

Diabetic Patients' Re-admission within 30 days Prediction

Team 16:

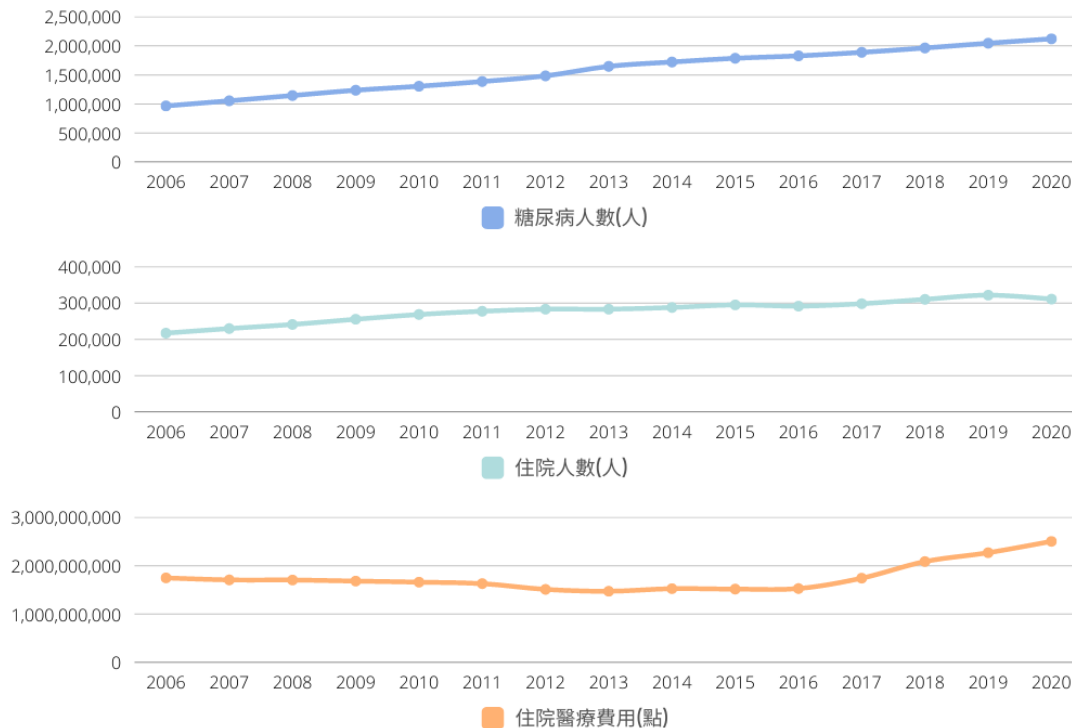
411551024 陳敏楨 411551030 吳羽佳
310581002 鄭乃心 0886007 李雅文

Report Date: 20221026



Background

資料來源:台灣衛生福利部 / 資料整理:本計畫



2,123,802

約212萬人



311,235

約31萬人



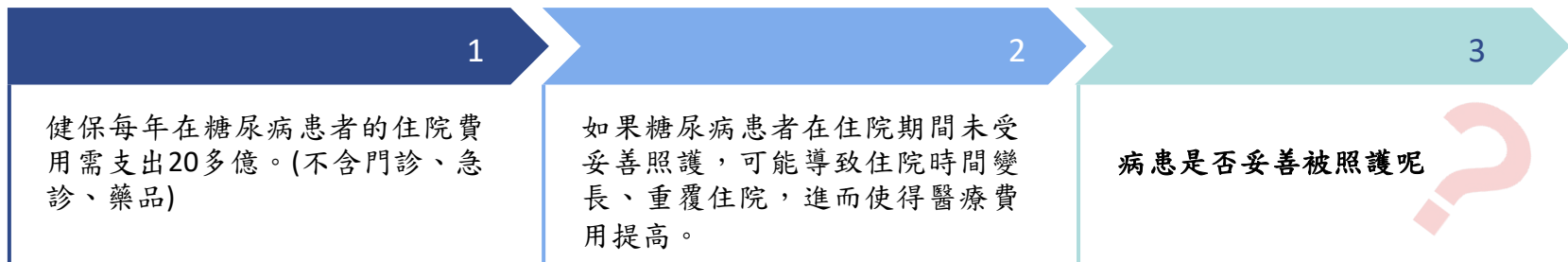
2,503,774,906

約25億元



Motivation and Research Aims

- 糖尿病住院人數多, 住院費用高



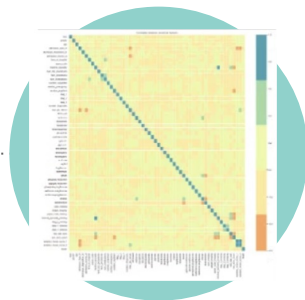
- 出院N天後再住院 (e.g., 30天)

- 為病人住院是否獲得妥善醫療照護的衡量指標。若病人出院N天後再住院，表示醫院對住院病人的照護可能需再加強。
- 藉由此指標，可以督促醫院更深入瞭解原因，並提升住院病人的醫療照護品質。

