



國立臺灣大學公共衛生學院

健康數據拓析統計研究所碩士論文

Institute of Health Data Analytics and Statistic

College of Public Health

National Taiwan University

Master's Thesis

大腸癌腺瘤及癌症進展之機器學習模式分析

Progression of Colorectal Adenoma and Cancer:

A Machine Learning Analysis

簡瑞伶

Chien, Ruei-Ling

指導教授：陳秀熙 教授

Supervisor : Chen, Hsiu – Hsi, Professor

中華民國 115 年 7 月

July 2026

致謝

感謝陳秀熙指導教授，嚴謹的學術態度使我在研究訓練過程中獲益良多。從研究架構建立、疾病進程模型之思考，到資料詮釋與公共衛生觀點之討論，皆給予我許多啟發與指導，也讓我逐漸理解，研究不只是將問題複雜化，而是學習如何在龐雜的資料與現象之中，找到真正重要的核心問題。

同時，也感謝嚴明芳教授、陳立昇教授，以及許辰陽教授，在討論碩論所提供之寶貴建議。三位老師從不同學術觀點切入，針對疾病進程建模、篩檢資料詮釋、模型假設與實務應用等議題提出許多重要啟發，使本研究之理論架構與分析邏輯得以更加完整與嚴謹。

此外，也特別感謝林庭瑀博士。在整個碩士班生活中，不僅在學術與研究方法上給予我許多幫助，更在我面對研究壓力與低潮時，始終願意耐心傾聽並陪伴我度過許多困難的時刻。研究所的過程並不只是知識與技術的累積，更多時候是不斷面對挫折與自我懷疑。在我懷疑自己能力時提醒我不要輕易否定自己，而在我因一時成果而感到過於自滿時，也會直接點出問題，提醒我重新回到研究本身。這些提醒與陪伴，讓我逐漸理解真正的研究並不只是追求結果，而是學習如何在反覆修正與不確定之中，仍然保持謙遜、理性與持續前進的態度。

研究所的兩幾年裡，自己也記不清看過多少次日出。從深夜的程式除錯與反覆修正，甚至回頭重新檢查資料，很多時候甚至不知道自己到底還在忙些什麼。那些往返彰化出差的日子、在研究室與高鐵之間奔波的生活，如今回想起來，也都成為了碩士班生活中難以忘記的回憶。

最後最感謝我的父母親和家人。謝謝你們始終無條件地支持我、包容我，並在我面對壓力與低潮時，給予我最大的理解與陪伴。謝謝你們給我一個溫暖而穩定的家庭，讓我能夠在求學的道路上無後顧之憂。

摘要



研究背景

大腸直腸癌 (colorectal cancer, CRC) 腺瘤-癌症發生序列 (adenoma-carcinoma sequence) 由正常黏膜歷經小腺瘤、進展型腺瘤、臨床症前可偵測期 (preclinical detectable phase, PCDP) 逐步演變為臨床期癌症，疾病進展屬連續且不可直接觀測之潛在過程。傳統疾病自然史模型多以族群平均 (population average) 描述疾病進展，難以反映個體間之疾病異質性；而機器學習方法雖廣泛應用於風險預測，卻多侷限於疾病分類，缺乏與疾病自然史之動態連結。本研究旨在整合群體層級疾病自然史推論與個人化風險預測，建立一套混合式疾病進展架構 (Hybrid Disease Progression Framework)。

研究方法

群體層級分析採用臺灣全國糞便免疫化學檢查 (fecal immunochemical test, FIT) 篩檢資料 (14,478,478 人次篩檢紀錄、5,640,906 名受檢者)，建立含九個潛在疾病狀態之連續時間隱藏狀態馬可夫模型 (continuous-time hidden-state Markov model)，並透過觀測映射機制 (observation mapping) 結合 FIT 篩檢敏感度連結潛在狀態與觀測結果，以最大概似估計法 (MLE) 推估生成矩陣 (generator matrix) 之各轉移強度，再以矩陣指數法模擬潛在疾病分布。個體層級分析則以彰化地區社區篩檢資料建立「正常 vs. 腺瘤」與「腺瘤 vs. 篩檢偵測癌」兩項二元分類模型，比較人工神經網路 (ANN) 與隨機森林 (Random Forest, RF)，並以 SHapley Additive exPlanations (SHAP) 辨識重要危險因子。最後將機器學習預測機率轉換為相對風險倍數，回饋調節群體生成矩陣以建立個人化生成矩陣 (personalized generator matrix)，推估個人化疾病進展光譜。



研究結果

生成矩陣估計顯示，正常進入腺瘤之轉移強度 ($q_{12} = 0.00112$) 遠低於後續各階段之進展速度 (如 $q_{23} = 0.5896$ 、 $q_{34} = 0.5889$)，且觀測結果與真實疾病狀態並非一對一對應，FIT 陰性個案之間隔癌反映篩檢敏感度之限制。在機器學習分類中，「正常 vs. 腺瘤」以 RF 表現最佳 (測試集 AUC = 0.935)，優於 ANN (AUC = 0.896)。然而，「腺瘤 vs. 篩檢偵測癌」因兩階段特徵重疊、分類難度較高，測試集 AUC 分別降至 0.716 (ANN) 與 0.694 (RF)。SHAP 分析一致指出誘導時間(induction time)、糞便血紅素濃度 (f-Hb)、年齡及代謝症候群相關指標為最重要之預測因子。個人化進展模擬顯示， q_{12} (是否容易形成腺瘤)對整體癌症風險之影響遠大於 q_{34} (腺瘤惡化速度)：高 q_{12} /低 q_{34} 個案之十年臨床期癌症風險 (2.46%) 明顯高於低 q_{12} /高 q_{34} 個案，顯示「正常→腺瘤」階段為整體風險之上游主閘門。

結論

本研究首次整合九狀態連續時間隱藏狀態疾病自然史模型與人工智慧個人化風險預測，建立由群體疾病進展延伸至個人化疾病進展之分析架構，使疾病自然史研究由「描述平均進展」提升至「推估個人進展」。此一兼具機制解釋能力與個人化預測能力之架構，可作為風險導向篩檢 (risk-adaptive screening)、個人化篩檢間隔設計與臨床共享決策之理論基礎，並有望成為 AI 數位雙生 (AI-enabled Digital Twin) 及精準公共衛生之方法學平台。辨識「容易形成腺瘤」之高風險個體較辨識「腺瘤惡化速度快」者更能於上游有效攔截大腸直腸癌之發生。

關鍵詞

大腸直腸癌、疾病自然史、連續時間隱藏狀態馬可夫模型、生成矩陣、機器學習、隨機森林、SHAP、個人化疾病進展、精準篩檢

Abstract

Background

Colorectal cancer (CRC) follows the adenoma–carcinoma sequence, progressing from normal mucosa through small and advanced adenomas and the preclinical detectable phase (PCDP) to clinical cancer. This progression is a continuous and latent process that cannot be directly observed. Conventional disease natural history models characterize progression at the population-average level and fail to capture individual heterogeneity, whereas machine learning approaches, though widely applied to risk prediction, are largely confined to disease classification and lack a dynamic link to the disease natural history. This study aimed to integrate population-level disease natural history inference with personalized risk prediction into a Hybrid Disease Progression Framework.

Methods

At the population level, nationwide fecal immunochemical test (FIT) screening data from Taiwan (14,478,478 screening records; 5,640,906 participants) were used to construct a nine-state continuous-time hidden-state Markov model. An observation mapping mechanism incorporating FIT screening sensitivity linked latent states to observed screening outcomes, and the transition intensities of the generator matrix were estimated by maximum likelihood estimation (MLE), with the latent disease distribution simulated via the matrix exponential. At the individual level, community screening data from Changhua were used to build two binary classifiers—Normal vs. Adenoma and Adenoma vs. Screen-detected CRC—comparing an artificial neural network (ANN) and a Random Forest (RF), with SHapley Additive exPlanations (SHAP) applied to identify important risk factors. Finally, RF-predicted probabilities



were converted into relative risk multipliers to modulate the population generator matrix, yielding a personalized generator matrix and a personalized disease progression spectrum.

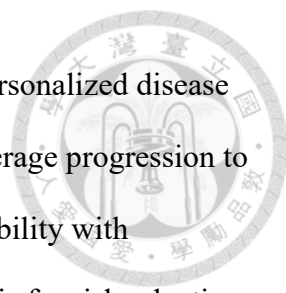


Results

The estimated generator matrix showed that the transition from normal to small adenoma ($q_{12} = 0.00112$) was markedly slower than subsequent progression steps (e.g., $q_{23} = 0.5896$, $q_{34} = 0.5889$), and observed outcomes did not correspond one-to-one with true disease states—interval cancers among FIT-negative cases reflected the limits of screening sensitivity. For the Normal vs. Adenoma classification, RF performed best (test AUC = 0.935), outperforming ANN (AUC = 0.896). The Adenoma vs. Screen-detected CRC task was more challenging owing to overlapping features, yielding test AUCs of 0.716 and 0.694 for ANN and RF, respectively. SHAP analyses consistently identified induction time, fecal hemoglobin concentration (f-Hb), age, and metabolic-syndrome-related indicators as the most important predictors. Personalized simulations indicated that the propensity to form an adenoma (q_{12}) influenced overall cancer risk far more than the speed of adenoma malignant progression (q_{34}): a high- q_{12} /low- q_{34} case carried a 10-year clinical-cancer risk of 2.46%, substantially higher than a low- q_{12} /high- q_{34} case, identifying the Normal→Adenoma step as the upstream gatekeeper of overall risk.

Conclusion

This study is the first to integrate a nine-state continuous-time hidden-state disease natural history model with AI-based personalized risk prediction, establishing an



analytical framework that extends population-level progression to personalized disease progression and elevates natural history research from describing average progression to estimating individual progression. Combining mechanistic interpretability with personalized prediction, the framework can serve as a theoretical basis for risk-adaptive screening, personalized screening-interval design, and shared clinical decision-making, and as a methodological platform for AI-enabled Digital Twins and precision public health. Identifying individuals prone to forming adenomas appears more effective for intercepting CRC upstream than identifying those with fast-progressing adenomas. Limitations include reliance on single-region data for the personalized model and the static nature of the generator matrix; external validation and integration of longitudinal and multi-omics data are warranted.

Keywords

Colorectal cancer; disease natural history; continuous-time hidden-state Markov model; generator matrix; machine learning; Random Forest; SHAP; personalized disease progression; precision screening

目次



致謝	i
摘要	ii
Abstract	iv
目次	vii
圖次	x
表次	xi
第一章 引言	1
1.1 研究背景	1
1.2 研究目的	3
1.3 研究流程	3
1.4 論文架構	4
第二章 文獻回顧	5
2.1 大腸直腸癌篩檢與疾病自然史	5
2.2 臺灣大腸直腸癌篩檢計畫與糞便免疫化學檢查	6
2.3 大腸直腸癌自然史模型	7
2.4 個人化因子與大腸直腸癌進展風險	9
2.5 機器學習於大腸直腸癌個人化風險預測之角色	11
2.6 現有研究限制與研究動機	13
第三章 研究架構	15
3.1 大腸直腸癌疾病自然史模型	15
3.2 連續時間隱藏狀態疾病進展模型	18



3.3 模型可識別性	21
3.3.1 基本假設	22
3.3.2 轉移矩陣之可識別性	24
3.3.3 生成矩陣之唯一性	26
3.3.4 全域可識別性	29
3.4 大樣本概似理論	30
3.4.1 一致性	32
3.4.2 漸近常態性	33
3.5 機器學習預測模型	35
3.5.1 研究對象與研究變項	35
3.5.2 資料處理與模型建立	38
3.5.3 人工神經網路模型	42
3.5.4 隨機森林模型	45
3.5.5 模型效能評估	47
3.5.6 模型解釋性分析	50
3.6 個人化風險分層之疾病進展模型	54
3.7 小結	56
第四章 資料分析結果	58
4.1 全國 FIT 篩檢資料庫之人口特徵與觀測結果分布	58
4.2 生成矩陣參數估計結果	64
4.3 發射矩陣估計結果	65
4.4 潛在疾病狀態分布模擬結果	66
4.5 機器學習模型分析	69

4.5.1 彰化社區篩檢資料之人口學與個人化特徵描述.....	69
4.5.2 正常與腺瘤分類器	81
4.5.3 腺瘤與篩檢癌分類器.....	89
4.5.4 綜合比較與研究意涵	97
4.6 個人化風險分層之疾病進展.....	98
4.6.1 風險分層相對風險倍數	98
4.6.2 各風險層之累積進展機率.....	101
4.6.3 四種代表性風險情境之十年疾病進展模擬.....	103
第五章 討論與結論.....	117
5.1 主要研究發現	117
5.2 群體層級疾病自然史模型之方法學意義.....	118
5.3 個人化風險預測與機器學習	118
5.4 群體疾病自然史與個人化疾病進展之整合	119
5.5 臨床與公共衛生意涵	121
5.6 研究限制與未來發展	122
5.7 結論	124



圖次

圖 3.1 大腸直腸癌疾病自然史之九狀態疾病進展模型	18
圖 4.1 ANN 模型於正常狀態與腺瘤分類之 ROC 曲線	82
圖 4.2 隨機森林模型於正常狀態與腺瘤分類之 ROC 曲線	82
圖 4.3 ANN 模型訓練集與測試集之 SHAP Beeswarm Plot	84
圖 4.4 隨機森林模型訓練集與測試集之 SHAP Beeswarm Plot	85
圖 4.5 ANN 模型 SHAP 平均絕對值重要性比較圖	86
圖 4.6 隨機森林模型 SHAP 平均絕對值重要性比較圖	87
圖 4.7 ANN 模型於腺瘤與大腸直腸癌分類之 ROC 曲線	90
圖 4.8 隨機森林模型於腺瘤與大腸直腸癌分類之 ROC 曲線	90
圖 4.9 ANN 模型訓練集與測試集之 SHAP Beeswarm Plot	92
圖 4.10 隨機森林模型訓練集與測試集之 SHAP Beeswarm Plot	92
圖 4.11 ANN 模型 SHAP 平均絕對值重要性比較圖	94
圖 4.12 隨機森林模型 SHAP 平均絕對值重要性比較圖	95
圖 4.13 兩段疾病轉移各風險層之相對風險倍數;虛線代表全域效應(加權 平均 $RR = 1$)	100



表次

表 4.1 台灣全國 FIT 篩檢資料之人口特徵與疾病分布	59
表 4.2 不同確診結果之人口特徵與 FIT 結果分布	61
表 4.3 生成矩陣轉移強度估計結果.....	64
表 4.4 發射矩陣之 FIT 敏感度估計結果.....	66
表 4.5 模擬潛在疾病分布	68
表 4.6 彰化糞便潛血描述性分析	71
表 4.7 彰化社區篩檢資料不同疾病組別之基本特徵分布.....	75
表 4.8 ANN 模型於正常狀態與腺瘤之分類效能	81
表 4.9 隨機森林模型於正常狀態與腺瘤之分類效能	81
表 4.10 整理 ANN 與隨機森林模型之 SHAP 重要變數排名	88
表 4.11 ANN 模型於腺瘤與篩檢癌之分類效能.....	89
表 4.12 隨機森林模型於腺瘤與篩檢癌之分類效能	89
表 4.13 ANN 與隨機森林模型於腺瘤與大腸篩檢癌分類之 SHAP 重要變數排名	96
表 4.14 Normal → Small Adenoma 各風險層之相對風險倍數與分層轉移強度	99
表 4.15 Advanced Adenoma → Early PCDP 各風險層之相對風險倍數與分層轉移強度	100
表 4.16 各風險層自正常狀態出發追蹤十年進展至「腺瘤」之累積機率	101
表 4.17 各風險層自 Advanced Adenoma 出發追蹤兩年進展至「癌前期以上」之累積機率.....	102
表 4.18 「分層 q_{12} 」與「分層 q_{34} 」模型低/高風險十分位族群之臨床	

特徵	103
表 4.19 四種代表性風險情境之配對代表個案臨床特徵	111
表 4.20 四種代表性特徵自正常狀態出發追蹤十年之疾病進展機率 ...	115





第一章 引言

1.1 研究背景

大腸直腸癌 (colorectal cancer, CRC) 為全球最重要之癌症之一，無論在癌症發生率或死亡率方面皆居於前列，對公共衛生造成沉重負擔。根據世界衛生組織 (World Health Organization, WHO)及國際癌症研究總署 (International Agency for Research on Cancer, IARC) 之統計資料，隨著人口高齡化、生活型態西化及代謝性疾病盛行，大腸直腸癌之疾病負擔仍持續增加，已成為世界各國癌症防治的重要課題。相較於其他惡性腫瘤，大腸直腸癌通常遵循 adenoma–carcinoma sequence，由正常腸黏膜逐步發展為小腺瘤、進展型腺瘤，再進一步演變為臨床症前可偵測期 (preclinical detectable phase, PCDP) 及臨床期癌症。由於整個疾病進展歷程通常歷經數年至十餘年，因此提供篩檢介入、早期診斷及預防癌症發生之重要契機。

在癌症防治策略中，次級預防 (secondary prevention) 透過篩檢於疾病尚未產生臨床症狀前及早發現臨床前可偵測癌症階段或早期癌症，已被證實可有效降低大腸直腸癌發生率及死亡率。其中，糞便免疫化學檢查 (fecal immunochemical test, FIT) 因具有較高敏感度、良好特異度、操作簡便及民眾接受度高等優點，已成為多數國家推動族群式大腸直腸癌篩檢之主要工具。臺灣自 2004 年開始推動社區型 FIT 篩檢，並自 2010 年正式納入國家癌症篩檢計畫，經多年推動後已累積數百萬名受檢者及長期追蹤資料，不僅提供篩檢效益評估的重要依據，也建立全球少見的大型縱向癌症篩檢資料庫，為疾病自然史研究與精準篩檢策略之發展奠定堅實基礎。

然而，大腸直腸癌疾病進展本質上屬於一個連續且不可直接觀測之生物學過



程。研究者於任何時間點均無法直接得知個體真實所處之疾病狀態，而實際可取得者多為 FIT 篩檢結果、大腸鏡檢查、病理診斷及癌症登記等觀測資料。換言之，篩檢結果僅為疾病自然史之間接表徵，而非真實疾病狀態本身，因此如何利用有限之觀測資訊推論潛在疾病狀態、估計疾病於各階段間之轉移速率，並重建疾病自然史，一直是癌症篩檢研究中最具挑戰性的問題之一。近年來，多狀態疾病自然史模型 (multistate disease progression model) 與隱藏狀態馬可夫模型 (hidden-state Markov model) 逐漸成為研究疾病自然史的重要工具，能以連續時間馬可夫鏈描述疾病由正常、腺瘤、臨床前可偵測癌症階段至臨床癌症之動態進展，並估計疾病轉移強度 (transition intensity)、停留時間 (sojourn time) 及生成矩陣 (generator matrix) 等關鍵參數。

然而，現有疾病自然史模型多以族群平均 (population average) 為分析對象，其估計所得之疾病進展速率反映的是整體族群之平均疾病歷程，而難以充分描述不同個體間之疾病異質性。事實上，即使具有相同年齡或相同篩檢結果，不同個體仍可能因糞便血紅素濃度、生活型態、健康狀況及其他危險因子不同，而具有截然不同之疾病進展速度。因此，如何由群體層級疾病自然史進一步發展至個體層級疾病進展 (personalized disease progression)，已成為近年精準醫療與個人化篩檢的重要研究方向。

另一方面，人工智慧 (Artificial Intelligence, AI) 及機器學習 (Machine Learning) 近年已廣泛應用於癌症風險預測。人工神經網路 (Artificial Neural Network, ANN) 及隨機森林 (Random Forest, RF) 等監督式學習方法，能有效整合人口學特徵、生化指標及篩檢資訊，建立個人化疾病風險預測模型。然而，多數機器學習研究主要著重於疾病分類 (classification) 或事件預測 (prediction)，其輸出通常為某一時間點之疾病機率，而缺乏疾病自然史中「疾病如何隨時間進展」之動態資訊，因此雖能回答「誰具有較高疾病風險」，卻無法回答「疾病將

如何演變」或「疾病將以何種速度進展」。此外，機器學習模型亦缺乏與疾病自然史模型之直接連結，使群體層級疾病進展與個體層級風險預測仍分屬兩個獨立研究領域。



1.2 研究目的

基於上述背景，本研究提出一套整合群體疾病自然史推論 (population-level disease progression inference) 與個人化疾病進展預測 (personalized disease progression prediction) 之混合式分析架構 (Hybrid Disease Progression Framework)。在群體層級方面，本研究利用臺灣全國 FIT 篩檢資料建立九狀態連續時間隱藏狀態疾病進展模型，結合 FIT 工具篩檢敏感度及生成矩陣推估疾病自然史；在個體層級方面，利用人工神經網路及隨機森林建立個人化疾病風險模型，並藉由 SHapley Additive exPlanations (SHAP) 分析重要危險因子，進一步建立個人化疾病進展光譜 (personalized disease progression spectrum)。本研究希望藉由整合疾病自然史模型與人工智慧方法，不僅提升疾病自然史推論能力，更進一步建立可支援風險分層 (risk stratification)、個人化篩檢 (personalized screening) 及精準篩檢 (precision screening) 之理論架構，作為未來智慧化癌症篩檢與臨床決策之重要基礎。

1.3 研究流程

本研究先利用臺灣全國大腸直腸癌糞便潛血篩檢資料，建立連續時間隱藏狀態(hidden-state)多狀態疾病進展模型。透過疾病真實狀態與實際觀測結果之對應關係，評估篩檢敏感度因素對疾病進展推論之影響，並進行探討。

除此之外，此外，本研究以彰化地區篩檢資料為基礎，依據疾病狀態採樣建立分析資料集，蒐集受檢者之基本特徵、FIT 濃度、代謝性指標及生活型態等資訊，分別建立正常與腺瘤及腺瘤與篩檢偵測癌兩項二元分類模型，比較人工神經

網路與隨機森林之預測效能，並透過 SHAP 方法評估各變數對模型預測之貢獻，以辨識不同疾病進展階段的重要風險因子。



1.4 論文架構

本論文共分為五章，各章內容如下：第一章為引言，說明研究背景、研究動機、研究目的、研究流程及論文架構。第二章為文獻回顧，整理大腸直腸癌疾病自然史、多狀態模型、篩檢機制以及機器學習風險預測的相關研究，作為本研究之理論基礎。第三章為研究方法，介紹研究資料來源、疾病自然史模型建構方式、以及人工神經網路、隨機森林、SHAP 分析及個人風險分層倍數流程。第四章為研究結果，首先呈現疾病自然史模型之估計結果，接著說明機器學習模型之預測表現與重要影響因子分析，最後整合兩者結果，建立個人化風險分層，並據以估計個體之疾病轉移率。第五章為討論與結論，針對研究發現進行討論，說明研究限制與未來研究方向，並提出相關建議。



第二章 文獻回顧

2.1 大腸直腸癌篩檢與疾病自然史

大腸直腸癌 (colorectal cancer, CRC) 為全球常見之惡性腫瘤之一，亦為臺灣癌症發生率及死亡率的重要疾病。根據世界衛生組織 (World Health Organization, WHO) 及國際癌症研究總署 (International Agency for Research on Cancer, IARC) 資料顯示，大腸直腸癌長期位居全球癌症發生率前列，對公共衛生造成重大負擔。

在公共衛生疾病防治架構中，預防策略可分為初級預防 (primary prevention)、次級預防 (secondary prevention) 及三級預防 (tertiary prevention)。其中，次級預防之核心目標在於透過篩檢於疾病尚未出現臨床症狀前及早發現病變，並藉由早期介入降低疾病發生率與死亡率。由於大腸直腸癌通常具有較長之疾病進展過程，因此被認為是最適合進行族群篩檢之癌症之一。

大腸直腸癌之發展普遍遵循 adenoma–carcinoma sequence 理論。Fearon 與 Vogelstein(1990)[1] 研究指出，大腸黏膜細胞於基因變異逐漸累積之過程中，可由正常黏膜逐步發展為腺瘤 (adenoma)、進展型腺瘤 (advanced adenoma)，最終演變為侵襲性癌症。此疾病進展過程通常需歷經多年，因此提供篩檢介入之機會。

Zelen 與 Feinleib [2] 首先提出疾病自然史中臨床症前可偵測期與臨床期之概念架構，指出疾病於出現臨床症狀前即可能存在一段可被篩檢工具偵測之潛在階段。其後，Cole 與 Morrison [3] 進一步發展癌症自然史理論，提出前導時間 (lead time) 與滯留時間 (sojourn time) 等關鍵概念，作為評估癌症篩檢效益的重要理論基礎。癌症於臨床診斷前存在一段無症狀但可被篩檢工具偵測之時期，稱

為臨床症前可偵測期 (preclinical detectable phase, PCDP)；當疾病進一步發展並出現臨床症狀後，則進入臨床期 (clinical phase, CP)。此一疾病進展歷程構成後續疾病自然史模型建構之重要理論基礎。



