**	• **Multicollinearity** check on list of possible variables • **Oversampling** to make up for data imbalance in dependent variable • **Data Transformation** • **Hyperparameter Tuning, Cross Validation**
4	**Model Evaluation**	• **P-value checks** on variables to finalise list in model • **Logical checks** on variable coefficients • **Optimal Threshold** for ROC-AUC curve • **Performance measurement** of model using ROC-AUC curve, Confusion Matrix and Prob. Density Plot
5	**Conclus-ions**	• **Recommendations** to management based on model focus

Analysing the dependent and independent variables

Dependent Variable - PoorCare

1 **Imbalance of data** in the data set, where the majority patients received good care.



2 **What does this mean for our model development and evaluation?**

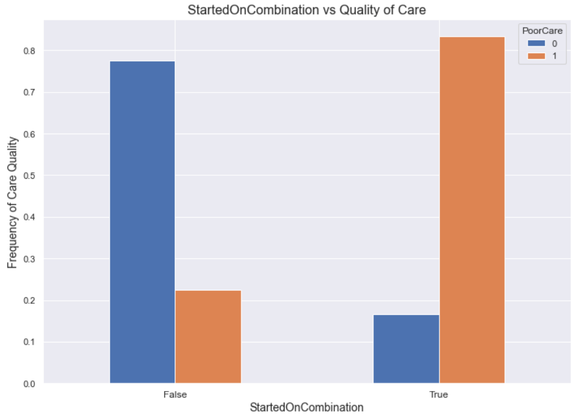
- **Model Development** - Consider **oversampling** to ensure that patients that received both good and poor-quality care are equally represented in the model and will be less biased towards predicting good quality care
- **Model Evaluation** - Since the percentage of patients receiving poor care is 25.2%, this means that the **baseline accuracy of the model is 75%**.

Qualitative assessment of independent variables on care quality

Variable	Description	Initial Hypothesis – Extent of affecting quality of care?
Member ID	Identifies the member	No
Inpatient Days	No. of days patient stayed in hospital	Medium; Inpatient overstay
ER Visits	No. of visits patient made to emergency department	Medium/High; Long wait times
Office Visits	No. of visits patient made to the office/clinic	High; Unnecessary follow-ups
Narcotics	No. of times patient was prescribed drugs	High; Unnecessary follow-ups
Days Since Last ER Visits	No. of days between patient's last emergency department visit and the time the data was collected	Low
Pain	No. of visits where patient complained of pain	Medium; Treatment modality
Total Visits	No. of times patient visited any healthcare facility for treatment	High; Unnecessary follow-ups
Provider Count	No. of unique healthcare providers that the patient visited	Medium; Possibility of poor care that leads to repeated visits or changing of physicians
Medical Claims	No. of days of which patient had a medical claim	
Claim Lines	Total number of medical claims	
Started On Combination	Whether the patient was given a combination of drugs	High; Over-prescription
Acute Drug Gap Small	Fraction of acute drugs refilled after prescription ran out	Low

Deep Dive: Analysing variables with high likelihood of affecting quality of care

Started on Combination



Hypothesis

- Patients on more **drug combinations are more likely to be over-prescribed**

Observations

- **Care was mostly poor for patients receiving drug combinations** (‘T’) and vice versa
- However, not many instances (6; 4.6%) where StartedOnCombination = ‘T’.

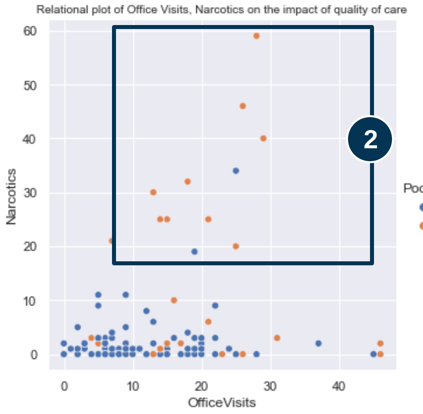
Therefore, **consider Started on Combination** as independent variable in the model

Narcotics, Office Visits

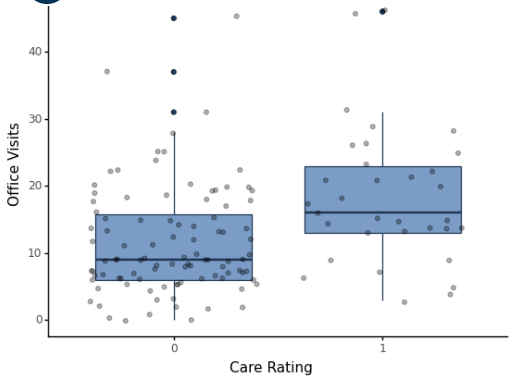
1

Correlation of Narcotics with Office Visits

0.27576



3



Hypothesis

- The relationship between narcotics and office visits is important as it helps us determine whether the no. of times a patient visits the hospital trends with the no. of times being prescribed with medication, or **whether we should think of them separately** in affecting poor quality of care (e.g. **narcotics – over prescription, office visits – misdiagnosis or unnecessary follow-up appointments**)