Challenge

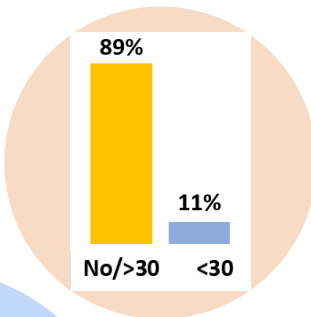
3. Feature Selection

特徵數多
關聯性不強



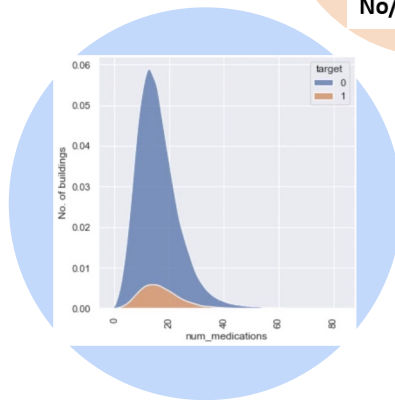
2. Imbalance

資料嚴重不平衡
89% : 11%
(No / >30 : <30)



1. Pattern

資料過於均勻
沒有明顯的pattern



Problem Description

- 使用臨床資料集訓練機器學習模型，用以糖尿病患者出院 30 天再住院之預測。
- **Input**：資料集包含人口統計學、相關的基本資料（例如：科別、檢驗次數、付款人與保險）、血糖的數值與指標、ICD-9 code、住院及用藥紀錄等欄位。
- **Process**：使用皮爾遜積差相關係數找到與 Label 較相關的特徵，並以 Apriori 演算法與決策樹找出糖尿病患者與再住院相關變數之間的關聯及再住院的風險群。選擇 5 種機器學習模型進行 ensemble。
- **Output**：預測糖尿病病人是否 30 天再住院，並得到再住院的關聯規則與高低風險群。可用於出院前準備，預警醫事人員找出可能尚未被完全診療的高危糖尿病患者，藉此減少可能產生的後續醫療照護支出與社會成本。

Target Performance

Previous Studies

- 參考過去的研究 [1,2,3]，AUROC 介於 [0.5, 0.7)。



Target I

- AUROC: 0.75 up
- Precision: 0.70 up
- Recall: 0.60 up
- F1 score: 0.65 up

Target II

- 找出再住院的關聯規則
- 分析再住院的高低風險群

[1] Mingle, D. (2017). Predicting diabetic readmission rates: moving beyond HbA1c. *Current Trends in Biomedical Engineering & Biosciences*, 7(3), 555707.

[2] Shang, Y., Jiang, K., Wang, L., Zhang, Z., Zhou, S., Liu, Y., ... & Wu, H. (2021). The 30-days hospital readmission risk in diabetic patients: predictive modeling with machine learning classifiers. *BMC medical informatics and decision making*, 21(2), 1-11.

[3] Using Machine Learning to Predict Hospital Readmission for Patients with Diabetes with Scikit-Learn, <https://towardsdatascience.com/predicting-hospital-readmission-for-patients-with-diabetes-using-scikit-learn-a2e359b15f0>

Data Description



Data Description (Cont.)

名稱 (英文)	名稱 (中文)	型態	值域	缺值	說明
encounter_id	住院識別碼	Numerical		None	共101,766筆
patient_nbr	患者識別碼	Numerical		None	共71,518人，其中3萬多人超過1次住院
race	種族	Categorical	5 類	2.23%	種族：白人、亞洲人、黑人、西班牙裔人
gender	性別	Categorical	3 類	0% (3筆)	性別：男、女、未知/無效
age	年齡	Categorical	10 類	None	各類間相差10歲：[0,10]、...、[90,100]
weight	體重	Numerical	9 類	96.86%	以磅為單位，各類間相差25磅
admission_type_id	住院類型	Categorical	8 類	5.20%	住院的類型，例如緊急、選擇性、新生兒...
discharge_disposition_id	出院後的安置地點	Categorical	29 類	3.63%	出院後去的地方，例如回家、另一家醫院...
admission_source_id	入院來源	Categorical	26 類	6.66%	住院的來源，例如醫生轉診、急診室...
time_in_hospital	住院天數	Numerical	1 ~ 14天	None	從住院到出院間的天數
payer_code	付款人代碼	Categorical	17類	39.56%	付款人代碼，例如藍十字、健保和自付
medical_specialty	醫療專業	Categorical	72類	49.08%	醫療專業，例如心臟病學、內科、...、外科
num_lab_procedures	檢驗次數	Numerical	1 ~ 132次	None	住院期間進行實驗室檢查的次數

Data Description (Cont.)

名稱 (英文)	名稱 (中文)	型態	值域	缺值	說明
num_procedures	程序次數	Numerical	0 ~ 6次	None	住院期間進行的程序次數(實驗室檢查除外)
num_medications	用藥次數	Numerical	1 ~ 81次	None	在住院期間管理的不同藥物名稱的數量
number_outpatient	門診次數	Numerical	0 ~ 42次	None	在住院前一年患者的門診次數
number_emergency	急診次數	Numerical	0 ~ 76次	None	在住院前一年患者的急診次數
number_inpatient	住院次數	Numerical	0 ~ 21次	None	在住院前一年患者的住院次數
diag_1	初步診斷	Categorical		0.02% (21筆)	初步診斷 (編碼為 ICD9 的前三個數字)
diag_2	輔助診斷	Categorical		0.35% (358筆)	輔助診斷 (編碼為 ICD9 的前三個數字)
diag_3	額外的輔助診斷	Categorical		1.40% (1,423筆)	額外的輔助診斷 (編碼為 ICD9 的前三位數字)
number_diagnoses	診斷次數	Numerical	1 ~ 16次	None	輸入到系統的診斷次數
max_glu_serum	血糖	Categorical	4 類	94.75%	>200, >300, Norm, None
A1Cresult	糖化血色素	Categorical	4 類	83.28%	>7, >8, Norm, None
change	改變糖尿病用藥	Categorical	2 類	None	糖尿病藥物是否有發生改變(劑量或藥物)
diabetesMed	使用糖尿病用藥	Categorical	2 類	None	是否有使用糖尿病藥物

Data Description (Cont.)