一般而言，PCDP 長度 (sojourn time) 越長，篩檢工具敏感度 (screening sensitivity) 越高，篩檢越有機會於疾病進入臨床症前可偵測期偵測病灶。然而，篩檢效益並不僅取決於 PCDP 長度與篩檢敏感度，亦受到篩檢間隔、篩檢參與率以及疾病進展速度等因素影響。因此，如何透過適當之統計模型描述疾病自然史及疾病進展過程，逐漸成為癌症篩檢研究的重要課題。

2.2 臺灣大腸直腸癌篩檢計畫與糞便免疫化學檢查

臺灣大腸直腸癌篩檢研究之發展可追溯至 1990 年代推動之 Taiwan Multicenter Cancer Screening (TAMCAS) 計畫。該計畫利用高風險族群篩檢資料建立大腸直腸癌自然史研究架構，並開始應用馬可夫模型 (Markov model) 推估疾病進展參數，奠定後續自然史模型研究之基礎 [4]。其後，研究方向逐漸由高風險族群篩檢擴展至社區族群之癌症防治與篩檢效益評估，並進一步發展社區整合性篩檢模式。

為提升社區癌症防治與慢性病管理效益，陳秀熙研究團隊於基隆市建置社區整合性篩檢計畫 (Keelung Community-based Integrated Screening, KCIS)，該計畫於 1999 至 2001 年間發展完成，整合五項癌症 (子宮頸癌、乳癌、大腸直腸癌、口腔癌及肝癌) 與三項慢性疾病 (高血壓、高血脂及第二型糖尿病) 之篩檢服務，涵蓋基隆市 30 至 79 歲居民共 217,895 人 [5]。KCIS 採社區主動邀請 (outreach screening) 模式，並結合癌症登記檔、死亡檔及全民健康保險資料庫等資訊來源，建立完整之篩檢、追蹤、轉介及確認診斷系統。此一大型社區篩檢平台不僅提供大規模縱向追蹤資料，亦成為後續大腸直腸癌篩檢研究與自然史模型

發展的重要基礎。

在 KCIS 計畫中，大腸直腸癌篩檢為重要組成項目之一。早期研究利用糞便潛血檢查 (fecal occult blood test, FOBT) 進行社區篩檢，並結合自然史模型評估篩檢效益與疾病進展過程 [6]。FOBT 可分為傳統化學法糞便潛血檢查 (guaiac-based fecal occult blood test, gFOBT) 與糞便免疫化學檢查 (fecal immunochemical test, FIT)。FIT 利用抗體偵測糞便中人類血紅素 (human hemoglobin)，藉由腸道微量出血作為疾病存在之指標。相較於傳統 gFOBT，FIT 具有較高敏感度、較佳特異度、無需飲食限制及民眾接受度較高等優勢，因此逐漸成為大腸直腸癌篩檢之主要工具。臺灣自 2004 年開始推動以 FIT 為主之大腸直腸癌篩檢計畫，並逐步建立全國性篩檢制度 [7]。

大量 FIT 篩檢資料的累積，使研究者得以進一步評估篩檢敏感度、間隔癌 (interval cancer) 發生率及不同疾病階段之偵測能力。Hsu 等人(2021)利用臺灣全國性 FIT 篩檢資料分析發現，FIT 間隔癌除可能源自檢測偽陰性 (false-negative) 外，亦可能來自篩檢後新發生之癌症 (newly developed colorectal cancer)，顯示篩檢結果未必能完全反映個體之潛在疾病狀態 [8]。此外，Chiu 等人 (2021) 研究指出，FIT 可有效提升進展型腺瘤 (advanced adenoma) 與早期大腸直腸癌之偵測率，並降低大腸直腸癌死亡率 [9]。

因此，如何根據篩檢結果推估個體潛在疾病狀態，並重建疾病由正常、腺瘤至大腸直腸癌之進展過程，逐漸成為大腸直腸癌篩檢研究的重要課題。由於疾病自然史無法直接觀測，研究者開始發展各類疾病自然史模型，以推估疾病進展過程及評估篩檢效益。

2.3 大腸直腸癌自然史模型

為量化描述疾病於不同階段間之進展過程，逐漸發展多狀態馬可夫模型

(multistate Markov model) 建構疾病自然史。馬可夫模型利用有限個疾病狀態 (finite states) 及其間之轉移強度 (transition intensity) 描述疾病進展歷程，並可進一步推估疾病發生率、停留時間及疾病風險等重要參數 [5]。相較於傳統單一疾病階段模型，多狀態模型能更完整呈現疾病於不同階段之進展特性，因此已成為癌症自然史研究中最常使用之建模方法之一。

隨著篩檢資料累積及統計方法發展，逐漸將傳統自然史模型擴展為多狀態疾病進展模型 (multistate disease progression model)，以更細緻描述疾病於不同階段間之進展過程。Yen 與 Chen 提出 Coxian phase-type multistate disease natural history model，以有限狀態馬可夫過程 (finite-state Markov process) 建構疾病自然史，允許疾病於臨床症前可偵測期 (preclinical detectable phase, PCDP) 及臨床期 (clinical phase, CP) 進一步細分為多個疾病階段，以反映疾病進展過程之複雜性與異質性 [10]。相較於傳統較為簡化之疾病自然史模型，此類多階段疾病進展模型能更完整描述疾病由臨床前可偵測癌症階段至臨床癌症之逐步進展歷程，並提供更細緻之疾病進展資訊。

除了疾病進展本身之外，近年研究亦開始重視過度診斷 (overdiagnosis) 對癌症篩檢效益評估之影響。而非進展型病變雖可能被篩檢工具偵測，但於個體一生中不會進展至臨床期癌症，因此可能造成過度診斷現象。為描述此現象，Yen 與 Chen 進一步將疾病自然史區分為進展型路徑 (progressive pathway) 與非進展型路徑 (nonprogressive pathway)，並利用潛在疾病狀態 (latent states) 描述兩種不同疾病歷程 [10]。其中，進展型病變可逐步發展為臨床期癌症，而非進展型病變雖可能被篩檢工具偵測，卻不會進一步進展至臨床期癌症，因此形成過度診斷現象。

後續 Lin 等人 (2023) 進一步延伸多階段疾病進展與潛在疾病路徑之概念，

將疾病進展過程視為由多個潛在疾病狀態 (latent disease states) 所構成之動態轉移系統，並利用多狀態模型描述不同疾病階段間之轉移歷程 [11]。此類方法不僅能反映疾病進展速度之差異，亦能考量不同潛在疾病路徑所造成之疾病異質性，使自然史模型更貼近實際疾病發展過程。

由於研究者實際所能取得之資料多為篩檢結果或臨床診斷結果，而非個體真實疾病狀態，因此近年研究逐漸採用隱藏狀態馬可夫模型 (Hidden Markov Model, HMM) 描述疾病進展過程。Yen 與 Chen 提出 Bayesian measurement-error-driven hidden Markov regression model，將真實疾病狀態視為潛在狀態 (hidden states)，而觀測結果則透過測量誤差機制 (measurement error mechanism) 由潛在狀態產生 [12]。此類模型能同時描述疾病自然史之動態轉移過程及觀測誤差所造成之狀態錯置 (misclassification)，因此更能反映實際篩檢資料中真實疾病狀態與觀測結果間之差異。

綜合上述文獻，大腸直腸癌自然史模型已逐漸由傳統單一路徑疾病進展架構，發展為兼具多階段疾病進展、潛在疾病狀態 (latent disease states)、過度診斷機制 (overdiagnosis) 及隱藏狀態觀測架構 (hidden-state framework) 之多狀態模型。現有研究多將大腸直腸癌疾病進展歷程區分為正常狀態、腺瘤階段、臨床症前可偵測期及臨床期癌症等疾病階段，部分研究亦進一步納入過度診斷路徑與潛在疾病狀態概念，使自然史模型得以同時描述疾病進展歷程與觀測結果之產生機制。

2.4 個人化因子與大腸直腸癌進展風險

然而，即使自然史模型能描述疾病於不同階段間之平均進展歷程，不同個體之疾病風險與進展速度仍可能存在顯著差異。因此，近年研究逐漸重視個人化因子於大腸直腸癌風險分層與篩檢策略中的角色。過去研究指出，年齡、性別、家



族病史、肥胖、吸菸、飲酒及身體活動不足等因素皆與大腸直腸癌風險增加有關 [13]。傳統大腸直腸癌篩檢策略多以年齡作為主要篩檢依據，例如針對 50 至 74 歲族群定期提供糞便免疫化學檢查 (fecal immunochemical test, FIT) 服務。然而，即使屬於相同年齡層之個體，其罹患大腸腫瘤及大腸直腸癌之風險仍可能存在顯著差異，因此近年研究逐漸由族群式篩檢 (population-based screening) 發展至個人化風險分層 (risk-stratified screening) 策略，以提供更適切之篩檢間隔與追蹤方案 [14]。

除傳統危險因子外，篩檢過程所獲得之資訊亦被證實具有重要的風險預測價值。Chen 等人 (2011) 利用臺灣社區大腸直腸癌篩檢資料發現，即使首次糞便免疫化學檢查結果未達傳統陽性判定標準，較高之糞便血紅素濃度 (fecal hemoglobin concentration, f-Hb) 仍與後續大腸腫瘤及大腸直腸癌發生風險增加有關，顯示 FIT 濃度本身可作為重要之風險分層指標 [15]。此研究首次提出 FIT 濃度不僅具有篩檢功能，更能反映個體未來疾病風險，為個人化篩檢策略奠定基礎。

其後，隨著研究逐漸證實 FIT 濃度與疾病風險具有連續性關係，研究者開始思考傳統固定陽性判定標準是否適用於所有族群。Chiu 等人發現若採用單一固定之 FIT 陽性判定切點 (one-size-fits-all cutoff)，可能造成不同族群間篩檢敏感度與特異度之差異，進而影響篩檢效能與醫療資源配置。因此研究者提出年齡及性別特異性 (age- and sex-specific) 之 FIT 陽性切點，藉由風險分層策略提升篩檢效益，同時兼顧大腸鏡檢查資源之合理運用 [9]。

近年來，個人化篩檢概念進一步由風險分層發展至精準篩檢 (precision screening)。Yen 等人利用臺灣超過三百萬名受檢者資料建立以糞便血紅素濃度為基礎之個人化篩檢模式，依據不同風險層級設定不同之篩檢間隔 [14]。研究結果



顯示，高風險族群可縮短篩檢週期，而低風險族群則可適度延長篩檢間隔，在維持篩檢效益的同時降低醫療資源耗用。此研究顯示，FIT 濃度已由單純篩檢指標逐步發展為個人化篩檢決策工具。

除篩檢策略個人化之外，近年研究亦開始探討如何將個體危險因子整合至疾病自然史模型之中。Hsu 等人進一步提出 Sampling-Based Markov Regression Model，將個體層級危險因子納入多狀態疾病進展架構中，使疾病轉移速率可隨個人特徵而改變，進而建立個人化疾病進展模型。研究結果顯示，不同個體特徵可能導致疾病進展路徑與疾病狀態分布產生顯著差異，顯示疾病自然史模型可進一步結合個體資訊，以提升疾病風險評估與疾病進展預測能力 [16]。此研究亦顯示，個人化因子不僅可應用於篩檢策略制定，亦可進一步整合至疾病自然史模型之中。

綜合上述研究可知，大腸直腸癌個人化風險評估之發展已由傳統危險因子分析逐漸擴展至 FIT 濃度風險分層、精準篩檢策略以及個人化疾病進展模型，顯示個體特徵資訊於疾病預測與篩檢決策中具有重要價值。

2.5 機器學習於大腸直腸癌個人化風險預測之角色

近年來，隨著個人化風險評估與風險分層篩檢策略逐漸受到重視，研究者開始嘗試利用大量篩檢資料與健康檢查資料建立疾病風險預測模型。由於大腸直腸癌之發生受到人口學特徵、生活型態、代謝指標及篩檢相關資訊等多重因素共同影響，變數間往往存在複雜且非線性之交互作用。傳統統計模型通常需預先設定變數與結果間之函數關係，而機器學習方法則可透過資料驅動方式學習複雜之非線性關係與交互作用[16]。

近年有多項研究將機器學習方法應用於大腸直腸癌風險預測、腺瘤辨識及癌症早期偵測等議題。其中，人工神經網路 (Artificial Neural Network, ANN) 與隨



機森林(Random Forest, RF)為最常使用之監督式學習方法。ANN 透過多層神經元結構模擬人類神經網路之資訊處理方式，能有效學習複雜之非線性關係，因此廣泛應用於疾病分類與風險預測問題。另一方面，隨機森林模型則透過拔靴抽樣法(Bootstrap sampling) 與隨機特徵選取機制建立多棵決策樹，並利用集成學習(Ensemble learning) 整合各決策樹之預測結果，以提升模型穩定性及泛化能力 [17]。

Burnett 等人(2023)針對大腸直腸癌機器學習風險預測研究進行範疇性文獻回顧，結果顯示隨機森林、梯度提升樹 (Gradient Boosting Machine) 及人工神經網路等方法皆具有良好之風險辨識能力，部分研究之預測效能甚至優於傳統統計模型。研究指出，機器學習模型能整合人口學特徵、生活型態、臨床指標及實驗室資料，提供更精細之個人化風險評估。然而，目前相關研究仍面臨外部驗證不足、研究方法異質性較高及臨床應用證據有限等問題，未來仍需大型族群研究進一步驗證其實務應用價值[18]。

然而，傳統機器學習模型常被視為黑盒子 (black box)，其預測結果雖具有良好效能，但模型內部決策過程不易解釋，因此可解釋人工智慧 (Explainable Artificial Intelligence, XAI) 逐漸受到重視。其中，SHapley Additive exPlanations (SHAP) 方法已被廣泛應用於醫療風險預測研究。SHAP 係基於合作賽局理論 (cooperative game theory) 所發展之模型解釋方法，可量化各變數對模型預測結果之貢獻程度，並提供全域 (global interpretation) 及局部 (local interpretation) 之模型解釋資訊。透過 SHAP 分析，可進一步辨識影響疾病風險預測之重要因子，並評估各變數對模型預測結果之正向或負向貢獻，以提升模型結果之臨床可解釋性 [19]。

綜合上述文獻可知，機器學習方法已逐漸成為大腸直腸癌個人化風險預測的

重要工具，而可解釋人工智慧技術則有助於提升模型之透明度與臨床應用價值。然而，目前尚缺乏以臺灣社區篩檢資料為基礎，整合人工神經網路、隨機森林及 SHAP 方法，同時建立不同疾病進展階段之風險預測模型與辨識重要危險因子之研究。

2.6 現有研究限制與研究動機

綜觀上述文獻，大腸直腸癌篩檢研究已由傳統族群式篩檢逐步發展至疾病自然史模型、個人化風險分層策略及機器學習風險預測模型。然而，現有研究仍存在若干限制，有待進一步探討。

首先，現有大腸直腸癌自然史模型雖可描述疾病由正常狀態逐步進展至臨床期癌症之歷程，但研究者實際所能取得之資料多為篩檢結果、病理診斷結果或癌症登記資料，而非個體於特定時間點之真實疾病狀態。由於疾病進展過程本質上屬於潛在過程 (latent process)，真實疾病狀態通常無法直接觀測，因此如何根據有限之觀測資料推估個體潛在疾病進展歷程，仍為疾病自然史研究的重要挑戰。

其次，在大腸直腸癌篩檢情境下，研究者實際所觀測之篩檢結果並不等同於個體真實疾病狀態。觀測結果受到篩檢敏感度 (screening sensitivity) 及疾病觀測機制等因素影響，因此如何建立潛在疾病狀態與觀測結果之對應關係，仍為疾病自然史研究中之重要議題。若未能適當考量疾病觀測機制，則可能導致疾病進展參數估計及篩檢效益評估結果產生偏差。

另一方面，近年機器學習方法已廣泛應用於大腸直腸癌風險預測研究，並展現良好之疾病辨識能力。然而，多數研究主要著重於疾病分類或風險預測效能之提升，較少進一步探討各項個人化因子對疾病進展風險之影響機制。此外，由於機器學習模型常被視為黑盒子模型，其決策過程缺乏透明性，因此如何透過模型解釋方法探討重要危險因子對疾病風險之影響，亦為近年研究關注之議題。



基於上述研究缺口，本研究從群體層級疾病進展推論與個體層級風險預測兩個面向進行探討。在群體層級分析方面，本研究採用連續時間隱藏狀態馬可夫模型（continuous-time hidden-state Markov model），區分潛在疾病狀態（latent states）與觀測疾病狀態（observed states），並透過觀測映射機制連結潛在疾病進展與篩檢結果，以最大概似估計法推估疾病進展之轉移強度與篩檢的敏感度，進而描述大腸直腸癌疾病自然史之動態進展過程。

在個體層級分析方面，本研究利用人工神經網路（ANN）與隨機森林（Random Forest）建立疾病風險預測模型，並結合 SHapley Additive exPlanations（SHAP）分析辨識重要危險因子，以提升模型之可解釋性，探討各項個人化因子對疾病風險之影響。

最後，本研究進一步整合群體層級疾病自然史模型與個體層級風險預測結果，依據機器學習模型之預測風險建立個人化風險分層，並以相對風險倍數調整疾病轉移強度，建立個人化疾病進展模型，使疾病自然史推估由群體平均進一步延伸至個體化風險評估。



第三章 研究架構

本研究以大腸直腸癌疾病自然史為基礎，結合篩檢觀測機制與統計推論方法，建立連續時間隱藏狀態疾病進展模型（continuous-time hidden-state disease progression model），建構包含九個潛在疾病狀態之疾病自然史架構，以描述群體層級下由正常狀態逐步進展至癌症之疾病歷程。

由於個體於特定時間點之真實疾病狀態通常無法直接觀測，研究者僅能透過篩檢結果、病理診斷及臨床診斷等觀測資料間接推論疾病進展，因此本研究將疾病進展視為潛在過程（latent process），並透過篩檢敏感度所建構之觀測機制，連結潛在疾病狀態與實際觀測結果，以最大概似估計法推估疾病自然史之轉移生成矩陣。此外，本研究透過四項模型假設、三項理論定理及大樣本理論，說明模型之可識別性，並驗證參數估計具有一致性與漸近常態性，以作為群體層級疾病自然史推論之理論基礎。

在建立群體層級疾病自然史模型後，本研究進一步整合個人化風險分層（individualized risk stratification）之概念，利用機器學習模型所估計之個人風險建立相對風險倍數，據以調整疾病轉移強度，使模型同時反映整體疾病進展趨勢與不同個體間之風險異質性，建立兼具群體層級與個體層級之疾病進展分析架構。

3.1 大腸直腸癌疾病自然史模型

大腸直腸癌之發展通常遵循腺瘤－癌症序列（adenoma-carcinoma sequence），即由正常大腸黏膜逐步發展為腺瘤，再進一步進展為癌症。由於疾病進展通常需歷經多年，因此篩檢有機會於臨床前可偵測癌症階段或早期癌症，進而降低疾病死亡率。



根據大腸直腸癌自然史理論，疾病於出現臨床症狀前通常存在一段可被篩檢工具偵測但尚未產生症狀之階段，稱為臨床症前可偵測期 (preclinical detectable phase, PCDP)；當疾病進一步進展並產生臨床症狀後，則進入臨床期 (clinical phase, CP)。因此，本研究將疾病進展過程區分為正常狀態、腺瘤階段、臨床症前可偵測期及臨床期癌症等不同疾病階段，以描述大腸直腸癌之疾病自然史。

此外，考量部分病灶可能因生長緩慢或惡性潛能較低，而於個體終生內不會進展為臨床期癌症，本研究進一步納入過度診斷 (overdiagnosis) 概念，建立獨立之非進展性病灶路徑 (indolent pathway)，以描述此類病灶之疾病進程。因此，本研究建構包含九個潛在疾病狀態之疾病自然史模型：

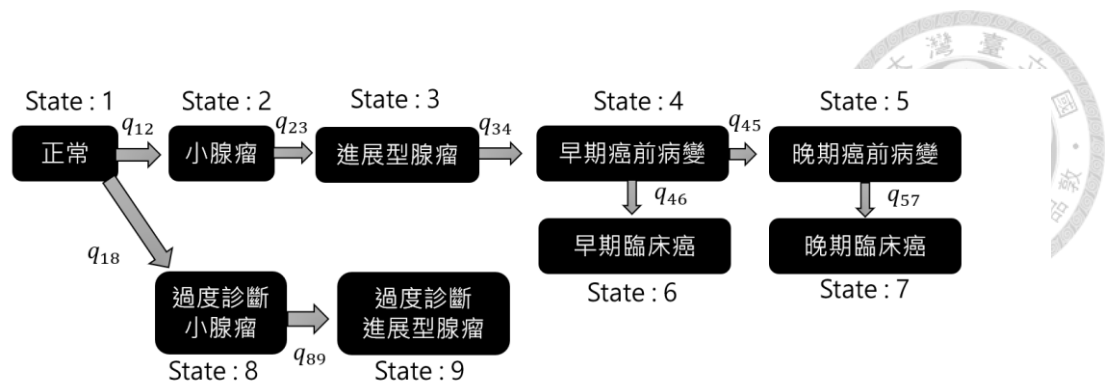
1. 正常 (Normal)
2. 小腺瘤 (Small Adenoma)
3. 進展型腺瘤 (Advanced Adenoma)
4. 早期臨床前可偵測癌症階段 (Early PCDP)
5. 晚期臨床前可偵測癌症階段 (Late PCDP)
6. 早期臨床癌 (Early Clinical Phase, Early CP)
7. 晚期臨床癌 (Late Clinical Phase, Late CP)
8. 過度診斷小腺瘤 (Overdiagnosed Small Adenoma)
9. 過度診斷進展型腺瘤 (Overdiagnosed Advanced Adenoma)

由於上述九個疾病狀態代表個體於疾病自然史中之潛在真實狀態，研究者於實際篩檢資料中無法直接觀測個體所處之潛在疾病狀態，而僅能取得個體之篩檢與診斷資訊，包括糞便免疫化學檢查 (FIT) 濃度、病理診斷結果及臨床診斷結果等

觀測資料。本研究依據上述觀測資訊，進一步將個體分類為對應之觀察狀態（observed states），以作為疾病進展模型之觀測結果。本研究所定義之觀察狀態包括：

1. 觀察正常（Observed Normal）
2. 觀察小腺瘤（Observed Small Adenoma）
3. 觀察進展型腺瘤（Observed Advanced Adenoma）
4. 觀察早期臨床症前可偵測期（Observed Early PCDP）
5. 觀察晚期臨床症前可偵測期（Observed Late PCDP）
6. 觀察早期臨床癌（Observed Early CP）
7. 觀察晚期臨床癌（Observed Late CP）

其中，潛在狀態與觀察狀態之間並非一對一對應。舉例而言，處於腺瘤或臨床症前可偵測期之個體，可能因篩檢敏感度不足或未參與篩檢而未被偵測，此時其觀察結果可能仍被記錄為正常；相反地，若病灶被篩檢或後續診斷成功偵測，則會被記錄為相對應之觀察疾病狀態。此外，過度診斷路徑中之病灶雖屬非進展性病灶，但若於篩檢過程中被偵測，仍可能在觀察資料中呈現為小腺瘤或進展型腺瘤。因此，本研究透過觀測機制連結潛在疾病狀態與實際觀察結果，作為後續建立連續時間隱藏狀態疾病進展模型之基礎。



- 註：State 1–7 代表疾病自然史主要進展路徑，State 8–9 為過度診斷 (overdiagnosis) 路徑。所有疾病狀態轉移皆假設為不可逆 (irreversible transition)，允許之狀態轉移率以 q_{ij} 表示。

圖 3.1 大腸直腸癌疾病自然史之九狀態疾病進展模型

圖 3.1 顯示本研究所建構之大腸直腸癌疾病自然史架構。在此架構下，個體可能沿主要疾病進展路徑，由正常狀態逐步進展至臨床期癌症；亦可能進入過度診斷路徑而停留於非進展性病灶狀態。

3.2 連續時間隱藏狀態疾病進展模型

為描述大腸直腸癌疾病自然史之動態進展過程，本研究建立連續時間隱藏狀態疾病進展模型 (continuous-time hidden-state disease progression model)，結合連續時間馬可夫鏈 (continuous-time Markov chain, CTMC) 與篩檢觀測機制，以描述個體潛在疾病狀態之轉移過程及其與實際篩檢結果間之關係。

令 $X(t)$ 表示個體於時間 t 之潛在疾病狀態，其中狀態空間由第 3.1 節所定義之九個疾病狀態所組成。另設

$$\mathcal{T} = \{t_1, t_2, \dots, t_{n_i}\}$$

表示個體接受篩檢之時間點，其中 t_k 為第 k 次篩檢時間， $k = 1, \dots, n_i$ ，而 n_i 表示個體於追蹤期間內之總篩檢次數。在各篩檢時間點，研究者無法直接觀測潛