Observations

1. We should think of office visits and narcotics as separate variables that directly affects poor quality of care
2. Prominent that quality of care is poorer when the patient was prescribed a greater amount/number of drugs
3. There is also a relative difference in office visits between patients receiving poor care and good care.

Over prescription (i.e., **narcotics**) **seems to be a bigger concern**, but **high no. of office visits are also concerning and could signal unnecessary follow-ups** incentivized by the hospital’s profit motives

Deep Dive: Analysing variables with high likelihood of affecting quality of care

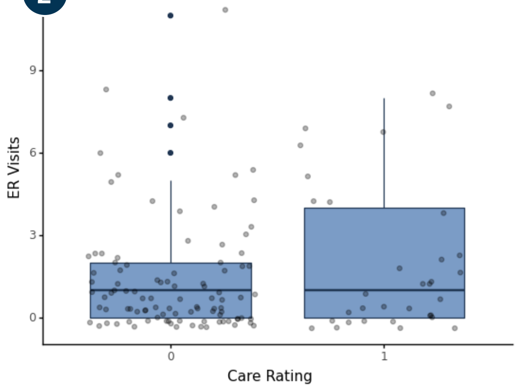
Breaking down the differences between ER Visits, Office Visits and Total Visits

Logically, ER Visits + Office Visits = Total Visits
However, total visits seem to include other visits besides office visits and ER visits.

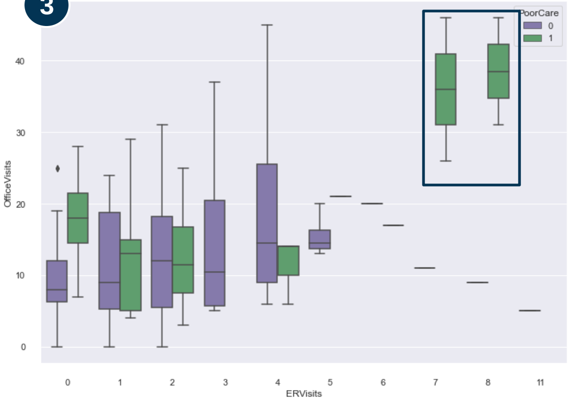
1

Corr	Total Visits	Office Visits	ER Visits
Total Visits	1.000000	0.865387	0.586439
Office Visits	0.865387	1.000000	0.308526
ER Visits	0.586439	0.308526	1.000000

2



3



Hypothesis

- Since high number of **total visits or office visits** may allude to **misdiagnosis, long-waiting times or unnecessary follow-ups**, we **suspect high correlation**.
- **ER Visits should be less correlated with the former two**, but **long waiting times in ER department** may ultimately drive poor ER care quality.

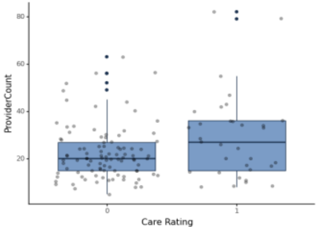
Observations

1. High correlation between total visits and office visits (0.865), but less for ER visits (w Total: 0.586; w Office: 0.309)
2. Poor care patients exhibits a larger distribution of ER Visits. The **median is however similar** between both poor and good care patients.
3. Previously, we mentioned that office visits could be a significant predictor of quality of care. Interestingly, patients who experience **high ER Visits (7-8) all report poor quality care**. At the same time, they are **also the ones who pay a significantly greater number of office visits compared to patients who only visited the ER for a fewer number of occasions**.

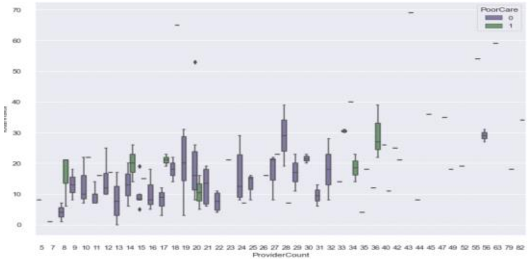
ER Visits alone does not seem to directly affect the quality of care (high distribution, same median). However, since people who had high no. of office and ER visits reported poor care, we want to consider the **summation of office visits and ER visits as a variable** (long waiting times during visits can affect the care quality). This approach is preferable to using Total Visits alone, as we do not know what other types of visits are factored and hence, our recommendation to management.