名稱 (英文)	名稱 (中文)	型態	值域	缺值	說明
metformin	二甲雙胍	Categorical	4 類	None	糖尿病的藥物名稱，其欄位裡的值若為： <ul style="list-style-type: none">「up」代表劑量增加「down」代表劑量減少「steady」代表劑量不變「no」代表沒有開此糖尿病藥物處方
repaglinide	瑞格列奈	Categorical	4 類	None	
nateglinide	那格列奈	Categorical	4 類	None	
chlorpropamide	氯磺丙脲	Categorical	4 類	None	
glimepiride	格列美脲	Categorical	4 類	None	
acetohexamide	醋磺己脲	Categorical	4 類	None	
glipizide	格列吡嗪	Categorical	4 類	None	
glyburide	格列本脲	Categorical	4 類	None	
tolbutamide	甲苯磺丁脲	Categorical	4 類	None	
pioglitazone	吡格列酮	Categorical	4 類	None	
rosiglitazone	羅格列酮	Categorical	4 類	None	
acarbose	阿卡波糖	Categorical	4 類	None	
miglitol	米格列醇	Categorical	4 類	None	

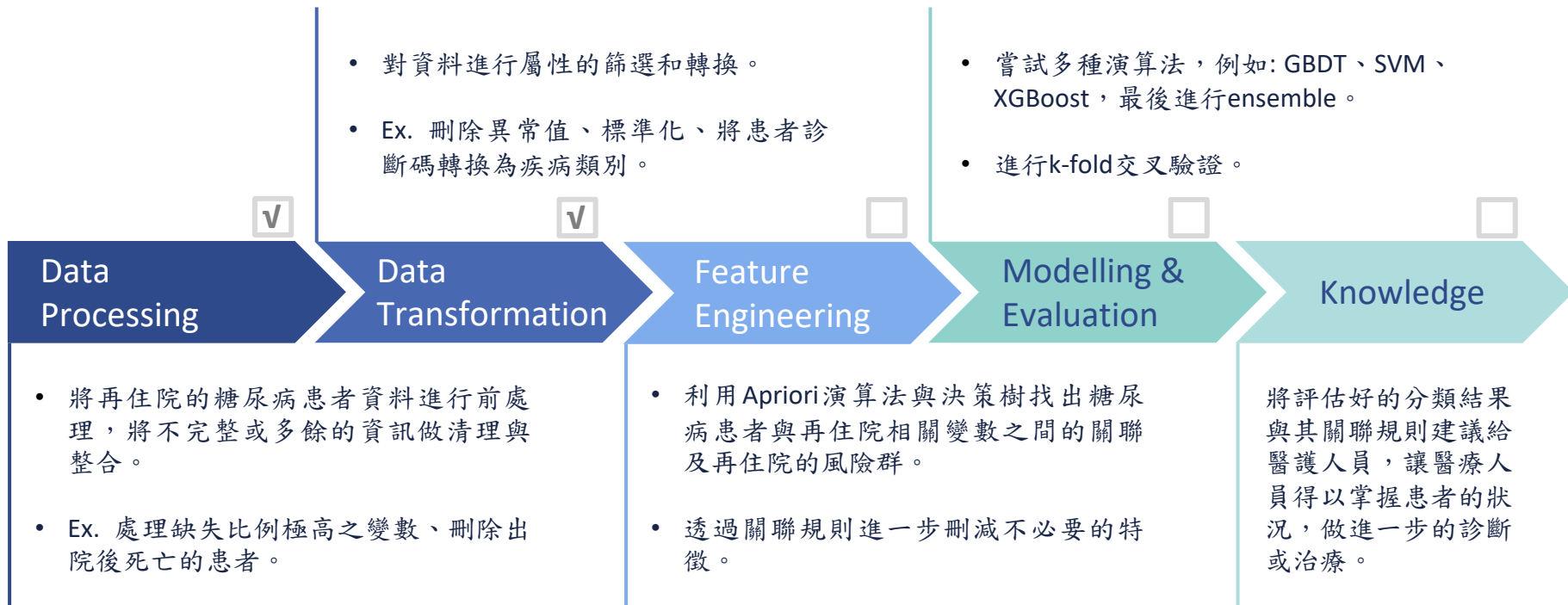
Data Description (Cont.)