在疾病狀態 $X(t_k)$ ，而僅能透過篩檢結果獲得部分疾病資訊。因此，本研究將疾病進展視為潛在狀態過程，而實際觀測結果則於後續觀測機制中加以描述。

本研究假設潛在疾病狀態過程 $X(t)$ 滿足馬可夫性 (Markov property)，亦即在已知目前疾病狀態之條件下，未來疾病狀態之分布僅與目前狀態有關，而與過去疾病歷史無關，其數學表示如下：

$$P(X(t+s)=j | X(t)=i, \mathcal{F}_t) = P(X(t+s)=j | X(t)=i)$$

其中 $\mathcal{F}_t$ 表示時間 t 前之所有疾病歷史資訊。於連續時間架構中，疾病狀態間之轉移可由生成矩陣 (generator matrix)

$$Q = (q_{ij})$$

加以描述，其中非對角元素 q_{ij} 表示個體由狀態 i 轉移至狀態 j 之瞬時轉移率 (transition intensity)，定義為：

$$q_{ij} = \lim_{\Delta \rightarrow 0^+} \frac{P(X(t+\Delta)=j | X(t)=i)}{\Delta}, i \neq j$$

為使各列元素總和為零，對角元素定義為

$$q_{ii} = -\sum_{j \neq i} q_{ij}$$

因此生成矩陣滿足

$$\sum_j q_{ij} = 0, \forall i$$

根據第 3.1 節所建立之疾病自然史架構，本研究假設疾病進展為不可逆轉移

(irreversible progression)，亦即疾病僅能沿自然史方向逐步進展，而不允許疾病狀態逆向轉移。故生成矩陣中僅允許下列轉移率為非零參數：

$$\mathcal{Q}_G = \{Q \in \mathbb{R}^{m \times m} : q_{ij} \geq 0 (i \neq j), \sum_{j=1}^m q_{ij} = 0, \forall i, q_{ij} = 0 \text{ 若 } (i, j) \notin \mathcal{E}\},$$

$$\mathcal{E} = \{ (1,2), (1,8), (2,3), (3,4), (4,5), (4,6), (5,7), (8,9) \}$$

由於 $\mathcal{E}$ 中所有允許之轉移邊均沿狀態編號遞增方向排列（即僅允許由較早期狀態轉移至較晚期狀態），此設定亦直接意味著生成矩陣 Q 呈現上三角結構，即所有非零轉移率皆位於主對角線上方。

$$Q = \begin{pmatrix} -(q_{12} + q_{18}) & q_{12} & 0 & 0 & 0 & 0 & 0 & q_{18} & 0 \\ 0 & -q_{23} & q_{23} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -q_{34} & q_{34} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -(q_{45} + q_{46}) & q_{45} & q_{46} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -q_{57} & 0 & q_{57} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -q_{89} & q_{89} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

潛在疾病狀態過程 $X(t)$ 滿足 Chapman-Kolmogorov 性質（Chapman-Kolmogorov property）：

$$P(\Delta_1 + \Delta_2) = P(\Delta_1)P(\Delta_2)$$

其中 $P(\Delta)$ 表示時間間隔 Δ 之轉移機率矩陣。根據 Kolmogorov forward equation：

$$\frac{dP(\Delta)}{d\Delta} = P(\Delta)Q$$

並滿足初始條件

$$P(0) = I$$

其中， I 為單位矩陣（identity matrix）。因此可得轉移機率矩陣之解析解為：

$$P(\Delta) = \exp(Q\Delta)$$

其中 $\exp(\cdot)$ 表示矩陣指數函數 (matrix exponential)。轉移機率矩陣 $P(\Delta)$ 描述個體於時間間隔 Δ 內由一潛在疾病狀態轉移至另一潛在疾病狀態之機率，並作為後續篩檢觀測機制與似然函數建構之基礎。

然而，上述潛在連續時間馬可夫鏈僅描述潛在疾病狀態之動態演化規律，研究者於實際篩檢情境下無法直接觀測個體於時間 t 之真實潛在疾病狀態 $X(t)$ ，而僅能於篩檢時間點獲得對應之篩檢結果。為連結潛在疾病進展過程與實際觀測資料，本研究進一步引入發射模型 (Emission model)，以描述潛在疾病狀態與實際觀測結果之間的機率關係。

令 Y_k 表示個體於第 k 次篩檢之觀測結果，則當個體於時間 t_k 之潛在疾病狀態為 s 時，其觀測結果之條件分布可表示為：

$$Pr(Y_k = y \mid X(t_k) = s) = g_s(y; \eta)$$

其中， $X(t_k)$ 為時間 t_k 之潛在疾病狀態； Y_k 為第 k 次篩檢之觀測結果； $g_s(y; \eta)$ 為狀態 s 所對應之觀測分布 (emission distribution)； η 為觀測模型參數，其包含不同疾病狀態下之篩檢偵測機率等與觀測機制相關之參數。

上述觀測機制描述潛在疾病狀態與篩檢結果之間的機率關係，使研究者得以利用實際篩檢資料推論個體之潛在疾病狀態及疾病進展過程。在本研究中，觀測結果主要受到篩檢敏感度之影響，因此不同潛在疾病狀態具有不同之被偵測機率。透過此觀測機制，潛在疾病進展過程得以與實際篩檢資料建立連結，並作為後續概似函數建構、模型可識別性分析及參數估計之基礎。

3.3 模型可識別性

模型可識別性 (identifiability) 係指模型參數是否能夠由觀測資料唯一辨識。

若存在不同參數組合產生相同之觀測分布，則無法保證參數估計結果之唯一性。因此，在進行模型推論與參數估計之前，必須先確認所建立之疾病自然史模型具有可識別性。本節依序建立三項結果：(1)轉移機率矩陣之可識別性、(2)生成矩陣之唯一性、(3)模型參數之全域可識別性，並說明三者間之邏輯遞進關係。

3.3.1 基本假設

在討論模型可識別性之前，首先將第 3.2 節所定義之觀測分布 $g_s(y; \eta)$ 表示為矩陣形式。設觀測結果空間為有限集合

$$\mathcal{Y} = \{y_1, y_2, \dots, y_r\}$$

對於每一潛在疾病狀態 s ，其觀測分布可表示為

$$\mathbf{g}_s = \begin{pmatrix} g_s(y_1; \eta) \\ g_s(y_2; \eta) \\ \vdots \\ g_s(y_r; \eta) \end{pmatrix}$$

將所有潛在疾病狀態所對應之觀測分布並列後，可形成觀測機率矩陣 (Emission Matrix)

$$\mathbf{G} = (\mathbf{g}_1, \dots, \mathbf{g}_m) = \begin{pmatrix} g_1(y_1; \eta) & g_2(y_1; \eta) & \cdots & g_m(y_1; \eta) \\ g_1(y_2; \eta) & g_2(y_2; \eta) & \cdots & g_m(y_2; \eta) \\ \vdots & \vdots & & \vdots \\ g_1(y_r; \eta) & g_2(y_r; \eta) & \cdots & g_m(y_r; \eta) \end{pmatrix} \in \mathbb{R}^{r \times m}$$

其中 m 為潛在疾病狀態數， r 表示觀測結果狀態之可能類別數。觀測機率矩陣 $\mathbf{G}$ 述各潛在疾病狀態對應之觀測分布，並作為後續模型可識別性分析之基礎。

假設一(觀測機率矩陣滿欄秩)

觀測機率矩陣 (Emission Matrix) $\mathbf{G}$ 具有完全欄秩 (full column rank)，即

$$\text{rank}(G) = m$$

其中 m 為潛在疾病狀態數。

若假設觀測機率矩陣完全攔秩，亦表示不同潛在疾病狀態所對應之觀測分布彼此線性獨立，因此不同潛在疾病狀態可透過觀測資料加以區分。換言之，不同潛在狀態不會產生完全相同之觀測分布，故觀測資料中包含足夠資訊以辨識潛在疾病狀態，為模型可識別性成立之必要條件。

假設二(多重篩檢間隔 $\Delta_1 \neq \Delta_2$)：

雖然臺灣 FIT 篩檢計畫原則上以兩年為篩檢間隔，惟實際受檢者之回診時間仍可能因個人就醫行為及篩檢參與情形而有所差異。因此，研究資料中存在多個不同之觀測時間間隔，記為

$$\Delta_1, \Delta_2, \dots, \Delta_L$$

且至少存在兩個不同之觀測時間間隔，即

$$\Delta_1 \neq \Delta_2$$

在連續時間馬可夫鏈架構下，觀測間隔 Δ 所對應之轉移機率矩陣滿足

$$P(\Delta) = \exp\{Q\Delta\}$$

因此，不同觀測時間間隔可提供疾病進展過程於不同時間尺度下之轉移資訊，使同一生成矩陣 Q 同時受到多組轉移機率矩陣之約束，進而有助於提升模型之可識別性。

假設三(生成矩陣之不可逆結構，irreversible progression)：

生成矩陣 Q 之轉移結構由疾病自然史理論給定，各潛在疾病狀態僅允許沿疾病進展方向發生轉移，不允許由後期疾病狀態返回早期疾病狀態。亦即，生成



矩陣中僅允許符合疾病自然史之非零轉移率，其餘不符合疾病進展方向之元素皆固定為零。



此假設反映大腸直腸癌由正常黏膜逐步進展至腺瘤、臨床症前可偵測期及臨床期癌症之單向演化特性，並使生成矩陣之零元素位置具有固定結構，避免不同生成矩陣因轉移方向不同而產生相同之觀測分布，進而提升模型之可識別性。

假設四(參數化之單射映射性)：

假設自由參數向量為

$$\theta = (q_{12}, q_{18}, q_{23}, q_{34}, q_{45}, q_{46}, q_{57}, q_{89}) \in \Theta$$

且定義參數化映射

$$\varphi: \Theta \rightarrow Q(\mathcal{E}), \varphi(\theta) = Q(\theta)$$

所以對於任意兩組不同的參數向量 $\theta_1 \neq \theta_2$ ，則 $Q(\theta_1) \neq Q(\theta_2)$ 。

3.3.2 轉移矩陣之可識別性

定理 3.1：

在假設一 (Emission Matrix 具有完全欄秩)、假設二 (存在不同觀測時間間隔)、假設三 (生成矩陣之不可逆結構) 及假設四 (參數化之單射映射性) 成立下，連續時間隱藏狀態疾病進展模型所產生之觀測序列聯合分布可唯一決定轉移機率矩陣

$$P_{\theta}(\Delta) = \exp(Q(\theta)\Delta)$$

其結果除潛在狀態標籤置換 (label permutation) 外為可識別 (identifiable)。

證明：

採反證法。假設存在兩組不同之參數向量 θ_1 與 θ_2 ，其所對應之轉移機率矩



陣不同，即

$$P_{\theta_1}(\Delta) \neq P_{\theta_2}(\Delta)$$

但其所產生之觀測序列聯合分布完全相同，即

$$g_{\theta_1}(Y_1, Y_2, Y_3) = g_{\theta_2}(Y_1, Y_2, Y_3)$$

由假設一，觀測機率矩陣

$$\mathbf{G} = (g_1, \dots, g_m)$$

具有完全欄秩，

$$\text{rank}(\mathbf{G}) = m$$

因此，觀測算子 (observation operator) 為單射映射 (injective)。換言之，不同潛在狀態機率結構不會映射為相同之觀測分布，故觀測序列聯合分布保留潛在狀態轉移結構之資訊。考慮三個連續觀測時間點所形成之三階聯合機率張量

$$\mathcal{T} = [\Pr(Y_1 = y_a, Y_2 = y_b, Y_3 = y_c)]_{a,b,c}$$

由 Allman 等人(2009)[20] 之隱藏馬可夫模型識別性結果可知，在觀測算子具有單射性之條件下，三階以上聯合機率張量之分解可唯一回復潛在狀態轉移結構，其唯一性至多差一個共同之潛在狀態標籤置換 (label permutation)。因此，存在置換矩陣 Π ，使得

$$P_{\theta_2}(\Delta) = \Pi^{-1} P_{\theta_1}(\Delta) \Pi$$

由假設三之疾病進展不可逆性 (irreversible progression)，生成矩陣 Q 之非零元素位置已由疾病自然史理論事先決定。疾病僅允許沿自然史方向由早期疾病狀態轉移至較晚期疾病狀態，而不允許逆向轉移。因此，各潛在疾病狀態具有固定之相對排序關係。設 P 為一置換矩陣。若存在非恆等置換使得



$$\tilde{Q} = P Q P^T$$

則部分疾病狀態之排序將被改變，進而改變生成矩陣中非零元素之位置配置，使原先所設定之疾病自然史轉移結構遭到破壞。換言之，置換後之矩陣 $\tilde{Q}$ 將不再滿足本研究所假設之不可逆疾病進展架構。因此，在本研究所建立之疾病自然史模型下，不存在保持生成矩陣結構不變之非平凡狀態標籤置換，故唯一允許之置換矩陣為恆等矩陣

$$P = I_m$$

代入上式可得

$$P_{\theta_2}(\Delta) = I_m^{-1} P_{\theta_1}(\Delta) I_m = P_{\theta_1}(\Delta)$$

即

$$P_{\theta_1}(\Delta) = P_{\theta_2}(\Delta)$$

此與反證假設

$$P_{\theta_1}(\Delta) \neq P_{\theta_2}(\Delta)$$

矛盾。故假設不成立。因此，不存在兩組不同之轉移機率矩陣能產生相同之觀測序列聯合分布。故觀測序列聯合分布可唯一決定轉移機率矩陣

$$P_{\theta}(\Delta)$$

亦即轉移機率矩陣 $P_{\theta}(\Delta)$ 為可識別 (identifiable)。證畢。

3.3.3 生成矩陣之唯一性

由定理 3.1 可知，在假設一至假設四成立下，篩檢觀測資料之聯合分布可識別轉移機率矩陣 $P_{\theta}(\Delta)$ 。因此，後續可進一步探討是否能由已識別之轉移機率矩陣唯一恢復其所對應之生成矩陣 Q 。下列定理說明在本研究所設定之疾病自然

史結構與連續時間馬可夫鏈條件下，生成矩陣具有唯一性。



定理 3.2：

在符合連續時間馬可夫鏈生成矩陣條件（即實數矩陣、非對角元素非負及各列元素總和為零）之類別中，設 $P = \exp(\Delta Q), \Delta > 0$ ，其中 Q 生成矩陣。則轉移機率矩陣 P 對應之生成矩陣 Q 唯一存在。

證明：

由假設三（疾病進展之不可逆性）可知，生成矩陣 Q 之非零元素位置由疾病自然史架構事先固定。經適當狀態排序後， Q 可表示為上三角矩陣，因此其特徵值即為對角線元素：

$$\lambda_i(Q) = q_{ii}, i = 1, \dots, m.$$

又由連續時間馬可夫鏈生成矩陣之性質，

$$q_{ii} = - \sum_{j \neq i} q_{ij} \leq 0$$

故 Q 之所有特徵值均為非正實數。考慮轉移機率矩陣

$$P = \exp(\Delta Q), \Delta > 0$$

由矩陣指數之譜映射性質可得

$$\lambda_i(P) = \exp(\Delta \lambda_i(Q)) = \exp(\Delta q_{ii}), i = 1, \dots, m$$

由於 $\Delta > 0$ 且 $q_{ii} \leq 0$ ，因此

$$0 < \exp(\Delta q_{ii}) \leq 1$$

故 P 之所有特徵值均為正實數，特別地， P 不具有位於負實軸上的特徵值。根據矩陣主對數之存在與唯一性結果，若一矩陣不具有負實軸上的特徵值，則其主

對數 $\log(P)$ 唯一存在。另一方面， ΔQ 的所有特徵值皆為非正實數，其虛部均為 0，落在主對數所對應之譜帶

$$\{z \in \mathbb{C}: -\pi < \text{Im}(z) < \pi\}$$

內。因此，

$$\log\{\exp(\Delta Q)\} = \Delta Q$$

由於 $P = \exp(\Delta Q)$ ，可得

$$\Delta Q = \log(P)$$

因此

$$Q = \frac{1}{\Delta} \log(P)$$

現在假設存在兩組生成矩陣 Q_1 與 Q_2 ，皆符合本研究所設定之生成矩陣條件與疾病自然史轉移結構，且滿足

$$\exp(\Delta Q_1) = \exp(\Delta Q_2) = P$$

由上述結果可知，

$$\Delta Q_1 = \log(P) = \Delta Q_2$$

因為 $\Delta > 0$ ，故

$$Q_1 = Q_2$$

因此，在本研究所設定之不可逆疾病進展結構下，轉移機率矩陣 P 唯一對應於生成矩陣 Q ，故生成矩陣 Q 可由 P 唯一識別。證畢。

3.3.4 全域可識別性

由定理 3.1 可知，觀測序列之聯合分布可識別轉移機率矩陣；再由定理 3.2 可得，已識別之轉移機率矩陣唯一對應於生成矩陣。此外，假設生成參數向量與生成矩陣之間具有一對一對應關係。因此，可進一步建立觀測資料聯合分布與模型參數之唯一對應關係，並證明本研究所建立之連續時間隱藏狀態疾病進展模型具有全域可識別性。

定理 3.3：

若兩組參數向量 θ_1, θ_2 所產生之篩檢觀測分布完全相同，則 $\theta_1 = \theta_2$ 。

證明(反證法)：

假設模型不具有全域可識別性，則存在兩組不同參數向量

$$\theta_1 \neq \theta_2$$

使其所誘導之觀測資料聯合分布完全相同，即

$$g_{\theta_1}(Y_1, \dots, Y_T) = g_{\theta_2}(Y_1, \dots, Y_T)$$

令

$$Q_1 = Q(\theta_1), Q_2 = Q(\theta_2)$$

並定義其對應之轉移機率矩陣為

$$P_1(\Delta) = \exp(\Delta Q_1), P_2(\Delta) = \exp(\Delta Q_2)$$

由假設四，參數化映射 $\theta \mapsto Q(\theta)$ 為單射映射。因此，若

$$\theta_1 \neq \theta_2$$

則必有





$$Q_1 \neq Q_2$$

另一方面，由於兩組參數所誘導之觀測資料聯合分布相同，依定理 3.1 可知其潛在轉移機率矩陣可由觀測分布唯一識別，因此

$$P_1(\Delta) = P_2(\Delta)$$

由定理 3.1 可知，兩組參數具有相同之觀測資料聯合分布時，其對應之轉移機率矩陣相同；再由定理 3.2 可知，在本研究所建立之不可逆疾病進展架構下，轉移機率矩陣唯一決定生成矩陣，因此

$$P_1(\Delta) = P_2(\Delta) \Rightarrow Q_1 = Q_2$$

此與前述由 $\theta_1 \neq \theta_2$ 及假設四所推出之

$$Q_1 \neq Q_2$$

相矛盾。因此，反證假設不成立，故不存在兩組不同參數向量能誘導相同之觀測資料聯合分布。換言之，若

$$g_{\theta_1}(Y_1, \dots, Y_T) = g_{\theta_2}(Y_1, \dots, Y_T)$$

則必有

$$\theta_1 = \theta_2$$

因此，本研究建立之連續時間隱藏狀態疾病進展模型具有全域可識別性。證畢。

3.4 大樣本概似理論

為建立最大概似估計量之大樣本性質，本節採用計數過程 (counting process) 與鞅理論 (martingale theory) 之架構進行說明。設 $X_k(t)$ 表示第 k 位個體於時間 t 之潛在疾病狀態。對任意狀態轉移 $i \rightarrow j$ ，定義計數過程



$$N_{ij}^{(k)}(t)$$

表示個體截至時間 t 由狀態 i 轉移至狀態 j 之累積轉移次數。進一步定義風險指標(at-risk indicator)

$$Y_i^{(k)}(t) = I\{X_k(t^-) = i\}$$

其中 $I(\cdot)$ 為指示函數 (indicator function)。當個體於時間 t 前瞬間處於狀態 i 時，其值為 1；否則為 0。因此，

$$Y_i^{(k)}(t)$$

表示個體於時間 t 是否具有由狀態 i 發生轉移之風險。在參數向量 θ 下，狀態 i 至狀態 j 之轉移強度函數 (transition intensity function) 定義為

$$\lambda_{ij}^{(k)}(t; \theta) = Y_i^{(k)}(t)q_{ij}(\theta)$$

其中 $q_{ij}(\theta)$ 表示由狀態 i 轉移至狀態 j 之瞬時轉移率 (instantaneous transition rate)。令 θ_0 表示真實參數值。根據 Doob-Meyer 分解定理 (Andersen et al., 1993)[21]，具有可預測強度函數之計數過程可唯一分解為可預測補償過程 (predictable compensator) 與鞅過程 (martingale process) 之和。此分解之目的在於將計數過程中的系統性結構與隨機變異加以區分，其中補償過程描述模型所預測之平均累積轉移行為，而鞅過程則描述實際觀測結果相對於理論期望之隨機偏離。透過此分解，可利用鞅理論研究計數過程之統計性質，並作為建立最大概似估計量大樣本理論之基礎。

令 θ_0 表示真實參數值，則其對應之可預測補償過程為

$$A_{ij}^{(k)}(t) = \int_0^t \lambda_{ij}^{(k)}(u; \theta_0) du$$



因此，計數過程可表示為

$$N_{ij}^{(k)}(t) = A_{ij}^{(k)}(t) + M_{ij}^{(k)}(t)$$

其中 $\mathcal{F}_t$ 表示截至時間 t 為止所累積之歷史資訊，而 $\{\mathcal{F}_t\}_{t \geq 0}$ 稱為自然過濾族 (natural filtration)。根據 Doob-Meyer 分解定理，

$$M_{ij}^{(k)}(t) = N_{ij}^{(k)}(t) - \int_0^t \lambda_{ij}^{(k)}(u; \theta_0) du$$

為相對於 $\{\mathcal{F}_t\}_{t \geq 0}$ 之鞅過程，並滿足

$$E \left[M_{ij}^{(k)}(t) - M_{ij}^{(k)}(s) \mid \mathcal{F}_s \right] = 0, 0 \leq s < t$$

此結果表示，計數過程可分解為理論期望累積轉移次數與隨機誤差之和。其中，補償過程描述模型所預測之累積轉移次數，而鞅項則反映實際觀測結果相對於理論期望之隨機波動。由於鞅增量之條件期望為零，因此不具有系統性偏差。此性質為得分函數 (Score function) 於真實參數下具有零期望值之理論依據，並構成建立最大概似估計量一致性與漸近常態性之重要基礎。

3.4.1 一致性

在模型可識別性成立且滿足一般正則條件 (Regular condition) 下，最大概似估計量 $\hat{\theta}_n$ 具有一致性，即

$$\hat{\theta}_n \xrightarrow{p} \theta_0, n \rightarrow \infty$$

證明：

由大數法則可得平均對數概似函數 $\frac{1}{n} \ell_n(\theta)$ 一致收斂至其期望函數

$$L(\theta) = E_{\theta_0}[\log f(Y; \theta)]$$



又由定理 3.3 之全域可識別性可知，

$$L(\theta) < L(\theta_0), \forall \theta \neq \theta_0$$

因此 θ_0 為唯一極大點。根據 Wald 一致性定理，

$$\hat{\theta}_n \xrightarrow{p} \theta_0$$

證畢。

3.4.2 漸近常態性

在定理 3.4.1 之條件下，若費雪資訊 (Fisher information) 矩陣

$$I(\theta_0) = -E \left[\frac{\partial^2 \ell(\theta_0)}{\partial \theta \partial \theta^\top} \right]$$

存在且為正定矩陣，則

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, I^{-1}(\theta_0))$$

證明

由得分函數 (Score function)

$$U_n(\theta) = \frac{\partial \ell_n(\theta)}{\partial \theta}$$

之 Taylor 展開，

$$0 = U_n(\hat{\theta}_n) = U_n(\theta_0) + \frac{\partial U_n(\tilde{\theta})}{\partial \theta^\top} (\hat{\theta}_n - \theta_0)$$

其中 $\tilde{\theta}$ 介於 $\hat{\theta}_n$ 與 θ_0 之間。由鞅中心極限定理 (Martingale Central Limit Theorem) 及可預測變異過程 (Predictable variation process) 之收斂性可得

$$\frac{1}{\sqrt{n}} U_n(\theta_0) \xrightarrow{d} N(0, I(\theta_0))$$



再由大數法則，

$$-\frac{1}{n} \frac{\partial U_n(\tilde{\theta})}{\partial \theta^\top} \xrightarrow{p} I(\theta_0)$$

最後利用 Slutsky 定理可得

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, I^{-1}(\theta_0))$$

證畢。

綜合上述結果，在模型可識別性成立且滿足一般正則條件下，本研究所估計之轉移率參數 θ 具有一致性與漸近常態性。換言之，隨著樣本數增加，估計值將收斂至真實轉移率，且其估計誤差可由費雪資訊矩陣加以量化。因此，由估計所得之生成矩陣

$$\hat{Q} = (\hat{q}_{ij})$$

可作為描述群體層級疾病自然史之基礎模型，並提供後續整合個人風險因子、建構個人化疾病進展模型及風險分層分析之理論基礎。

綜上所述，在模型可識別性成立及相關正則條件滿足下，本研究所提出之最大概似估計量具有一致性（consistency）與漸近常態性（asymptotic normality），顯示所建立之連續時間隱藏狀態疾病進展模型具有穩定且可靠之統計推論基礎。然而，上述模型所估計之轉移強度主要反映整體族群平均之疾病進展特性，尚未考量不同個體因臨床特徵、代謝指標及生活型態等因素所造成之疾病進展異質性。因此，為進一步建立個體化疾病風險評估架構，本研究於後續採用機器學習方法，利用受檢者之個人特徵建立疾病進展風險預測模型，並將預測結果整合至疾病自然史模型，以建立兼具群體層級疾病進展推論與個體層級風險預測之整合架構。