Deep Dive: Analysing variables with lower likelihood of affecting quality of care

Provider Count



Poor care patients experienced a **higher median** of provider counts, suggesting that some of them had to **switch between multiple providers**

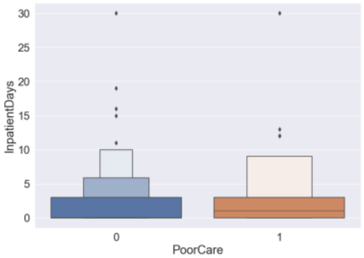


Hypothesis
More visits at different providers, could suggest that their **cumulative visit experiences and care quality received may not be up to par**

Observations
- Higher provider counts hints at poor quality of care
- However, the same could be said for those who visited a low no. of healthcare providers and still received poor care

Provider count **should not be included** in the model

Inpatient Days



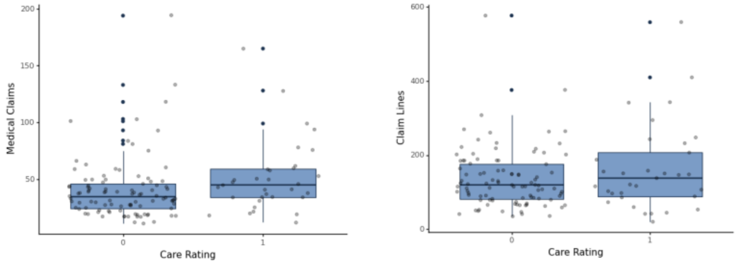
Hypothesis
High inpatient days may signal unnecessary overstay

Observations
No significant impact on quality of care

Hence, **inpatient days should not be included** in the model

Medical Claims, Claim Lines

1



2

Corr	Medical Claims	Claim Lines	Office Visits
Medical Claims	1.000000	0.813935	0.498513
Claim Lines	0.813935	1.000000	0.424953
Office Visits	0.498513	0.424953	1.000000

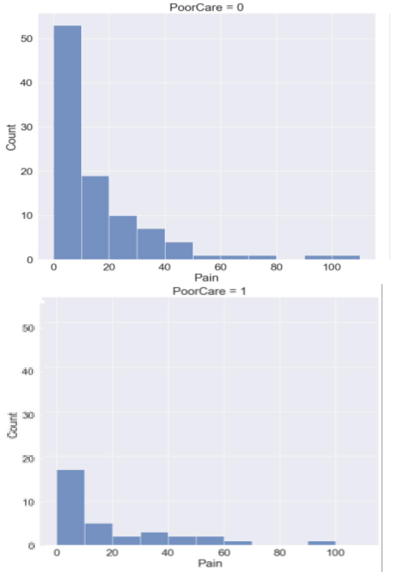
Hypothesis
Medical claims is highly correlated with office visits, as medical claims to insurers can only be made post office/ER visits.

Observations
1. Poor quality care patients reporting a **slightly higher** median number of medical claims and claim lines
2. Corr. between Medical Claims and Office Visits is relatively high (0.5)

Medical claims and claim lines are **reported by insurance companies and might not be fully accurate** since patients might not have fully/accurately declared all of their medical claims. Hence, we **do not include** this variable.

Deep Dive: Analysing variables with lower likelihood of affecting quality of care

Pain



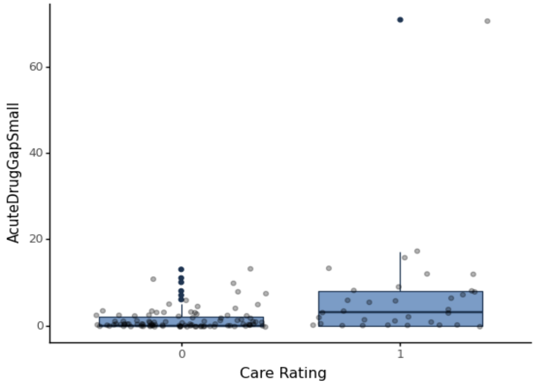
Hypothesis
More painful experiences = Poor Care

Observations
Most patients in both care groups experienced **minimal painful visits**

Pain **not included** in the model

Acute Drug Gap Small (ADGS)

