以縱貫性資料預測高血糖罹患風險
Predicting the Risk of Hyperglycemia Incidence Using Longitudinal Data

實習單位：中研院資料科學統計合作社

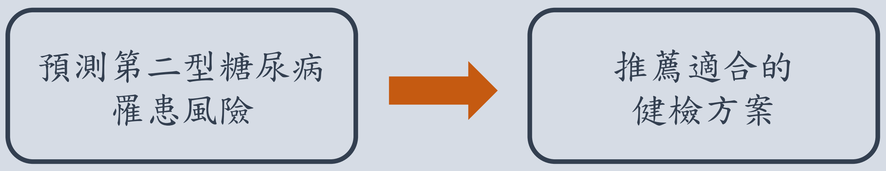
學生：吳俞萱 指導老師：王彥雯老師 單位指導：陳淑君博士

研究背景

代謝症候群(Metabolic Syndrome)不是一個特定的疾病，而是健康的危險訊號，亦是心血管疾病以及糖尿病的前身。若不及早檢查治療，將比一般人增加6倍得到糖尿病的風險、4倍高血壓風險，易變成慢性病病人。臺灣的糖尿病是國人的十大死因之一，根據國健署統計，全國每年近萬人因糖尿病死亡，目前約有200多萬名糖尿病的病友，且每年以25,000名的速度持續增加，因此糖尿病及隨之引發的併發症影響國人健康的程度不容小覷、醫療負擔相當龐大。

研究動機與問題

根據衛福部國健署的代謝症候群手冊中，避免成為代謝症候群的方法中，其中之一是健檢。因此，我想藉由過往的健檢資料，預測下次來健檢時的空腹血糖值、推測該名客戶罹患糖尿病的風險，再推薦適合的健檢方案。如此一來，對需求方而言，可以根據需求有更精密的健檢，早發現早治療，避免往後演變成難以康復的慢性病。



極大化社會利潤、減少無謂損失(deadweight loss)

根據經濟學，在供給方可以供應的情況下，若知道每個人的需求，以整體社會而言，就可以極大化社會利潤、減少無謂損失，社會資源達到最好的配置。

材料與方法

資料來自美兆健康管理機構，是一間私人的健檢中心。資料納入了2005、2008、2011、2014、2017年，這五年份為切點的資料，每人有三至五筆不等的紀錄。一般判定是否有糖尿病，是看糖化血色素(HbA1c)、空腹血糖值，或是口服葡萄糖耐受試驗之血漿血糖。有鑑於在本次的資料中只有空腹血糖值這個評斷標準，因此在此定義高風險罹患糖尿病是空腹血糖≧100。

本次研究在刪去缺失值、篩選出固定時間間隔來的個案以後，剩下9595名被納入做後續分析，女性為4420名，男女比近一比一。除了代謝症候群的五個判斷因子以外，還放入了衛福部國健署及國內外文獻提及罹患高血醣的可能危險因子，並同時有在此份數據中有的變項，包含連續型的生化數據以及問卷資料(生活習慣、飲食習慣等等)。

分析前，首先對於共線性的問題，使用方差膨脹係數 (variance inflation factor，VIF) 來處理。在盡可能保留代謝症候群的指標，並且刪除越少變項越好的前提之下，拿掉了 BMI 以及膽固醇此兩項變項。接著確認重複測量之間的相關性，利用了下列兩個方法：Covariance Parameter Estimates 以及 Null Model Likelihood Ratio Test，若無法拒絕重複測量 covariance 為0之虛無假說，則可忽略重複測量間的相關性、用一般線性迴歸來分析即可。結果不論用何者，均拒絕虛無假說，可得出無法忽略重複測量之間的相關性之結論。

為了解決空腹血糖值非常態分佈的問題，無法符合 Mixed effect model 的假設。且一開始的研究目的是分類高/低風險，因此Y設定為 binary variable，同時要考慮 Population-averaged effect 以及Subject-specific effect，因此採用Generalized linear mixed model (GLMM)以及 Transition model 較為合適。為了預測模型的好壞，所以將資料切割 testing 和validation，資料筆數的比例為7:3。後續的分析是使用SAS，並將 alpha 值設定為0.05。