3.5 機器學習預測模型

本研究進一步利用機器學習方法建立個人化疾病預測模型 (individual-level disease prediction model)，以評估個人健康特徵與疾病狀態之關聯性。與前述隱藏狀態模型著重於疾病進展機制推論不同，機器學習模型主要利用個人層級之人口學、生活型態及臨床檢驗資料預測疾病狀態，因此本研究將其視為疾病風險預測之補充分析。

為反映大腸直腸疾病之階段性進展，本研究分別建立兩個二元分類模型。第一個模型用於區分正常個體 (Normal) 與腺瘤個體 (Adenoma)，以探討腺瘤形成之相關危險因子；第二個模型則用於區分腺瘤個體 (Adenoma) 與篩檢檢出癌症 (screen-detected cancer)，以評估影響疾病進一步惡化為癌症之重要特徵。

本研究分別建立人工神經網路 (Artificial Neural Network, ANN) 與隨機森林 (Random Forest, RF) 模型進行疾病分類，並利用模型解釋方法探討各預測變數對疾病進展之影響。

3.5.1 研究對象與研究變項

本研究使用包含個人健康資訊之篩檢資料建立機器學習疾病預測模型，以評估個人特徵對腺瘤形成及大腸直腸癌發生之預測能力。研究資料包含人口學特徵、家族病史、生活型態、身體測量指標及臨床生化檢驗結果等資訊。本研究共建立兩個二元分類模型，其疾病結果變數定義如下：

(一) Normal versus Adenoma 分類模型

$$Y = \begin{cases} 0, & \text{Normal} \\ 1, & \text{Adenoma} \end{cases}$$

此模型主要探討健康個體與腺瘤個體之差異，評估各項危險因子對腺瘤形成之預測能力。

(二)Adenoma versus Screen-detected CRC 分類模型

$$Y = \begin{cases} 0, & \text{Adenoma} \\ 1, & \text{Screen - detected CRC} \end{cases}$$



此模型主要探討腺瘤患者進一步發展為大腸直腸癌之相關危險因子，評估各項臨床特徵對癌症發生之預測能力。本研究所使用之特徵變相可分為以下四類：

(1)人口學特徵 (Demographic Variables)

包括：

- 性別 (Gender)
- 進入研究年齡 (Entry Age)
- 教育程度 (Education Level)
- 大腸癌家族病史 (Family History of Colorectal Cancer)

人口學特徵可反映個體基本社會背景及遺傳風險，其中家族病史已被證實與大腸直腸癌風險增加具有顯著相關性。

(2)生活型態變數 (Lifestyle Variables)

包括：

- 抽菸 (Smoking)
- 喝酒 (Alcohol Drinking)
- 嚼檳榔 (Betel Nut Chewing)
- 運動習慣 (Exercise Habit)
- 吃蔬菜習慣 (Vegetable Consumption)



- 吃葷食習慣 (Meat Consumption)
- 吃水果習慣 (Fruit Consumption)
- 喝咖啡習慣 (Coffee Consumption)

上述變數反映個體長期健康行為與飲食習慣。過去研究指出吸菸、飲酒、高紅肉攝取及缺乏運動等因素可能增加大腸腺瘤及大腸直腸癌之發生風險。

(3)身體測量指標 (Anthropometric Variables)

包括：

- 腰圍 (Waist)
- 空腹血糖 (AC)

此類變數可反映肥胖程度與腹部脂肪堆積情形。肥胖與代謝症候群已被認為與大腸腺瘤形成及大腸直腸癌風險增加有關。

(4)臨床與生化檢驗變數 (Clinical and Biochemical Variables)

包括：

- 糞便免疫化學檢查數值 (Fecal Immunochemical Test Level, FIT)
- 糞便免疫化學檢查分組 (Fecal Immunochemical Test Group, FITGP)
- 收縮壓 (Systolic Blood Pressure, SBP)
- 舒張壓 (Diastolic Blood Pressure, DBP)
- 三酸甘油酯 (Triglyceride, TG)
- 高密度脂蛋白膽固醇 (High-density Lipoprotein Cholesterol, HDL)
- 麩草醋酸轉胺酶 (Glutamate Oxaloacetate Transaminase, GOT)
- 麩草丙酮酸轉胺酶 (Glutamate Pyruvate Transaminase, GPT)



- 尿酸 (Uric Acid, UA)
- 血中尿素氮 (Blood Urea Nitrogen, BUN)

上述變數反映個體心血管代謝狀態、肝功能與腎功能狀況。近年研究指出，代謝異常、慢性發炎及胰島素阻抗可能參與大腸腺瘤形成及大腸直腸癌發生之病理機制，因此本研究將其納入預測模型分析。其中，資料庫中糞便免疫化學檢查值之變數名稱為 FBT，代表糞便免疫化學檢查 (Fecal Immunochemical Test, FIT) 之定量檢測值。為與資料庫及 SHAP 視覺化分析結果一致，後續涉及 SHAP 圖形時保留原始變數名稱 FBT；其餘本文內容則統一以 FIT 表示。

3.5.2 資料處理與模型建立

於建立機器學習模型前，必須先進行資料前處理 (data preprocessing)，以提升資料品質並改善模型訓練效果。由於本研究所使用之資料同時包含連續型變數與類別型變數，且不同變數之量測尺度差異較大，因此需進行適當之前處理程序。此外，部分疾病類別樣本數相對較少，可能導致模型產生類別偏倚 (class bias)，因此亦需進行類別不平衡處理。本研究之資料前處理流程包括資料清理、變數標準化、類別變數編碼、交叉驗證及類別不平衡處理等步驟，說明如下。

(一) 資料清理與變數轉換

於模型建立前，首先進行缺失值 (missing values) 與異常值 (outliers) 檢查，以確保資料品質。對於缺失值、無限值 (infinite values) 及異常資料，先進行篩檢與清理，以避免影響模型訓練結果。

對於連續型變數，包括初次研究年齡 (Entry Age)、糞便免疫化學檢查值 (FIT)、誘導時間 (Induction Time)、診斷年齡 (Diagnosed Age)、收縮壓 (SBP)、舒張壓 (DBP)、腰圍 (Waist)、腹圍 (AC)、三酸甘油酯 (TG)、高密度脂蛋白膽固醇 (HDL)、肝功能指標 (GOT、GPT)、尿酸 (UA)、血中尿素氮

(BUN) 及肌酸酐 (Creatinine) 等，於人工神經網路 (Artificial Neural Network, ANN) 模型建構前皆進行標準化處理，以降低不同量測尺度對模型訓練之影響。

標準化後之變數表示如下：

$$Z_i = \frac{X_i - \mu}{\sigma}$$

其中：

- X_i 為原始變數值；
- μ 為該變數平均數；
- σ 為該變數標準差；
- Z_i 為標準化後之數值。

經標準化後，各變數平均數接近 0，標準差接近 1，有助於提升人工神經網路模型之收斂效率與訓練穩定性。另一方面，隨機森林 (Random Forest, RF) 模型以決策樹為基礎，其節點分裂係依據變數排序及切分點進行，不受變數量測尺度影響，因此本研究於隨機森林模型分析時未進行標準化處理，以保留原始資料之數值資訊。

對於類別型變數，包括性別 (Gender)、教育程度 (Education)、家族病史 (CRC Family History)、吸菸 (Smoking)、飲酒 (Alcohol Drinking)、檳榔使用 (Betel Nut Chewing)、運動習慣 (Exercise) 及飲食相關變數等，則轉換為適當之數值編碼格式，以利後續模型訓練。

(二)分層十折交叉驗證 (Stratified 10-Fold Cross-Validation)

為評估模型之泛化能力 (Generalization ability) 並降低單次資料切分所造成之隨機誤差，本研究採用分層十折交叉驗證 (Stratified 10-fold cross-validation) 進行



模型訓練與驗證。設完整資料集表示為：

$$D = \{(x_i, y_i)\}_{i=1}^n$$

其中：

- x_i 表示第 i 位個體之特徵向量；
- y_i 表示其疾病分類結果；
- n 為樣本總數。

本研究首先依疾病類別比例將資料分為十個互斥子集合 (10 - folds)，每次選取其中九個子集合作為訓練資料集 (Training dataset)，其餘一個子集合作為測試資料集 (Testing dataset)，並重複進行十次模型訓練與驗證，使每筆資料皆有一次作為測試樣本之機會。最後將十次驗證結果取平均作為模型最終效能估計。由於各 fold 均維持原始疾病類別比例，因此可降低類別不平衡對模型評估結果之影響，並提供較穩定且具代表性之模型效能指標。

(三)類別不平衡處理

由於本研究建立兩個二元分類模型，包括 Normal versus Adenoma 與 Adenoma versus Screen-detected CRC，兩者之疾病分布情形不同，因此採用不同策略處理類別不平衡問題。

1. Normal versus Adenoma 模型：SMOTE

在 Normal versus Adenoma 分類模型中，由於 Adenoma 個案數明顯少於 Normal 個案，因此本研究採用 Synthetic Minority Over-sampling Technique (SMOTE) 進行少數類別過抽樣。設少數類別樣本集合表示為：

$$S_m = \{x_1, x_2, \dots, x_n\}$$



其中 n 為少數類別樣本數。對於任一少數類別樣本 x_i ，首先利用歐氏距離 (Euclidean distance) 計算其與其他少數類別樣本之距離：

$$d(x_i, x_j) = \sqrt{\sum_{l=1}^p (x_{il} - x_{jl})^2}$$

並選取其 k 個最近鄰樣本 (k-nearest neighbors)。隨後隨機選取其中一個鄰近樣本 x_{NN} ，產生新的合成樣本：

$$x_{new} = x_i + \lambda(x_{NN} - x_i)$$

其中：

$$\lambda \sim U(0,1)$$

因此，新生成樣本位於原始樣本與其鄰近樣本之間之特徵空間，可增加少數類別樣本之多樣性，降低模型對多數類別之偏倚。

2. Adenoma versus Screen-detected CRC 模型：Class Weight

在 Adenoma versus CRC 分類模型中，由於 CRC 個案數相對較少，因此本研究於人工神經網路模型中採用類別權重 (class weight) 方式處理類別不平衡問題。設第 c 類別之權重為 w_c ，則加權交叉熵損失函數可表示為：

$$L = -\frac{1}{n} \sum_{i=1}^n \sum_{c=0}^1 w_c I(y_i = c) \log(\hat{p}_{ic})$$

其中：

- w_c 為第 c 類別之權重；



- $\hat{p}_{ic}$ 為樣本 i 被預測為類別 c 之機率；
- $I(\cdot)$ 為指示函數。

當某類別樣本數較少時，其權重將提高，使模型在訓練過程中更加重視少數類別個案，以提升 Screen-detected CRC 之辨識能力。此外，所有類別不平衡處理均僅於訓練資料中進行，以避免測試資料資訊洩漏 (data leakage)，並確保模型效能評估之客觀性。

3.5.3 人工神經網路模型

人工神經網路 (Artificial Neural Network, ANN) 是一種模仿生物神經系統資訊處理機制所建立之監督式機器學習模型，具有良好的非線性擬合能力 (Nonlinear fitting capability)，能夠有效捕捉複雜變數間之交互作用 (Interaction effects) 及非線性關係 (Nonlinear relationships)。由於大腸直腸疾病之發生通常受到多種危險因子共同影響，且各危險因子間可能存在複雜關聯，因此本研究採用人工神經網路建立疾病分類模型，以評估個人健康特徵對疾病狀態之預測能力。本研究所建立之人工神經網路模型，其輸出層僅設置單一神經元，用以輸出個體屬於目標疾病類別之預測機率。

(一)模型架構

人工神經網路主要由輸入層 (Input Layer)、隱藏層 (Hidden Layer) 及輸出層 (Output Layer) 所組成。輸入層負責接收個體之臨床特徵資料，包括人口學變數、生活型態變數、身體測量指標及生化檢驗變數等。設輸入向量表示為：

$$X = (x_1, x_2, \dots, x_p)^T$$

其中：

- p 為輸入特徵數目；



- x_i 為第 i 個輸入變數。

對於隱藏層中第 j 個神經元，其線性組合可表示為：

$$z_j = \sum_{i=1}^p w_{ij} x_i + b_j$$

其中：

- w_{ij} 為連接權重 (weight)；
- b_j 為偏差項 (bias)。

隱藏層輸出則透過活化函數 (activation function) 進行非線性轉換：

$$h_j = g(z_j)$$

本研究採用 Rectified Linear Unit (ReLU) 作為活化函數：

$$ReLU(z) = \max(0, z)$$

ReLU 函數具有計算效率高及可降低梯度消失 (vanishing gradient) 問題等優點，因此廣泛應用於現代神經網路模型。

(二)輸出層與疾病預測機率

由於本研究屬於二元分類問題，因此輸出層僅包含一個神經元，其輸出值表示個體屬於目標疾病類別之機率。設輸出層線性組合為：

$$z = \sum_{j=1}^m v_j h_j + b$$

其中：

- m 為隱藏層神經元數目；



- v_j 為隱藏層至輸出層之權重；
- b 為輸出層偏差項。

為將輸出值轉換為介於 0 至 1 間之預測機率，本研究採用 Sigmoid 函數：

$$\hat{p} = \frac{1}{1 + e^{-z}}$$

其中：

- $\hat{p}$ 表示模型預測個體屬於目標疾病類別之機率。

(三)損失函數

為衡量模型預測結果與真實疾病狀態間之差異，本研究採用二元交叉熵損失函數 (Binary Cross-Entropy Loss) 作為最佳化目標：

$$L = -\frac{1}{n} \sum_{i=1}^n [y_i \log(\hat{p}_i) + (1 - y_i) \log(1 - \hat{p}_i)]$$

其中：

- n 為樣本數；
- y_i 為實際疾病狀態；
- $\hat{p}_i$ 為模型預測機率。

當模型預測結果越接近真實疾病狀態時，損失函數值越小。模型訓練過程中，利用反向傳播演算法 (Backpropagation Algorithm) 計算損失函數對各權重參數之梯度，並採用 Adam 最佳化演算法 (Adaptive Moment Estimation, Adam) 更新模型權重。Adam 結合梯度之一階動量與二階動量資訊，自動調整各參數之學習率，以提升模型收斂速度及訓練穩定性。經由多次迭代訓練 (iterations)，模型逐步更

新權重與偏差參數，使二元交叉熵損失函數逐漸收斂，進而提升疾病分類之預測能力。



3.5.4 隨機森林模型

除採用人工神經網路進行疾病分類外，本研究亦建立隨機森林 (Random Forest, RF) 模型作為另一種監督式機器學習方法，以評估個人臨床特徵對疾病狀態之預測能力。隨機森林係由 Breiman 提出之集成學習 (Ensemble learning) 方法，其核心概念為透過建立多棵決策樹 (Decision trees)，並整合各決策樹之預測結果，以提升模型之穩定性與分類效能。相較於單一決策樹容易受到資料波動影響而產生過度擬合 (Overfitting) 問題，隨機森林藉由 Bootstrap 抽樣及隨機特徵選取 (Random feature selection) 機制，可有效降低模型變異性 (Variance)，並提升模型對未知資料之泛化能力 (Generalization ability)。

(一) Bootstrap Sampling

設訓練資料集表示為：

$$D = \{(x_i, y_i)\}_{i=1}^n$$

其中：

- x_i 表示第 i 位個體之特徵向量；
- y_i 表示其疾病分類結果；
- n 為樣本數。

隨機森林首先利用拔靴抽樣法自原始訓練資料集中進行有放回抽樣，建立多組子訓練資料集：

$$D_1, D_2, \dots, D_B$$



其中：

- B 為決策樹總數。

由於 Bootstrap 抽樣允許重複抽取，因此部分樣本可能於同一訓練資料集中重複出現，而部分樣本則未被抽中。未被某棵決策樹使用之樣本稱為 Out-of-Bag(OOB)樣本，可作為模型內部驗證資料，用以估計模型分類誤差。

(二)節點分裂與 Gini Impurity

於每一棵決策樹建立過程中，模型並非使用全部特徵進行節點分裂，而是在每次分裂時隨機抽取部分特徵作為候選變數集合。此機制可降低各決策樹間之相關性，提高森林內各決策樹之異質性，進而提升模型之泛化能力。

本研究以 Gini impurity 作為節點純度之衡量指標。設節點中第 k 類樣本所占比例為 p_k ，則 Gini impurity 定義為：

$$G = 1 - \sum_{k=1}^2 p_k^2$$

其中 p_k 為第 k 類樣本於該節點之比例。當節點內所有樣本皆屬於同一類別時， $G = 0$ ，表示節點已完全純化；當各類別比例越接近時，Gini impurity 越大，表示節點分類不確定性越高。因此模型會選擇使分裂後加權 Gini impurity 最小（或節點純度提升最多）之變數及切點作為最佳分裂規則。。

(三)分類決策機制

完成 B 棵決策樹建構後，每棵決策樹均會對個體產生一組分類結果。設第 b 棵決策樹對個體 i 之預測結果為：

$$\hat{y}_i^{(b)}$$



則隨機森林最終分類結果可表示為：

$$\hat{y}_i = \text{mode}(\hat{y}_i^{(1)}, \hat{y}_i^{(2)}, \dots, \hat{y}_i^{(B)})$$

即以多數決(majority voting)方式決定最終分類結果。

3.5.5 模型效能評估

為評估人工神經網路 (Artificial Neural Network, ANN) 與隨機森林 (Random Forest, RF) 模型之疾病分類能力，本研究於模型訓練完成後，利用測試資料集 (testing dataset) 進行模型驗證，並以混淆矩陣 (Confusion Matrix) 作為分類結果評估之基礎。本研究分別建立兩個二元分類模型，包括：

1. Normal versus Adenoma
2. Adenoma versus Screen-detected CRC

因此，各模型均屬於二元分類問題 (binary classification problem)。設：

- True Positive (TP) 表示實際為陽性且預測為陽性之個案數；
- True Negative (TN) 表示實際為陰性且預測為陰性之個案數；
- False Positive (FP) 表示實際為陰性但預測為陽性之個案數；
- False Negative (FN) 表示實際為陽性但預測為陰性之個案數。

根據上述定義，本研究計算準確率 (Accuracy)、敏感度 (Sensitivity)、特異度 (Specificity)、精確率 (Precision)、F1-score 及 ROC 曲線下面積 (Area Under the Curve, AUC) 等指標，以綜合評估模型分類效能。

(一)混淆矩陣 (Confusion Matrix)

混淆矩陣用以呈現模型預測結果與實際疾病狀態之對應關係，其形式如下：



	實際陽性 實際陰性	
預測陽性	TP	FP
預測陰性	FN	TN

透過混淆矩陣可進一步計算各項分類效能指標，以評估模型對疾病狀態之辨識能力。

(二)準確率 (Accuracy)

準確率表示模型正確分類之樣本比例，其計算公式如下：

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

準確率可反映模型整體分類正確程度，但當資料存在類別不平衡 (Class imbalance) 時，可能高估模型表現，因此需搭配其他評估指標共同解釋。

(三)敏感度 (Sensitivity)

敏感度又稱召回率 (Recall)，表示模型正確辨識疾病個案之能力，其定義如下：

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

敏感度越高，表示模型漏判疾病個案(false negative)之比例越低。在本研究中，較高之敏感度代表模型能有效辨識腺瘤個案或大腸直腸癌個案，對疾病早期發現具有重要意義。

(四)特異度 (Specificity)

特異度表示模型正確辨識非疾病個案之能力，其計算公式如下：

$$Specificity = \frac{TN}{TN + FP}$$

特異度越高，表示模型較不容易將健康個體或較低風險個體誤判為疾病個案，可降低不必要之後續檢查與醫療資源浪費。



(五)精確率 (Precision)

精確率表示模型預測為疾病個案中，實際確實為疾病個案之比例，其定義如下：

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

精確率越高，表示模型預測結果之可信度越高。

(六)F1-score

由於 Precision 與 Recall 間通常存在權衡關係，因此本研究進一步利用 F1-score 作為綜合評估指標。F1-score 為 Precision 與 Recall 之調和平均數，其定義如下：

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

F1-score 同時考量模型之敏感度與精確率，特別適用於疾病資料存在類別不平衡情況下之模型評估。

(七)ROC 曲線與 AUC

除上述指標外，本研究亦利用受試者操作特徵曲線 (Receiver Operating Characteristic Curve, ROC Curve) 評估模型之區辨能力。ROC 曲線係以真陽性率 (True Positive Rate, TPR) 為縱軸，假陽性率 (False Positive Rate, FPR) 為橫軸繪製而成：



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

$$FPR = \frac{FP}{FP + TN}$$

透過改變分類閾值 (Classification threshold)，可獲得不同之 TPR 與 FPR 組合，進而形成 ROC 曲線。ROC 曲線下面積 (Area Under the Curve, AUC) 可作為模型整體區辨能力之量化指標。AUC 值介於 0.5 至 1.0 之間：

- AUC = 0.5 表示模型分類能力接近隨機猜測；
- AUC = 1.0 表示模型具有完美區辨能力。

一般而言，AUC 值越接近 1，表示模型越能有效區分不同疾病狀態。

(八)模型比較

本研究將依據 Accuracy、Sensitivity、Specificity、Precision、F1-score 與 AUC 等指標，比較人工神經網路與隨機森林模型之分類效能。此外，本研究將分別比較：

1. Normal versus Adenoma 模型之預測表現；
2. Adenoma versus Screen-detected CRC 模型之預測表現。

藉此評估不同機器學習方法於大腸直腸疾病不同進展階段之分類能力，並作為最佳預測模型選擇之依據。

3.5.6 模型解釋性分析

由於人工神經網路 (Artificial Neural Network, ANN) 與隨機森林 (Random Forest, RF) 等機器學習模型具有良好的預測能力，但其內部決策機制較難直接解釋，因此本研究進一步採用模型解釋性分析 (Model Interpretability Analysis)，探

討各臨床特徵對疾病分類結果之影響程度，以提升模型之可解釋性 (interpretability)。



本研究採用 SHapley Additive exPlanations (SHAP) 方法進行模型解釋。SHAP 建立於合作賽局理論之 Shapley Value 概念，可量化各特徵變數對模型預測結果之貢獻，並提供變數重要性排序。本研究分別針對 Normal versus Adenoma 與 Adenoma versus Screen-detected CRC 兩個二元分類模型進行 SHAP 分析，以探討不同疾病進展階段之重要預測因子。

(一)Shapley Value

SHAP 方法係基於合作賽局理論中的 Shapley Value 所建立。設模型包含 p 個輸入特徵，令 $N = \{1, 2, \dots, p\}$ 表示所有特徵之集合。則第 i 個特徵之 Shapley Value 定義如下：

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(p-|S|-1)!}{p!} [f(S \cup \{i\}) - f(S)]$$

其中：

- N 為所有特徵集合；
- S 為不包含特徵 i 之特徵子集合；
- $f(S)$ 為模型於特徵集合 S 下之預測結果；
- ϕ_i 為特徵 i 對模型預測結果之邊際貢獻 (marginal contribution)。

Shapley Value 可視為在所有可能特徵組合下，某一特徵對模型預測結果之平均貢獻，因此能夠公平地衡量各特徵於模型決策中的相對重要程度。

(二)SHAP 值之表示

對於第 j 位受試者，其模型預測結果可表示為：

$$f(x_j) = \phi_0 + \sum_{i=1}^p \phi_{ij}$$



其中：

- $f(x_j)$ 為第 j 位個體之模型預測值；
- ϕ_0 為所有樣本之平均預測值 (Baseline prediction)；
- ϕ_{ij} 為第 j 位個體於第 i 個特徵之 SHAP 值。

由於本研究採用二元分類模型，因此 SHAP 值可解釋為某一特徵對模型預測機率之影響方向與影響程度。當 SHAP 值為正時，表示該特徵會提高模型預測個體屬於較進展疾病狀態之機率；反之，若 SHAP 值為負，則表示該特徵降低模型預測為較進展疾病狀態之機率。例如：

- 在 Normal versus Adenoma 模型中，正向 SHAP 值代表提高個體被分類為 Adenoma 之傾向；
- 在 Adenoma versus Screen-detected CRC 模型中，正向 SHAP 值則代表提高個體被分類為 Screen-detected CRC 之傾向。

(三)SHAP 蜂群圖(SHAP Beeswarm Plot)

為評估各特徵對模型預測結果之整體影響，本研究利用 SHAP Beeswarm Plot 進行全域解釋分析 (Global interpretation)。SHAP Beeswarm Plot 將所有受試者之 SHAP 值分布繪製於同一圖中：

- 橫軸表示 SHAP value；
- 縱軸表示特徵名稱；



- 顏色表示特徵數值大小。

藉由 SHAP Beeswarm Plot 可同時觀察：

1. 特徵重要性大小；
2. 特徵對疾病風險之影響方向；
3. 特徵影響之個體間變異程度。

本研究利用平均絕對 SHAP 值 (Mean Absolute SHAP Value) 作為變數重要性指標，其定義如下：

$$Importance_i = \frac{1}{n} \sum_{j=1}^n |\phi_{ij}|$$

其中：

- n 為樣本數；
- ϕ_{ij} 為第 j 位個體於第 i 個特徵之 SHAP 值。

平均絕對 SHAP 值越大，表示該特徵對模型預測結果之影響越重要。

(四)SHAP 蜂群摘要圖 (SHAP Beeswarm Plot)