1



2

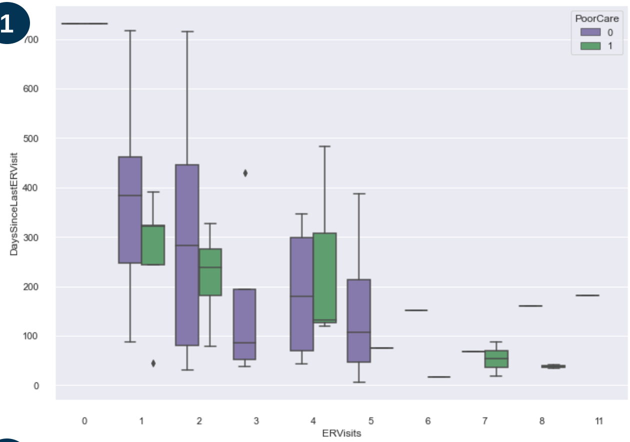
Correlation of ADGS with Narcotics
0.71089

Hypothesis
Slower refills after drug ran out = Poor Care
Observations
1. On the contrary, patients who had poor care had more instances of immediate drug refills after they ran out.
2. Correlation of narcotics with ADGS is high (0.71) as more narcotics could increase the need for refill.

Acute Drug Gap Small **not included** in the model

Days Since Last ER Visit (DSLEV)

1



2

Correlation with ER Visits
-0.73525

Hypothesis
Negative correlation with ER Visits and minimal impact on quality of care
Observations
1. As ER visits increase, the median for DSLEV largely decreases for both good and poor care patients
2. Negative correlation with ER Visits (-0.74)

DSLEV **not included** in the model

Model Development

Multicollinearity, Logical and P-Value Checks

1

Corr	ER + Office Visits	Narcotics	StartedOnCombination_False
ER + Office Visits	1.000000	0.250464	-0.175332
Narcotics	0.250464	1.000000	-0.043641
StartedOnCombination_False	-0.175332	-0.043641	1.000000

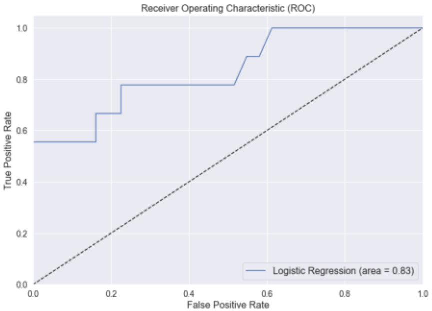
```
                           Logit Regression Results
==============================================================================
Dep. Variable:               PoorCare   No. Observations:                  134
Model:                          Logit   Df Residuals:                      131
Method:                           MLE   Df Model:                            2
Date:                Sat, 12 Feb 2022   Pseudo R-squ.:                  0.3172
Time:                        23:48:36   Log-Likelihood:                -63.422
converged:                       True   LL-Null:                       -92.882
Covariance Type:            nonrobust   LLR p-value:                 1.606e-13
==============================================================================================
                                 coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------
Total_ER_Office_Visits         0.0674      0.018      3.659      0.000       0.031       0.104
Narcotics                      0.1101      0.033      3.339      0.001       0.045       0.175
StartedOnCombination_False    -2.1036      0.381     -5.520      0.000      -2.850      -1.357
==============================================================================================
```

1 All variables are weakly correlated, **multicollinearity is absent**

2 **Logical check:** +ve Increase in Narcotics and Total ER Office Visits lead to greater likelihood of poor care. Started on combination false, leads to lower likelihood of poor care

3 **P-value** for all variables is less than 0.05, and hence ind. variables are likely statistically significant at our chosen confidence level of 95%

Finding the optimal threshold

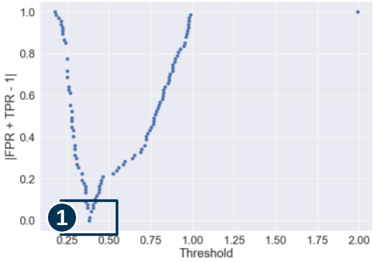


AUC of **0.83** which is close to 1

This means our model has **83% chance of distinguishing between patients who received poor care and good care** and is highly effective.