名稱 (英文)	名稱 (中文)	型態	值域	缺值	說明
troglitazone	曲格列酮	Categorical	4 類	None	糖尿病的藥物名稱，其欄位裡的值若為： <ul style="list-style-type: none">「up」代表劑量增加「down」代表劑量減少「steady」代表劑量不變「no」代表沒有開此糖尿病藥物處方
tolazamide	妥拉磺脲	Categorical	4 類	None	
examide	醋酸己脲	Categorical	4 類	None	
citoglipton	西格列汀	Categorical	4 類	None	
insulin	胰島素	Categorical	4 類	None	
Glyburide-metformin	格列本脲二甲雙胍	Categorical	4 類	None	
glipizide-metformin	格列吡嗪二甲雙胍	Categorical	4 類	None	
glimepiride-pioglitazone	格列美脲吡格列酮	Categorical	4 類	None	
metformin-rosiglitazone	二甲雙胍羅格列酮	Categorical	4 類	None	
metformin-pioglitazone	二甲雙胍吡格列酮	Categorical	4 類	None	
readmitted	Y值，再次住院	Categorical	3 類	None	No , > 30天 , < 30天

Environment

項目	說明
作業系統 (OS)	Windows 10
程式語言	Python
工具	Jupyter Notebook
函式庫	scikit-learn

Analysis Workflow



Evaluation Metrics

01. Confusion Matrix

		Actual	
		30 天再住院	非 30 天再住院
Predicted	30 天再住院	True Positive (TP)	False Positive (FP)
	非 30 天再住院	False Negative (FN)	True Negative (TN)

02. Area Under the Receiver Operating Characteristic (AUROC)

- Receiver operator characteristic curve (ROC curve)下的面積，False positive rate (FPR) 為X軸、True positive rate (TPR) 為Y軸。
- 表示分類模型的預測能力。

Evaluation Metrics (Cont.)

*因涉及不平衡資料集，故採用 weighted average 作為參考標準。

03. Accuracy

模型的正確的分類率

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

05. Recall

真實再住院的病患中，有多少被預測出來

$$Recall = \frac{TP}{TP + FN}$$

04. Precision

被預測為再住院患者中，有多少為真實再住院

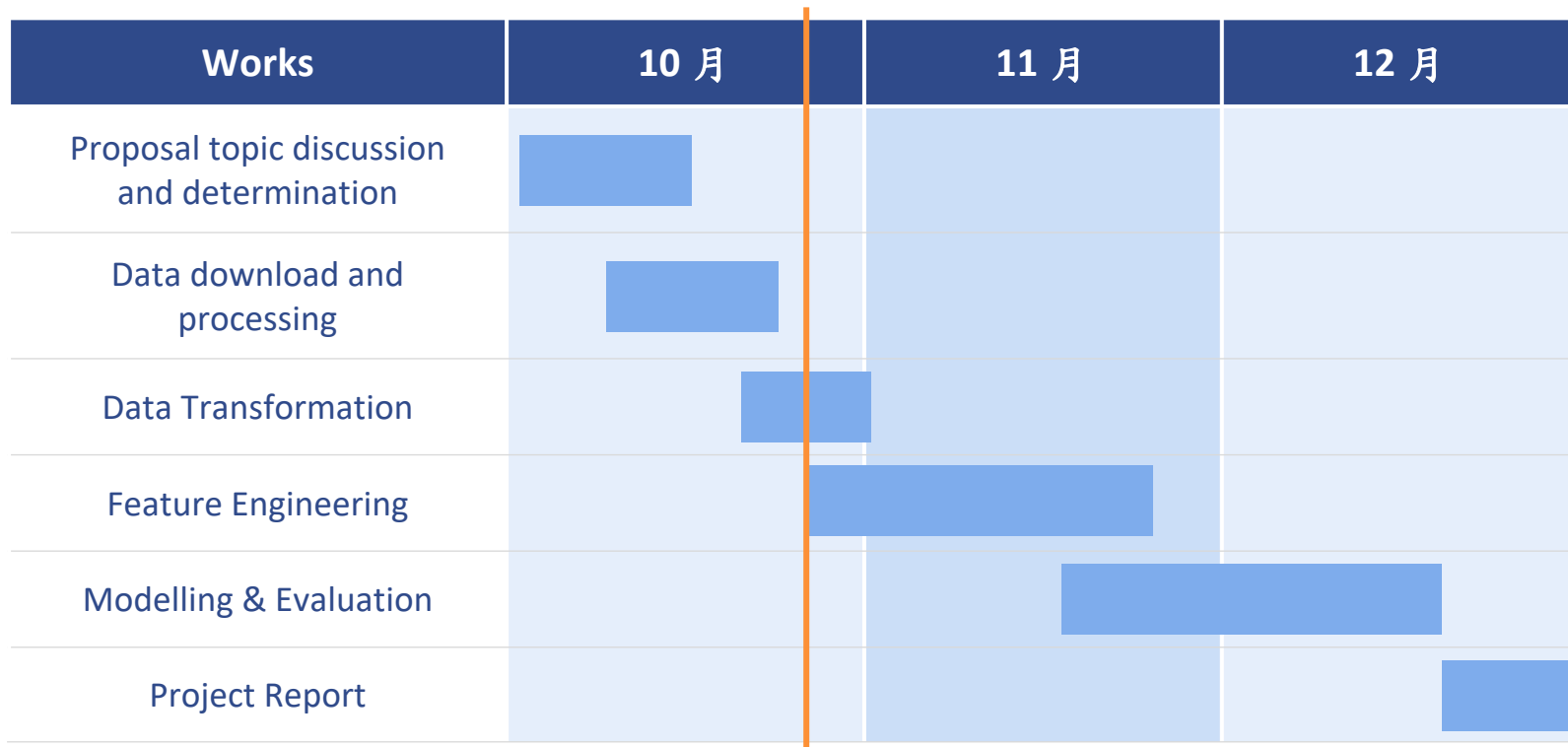
$$Precision = \frac{TP}{TP + FP}$$

06. F1 score

Precision 與 Recall 的調和平均數

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Schedule



Thank you for
your attention!