為進一步探討重要變數與疾病風險間之關聯，本研究繪製 SHAP 蜂群摘要圖，以呈現特徵數值與 SHAP 值間之變化關係。SHAP 蜂群摘要圖以特徵實際數值作為橫軸，以對應之 SHAP 值作為縱軸，可用以評估特徵與疾病風險間的特徵影響。例如，本研究可利用 SHAP 蜂群摘要圖探討：

- 初次研究年齡 (Entry Age) 增加是否提高 Adenoma 或 Screen-detected CRC 發生風險；



- 腰圍 (Waist) 增加是否提高疾病預測機率；
- 收縮壓 (SBP) 與舒張壓 (DBP) 對疾病風險之影響模式；
- 三酸甘油酯 (TG)、高密度脂蛋白膽固醇 (HDL) 等代謝指標與疾病風險之非線性關係。

透過 SHAP 蜂群摘要圖，可進一步解釋機器學習模型之決策邏輯，並協助辨識不同疾病進展階段的重要危險因子。

3.6 個人化風險分層之疾病進展模型

本研究前述所建立之連續時間隱藏狀態疾病進展模型，可估計大腸直腸癌自然史中各疾病狀態間之平均轉移強度 (Transition intensity)，其所得之生成矩陣 (Generator matrix) 代表整體族群 (Population level) 之疾病自然史，亦即描述平均受檢者由正常狀態逐步進展至腺瘤、臨床症前可偵測期階段及臨床癌症之疾病進展速度。然而，不同個體因年齡、糞便血紅素濃度 (f-Hb)、生活型態及健康特徵不同，其疾病進展風險可能具有顯著異質性，因此僅利用群體平均生成矩陣，尚不足以反映個體實際疾病進展歷程。

因此，本研究進一步整合機器學習所得之個人化疾病風險資訊，建立個人化風險分層之疾病進展模型 (Personalized risk-stratified disease progression model)，將群體層級之疾病自然史模型延伸至個體層級之疾病進展推估。其基本概念為利用疾病自然史模型所估計之群體平均生成矩陣作為基準 (Generator matrix at population level)，再依據機器學習模型所估計之個體風險，調整各疾病轉移率，使不同個體具有不同之疾病進展速度。

本研究於第四章首先建立兩項二元分類模型，包括 Normal versus Adenoma 以及 Adenoma versus Screen-detected CRC，並分別利用人工神經網路 (Artificial

Neural Network, ANN) 與隨機森林 (Random Forest, RF) 進行分類分析。經比較模型分類能力後，以具有最佳預測效能之模型作為個人化風險估計模型，並利用其預測機率 (Predicted probability) 作為個體疾病風險之量化指標。同時，利用 SHapley Additive exPlanations (SHAP) 分析重要預測變數，以了解各危險因子對個人疾病風險之相對貢獻。

設第 i 位個體於某疾病進展階段之機器學習預測機率為 p_i ，全體樣本之平均預測機率為 $\bar{p}$ ，則定義其相對風險倍數 (Relative risk multiplier) 為

$$RR_i = \frac{p_i}{\bar{p}}.$$

其中，($RR_i > 1$)表示該個體疾病進展風險高於平均族群；($RR_i < 1$)則表示其疾病進展風險低於平均水準。

由於本研究主要關注兩段關鍵疾病進展，包括正常狀態進展至腺瘤 (Normal → Adenoma) 以及腺瘤進展至篩檢癌 (Adenoma → Screen-detected CRC)，因此分別建立兩組相對風險倍數，記為

$$RR_i^{(NA)}$$

及

$$RR_i^{(AC)}$$

再利用群體層級生成矩陣所估計之平均轉移率，建構個人化生成矩陣 (Personalized generator matrix)。若群體平均生成矩陣中 Normal 至 Adenoma 之轉移率記為

$$q_{12}$$

Adenoma 至臨床症前可偵測期階段(或癌症)之轉移率記為



$$q_{23}$$

則第 (i) 位個體之個人化轉移率定義為

$$q_{12}(i) = RR_i(NA)q_{12}$$

$$q_{23}(i) = RR_i(AC)q_{23}$$

其餘疾病轉移率仍維持群體模型所估計之平均值，因此可建立第 i 位個體之生成矩陣

$$Q_i$$

根據連續時間馬可夫鏈理論，第 i 位個體於時間 t 之轉移機率矩陣可表示為

$$P_i(t) = \exp(Q_i t)$$

其中 $\exp(\cdot)$ 表示矩陣指數 (Matrix exponential)。利用此個人化轉移機率矩陣，即可計算不同個體於各疾病狀態之累積進展機率 (Cumulative progression probability)，進而建立個人化疾病進展光譜 (Personalized disease progression spectrum)。

此外，為提升臨床解釋性，本研究依據隨機森林模型所估計之預測機率，將受檢者依十分位數(decile)劃分為十個風險層級。各風險層皆具有不同之相對風險倍數及對應之疾病轉移率，因此可呈現由低風險至高風險之疾病進展梯度 (Risk spectrum)。最後，本研究選取不同風險層代表個案，利用個人化生成矩陣推估其疾病自然史，呈現不同個體於相同追蹤期間內之疾病進展軌跡，以說明個人化疾病進展模型於精準風險分層及個人化篩檢策略中的應用價值

3.7 小結

本章首先建立連續時間隱藏狀態疾病模型 (Continuous-time hidden-state



disease model)，以描述大腸直腸癌由正常狀態、腺瘤、臨床症前可偵測期至臨床癌症之疾病進展過程。透過連續時間馬可夫過程 (Continuous-time Markov process) 與生成矩陣 (Generator matrix) 和發射矩陣 (Emission model) 之建構，刻畫潛在疾病狀態 (Latent disease states) 與 FIT 工具篩檢敏感度間之轉移動態。此外，本研究利用計數過程 (Counting process) 與概似函數 (Likelihood function) 建立模型之統計推論架構，並說明模型可辨識性及參數估計方法。

除疾病自然史模型外，本研究亦建立機器學習疾病預測模型，以評估個人健康特徵對不同疾病進展階段之預測能力。為反映大腸直腸癌疾病自然史中兩個重要疾病轉移階段，本研究分別建立 Normal versus Adenoma 與 Adenoma versus Screen-detected CRC 兩個二元分類模型，並採用人工神經網路 (Artificial Neural Network, ANN) 與隨機森林 (Random Forest, RF) 進行疾病分類分析。模型建立前先完成資料前處理與類別不平衡處理，以提升模型訓練品質；模型建立後則以 Accuracy、Sensitivity、Specificity、Precision、F1-score 及 AUC 等指標評估模型效能，並利用 SHapley Additive exPlanations (SHAP) 分析各預測變數對模型分類結果之貢獻。

最後，本研究進一步整合群體層級疾病自然史模型與機器學習預測結果，依據個體預測機率建立相對風險倍數 (Relative risk multiplier)，並將其導入生成矩陣之轉移強度，以建構個人化風險分層疾病進展模型，作為後續疾病進展模擬與個人化風險評估之基礎。



第四章 資料分析結果

4.1 全國 FIT 篩檢資料庫之人口特徵與觀測結果分布

本研究利用臺灣全國大腸直腸癌 FIT 篩檢資料庫，納入西元 2004 年至 2021 年（共 18 年）之篩檢資料，共包含 14,478,478 人次篩檢紀錄及 5,640,906 名受檢者（表 4.1）。受檢者基本特徵係以首次篩檢資料進行統計；疾病狀態則依據病理診斷及癌症登記資料判定之最終疾病診斷。腺瘤依病灶大小區分為小型腺瘤 (Small adenoma) 與進展型腺瘤 (Advanced adenoma)；臨床症前可偵測期癌症則依 TNM 分期分為早期 (Stage 0–2) 及晚期 (Stage 3–4)。

而確診結果分類則依據篩檢流程進行分類，包括未確診、篩檢偵測癌、間隔癌、延遲診斷癌及拒絕者。因此，未確診係指於追蹤期間內未被診斷為大腸癌之受檢者，因此這裡未確診類別項包含腺瘤；首次篩檢偵測癌 (PSD) 係指個案於第一次參與篩檢後 2 年內確診之癌症；後續篩檢偵測癌 (SSD) 為後續篩檢輪次偵測之癌症；間隔癌 (Interval cancer) 定義為 FIT 陰性後 2 年內確診之癌症；延遲診斷癌係指 FIT 陽性但未於 2 年內完成確診之個案；拒絕者 (Refuser) 係指具有大腸直腸癌確診紀錄，但未完成對應 FIT 篩檢流程之個案。

為保留疾病自然進程之觀測資訊，本研究於個案首次出現非正常結果後即終止追蹤，以排除後續醫療介入對疾病狀態判定之干擾，並取該筆紀錄作為個案之最終觀測狀態。

表 4.1 台灣全國 FIT 篩檢資料之人口特徵與疾病分布



變數	分類	人次 n	人次 %	人數 n	人數 %
		14,478,478		5,640,906	
性別	男	6,082,551	42.01	2,585,140	45.83
	女	8,395,927	57.99	3,055,766	54.17
年齡	50–54	2,942,879	20.33	2,301,959	40.81
	55–59	3,869,577	26.73	1,429,419	25.34
	60–64	3,522,533	24.33	971,067	17.21
	65–69	2,938,653	20.30	729,870	12.94
	≥70	1,204,836	8.32	208,591	3.70
FIT 結果	陰性	13,420,875	93.09	5,168,497	91.62
	陽性	995,711	6.88	412,528	7.31
	未檢測	61,892	0.43	59,881	1.06
最終疾病診斷	Normal	14,153,469	97.76	5,482,396	97.19
	Small adenoma	157,514	1.09	53,642	0.95
	Advanced adenoma	71,500	0.49	31,495	0.56
	Early PCDP	23,527	0.16	12,246	0.22



	Late PCDP	7,841	0.05	4,639	0.08
	Early CP	36,214	0.25	30,757	0.55
	Late CP	28,413	0.20	25,731	0.46
確診結果分類	未確診	14,382,483	99.34	5,544,911	98.3
	首次篩檢偵測癌	16,885	0.12	16,885	0.3
	後續篩檢偵測癌	14,483	0.10	14,483	0.26
	間隔癌	8,383	0.06	8,383	0.15
	延遲診斷癌	14,021	0.10	14,021	0.25
	拒絕者	42,223	0.29	42,223	0.75

註：人次資料之年齡係依各次篩檢時之年齡進行統計；受檢人資料之年齡則以首次進入研究年齡進行分類。

受檢者以女性略多 (54.17%)，年齡分布以 50–54 歲最為集中 (40.81%)，反映 FIT 篩檢以 50 歲作為起始篩檢年齡之政策設計，該年齡層多為首次進入篩檢系統之新進受檢者。

在最終疾病狀態分布方面，絕大多數受檢者 (97.19%) 於篩檢時屬於正常狀態，腺瘤相關病灶 (Small adenoma 0.95%、Advanced adenoma 0.56%)及臨床症前可偵測期階段 (Early PCDP 0.22%、Late PCDP 0.08%) 合計約為 0.3 %，臨床期癌症 (Early CP 0.55%、Late CP 0.46%)則約占 1%。

依人次計算，未確診比例為 99.34%，篩檢偵測癌 (首次 0.12%、後續 0.10%)、間隔癌 (0.06%) 及延遲診斷癌 (0.10%) 合計約 0.38%；依人數計算，上述比例分別為 98.30%、0.30%、0.26%、0.15%及 0.25%，兩者差異主要來自未確診受檢者之重複篩檢。



表 4.2 不同確診結果之人口特徵與 FIT 結果分布

變數	分類	未確診	首次篩檢偵測癌	後續篩檢偵測癌	間隔癌	延遲診斷癌	拒絕者
		5,544,911	16,885	14,483	8,383	14,021	42,223
		N (%)	N (%)	N (%)	N (%)	N (%)	N (%)
性別	男	2,528,097 (45.59%)	10,676 (63.23%)	7,964 (54.99%)	4,594 (54.80%)	7,778 (55.47%)	26,031 (61.65%)
	女	3,016,814 (54.41%)	6,209 (36.77%)	6,519 (45.01%)	3,789 (45.20%)	6,243 (44.53%)	16,192 (38.35%)
初次研究 年齡	50–54	2,284,683 (41.20%)	3,552 (21.04%)	3,788 (26.15%)	1,969 (23.49%)	2,235 (15.94%)	5,732 (13.58%)
	55–59	1,408,357 (25.40%)	3,970 (23.51%)	4,416 (30.49%)	2,201 (26.26%)	2,786 (19.87%)	7,689 (18.21%)
	60–64	950,310 (17.14%)	4,110 (24.34%)	3,935 (27.17%)	2,028 (24.19%)	2,842 (20.27%)	7,842 (18.57%)
	65–69	709,071 (12.79%)	3,651 (21.62%)	2,163 (14.93%)	1,659 (19.79%)	4,970 (35.45%)	8,356 (19.79%)
	≥70	192,490 (3.47%)	1,602 (9.49%)	181 (1.25%)	526 (6.27%)	1,188 (8.47%)	12,604 (29.85%)

診斷年齡	50–54	960,902 (17.33)	3,552 (21.04)	566 (3.91)	1,067 (12.73)	1,713 (12.22)	5,732 (13.58)
	55–59	1,258,587 (22.70)	3,970 (23.51)	2,890 (19.95)	1,827 (21.79)	2,606 (18.59)	7,689 (18.21)
	60–64	1,193,417 (21.52)	4,110 (24.34)	4,153 (28.67)	2,097 (25.01)	2,634 (18.79)	7,842 (18.57)
	65–69	1,269,944 (22.90)	3,651 (21.62)	4,527 (31.26)	2,173 (25.92)	5,158 (36.79)	8,356 (19.79)
	≥70	862,061 (15.55)	1,602 (9.49)	2,347 (16.21)	1,219 (14.54)	1,910 (13.62)	12,604 (29.85)
FIT	陰性	4,951,919 (89.31%)	2 (0.01)	12 (0.08%)	6,218 (74.17%)	9,940 (70.89%)	—
檢驗結果	陽性	583,979 (10.53%)	16,877 (99.95%)	14,470 (99.91%)	2,111 (25.18%)	3,855 (27.49%)	—
	缺失	9,013 (0.16%)	6 (0.04%)	1 (0.01%)	54 (0.64%)	226 (1.61%)	—

註：FIT 結果缺失主要來自原始篩檢該資料缺漏；部分個案仍具有篩檢日期及後續診斷資訊，因此可完成確診結果分類。

表 4.2 呈現不同確診結果分類之受檢者基本特徵。整體而言，除未確診個案以女性占多數（54.41%）外，其餘各類篩檢結果均以男性比例較高，其中首次篩檢偵測癌（63.23%）及拒絕者（61.65%）之男性比例最高，後續篩檢偵測癌（54.99%）、間隔癌（54.80%）及延遲診斷癌（55.47%）亦皆以男性居多，顯示男性於大腸直腸癌篩檢相關異常結果中所占比例普遍較高。

就初次研究年齡而言，未確診個案主要集中於 50–54 歲族群（41.20%），隨確診結果不同，年齡分布逐漸往較高年齡層移動。首次篩檢偵測癌主要分布於 55–64 歲，其中 60–64 歲占 24.34%；後續篩檢偵測癌亦以 55–64 歲為主（57.66%）；延遲診斷癌則以 65–69 歲占比最高（35.45%）；拒絕者於 70 歲以上之比例最高（29.85%）。

診斷年齡方面，首次篩檢偵測癌主要集中於 55–64 歲，而後續篩檢偵測癌及延遲診斷癌則多分布於 60–69 歲，其中後續篩檢偵測癌於 65–69 歲占 31.26%，延遲診斷癌於 65–69 歲占 36.79%，顯示隨追蹤時間增加，疾病診斷年齡逐漸向較高年齡層分布。

FIT 檢驗結果方面，未確診個案主要為 FIT 陰性（89.31%）；首次篩檢偵測癌及後續篩檢偵測癌幾乎皆來自 FIT 陽性個案，分別占 99.95% 及 99.91%，符合 FIT 篩檢流程。間隔癌則主要發生於 FIT 陰性個案（74.17%），顯示部分癌症個案於篩檢時未被 FIT 檢出，而於後續追蹤期間確診。而拒絕者的定義為該患者從未參與（或拒絕參加）FIT 糞便潛血篩檢，因此在篩檢資料庫中必然沒有任何檢驗紀錄，因此 FIT 檢驗結果皆為無紀錄。

整體而言，不同確診結果於性別、初次研究年齡、診斷年齡及 FIT 結果之分布均存在差異，顯示受檢者之人口學特徵及確診結果與疾病發展具有一定關聯性，並為後續群體層級疾病自然史模型之建立提供基礎。惟本研究資料係依疾病



模式進行抽樣，各確診結果類別之比例僅反映研究樣本之組成，並不代表一般篩檢族群之疾病盛行率，因此相關比例之解讀應審慎。

4.2 生成矩陣參數估計結果

本研究利用連續時間多狀態疾病模型 (Continuous-time multistate disease framework) 建立九狀態隱藏狀態疾病模型 (9-state hidden-state disease model)，並採用最大概似估計法 (Maximum likelihood estimation, MLE) 估計生成矩陣 (generator matrix) Q 中之各疾病狀態轉移強度 (Transition intensities)。

表 4.3 生成矩陣轉移強度估計結果

疾病轉移路徑	轉移說明	估計轉移強度
q_{12}	Normal $\rightarrow$ Small Adenoma	0.001120
q_{18}	Normal $\rightarrow$ OD Small Adenoma	0.000528
q_{23}	Small Adenoma $\rightarrow$ Advanced Adenoma	0.589600
q_{34}	Advanced Adenoma $\rightarrow$ Early PCDP	0.588900
q_{45}	Early PCDP $\rightarrow$ Late PCDP	0.202200
q_{46}	Early PCDP $\rightarrow$ Early Clinical Phase	0.241100
q_{57}	Late PCDP $\rightarrow$ Late Clinical Phase	0.444700
q_{89}	OD Small Adenoma $\rightarrow$ OD Advanced Adenoma	0.015300

註：OD 為 Overdiagnosed (過度診斷) 之縮寫。

由估計結果可觀察到幾個值得關注的地方。首先，Normal 狀態進入腺瘤路徑之轉移強度分為兩條：正常腺瘤進展路徑 ($q_{12} = 0.001120$) 與過度診斷路徑

($q_{18} = 0.000528$)，顯示模型估計下仍存在部分個案可能進入過度診斷路徑，支持本研究於疾病自然史架構中納入過度診斷狀態之必要性。

其次，Small Adenoma 至 Advanced Adenoma ($q_{23} = 0.5896$) 與 Advanced Adenoma 至 Early PCDP ($q_{34} = 0.5889$) 之轉移強度相近，顯示腺瘤形成後仍具有持續向後續疾病階段進展之趨勢。Early PCDP 則存在兩條可能之疾病進展路徑，分別為進入 Late PCDP ($q_{45} = 0.2022$) 及直接進入 Early CP ($q_{46} = 0.2411$)。其中， q_{46} 略高於 q_{45} ，顯示部分處於早期臨床前可偵測癌症階段之個案，可能於尚未進展至晚期臨床前可偵測癌症 (Late PCDP) 前，即已發展至臨床可診斷階段 (Early CP)。

4.3 發射矩陣估計結果

發射矩陣描述潛在疾病狀態與實際篩檢觀測結果之間的對應關係，即為第 3.2 中的 η ，用以描述不同潛在疾病狀態於 FIT 篩檢下被成功偵測之機率(表 4.4)。過度診斷狀態 (Overdiagnosed Small Adenoma、Overdiagnosed Advanced Adenoma) 之篩檢敏感度與對應之非過度診斷狀態相同，反映兩者在形態學上難以區分，FIT 篩檢之偵測能力亦假設一致。對於 Normal 狀態，由於不存在可被篩檢偵測之病灶，觀測映射矩陣中不設定篩檢陽性之機率。臨床期癌症 (Early CP、Late CP) 通常已出現明顯病理病灶或臨床症狀，因此假設可被完全辨識，其觀測機率設定為 1。

表 4.4 發射矩陣之 FIT 敏感度估計結果

潛在疾病狀態	篩檢敏感度
Small Adenoma / Overdiagnosed Small Adenoma	0.44
Advanced Adenoma / Overdiagnosed Advanced Adenoma	0.52
Early PCDP	0.61
Late PCDP	0.69

敏感度數值隨疾病進展階段遞增，反映病灶越大、出血量越多時，FIT 篩檢之偵測能力越高，具有一定之生物學合理性。相較之下，早期腺瘤階段之篩檢敏感度相對較低，因此部分病灶即使已存在，仍可能無法於篩檢時被成功偵測，進而於後續追蹤期間被觀察為間隔癌，此現象亦與 4.1 節所觀察之 FIT 陰性個案間隔癌分布結果相互呼應。

4.4 潛在疾病狀態分布模擬結果

基於 4.2 節所估計之生成矩陣與 4.3 節 FIT 篩檢敏感度，本研究利用矩陣指數方法推算兩年追蹤期間下之模擬潛在疾病分布(表 4.5)。模擬過程說明如下：

首先，設定初始潛在疾病狀態分布向量 $\mu(0)$ 。由於本研究模擬對象為初次進入篩檢之族群，假設其於追蹤起始時點均處於 Normal 狀態，故

$$\mu(0) = (1,0,0,0,0,0,0,0,0)$$

對應第 3.1 節所定義之九個潛在疾病狀態順序 (Normal、Small Adenoma、Advanced Adenoma、Early PCDP、Late PCDP、Early CP、Late CP、

Overdiagnosed Small Adenoma、Overdiagnosed Advanced Adenoma)。

其次，依表 4.3 所估計之轉移強度建立生成矩陣 Q 。其中，非零非對角元素包括 q_{12} 、 q_{18} 、 q_{23} 、 q_{34} 、 q_{45} 、 q_{46} 、 q_{57} 及 q_{89} ，其餘未設定之轉移路徑其轉移強度均設為 0。對角元素則根據生成矩陣各列元素總和為 0 之性質計算，即

$$q_{ii} = - \sum_{j \neq i} q_{ij}$$

本研究所使用之轉移強度參數，均採用第 4.2 節以最大概似估計法所得到之估計值。此外，根據第 3.3 節之定理 3.3 之全域可識別性結果，在假設一至假設四成立下，模型參數與觀測資料之聯合分布具有一對一對應關係。因此，第 4.2 節所估計之轉移強度參數可視為由觀測資料唯一決定之估計結果，而非多組等價參數解之一。據此建立之生成矩陣 Q 具有明確之統計意義，後續模擬所得之疾病進展路徑與狀態分布，亦可視為本研究模型設定下之推估結果。

接著，設定追蹤時間間隔 $\Delta = 2$ (年)。此設定之依據有二：其一，臺灣大腸直腸癌篩檢政策原則上以兩年為一次篩檢週期，受檢者於完成一次篩檢後，下次篩檢通常於約兩年後進行，故以兩年作為追蹤時間單位，符合實際篩檢制度之運作方式；其二，此設定與第 3.3.1 節假設二所述之觀測時間間隔一致，使本節模擬結果可與後續觀測分布比較 (4.5 節) 建立於相同之時間尺度上，避免追蹤時間定義不一致所造成之比較偏誤。需說明的是，實際受檢者之回診間隔受個人就醫行為影響，未必恰好為兩年，本研究於此係以政策建議之篩檢週期作為追蹤時間之近似設定，而非依據個案實際曆年間隔進行模擬。

依第 3.2 節所推導之連續時間馬可夫鏈轉移機率矩陣解析解

$$P(\Delta) = \exp(Q\Delta)$$

利用矩陣指數運算求得對應之轉移機率矩陣 $P(2)$ 。本研究透過泰勒展開 (Taylor series expansion) 逐項疊加 $(Q\Delta)^k/k!$ 以數值方式逼近矩陣指數，並設定收斂門檻 (新增項之最大絕對值小於 10^{-12}) 作為疊代停止條件。最後，將初始分布向量與轉移機率矩陣相乘，即可推算兩年後之潛在疾病狀態分布：

$$\mu(2) = \mu(0) \cdot P(2)$$

上式所得之 $\mu(2)$ 即為表 4.5 所呈現之模擬潛在疾病分布結果，反映在僅依據疾病自然史轉移強度之假設下，個體於兩年追蹤期間後分布於各潛在疾病狀態之比例。

表 4.5 模擬潛在疾病分布

潛在疾病狀態	模擬結果比例(%)
Normal	99.6709
Small Adenoma	0.1313
Advanced Adenoma	0.0626
Early PCDP	0.0236
Late PCDP	0.0023
Early Clinical Phase	0.0033
Late Clinical Phase	0.0005
Overdiagnosed Small Adenoma	0.1038
Overdiagnosed Advanced Adenoma	0.0016