Threshold tuning based on ROC curve

Without Optimal Threshold	
Accuracy	0.825
Precision	0.625
Recall	0.556
Specificity	0.556

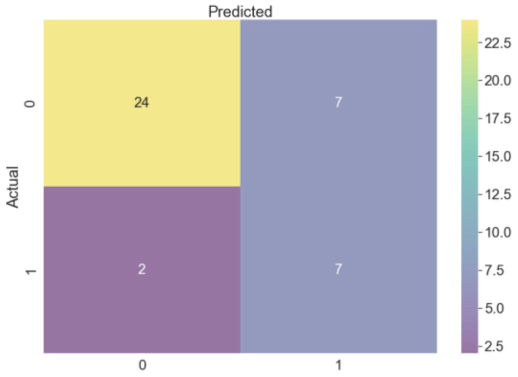


Why Threshold Tuning?
Accuracy may be misleading because of the imbalanced class distribution

Threshold Tuning – optimised for high TPR, low FPR
- **Threshold achieved: 0.38312**
- Sense check: threshold < 0.5, it is a low threshold that prefers sensitivity which ties into our error bias

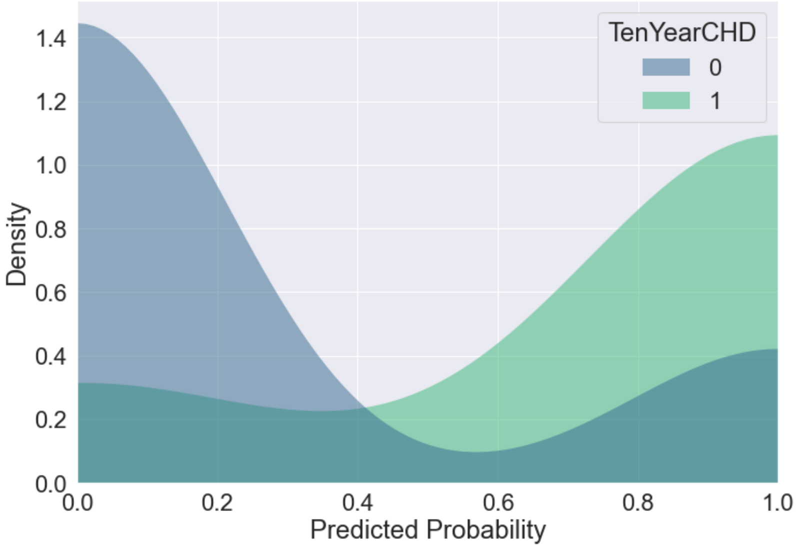
Model Evaluation

Confusion Matrix



At optimal threshold	
Accuracy	0.825
Precision	0.500
Recall	0.778
Specificity	0.556

Probability Density Graph



- ✓ **Out-of-sample accuracy of 0.825 is above baseline accuracy of 0.75**. Model generalises well and is more accurate than just predicting the most-frequent class – in this case, everyone as having good care.
- ✓ **Recall / Sensitivity is 0.778**. By lowering threshold, we increased sensitivity while decreasing specificity. As mentioned, the error cost for higher sensitivity is higher operating costs but the error cost for higher specificity is misidentification of poor care patients as having good care.

- **Threshold adjustment** – we set the prediction = 0 if the prediction < threshold, and the prediction = 1 for prediction > threshold
- Probability Density Graph shows that our model will be **effective in identifying patients that has suffered from poor care** (i.e., Green area – 1)

Conclusion

Our logistic regression model

Variable	What it means	What the hospital can look into
ER + Office Visits	**Higher number of ER and Office visits**, the more likely the patient would receive poor care	• Patient throughput • Waiting Times • Consultation Times • Appointment Schedules
Narcotics	**Higher number of times the patient was prescribed drugs**, the more likely the patient would receive poor care	• Over-prescription
Started On Combination _False	If the patient **was given a combination of drugs**, the more likely the patient would receive poor care	• Over-prescription

<u>Why we believe prioritising sensitivity make sense not just in terms of short-term objective, but also in terms of long-term strategic outlook</u>

- ✓ **Operational efficiency can align with patient care objectives when maximising resources** (e.g. Shorter waiting times, having a flexible e-appt system improve patient experience and prioritise patients in-need)
- ✓ **Providing good healthcare strengthens the hospital's reputation**, which in turn, helps increase the patient base and revenue opportunities in the long-term

Model Implementation

Limitations	Recommendations
• **Imbalanced** dataset • Dataset **not fully representative** of all patient experience - only those that are declared to the insurers	Could obtain data from **National Electronic Health Records** (NEHR) in Singapore, alongside survey and feedback forms, to obtain a **larger and more balanced** data set of patients who have either received "poor care" and "good care" to train the model

<u>Model Lifecycle and Management</u>

Why is it important?

- Capture new data for continuous learning
- Retrain models so they continually adapt to changing conditions

Contextualised to problem

In this context, re-evaluating the models from time to time is not only important in understanding where else to improve, but can also be an indicator on how the hospital has improved.

Future models can also consider time-weighting the data-set, to prioritise hospital resources to the pressing problems of the day.