411551024 陳敏楨
310581002 鄭乃心


411551030 吳羽佳
0886007 李雅文



Q A



附件



附件一：選題的評估

✓ Kaggle Dataset (Keyword: Medical cost)

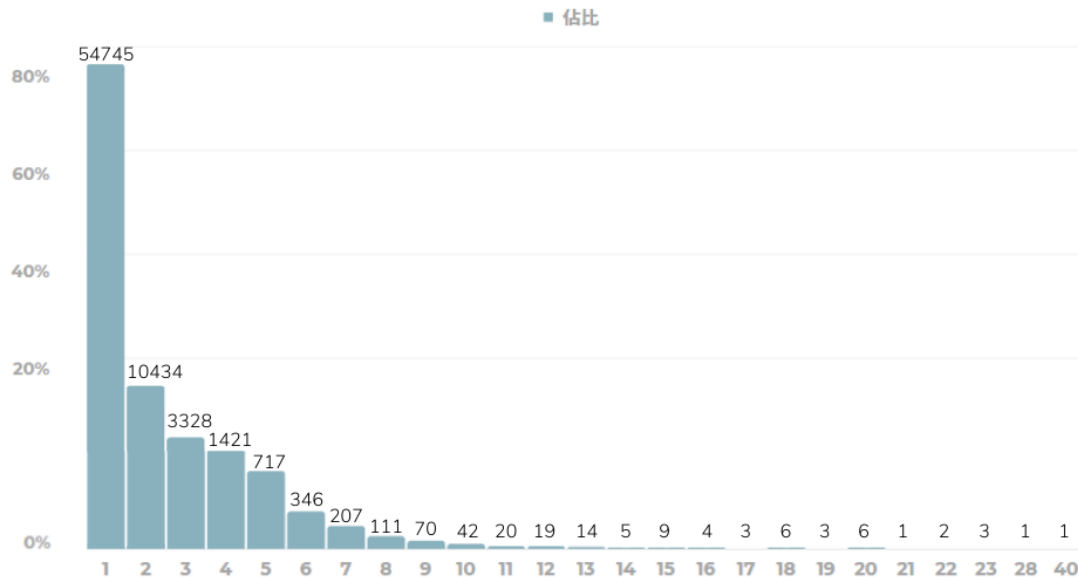
01. 資料集的大小

02. 問題的重要性

03. 解決方式的挑戰性

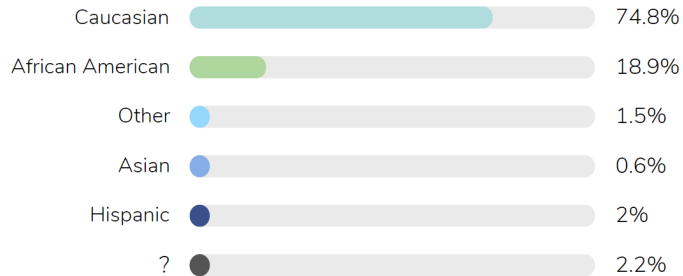
附件二：資料欄位

✓ 同一病患出現次數

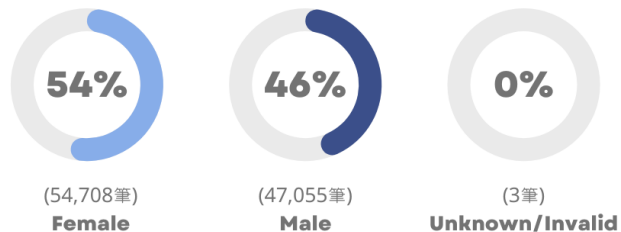


附件二：資料欄位 (Cont.)