結果顯示，在兩年追蹤期間內，大多數個體仍維持於 Normal 狀態 (99.6709%)，主要係因 Normal 至 Small Adenoma 之轉移強度 ($q_{12} = 0.001120$) 相對較低，且模擬期間有限所致。隨疾病進展，Small Adenoma (0.1313%)、Advanced Adenoma (0.0626%)、Early PCDP (0.0236%) 及 Late PCDP (0.0023%) 之比例逐步下降，呈現與大腸直腸癌腺瘤—癌症序列一致之遞進式分布型態。

此外，過度診斷路徑之病灶主要集中於 Overdiagnosed Small Adenoma (0.1038%)，而 Overdiagnosed Advanced Adenoma 僅占 0.0016%，顯示模型設定下之過度診斷病灶多停留於早期階段，與本研究所假設之低進展特性一致。

4.5 機器學習模型分析

前述 4.1 至 4.4 節之分析以群體層級疾病自然史模型為核心，透過生成矩陣估計，探討大腸直腸癌自然進程之動態特性。然而，群體層級模型主要描述整體族群之平均進展趨勢，仍無法直接回答個體層級之風險異質性問題。因此，為補充自然史模型在個體層級解釋能力之不足，本研究進一步採用彰化地區社區為基礎之篩檢世代資料，結合受檢者之人口學特徵、生活型態變數、FIT 篩檢結果及健康檢查生化指標，建立機器學習模型，以探討個體臨床與生化特徵與疾病狀態之關聯性。然而，考量各疾病類別樣本數差異及後續模型建立需求，本研究採用依疾病模式抽樣 (disease mode sampling) 建立分析資料集，事先設定各疾病類別之納入樣本數，再由各疾病類別中抽取相應受試者。最終納入 6,600 名受試者，包括正常 (Normal, $n = 6,000$)、腺瘤 (Adenoma, $n = 300$)、篩檢偵測癌 (Screen-detected CRC, $n = 100$) 及間隔癌 (Interval CRC, $n = 200$) 四類。

4.5.1 彰化社區篩檢資料之人口學與個人化特徵描述

Normal 組為篩檢期間未發現腺瘤或大腸直腸癌之個體；Adenoma 組為經篩檢或後續診斷確認具有腺瘤病變之個體；Screen-detected CRC 組為經篩檢流程偵

測並確診之大腸直腸癌個案；Interval CRC 組則為篩檢陰性後至下一次篩檢期間發生並確診之大腸直腸癌個案。表 4.6 呈現各疾病類別受試者之人口學特徵、生活型態及個人化風險因子分布情形。

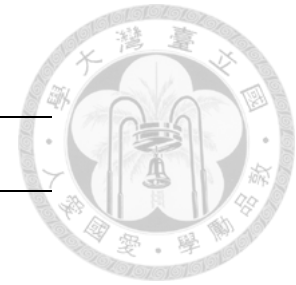
依據上述分類，本研究建構兩項二元分類任務。Classifier 1 為 Normal versus Adenoma，用以探討疾病由正常狀態進展至腺瘤階段之相關特徵；Classifier 2 為 Adenoma versus Screen-detected CRC，用以探討疾病由腺瘤進一步進展至可經篩檢偵測之大腸直腸癌階段之相關特徵。本研究利用機器學習方法辨識不同疾病進展階段之重要風險因子，藉以從個體層級角度探討大腸直腸癌疾病自然史之進展機制。

此外，確診年齡 (Diagnosed age) 於不同疾病狀態間之定義並不一致。對於正常個案，其代表最後一次篩檢年齡；對於疾病個案，則代表病灶確診年齡。為避免模型受到變數定義差異之影響而產生偏差，本研究於第一項分類任務中不納入該變數進行分析。另一方面，誘導時間 (Induction time) 作為疾病自然史相關之時間特徵，反映個體自進入篩檢系統至觀察終點之追蹤期間，因此納入後續模型建構。

本研究利用機器學習方法探討不同疾病進展階段之個人化風險特徵，藉由辨識與疾病狀態相關之重要因子，以個體層級角度理解大腸直腸癌疾病自然史之進展機制，並作為前述群體層級疾病自然史模型分析結果之補充。

表 4.6 彰化糞便潛血描述性分析

彰化糞便潛血人數 (n=6,600)	人數 (%)
篩檢結果分類	
正常	6,000 (90.9%)
腺瘤	300 (4.5%)
篩檢偵測癌	100 (1.5%)
間隔癌	200 (3.0%)
性別	
女性	1,726 (26.2%)
男性	4,874 (73.8%)
年齡分類	
50-54	2,057 (31.2%)
55-59	1,791 (27.1%)
60-64	1,370 (20.8%)
65-69	1,363 (20.7%)
70-74	14 (0.2%)
75-79	1 (0.0%)
80+	4 (0.1%)
教育程度	
低	4,522 (68.5%)
中	1,636 (24.8%)
高	442 (6.7%)
大腸癌家族病史	
無	6,333 (96.0%)



有	267 (4.0%)
抽菸狀態	
從不	5,677 (86.0%)
已戒	324 (4.9%)
仍在抽菸中	599 (9.1%)
喝酒狀態	
從不	6,119 (92.7%)
已戒	106 (1.6%)
仍在喝酒中	375 (5.7%)
嚼檳榔	
從不	6,106 (92.5%)
已戒	302 (4.6%)
仍在嚼檳榔中	192 (2.9%)
運動習慣	
無	2,705 (41.0%)
有	3,895 (59.0%)
攝取蔬菜程度	
低	1,071 (16.2%)
中	2,656 (40.2%)
高	2,873 (43.5%)
肉類攝取程度	
低	2,508 (38.0%)
中	3,152 (47.8%)
高	940 (14.2%)





水果攝取程度

低	1,227 (18.6%)
中	2,536 (38.4%)
高	2,837 (43.0%)

咖啡攝取程度

低	5,181 (78.5%)
中	723 (11.0%)
高	696 (10.5%)

FIT 濃度分類 (ng/ml)

0	3,506 (53.1%)
1-9	2,446 (37.1%)
10-19	224 (3.4%)
20-29	102 (1.5%)
30-39	96 (1.5%)
40-49	67 (1.0%)
50+	159 (2.4%)

Mean ± SD

第一次篩檢年齡 (年)	58.52 ± 5.81
FBT (糞便潛血檢查濃度 ng/ml)	59.69 ± 539.34
誘導時間 (年)	12.47 ± 2.55
診斷年齡 (歲)	70.99 ± 6.23
腰圍 (公分)	81.21 ± 9.48
收縮壓 (mmHg)	132.31 ± 19.92
舒張壓 (mmHg)	84.03 ± 12.24

空腹血糖 (mg/dL)	103.97 ± 32.81
三酸甘油脂 (mg/dL)	126.52 ± 111.98
高密度脂蛋白膽固醇 (mg/dL)	53.66 ± 15.30
GOT/AST (U/L)	26.31 ± 23.55
GPT/ALT (U/L)	26.79 ± 24.59
尿酸 (mg/dL)	5.33 ± 1.32
血尿素氮 (mg/dL)	14.05 ± 4.35
肌酸酐 (mg/dL)	0.82 ± 0.35



表 4.6 呈現彰化社區篩檢樣本之整體人口學特徵、生活型態、FIT 濃度及生化指標分布。整體樣本以男性為主，占 73.8%；年齡主要集中於 50 至 69 歲，其中以 50–54 歲族群最多，占 31.2%，其次為 55–59 歲族群，占 27.1%。疾病狀態分布方面，正常個案占 90.9%，腺瘤個案占 4.5%，篩檢偵測癌占 1.5%，間隔癌占 3.0%。FIT 濃度為 0 (ng/ml) 者最多，占 53.1%，而 FIT 濃度 ≥ 50 (ng/ml) 者占 2.4%。連續型變數方面，第一次篩檢年齡平均為 58.52 ± 5.81 歲，誘導時間平均為 12.47 ± 2.55 年。

由於本研究樣本係依疾病模式抽樣建立，正常組受試者占多數，因此整體生活型態分布主要受到正常組影響，相關比例僅反映研究樣本特性，不代表一般社區篩檢族群之實際生活型態分布。後續分析將透過機器學習模型進一步評估各生活型態變數對疾病風險之影響，而非僅依據整體比例進行推論。

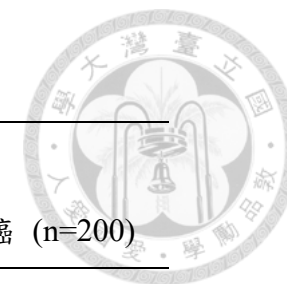
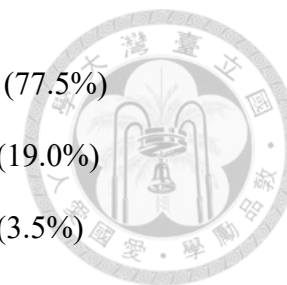


表 4.7 彰化社區篩檢資料不同疾病組別之基本特徵分布

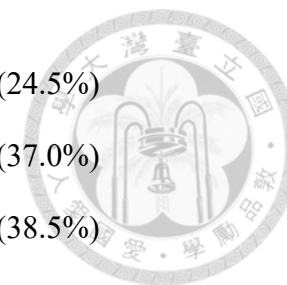
特徵變數 (n=6,600)	人數 N (%)			
	正常(n=6,000)	腺瘤(n=300)	篩檢偵測癌(n=100)	間隔癌 (n=200)
性別				
女性	1,507 (25.1%)	116 (38.7%)	34 (34.0%)	69 (34.5%)
男性	4,493 (74.9%)	184 (61.3%)	66 (66.0%)	131 (65.5%)
年齡分類				
50-54	1,921 (32.0%)	69 (23.0%)	38 (38.0%)	29 (14.5%)
55-59	1,648 (27.5%)	79 (26.3%)	28 (28.0%)	36 (18.0%)
60-64	1,242 (20.7%)	73 (24.3%)	16 (16.0%)	39 (19.5%)
65-69	1,182 (19.7%)	77 (25.7%)	17 (17.0%)	87 (43.5%)
70-74	5 (0.1%)	2 (0.7%)	1 (1.0%)	6 (3.0%)
75-79	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (0.5%)
80+	2 (0.0%)	0 (0.0%)	0 (0.0%)	2 (1.0%)
教育程度				



低	4,088 (68.1%)	209 (69.7%)	70 (70.0%)	155 (77.5%)
中	1,506 (25.1%)	71 (23.7%)	21 (21.0%)	38 (19.0%)
高	406 (6.8%)	20 (6.7%)	9 (9.0%)	7 (3.5%)
大腸癌家族病史				
無	5,766 (96.1%)	285 (95.0%)	91 (91.0%)	191 (95.5%)
有	234 (3.9%)	15 (5.0%)	9 (9.0%)	9 (4.5%)
抽菸狀態				
從不	5,197 (86.6%)	236 (78.7%)	79 (79.0%)	165 (82.5%)
已戒	291 (4.9%)	15 (5.0%)	6 (6.0%)	12 (6.0%)
仍在抽菸中	512 (8.5%)	49 (16.3%)	15 (15.0%)	23 (11.5%)
喝酒狀態				
從不	5,598 (93.3%)	256 (85.3%)	89 (89.0%)	176 (88.0%)
已戒	95 (1.6%)	7 (2.3%)	0 (0.0%)	4 (2.0%)
仍在喝酒中	307 (5.1%)	37 (12.3%)	11 (11.0%)	20 (10.0%)
嚼檳榔				



從不	5,578 (93.0%)	259 (86.3%)	89 (89.0%)	180 (90.0%)
已戒	261 (4.3%)	24 (8.0%)	7 (7.0%)	10 (5.0%)
仍在嚼檳榔中	161 (2.7%)	17 (5.7%)	4 (4.0%)	10 (5.0%)
運動習慣				
無	2,469 (41.1%)	114 (38.0%)	39 (39.0%)	83 (41.5%)
有	3,531 (58.9%)	186 (62.0%)	61 (61.0%)	117 (58.5%)
攝取蔬菜程度				
低	965 (16.1%)	48 (16.0%)	21 (21.0%)	37 (18.5%)
中	2,406 (40.1%)	120 (40.0%)	38 (38.0%)	92 (46.0%)
高	2,629 (43.8%)	132 (44.0%)	41 (41.0%)	71 (35.5%)
肉類攝取程度				
低	2,289 (38.1%)	96 (32.0%)	37 (37.0%)	86 (43.0%)
中	2,864 (47.7%)	153 (51.0%)	46 (46.0%)	89 (44.5%)
高	847 (14.1%)	51 (17.0%)	17 (17.0%)	25 (12.5%)
水果攝取程度				



低	1,096 (18.3%)	63 (21.0%)	19 (19.0%)	49 (24.5%)
中	2,311 (38.5%)	113 (37.7%)	38 (38.0%)	74 (37.0%)
高	2,593 (43.2%)	124 (41.3%)	43 (43.0%)	77 (38.5%)
咖啡攝取程度				
低	4,689 (78.1%)	239 (79.7%)	78 (78.0%)	175 (87.5%)
中	661 (11.0%)	33 (11.0%)	11 (11.0%)	18 (9.0%)
高	650 (10.8%)	28 (9.3%)	11 (11.0%)	7 (3.5%)
FIT 濃度分類 (ng/ml)				
0	3,313 (55.2%)	103 (34.3%)	30 (30.0%)	60 (30.0%)
1-9	2,224 (37.1%)	105 (35.0%)	25 (25.0%)	92 (46.0%)
10-19	186 (3.1%)	17 (5.7%)	5 (5.0%)	16 (8.0%)
20-29	79 (1.3%)	16 (5.3%)	4 (4.0%)	3 (1.5%)
30-39	67 (1.1%)	19 (6.3%)	5 (5.0%)	5 (2.5%)
40-49	46 (0.8%)	11 (3.7%)	6 (6.0%)	4 (2.0%)
50+	85 (1.4%)	29 (9.7%)	25 (25.0%)	20 (10.0%)



	平均值 ± 標準差			
初始篩檢年齡 (年)	58.36 ± 5.74	59.56 ± 5.78	57.59 ± 5.66	62.45 ± 6.39
FBT (糞便潛血檢查濃度 ng/ml)	44.73 ± 535.12	170.85 ± 502.75	380.14 ± 691.81	181.47 ± 551.14
誘導時間(年)	12.96 ± 1.62	7.53 ± 4.49	5.32 ± 4.38	8.71 ± 3.59
診斷年齡(歲)	71.32 ± 6.07	67.10 ± 5.17	62.91 ± 6.41	71.16 ± 7.47
腰圍(公分)	80.90 ± 9.30	85.01 ± 11.33	84.37 ± 9.95	83.45 ± 9.92
收縮壓(mmHg)	131.89 ± 19.74	137.19 ± 21.44	134.67 ± 20.59	136.66 ± 21.28
舒張壓(mmHg)	83.81 ± 12.17	87.45 ± 13.21	85.56 ± 11.60	84.84 ± 12.28
空腹血糖(mg/dL)	103.62 ± 32.12	106.77 ± 38.24	111.37 ± 45.50	106.72 ± 36.29
三酸甘油脂(mg/dL)	124.76 ± 107.57	151.12 ± 138.09	138.94 ± 117.06	136.13 ± 174.17
高密度脂蛋白膽固醇(mg/dL)	53.74 ± 15.29	51.99 ± 15.06	53.78 ± 15.11	53.72 ± 15.93
GOT/AST (U/L)	26.04 ± 15.39	26.36 ± 11.32	42.02 ± 147.20	26.41 ± 12.57
GPT/ALT (U/L)	26.49 ± 19.95	27.87 ± 16.91	41.15 ± 120.11	26.91 ± 18.49
尿酸 (mg/dL)	5.31 ± 1.31	5.64 ± 1.41	5.56 ± 1.59	5.28 ± 1.27
血尿素氮 (mg/dL)	13.98 ± 4.24	14.65 ± 5.23	14.67 ± 5.09	15.17 ± 5.41

肌酸酐 (mg/dL)

0.81 ± 0.33

0.90 ± 0.46

0.92 ± 0.59

0.88 ± 0.48

表 4.7 進一步比較不同疾病狀態下各項特徵之分布。誘導時間方面，正常組平均為 12.96 ± 1.62 年，腺瘤組為 7.53 ± 4.49 年，篩檢偵測癌組為 5.32 ± 4.38 年，間隔癌組為 8.71 ± 3.59 年，顯示誘導時間在不同疾病狀態間具有明顯差異，且疾病狀態較嚴重者整體誘導時間較短。此一結果亦為後續機器學習模型中誘導時間 (年) 成為重要特徵提供描述性基礎。

在 FIT 濃度分布方面，正常組 FIT 濃度為 0 (ng/ml) 者占 55.2%，而腺瘤組、篩檢偵測癌組及間隔癌組分別為 34.3%、30.0% 及 30.0%。相對地，FIT 濃度 ≥ 50 (ng/ml) 者於篩檢偵測癌組占 25.0%，明顯高於正常組之 1.4% 及腺瘤組之 9.7%，顯示 FIT 濃度與疾病嚴重程度具有關聯性。

代謝相關指標方面，腺瘤組之腰圍、收縮壓、舒張壓、三酸甘油脂及尿酸平均值皆高於正常組，顯示代謝異常可能與腺瘤發生風險相關。生活型態方面，腺瘤組中仍在抽菸者占 16.3%，高於正常組之 8.5%；仍在喝酒者占 12.3%，亦高於正常組之 5.1%，顯示抽菸與飲酒等生活型態因子可能與腺瘤風險具有關聯。

此外，間隔癌組之年齡分布明顯偏高，其中 65–69 歲者占 43.5%，高於正常組、腺瘤組及篩檢偵測癌組，顯示年長受檢者可能具有較高之間隔癌風險。整體而言，本研究樣本於誘導時間、FIT 濃度、代謝指標及生活型態變數上皆呈現不同疾病狀態之分布差異，提供後續機器學習模型建立及個體層級疾病風險分析之基礎。惟本研究樣本係依疾病模式進行抽樣，各疾病類別之比例僅反映研

究樣本組成，不代表實際社區篩檢族群之疾病分布。



4.5.2 正常與腺瘤分類器

本節建立 Classifier 1(Normal versus Adenoma) 之二元分類模型，探討正常個體與腺瘤個案之個人化風險特徵差異。表 4.8 呈現 ANN 模型之分類效能。

表 4.8 ANN 模型於正常狀態與腺瘤之分類效能

指標	訓練集	測試集
敏感度 (Sensitivity)	0.839	0.771
精確率 (Precision)	0.243	0.213
F1 分數 (F1 Score)	0.389	0.334
特異度 (Specificity)	0.873	0.868

表 4.9 隨機森林模型於正常狀態與腺瘤之分類效能

指標	訓練集	測試集
敏感度 (Sensitivity)	0.735	0.700
精確率 (Precision)	0.988	0.980
F1 分數 (F1 Score)	0.843	0.817
特異度 (Specificity)	1.000	0.999



ANN 模型於正常個體與腺瘤個體分類之結果如表 4.8 所示。於訓練集資料中，模型之敏感度、精確率、F1 分數及特異度分別為 0.839、0.243、0.389 及 0.873；於測試集資料中則分別為 0.771、0.213、0.334 及 0.868。

隨機森林模型於正常個體與腺瘤個體分類之結果如表 4.9 所示。於訓練集資料中，模型之敏感度、精確率、F1 分數及特異度分別為 0.735、0.988、0.843 及 1.000；於測試集資料中則分別為 0.700、0.980、0.817 及 0.999。

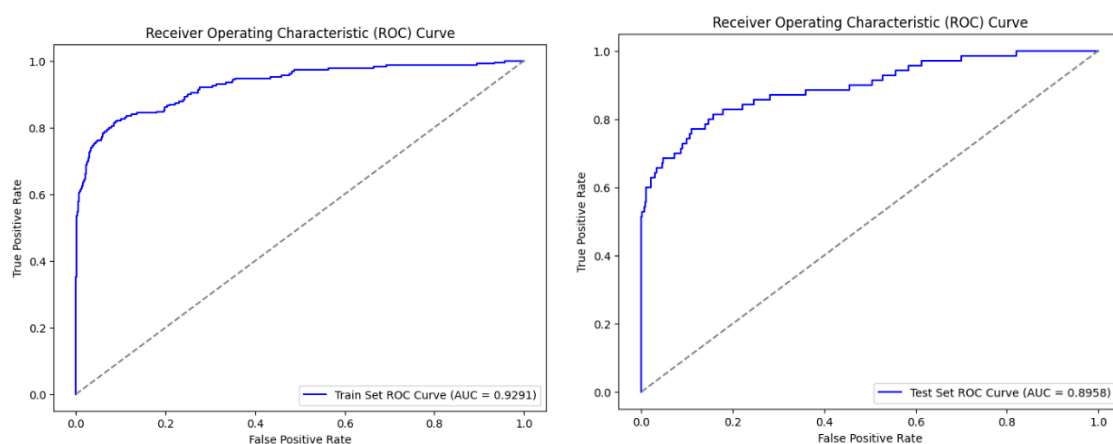


圖 4.1 ANN 模型於正常狀態與腺瘤分類之 ROC 曲線

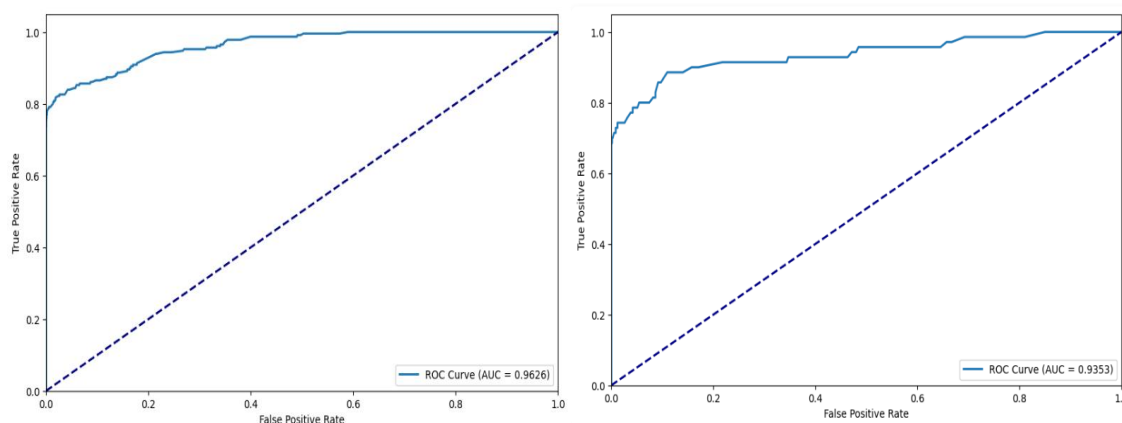


圖 4.2 隨機森林模型於正常狀態與腺瘤分類之 ROC 曲線

圖 4.1 顯示 ANN 模型於正常狀態與腺瘤分類之 ROC 曲線。訓練集與測試集之 AUC 分別約為 0.929 與 0.896，兩者結果相近，顯示模型具有一定程度之泛化

能力，未出現明顯過度擬合現象。

圖 4.2 顯示隨機森林模型於正常狀態與腺瘤分類之 ROC 曲線。訓練集與測試集之 AUC 分別約為 0.963 與 0.935，均高於 0.9，顯示模型具有優異之分類能力。相較於 ANN 模型之測試集 AUC 約為 0.896，隨機森林模型對正常個體與腺瘤個體具有更佳之區辨能力，且訓練集與測試集之 AUC 差異僅約為 0.028，顯示模型具有良好之泛化能力，未出現明顯過度擬合現象。

為了解 ANN 與隨機森林模型於正常個體與腺瘤個體分類過程中各變數對模型輸出之貢獻程度，本研究利用 SHAP (SHapley Additive exPlanations) 方法評估特徵重要性，並以 SHAP beeswarm plot 呈現各變數之影響方向與分布情形，結果如圖 4.3 與圖 4.4 所示。