Table1. Descriptive statistics of Fasting Plasma Glucose

Mean (median ; sd)		Fasting Plasma Glucose
Training data set	Male	102(99;16.9)
	Female	111(108;29.7)
	Total	**100(97;16.4)**
Validation data set	Male	102(99;17.3)
	Female	97.2(95;14.1)
	Total	**100.1(98;16.1)**

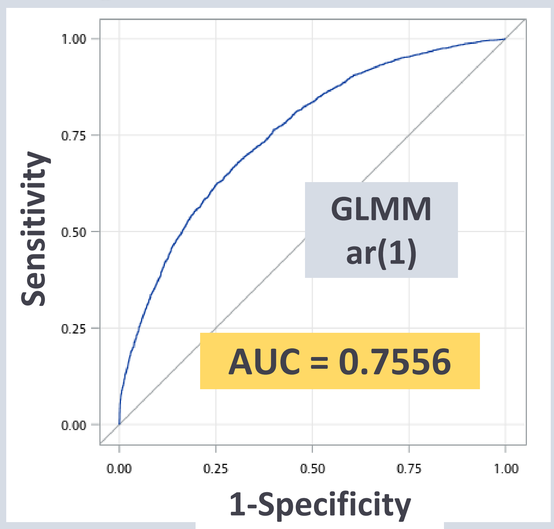
結果與討論

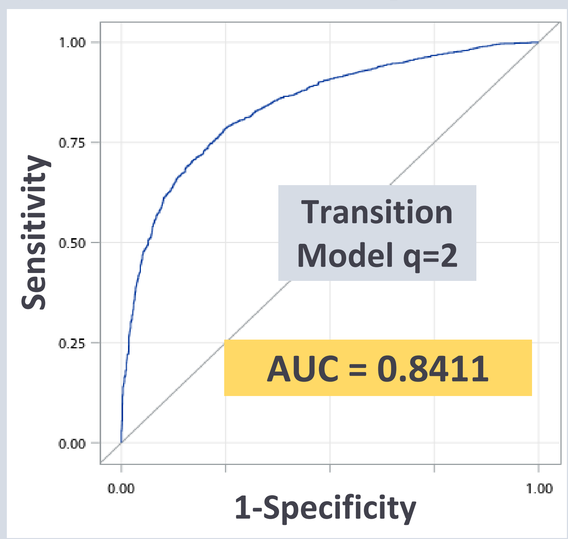
根據數據結果，我們可以得知 transition model 和 GLMM 相比之下，以AUC最大之 transition model 提供了較好的預測。

Table2. AUC, Sensitivity,Specivity, accuracy values for different models.

	AUC		Sen / Spe		accuracy	
	train	valid	train	valid	train	valid
Generalized Mixed Linear Model (GLMM)						
G-side	0.9344	0.7345	Sen： 0.8501 Spe： 0.8523	Sen： 0.6585 Spe： 0.6846	0.8561	0.6933
un	0.7288	0.7334	Sen： 0.6437 Spe： 0.6878	Sen：0.7170 Spe：0.6300	0.6929	0.6944
cs	0.7589	0.7551	Sen： 0.6701 Spe： 0.6599	Sen： 0.7201 Spe： 0.6291	0.6936	0.6951
ar(1)	0.7597	0.7556	Sen： 0.6477 Spe： 0.6857	Sen： 0.7344 Spe： 0.6159	0.6944	0.6946
Transition Model						
q=1	0.8279	0.8187	Sen：0.7003 Spe：0.7686	Sen：0.7102 Spe：0.7312	0.7493	0.7386
q=2	0.8438	0.8411	Sen：0.7527 Spe：0.7640	Sen：0.7366 Spe：0.7390	0.7609	0.7452

Figure1. ROC curves for GLMM ar(1) and Transition model q=2..





研究限制

- 具有較強健康意識者才會主動進行健檢，因此直接將模型外推至國人全體可能造成推論上的謬誤
- 納入的變項未能涵蓋所有已知罹患高血醣的可能危險因子

未來可繼續進行

- 結合個人配戴健康裝置的數據，收集到更多變項數據，以精準推估個人疾病罹患風險
- 考慮預測模型的外推性，使各大健檢中心均可以使用
- 將罹病風險細分為三組，看不同組之間的生化數據是否有變化模式之規續性