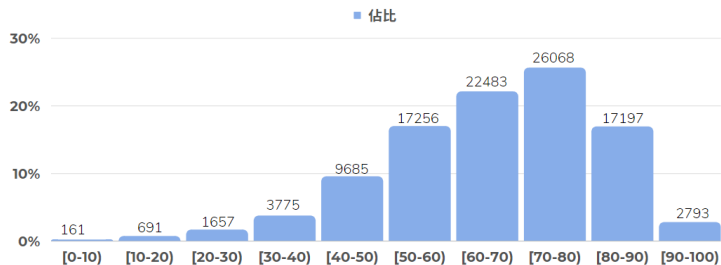
✓ 種族



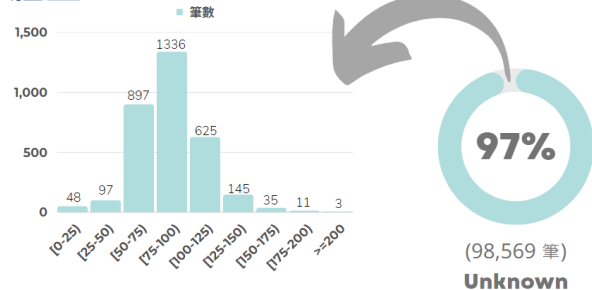
✓ 性別



✓ 年齡

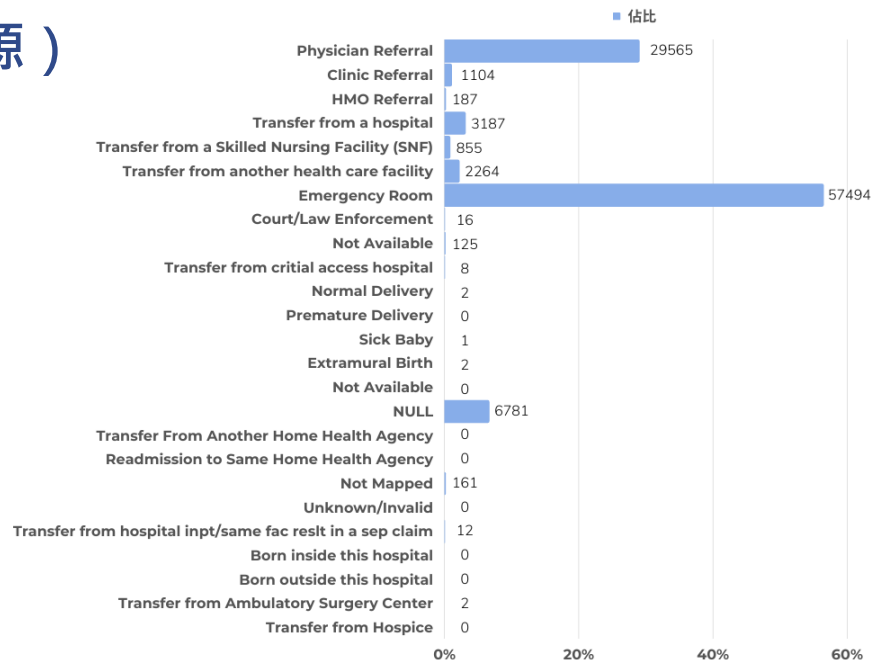


✓ 體重



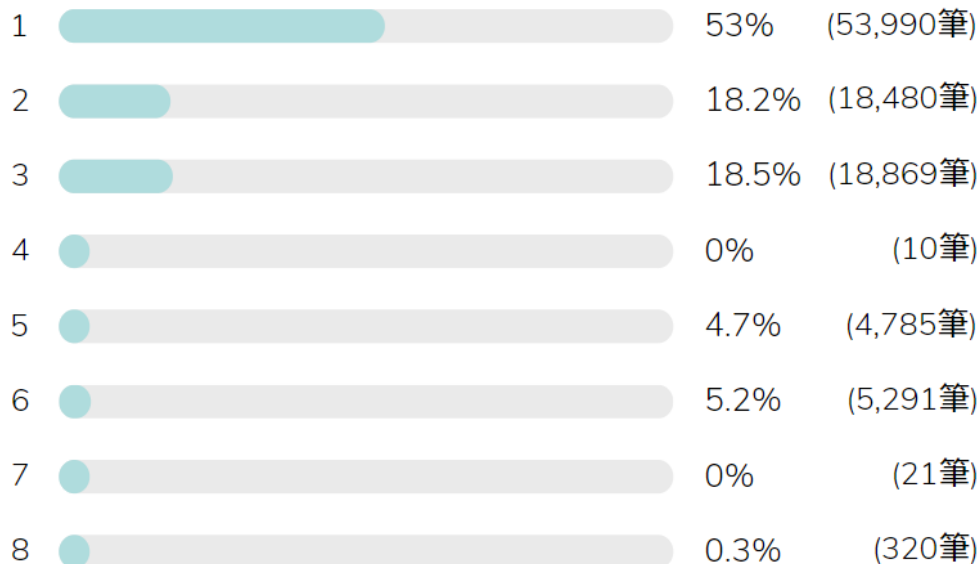
附件二：資料欄位 (Cont.)

✓ 入院 (來源)



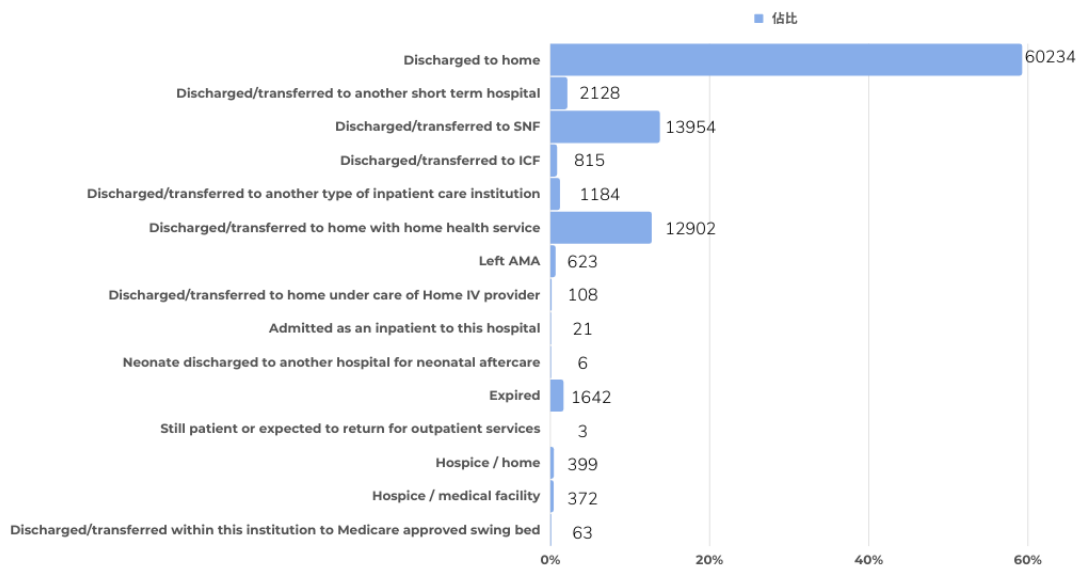
附件二：資料欄位 (Cont.)