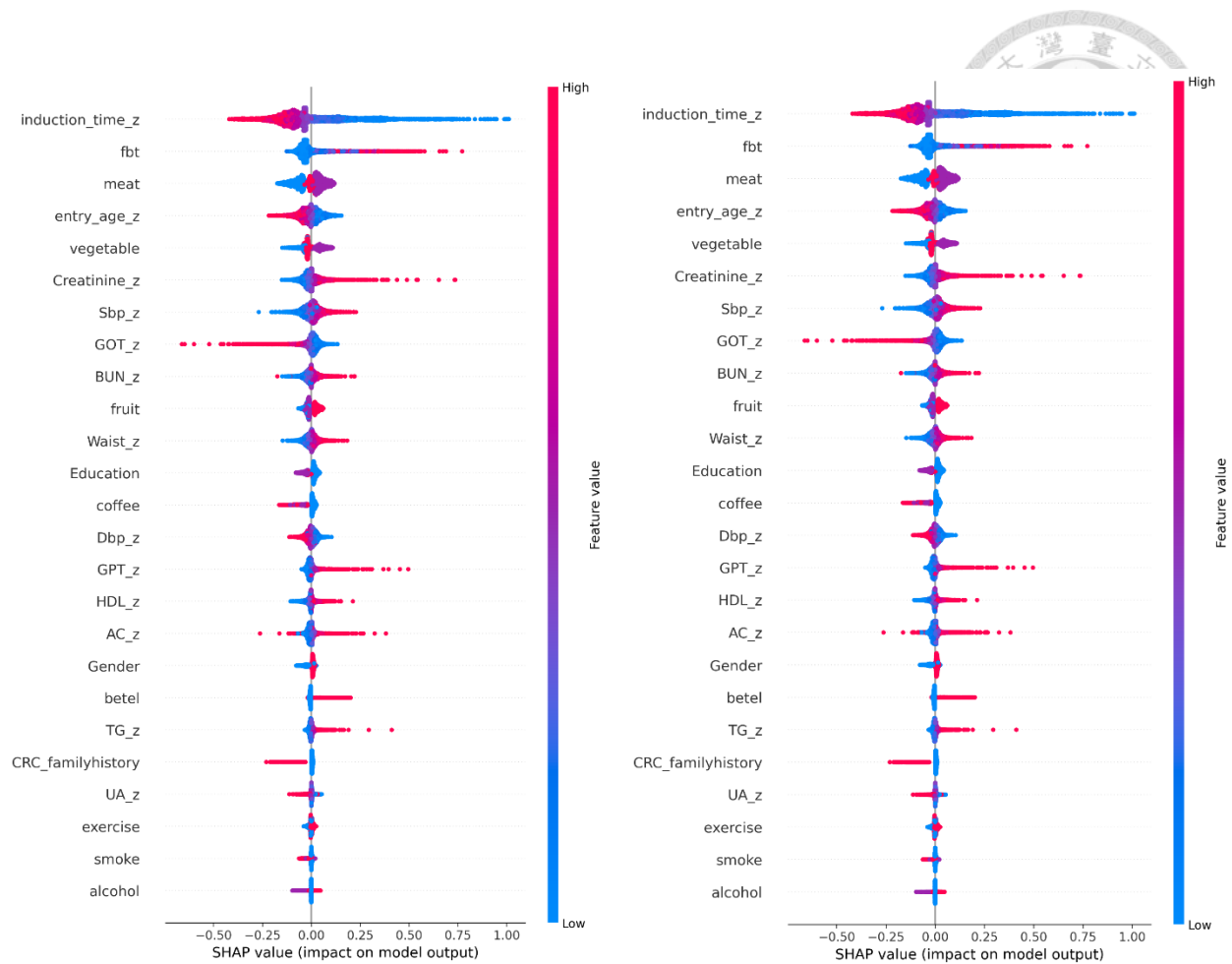


圖 4.3 ANN 模型訓練集與測試集之 SHAP Beeswarm Plot

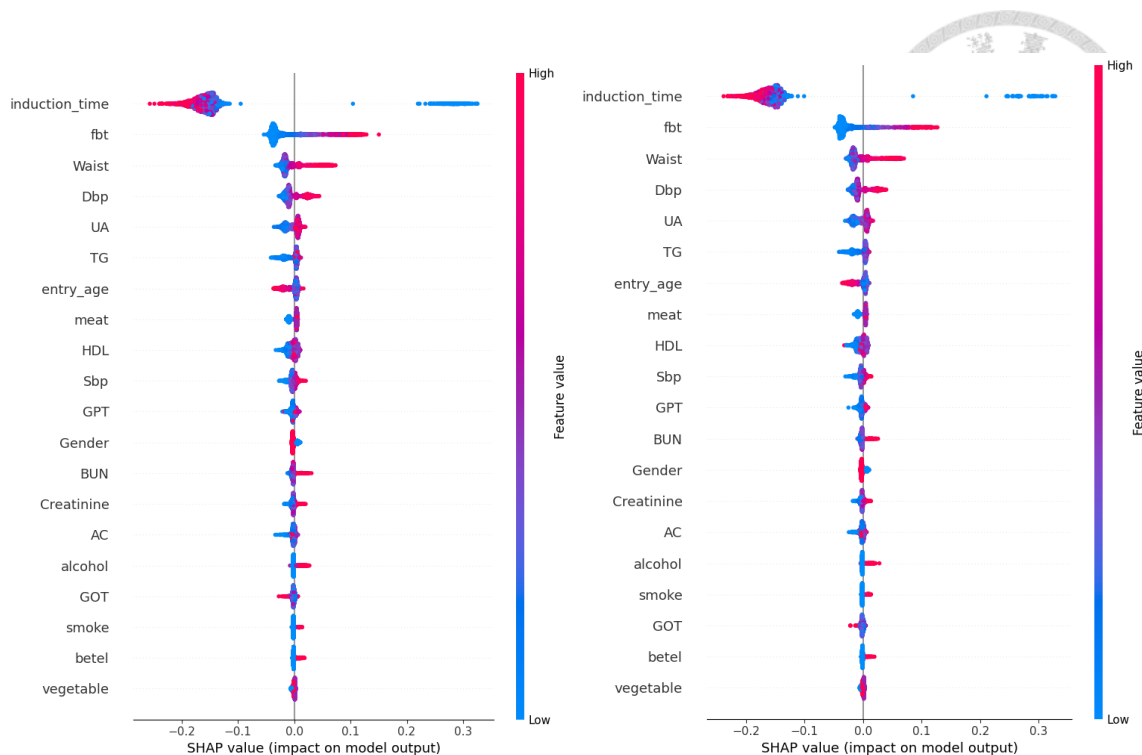


圖 4.4 隨機森林模型訓練集與測試集之 SHAP Beeswarm Plot

由圖 4.3 可知，ANN 模型在區分正常與腺瘤時，主要受到誘導時間（induction_time_z）、糞便血紅素濃度（fbt）、肉類攝取（meat）、進入研究年齡（entry age）及蔬菜攝取（vegetable）等變數影響。其中，誘導時間與糞便血紅素濃度之 SHAP 值分布範圍較廣，表示其對模型輸出具有較大影響。進一步觀察 SHAP 值方向可發現，較短之誘導時間與較高之糞便血紅素濃度會使模型輸出較傾向判定為腺瘤；相反地，較長之誘導時間與較低之糞便血紅素濃度則較傾向判定為正常。

而圖 4.4 顯示隨機森林區分正常與腺瘤時，主要受到誘導時間（Induction time）、糞便血紅素濃度（fbt）、腰圍（waist）、舒張壓（Dbp）及尿酸（UA）等變數影響。而低誘導時間、高糞便潛血濃度、低腰圍、高舒張壓等等較容易判定為腺瘤；相反地，較高之誘導時間與較低之糞便血紅素濃度等等則較傾向判定為正常。

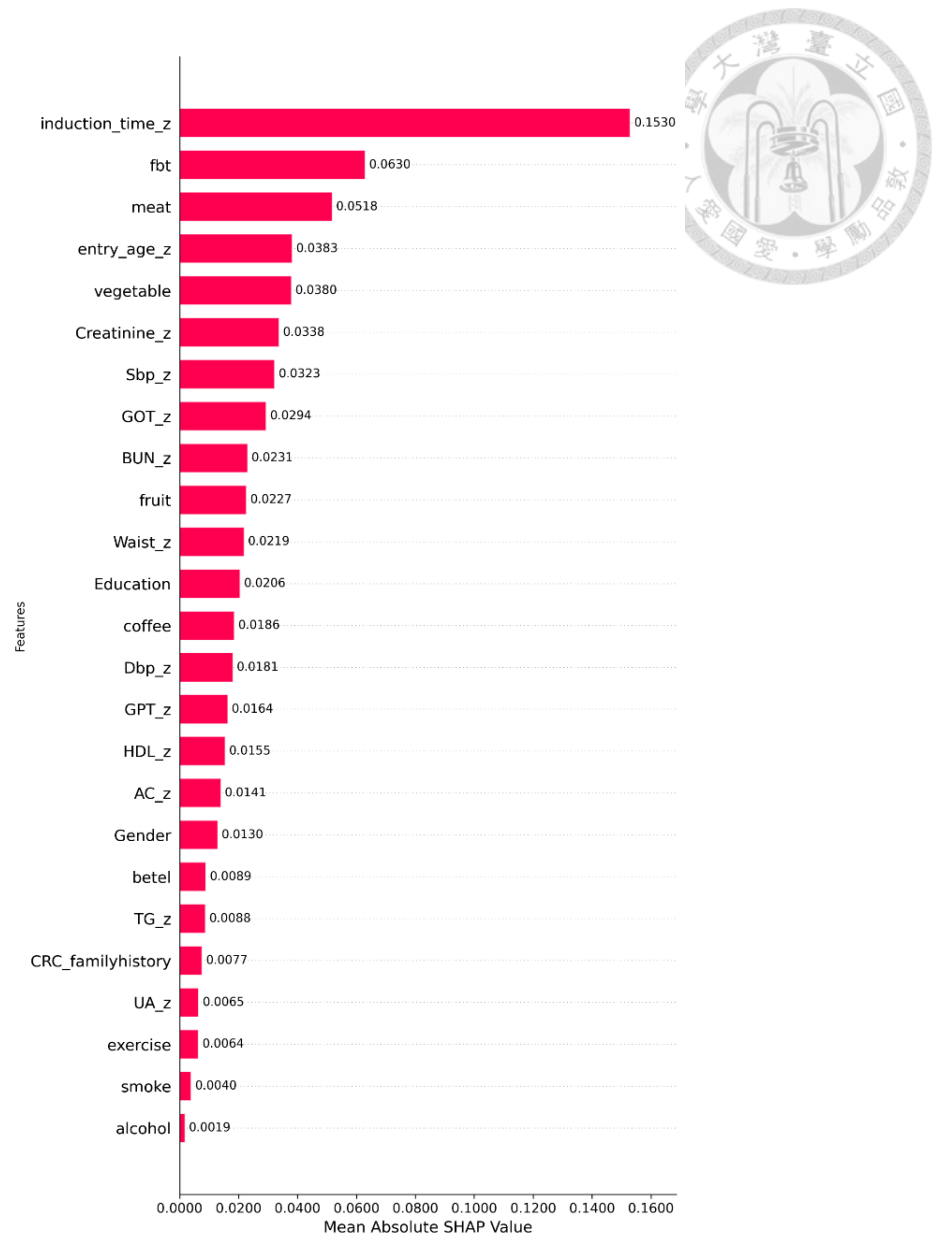


圖 4. 5ANN 模型 SHAP 平均絕對值重要性比較圖

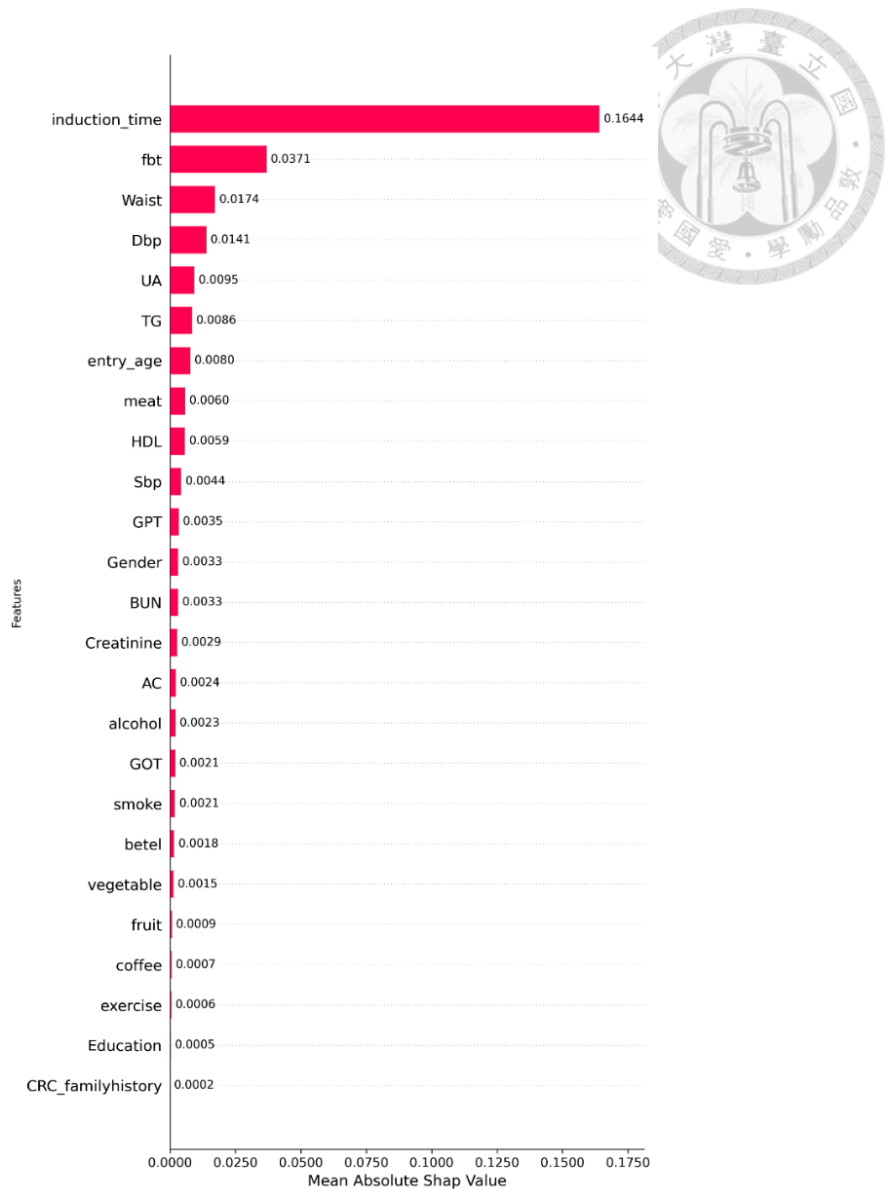


圖 4.6 隨機森林模型 SHAP 平均絕對值重要性比較圖

由圖 4.5 顯示，誘導時間（induction_time_z）為模型中最重要之特徵，其平均絕對 SHAP 值為 0.1530，明顯高於其他變數。其次依序為糞便血紅素濃度（fbt，0.0630）、肉類攝取（meat，0.0518）、進入研究年齡（entry_age_z，0.0383）及蔬菜攝取（vegetable，0.0380）。其餘如肌酸酐（Creatinine）、收縮壓（Sbp）、天門冬胺酸轉胺酶（Aspartate Aminotransferase, AST；GOT）、血尿素氮（Blood Urea Nitrogen, BUN）等變數亦對模型具有一定程度之貢獻，但重要性相對較低。整體而言，ANN 模型主要依賴誘導時間及糞便血紅素濃度等與疾病進展及篩檢結果相



關之資訊進行判別。

圖 4.6 為隨機森林模型之 SHAP 平均絕對值重要性比較圖。結果顯示，誘導時間 (induction_time) 仍為最重要之特徵，其平均絕對 SHAP 值為 0.1644，其次依序為糞便血紅素濃度 (fbt, 0.0371)、腰圍 (Waist, 0.0174)、舒張壓 (Dbp, 0.0141) 及尿酸 (UA, 0.0095)。此外，三酸甘油酯 (TG)、進入研究年齡 (entry_age) 及肉類攝取 (meat) 亦具有一定程度之貢獻。

整體而言，隨機森林模型與 ANN 模型皆將誘導時間及糞便血紅素濃度視為最重要之預測特徵，但隨機森林模型更重視代謝相關指標，如腰圍、血壓及尿酸等變數，而 ANN 較重視食物的攝取，如肉類攝取量 (Meat) 與蔬菜攝取量 (Vegetable)。表 4.10 整理 ANN 模型與隨機森林模型於正常個體與腺瘤個體分類之 SHAP 重要變數排名。

表 4.10 整理 ANN 與隨機森林模型之 SHAP 重要變數排名

排名	ANN 模型	隨機森林模型
1	誘導時間 (Induction Time)	誘導時間 (Induction Time)
2	糞便血紅素 (f-Hb)	糞便血紅素 (f-Hb)
3	肉類攝取量 (Meat)	腰圍 (Waist)
4	進入年齡 (Entry Age)	舒張壓 (DBP)
5	蔬菜攝取量 (Vegetable)	尿酸 (UA)
6	肌酸酐 (Creatinine)	三酸甘油酯 (TG)
7	收縮壓 (SBP)	進入年齡 (Entry Age)



4.5.3 腺瘤與篩檢癌分類器

相較於 Classifier 1，Classifier 2 (Adenoma versus Screen-detected CRC) 之分類問題具有較高之挑戰性。腺瘤與篩檢癌皆屬於疾病自然史進展路徑中之非正常疾病狀態，於可觀測之人口學特徵、臨床指標及生化特徵上具有較高程度之重疊性。兩者之差異主要反映疾病進展程度，而非截然不同之風險因子組合，因此理論上具有較高之分類困難度。表 4.11 與表 4.12 分別呈現 ANN 與隨機森林模型之分類效能。

表 4.11 ANN 模型於腺瘤與篩檢癌之分類效能

指標	訓練集	測試集
敏感度 (Sensitivity)	0.500	0.500
精確率 (Precision)	0.520	0.385
F1 分數 (F1 Score)	0.510	0.435
特異度 (Specificity)	0.832	0.800

表 4.12 隨機森林模型於腺瘤與篩檢癌之分類效能

指標	訓練集	測試集
敏感度 (Sensitivity)	0.712	0.650
精確率 (Precision)	0.538	0.433
F1 分數 (F1 Score)	0.613	0.520
特異度 (Specificity)	0.777	0.788

ANN 模型於腺瘤個體與篩檢發現大腸直腸癌個體分類之結果如表 4.11 所示。於訓練集資料中，模型之敏感度、精確率、F1 分數及特異度分別為 0.500、0.520、0.510 及 0.832；於測試集資料中則分別為 0.500、0.385、0.435 及 0.800。

而隨機森林模型於腺瘤個體與篩檢發現大腸直腸癌個體分類之結果如表 4.12 所示。於訓練集資料中，模型之敏感度、精確率、F1 分數及特異度分別為 0.712、0.538、0.613 及 0.777；於測試集資料中則分別為 0.65、0.433、0.52 及 0.788。

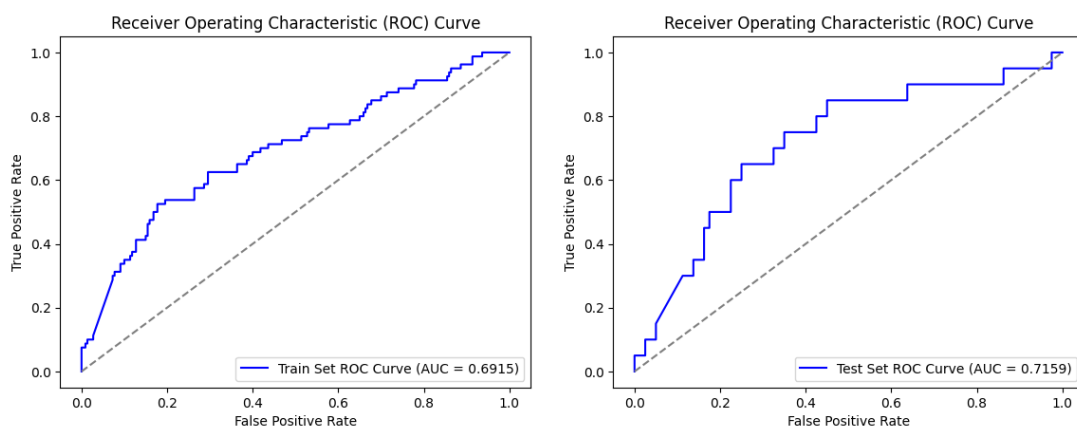


圖 4.7 ANN 模型於腺瘤與大腸直腸癌分類之 ROC 曲線

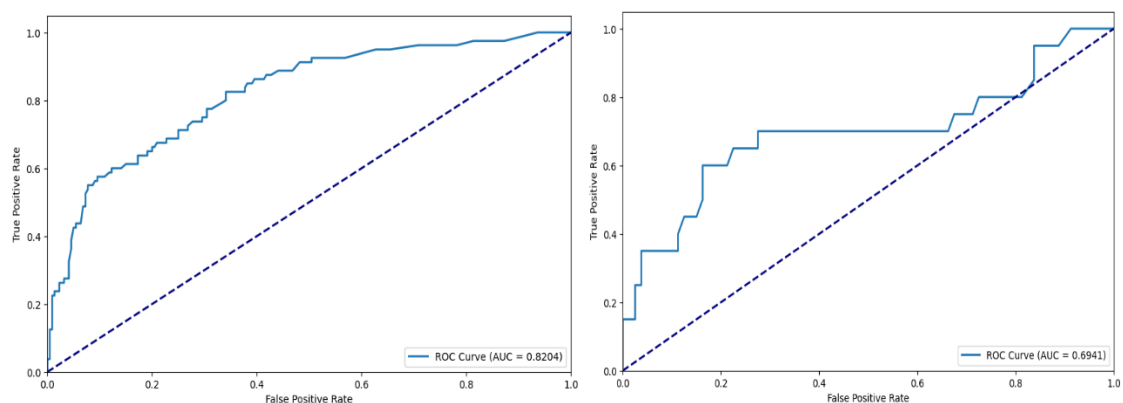


圖 4.8 隨機森林模型於腺瘤與大腸直腸癌分類之 ROC 曲線

圖 4.7 為 ANN 模型於腺瘤與大腸直腸癌分類之 ROC 曲線。訓練集與測試集之 AUC 分別約為 0.692 與 0.716，兩者結果相近，顯示模型於不同資料集間具有

一定程度之泛化能力，未出現明顯過度擬合。然而，整體 AUC 約為 0.7，表示 ANN 模型對腺瘤與大腸直腸癌之辨識能力屬中等，仍有提升空間。

圖 4.8 為隨機森林模型於腺瘤與大腸直腸癌分類之 ROC 曲線。訓練集與測試集之 AUC 分別約為 0.820 與 0.694。訓練集之辨識能力優於測試集，顯示模型於訓練資料具有較佳之分類效果，但於測試資料之辨識能力下降，可能存在一定程度之過度擬合。

綜合比較兩種模型之 ROC 曲線可發現，ANN 與隨機森林模型於測試集之 AUC 指標差異不大，顯示兩種模型於腺瘤與大腸直腸癌分類任務之整體辨識能力相近。由於兩者效能皆未達高度準確，顯示腺瘤與大腸直腸癌之分類較正常與腺瘤分類更具挑戰性，可能與兩類疾病之臨床特徵重疊及樣本數有限有關。為進一步探討模型於腺瘤與篩檢癌分類中之重要影響因子，本研究利用 SHAP 方法評估各變數之貢獻程度。