✓ 入院 (型態)



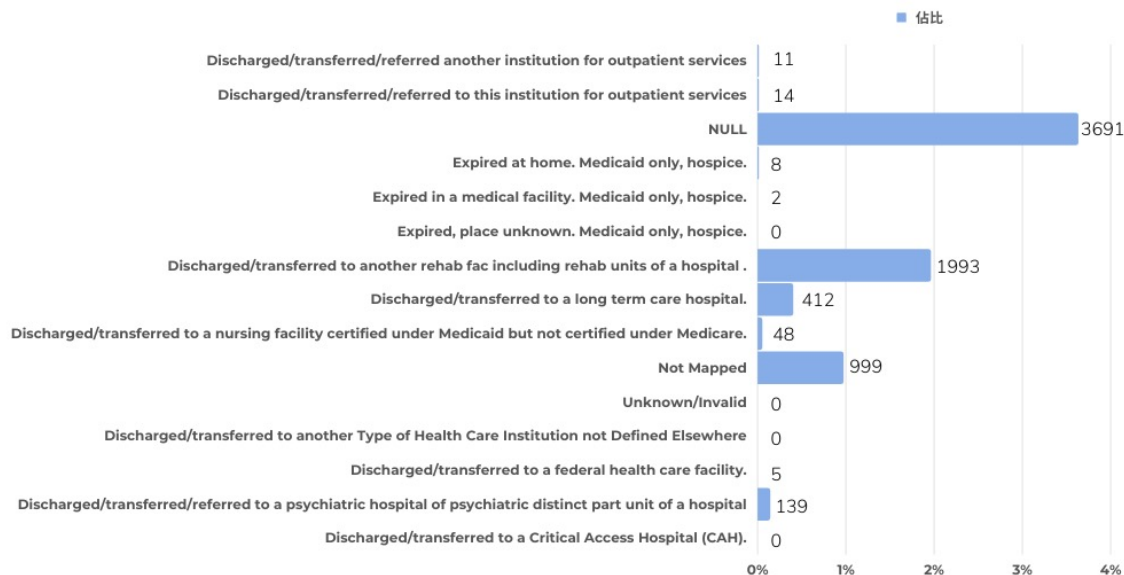
附件二：資料欄位 (Cont.)

✓ 出院 (出院後的安置地點)



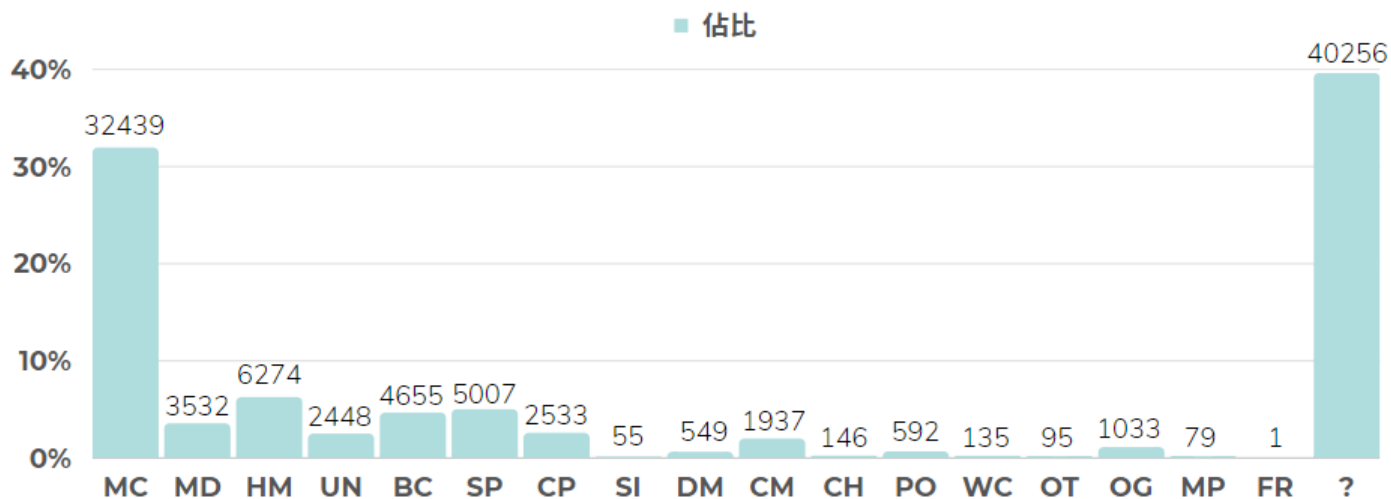
附件二：資料欄位 (Cont.)

✓ 出院 (出院後的安置地點)



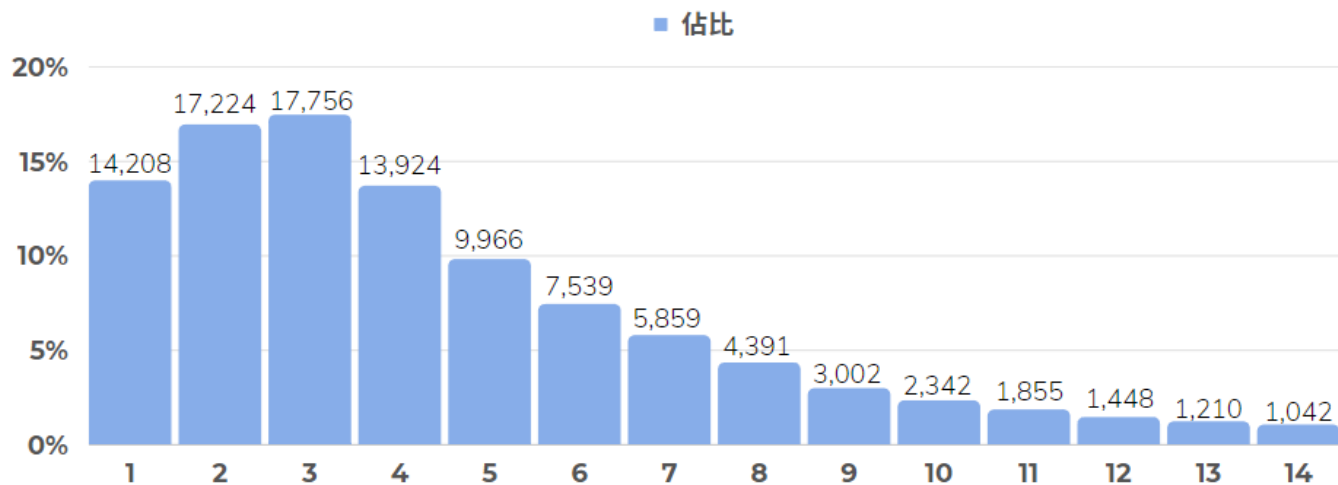
附件二：資料欄位 (Cont.)

✓ 支付系統



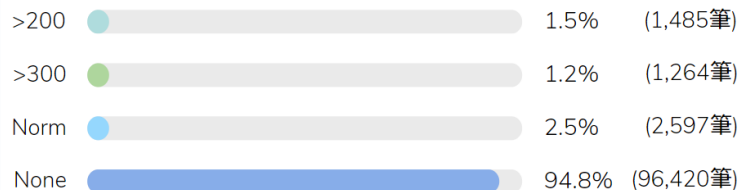
附件二：資料欄位 (Cont.)

✓ 住院天數

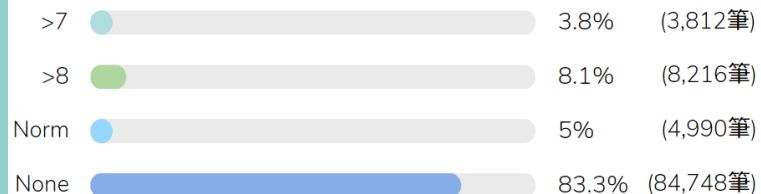


附件二：資料欄位 (Cont.)

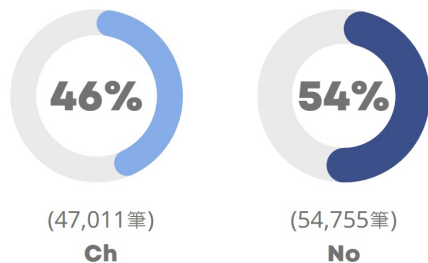
✓ 血糖值



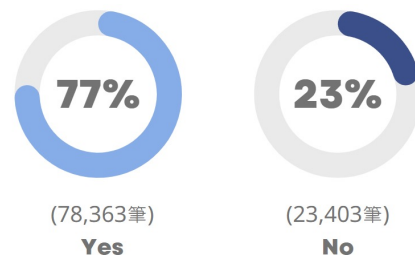
✓ 醣化血色素



✓ 改變糖尿病用藥

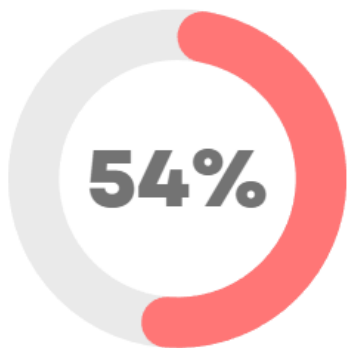


✓ 有糖尿病用藥



附件二：資料欄位 (Cont.)

✓ 再住院 (Label)



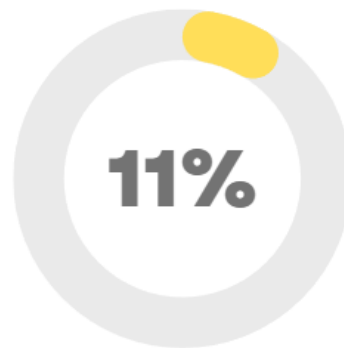
(54,864筆)

No



(35,545筆)

>30天



(11,357筆)

<30天

附件三：初步實驗結果

✓ Code: <https://reurl.cc/2m5Z1a>

AUC: 0.6579

Accuracy: 0.6385

Precision: 0.1751

Sensitivity: 0.5800

Specificity: 0.6460