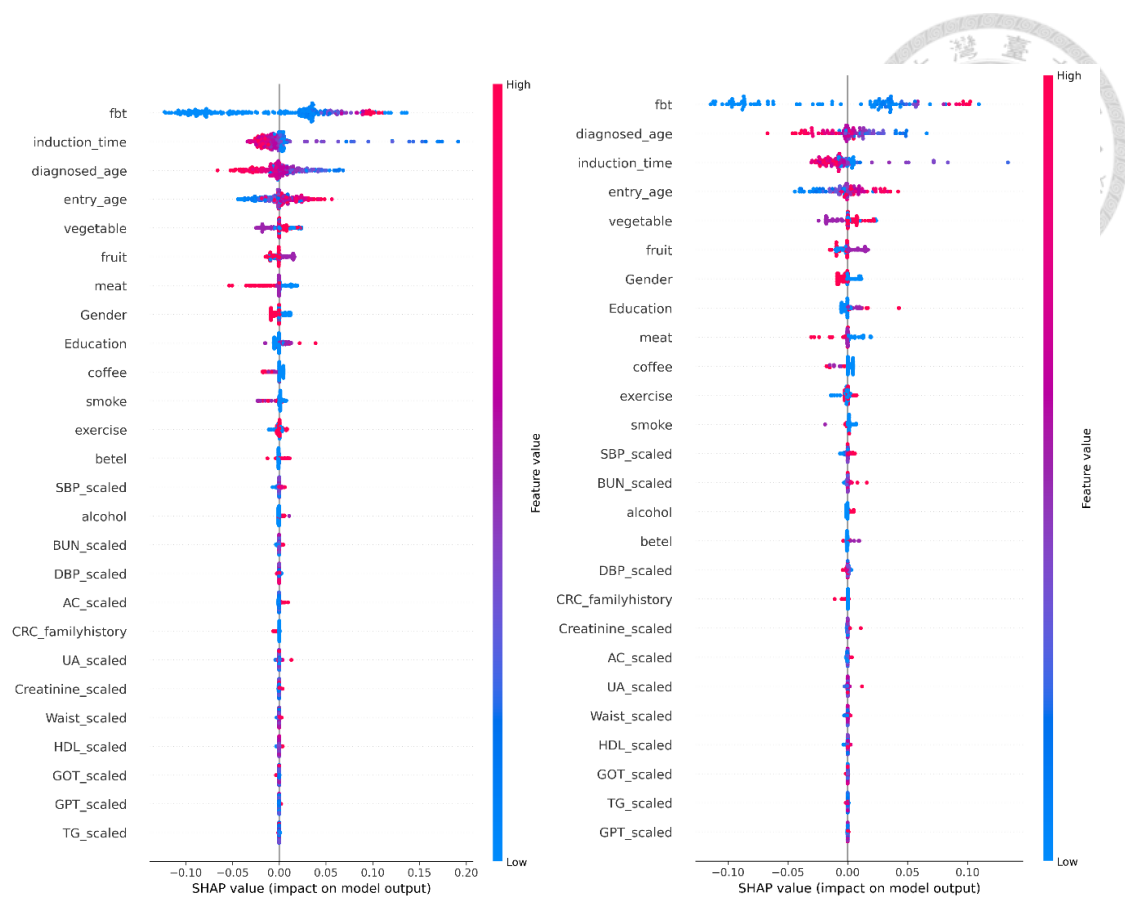


圖 4.9 ANN 模型訓練集與測試集之 SHAP Beeswarm Plot

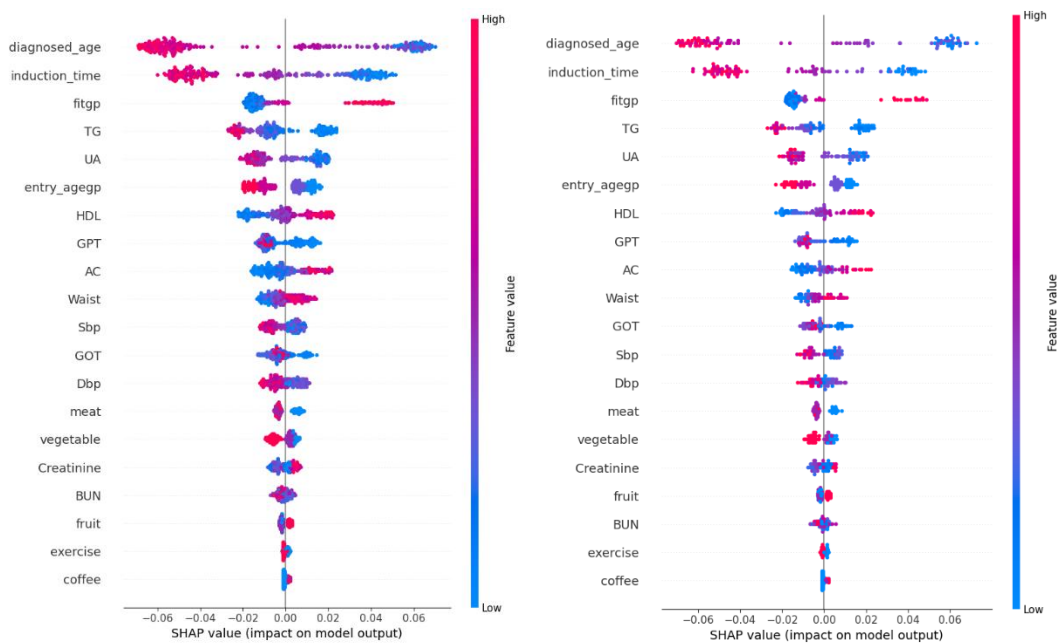


圖 4.10 隨機森林模型訓練集與測試集之 SHAP Beeswarm Plot



由圖 4.9 可知，ANN 模型在區分腺瘤與篩檢癌時，主要受到糞便血紅素濃度 (fbt)、誘導時間 (induction_time_z)、診斷年齡 (diagnosed_age) 等變數影響。其中，糞便血紅素濃度與誘導時間之 SHAP 值分布範圍較廣，表示其對模型輸出具有較大影響。進一步觀察 SHAP 值方向可發現，較短之誘導時間與較高之糞便血紅素濃度會使模型輸出較傾向判定為篩檢癌；相反地，較長之誘導時間與較低之糞便血紅素濃度則較傾向判定為腺瘤。

而圖 4.10 顯示隨機森林區分腺瘤與篩檢癌時，主要受到糞便血紅素濃度 (fbt)、誘導時間 (induction time)、診斷年齡 (diagnosed age)、進入年齡 (entry age) 等變數影響。而高糞便潛血濃度、低診斷年齡、低誘導時間等等較容易判定為篩檢癌；相反地，較高之誘導時間與較低之糞便血紅素濃度等等則較傾向判定為腺瘤。

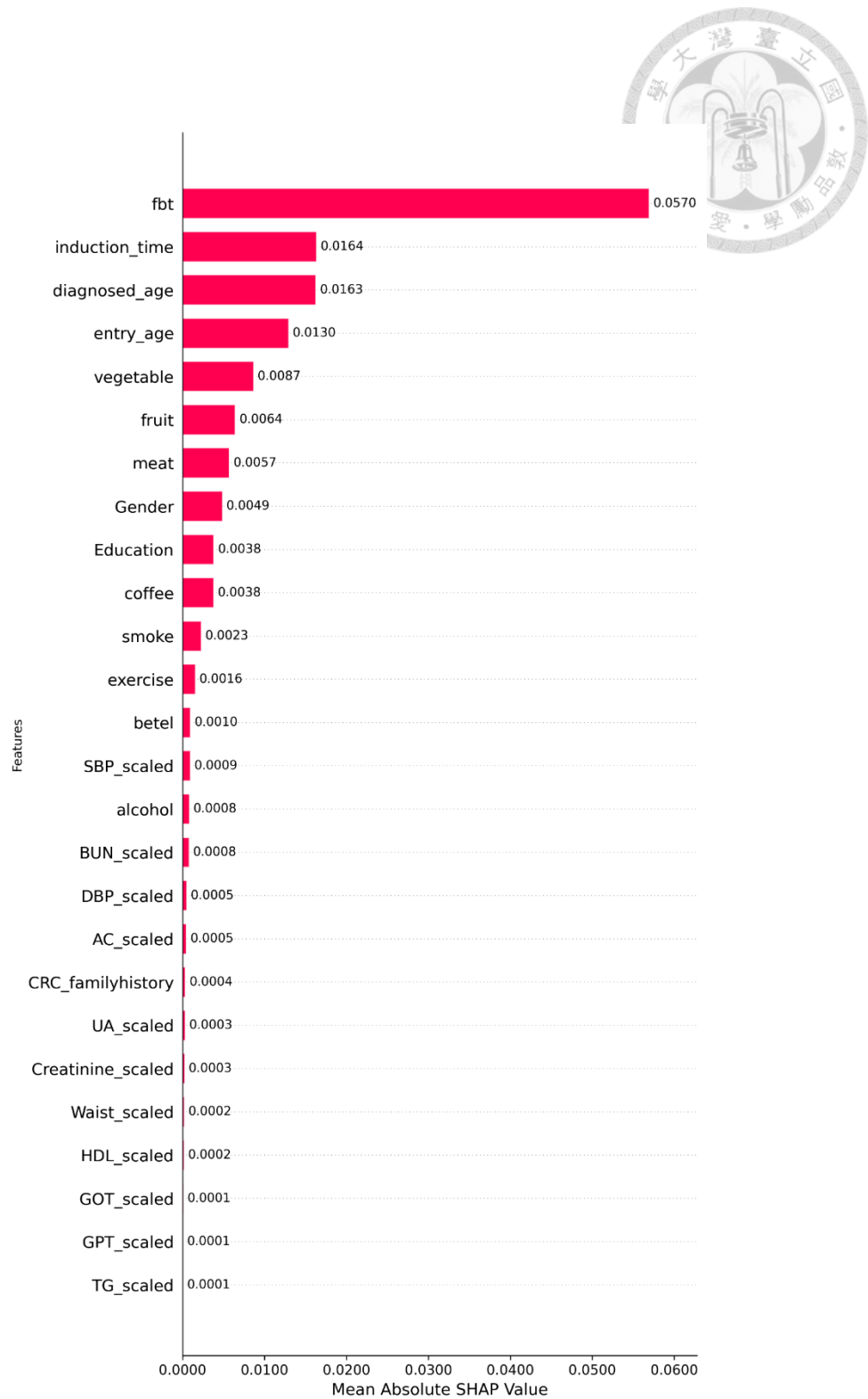


圖 4. 11 ANN 模型 SHAP 平均絕對值重要性比較圖

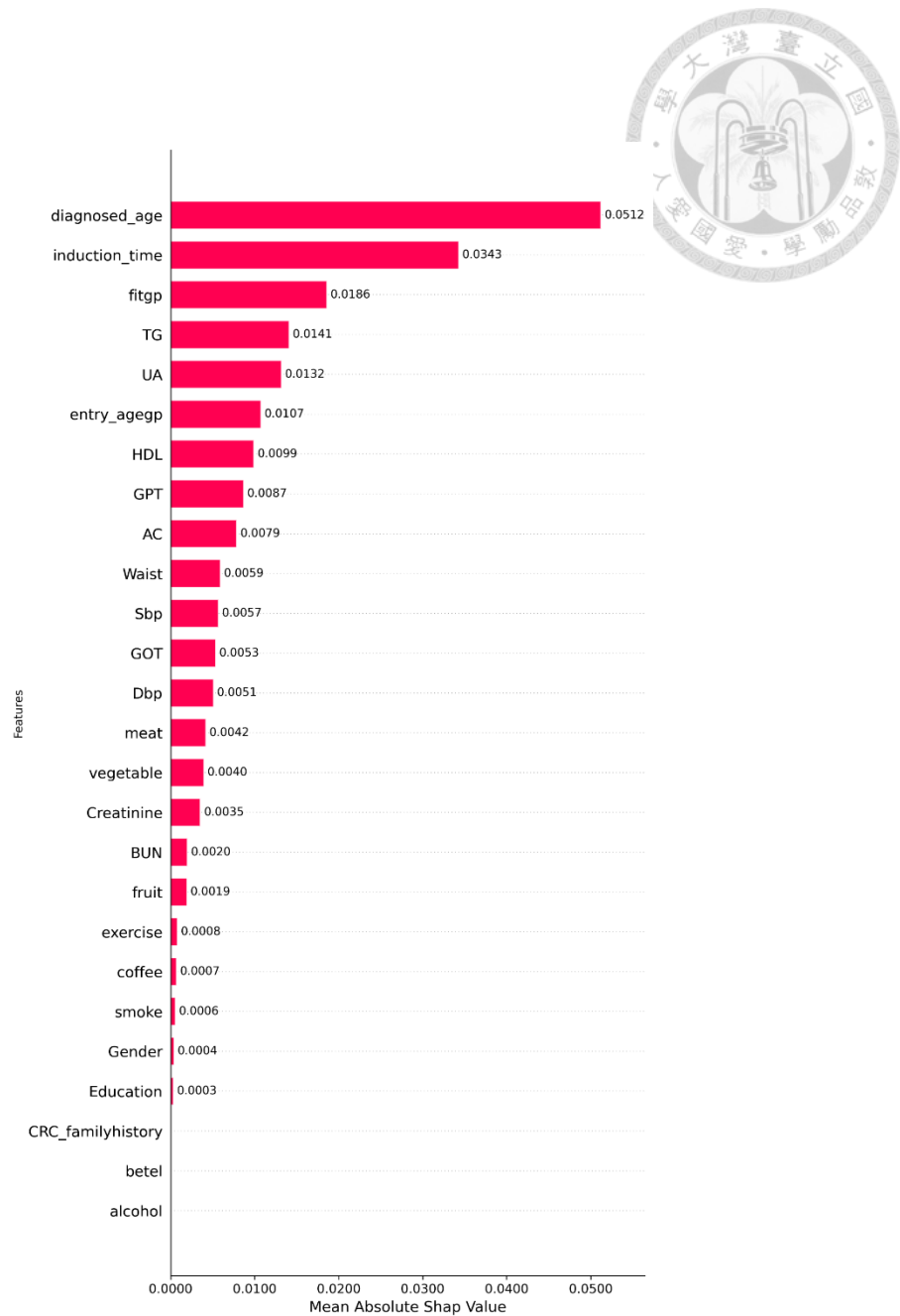


圖 4.12 隨機森林模型 SHAP 平均絕對值重要性比較圖

圖 4.11 為 ANN 模型之 SHAP 平均絕對值重要性比較圖。結果顯示，糞便血紅素濃度 (fbt) 為模型中最重要之特徵，其平均絕對 SHAP 值為 0.0570，明顯高於其他變數。其次依序為誘導時間 (induction_time, 0.0164)、確診年齡 (diagnosed_age, 0.0163)、進入研究年齡 (entry_age, 0.0130) 及蔬菜攝取 (vegetable, 0.0087)。此外，水果攝取、肉類攝取及性別等變數亦對模型具有一定

程度之貢獻，但其重要性相對較低。整體而言，ANN 模型主要依賴糞便血紅素濃度、疾病進展時間及年齡相關特徵進行腺瘤與大腸直腸癌之分類。



圖 4.12 為隨機森林模型之 SHAP 平均絕對值重要性比較圖。結果顯示，確診年齡 (diagnosed_age) 為模型中最重要之特徵，其平均絕對 SHAP 值為 0.0512；其次依序為誘導時間 (induction_time, 0.0343)、FIT 分級 (fitgp, 0.0186)、三酸甘油酯 (TG, 0.0141) 及尿酸 (UA, 0.0132)。此外，進入研究年齡、HDL、GPT、腰圍及血壓等代謝相關指標亦具有一定程度之重要性。顯示隨機森林模型除考量篩檢資訊外，亦納入受試者之年齡及臨床與生化檢驗相關特徵，以提升腺瘤與大腸直腸癌之分類能力。表 4.13 整理 ANN 模型與隨機森林模型於正常腺瘤與篩檢癌個體分類之 SHAP 重要變數排名。

表 4.13 ANN 與隨機森林模型於腺瘤與大腸篩檢癌分類之 SHAP 重要變數排名

排名	ANN 模型	隨機森林模型
1	糞便血紅素 (f-Hb)	診斷年齡 (Diagnosed Age)
2	誘導時間 (Induction Time)	誘導時間 (Induction Time)
3	診斷年齡 (Diagnosed Age)	FIT 值 (FIT-GP)
4	進入年齡 (Entry Age)	三酸甘油酯 (TG)
5	蔬菜攝取 (Vegetable)	尿酸 (UA)
6	水果攝取 (Fruit)	進入年齡 (Entry Age)
7	肉類攝取 (Meat)	高密度脂蛋白膽固醇 (HDL)



4.5.4 綜合比較與研究意涵

綜合 Classifier 1 (Normal versus Adenoma) 與 Classifier 2 (Adenoma versus Screen-detected CRC) 之機器學習分析結果，本節進一步比較兩項分類任務之模型效能差異，並探討個體層級分析結果與群體層級疾病自然史模型之關聯性。

一、模型效能比較

就整體分類效能而言，Classifier 1 (Normal versus Adenoma) 之分類結果明顯優於 Classifier 2 (Adenoma versus Screen-detected CRC)。於 Classifier 1 中，隨機森林模型於測試集之 AUC 達 0.935，而 ANN 模型之測試集 AUC 為 0.896，顯示兩種模型皆能有效區分正常個體與腺瘤個案。

相較之下，Classifier 2 之分類難度明顯提高。ANN 模型於測試集之 AUC 為 0.716，隨機森林模型則為 0.694，均低於 Classifier 1 之結果。此現象顯示腺瘤與篩檢發現大腸直腸癌個案於可觀測特徵上具有較高程度之重疊性，使模型較難建立明確之分類邊界。換言之，正常個體與腺瘤個案之差異相對明顯，而腺瘤與癌症個案則可能處於疾病進展之連續譜系中，因此增加分類之困難度。

就模型比較而言，隨機森林模型於 Classifier 1 表現最佳，不僅具有最高之測試集 AUC，亦呈現較佳之穩定性。然而於 Classifier 2 中，ANN 模型之測試集 AUC 略高於隨機森林模型 (0.716 vs 0.694)，顯示於較複雜之疾病階段分類問題中，ANN 模型可能較能捕捉變數間之非線性關係。因此，兩種模型於不同分類任務中各具優勢，但皆呈現相同趨勢，即疾病早期辨識遠較疾病進一步分期區辨容易。

二、特徵重要性之跨任務比較

SHAP 分析結果顯示，兩項分類重要變數具有部分一致性。Classifier 1 (Normal versus Adenoma) 中，ANN 與隨機森林模型皆將誘導時間 (Induction Time) 列為最

重要變數之一，並同時強調進入年齡 (Entry Age)、糞便血紅素 (f-Hb) 等特徵的重要性。此結果顯示，疾病發生早期階段與個體累積暴露時間具有密切關聯。

於 Classifier 2 (Adenoma versus Screen-detected CRC) 中，重要變數則轉為以糞便血紅素、誘導時間、診斷年齡及進入年齡等因素為主。ANN 模型將糞便血紅素 (f-Hb) 列為最重要變數，而隨機森林模型則將診斷年齡 (Diagnosed Age) 列為首要因子。兩種模型均指出年齡相關變數與疾病進展時間為區分腺瘤與癌症個案之關鍵資訊。

此結果顯示，不論由正常狀態進展至腺瘤階段或由腺瘤進一步發展為大腸直腸癌時，時間相關因素與糞便血紅素濃度之變化為重要之判別依據，雖然後者之分類效能仍有提升空間，但重要變數分析結果已提供潛在風險因子之資訊，可作為未來發展風險分層篩檢策略及個人化篩檢模式之參考依據。

4.6 個人化風險分層之疾病進展

依第 3.7 節所述之風險分層轉移強度建構方法，呈現以隨機森林模型預測機率所建立之個人化風險分層結果。考量隨機森林模型於第一項分類任務之區辨能力優於人工神經網路模型，第二項分類兩個模型平分秋色。因此，個人化風險分層之疾病進展分析以隨機森林模型之預測結果為依據，以下分別呈現兩段關鍵轉移(腺瘤發生以及腺瘤進展至癌症)之風險分層相對風險倍數、各風險層之累積進展機率，並透過四個不同風險個案呈現其個人化疾病進展。

4.6.1 風險分層相對風險倍數

表 4.14 呈現 Normal versus Adenoma 分類器之十分位風險分層結果。全體測試集之預測機率平均值為 0.0445，與實際腺瘤事件率 (0.0444) 高度吻合，顯示隨機森林模型具有良好之機率校準能力。各風險層 (Decile) 之預測機率均值自第 1 層之 0.0008 遞增至第 10 層之 0.3434，對應之相對風險倍數則由 0.017 增至 7.719，利

用此相對風險倍數，進一步使轉移速率 q_{12} 亦隨之變化，自第 1 層之 0.000019 提
升至第 10 層之 0.008645，相較於全域值 0.001120，最高風險層約為其 7.7 倍。

表 4. 14 Normal $\rightarrow$ Small Adenoma 各風險層之相對風險倍數與分層轉移強度

風險層	預測機率範圍	預測機率均值	相對風險倍數	分層 q_{12}
1	0.0000–0.0024	0.0008	0.017	0.000019
2	0.0024–0.0048	0.0035	0.079	0.000089
3	0.0048–0.0070	0.0058	0.130	0.000146
4	0.0070–0.0091	0.0077	0.174	0.000194
5	0.0091–0.0119	0.0102	0.230	0.000258
6	0.0119–0.0145	0.0131	0.295	0.000331
7	0.0147–0.0171	0.0161	0.361	0.000405
8	0.0171–0.0197	0.0183	0.411	0.000461
9	0.0197–0.0322	0.0249	0.561	0.000628
10	0.0326–1.0000	0.3434	7.719	0.008645

表 4.15 呈現 Adenoma versus Screen-detected CRC 分類器之風險分層結果。全體預測機率均值為 0.2553。相較於表 4.14，本任務之風險分層幅度較為平緩，相對風險倍數自第 1 層之 0.187 平滑遞增至第 10 層之 2.763，最高與最低風險層相差約 14.8 倍，明顯小於 Normal versus Adenoma 任務。對應之分層轉移強度 q_{34} 自第 1 層之 0.1101 增至第 10 層之 1.6270 (全域值為 0.5889)。此一較窄之風險梯度反映腺瘤與篩檢偵測癌於可觀測特徵上之重疊度較高，致使個體層級之風險區辨能力受限，與 4.5 節之發現相互呼應。

表 4. 15 Advanced Adenoma → Early PCDP 各風險層之相對風險倍數與分層轉移強度

風險層	預測機率範圍	預測機率均值	相對風險倍數	分層 q^{34}
1	0.0212–0.0700	0.0477	0.187	0.1101
2	0.0799–0.1218	0.0993	0.389	0.2291
3	0.1280–0.1427	0.1352	0.530	0.3120
4	0.1479–0.1861	0.1599	0.626	0.3689
5	0.1861–0.1976	0.1933	0.757	0.4459
6	0.1976–0.2174	0.2013	0.789	0.4644
7	0.2209–0.2706	0.2346	0.919	0.5411
8	0.2706–0.3911	0.3433	1.345	0.7919
9	0.3917–0.4554	0.4330	1.696	0.9987
10	0.5737–0.9333	0.7053	2.763	1.6270

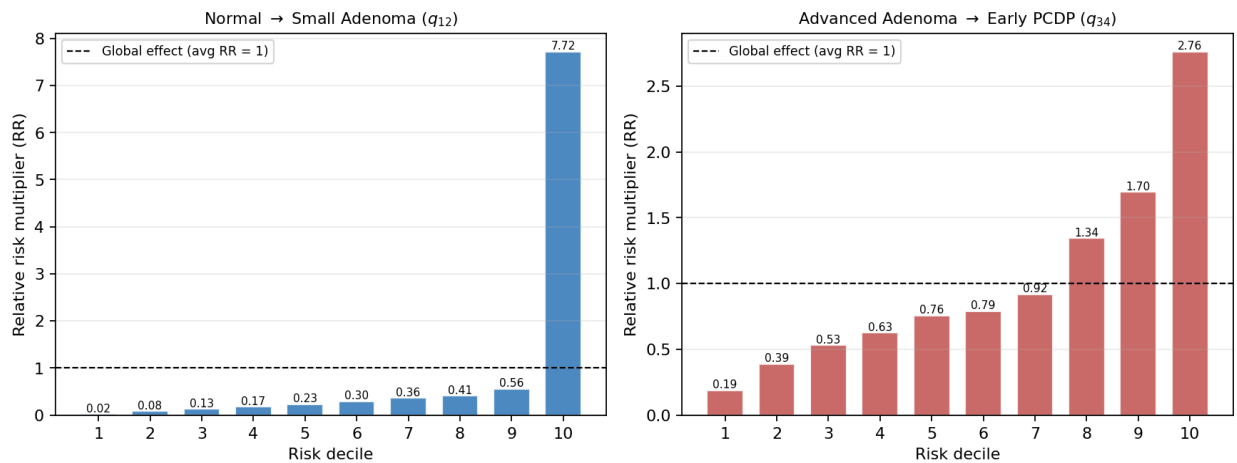


圖 4. 13 兩段疾病轉移各風險層之相對風險倍數;虛線代表全域效應(加權平均 $RR = 1$)

4.6.2 各風險層之累積進展機率

將各風險層之分層轉移強度代入連續時間多狀態模型，可推估不同風險層個體之累積進展機率。表 4.16 呈現於 Normal versus Adenoma 任務中，假設個體自正常狀態出發、追蹤十年後進展至腺瘤或更後期狀態之累積機率。結果顯示，第 1 風險層十年內進展至腺瘤以上之機率僅 0.0190%，第 5 層為 0.2569%，至第 9 層仍低於 1%(0.6244%)，惟第 10 風險層驟增至 8.2606%，約為全域平均 (1.1108%) 之 7.4 倍。此結果以累積進展機率之形式，再次印證腺瘤發生風險高度集中於最高風險層之特徵。

表 4.16 各風險層自正常狀態出發追蹤十年進展至「腺瘤」之累積機率

風險層	相對風險倍數	分層 q_{12}	十年進展腺瘤機率	相對全域之倍數
1	0.017	0.000019	0.0190%	0.02
2	0.079	0.000089	0.0887%	0.08
3	0.130	0.000146	0.1453%	0.13
4	0.174	0.000194	0.1937%	0.17
5	0.230	0.000258	0.2569%	0.23
6	0.295	0.000331	0.3292%	0.30
7	0.361	0.000405	0.4026%	0.36
8	0.411	0.000461	0.4584%	0.41
9	0.561	0.000628	0.6244%	0.56
10	7.719	0.008645	8.2606%	7.44
全域	1.000	0.001120	1.1108%	1.00

表 4.17 呈現於 Adenoma versus CRC 任務中，假設個體自 Advanced Adenoma 出發、追蹤兩年後進展至 Early PCDP 或更後期狀態之累積機率。由於 q_{34} 之數值較大，進展較為迅速，各風險層之兩年進展機率明顯高於表 4.16。第 1 風險層之進展機率為 19.7599%，第 5 層為 59.01%，至第 10 層達 96.14%，全域平均為 69.20%。最高風險層約為全域之 1.39 倍、最低風險層之 4.9 倍，風險梯度明顯較表 4.16 平緩，再次反映腺瘤至癌症階段之風險區辨能力相對有限。

表 4.17 各風險層自 Advanced Adenoma 出發追蹤兩年進展至「癌前期以上」之累積機率

風險層	相對風險倍數	分層 q_{34}	兩年從腺瘤進展 至癌症機率	相對全域之倍數
1	0.187	0.1101	19.76%	0.29
2	0.389	0.2291	36.76%	0.53
3	0.530	0.3120	46.42%	0.67
4	0.626	0.3689	52.19%	0.75
5	0.757	0.4459	59.01%	0.85
6	0.789	0.4644	60.50%	0.87
7	0.919	0.5411	66.11%	0.96
8	1.345	0.7919	79.48%	1.15
9	1.696	0.9987	86.43%	1.25
10	2.763	1.6270	96.14%	1.39
全域	1.000	0.588900	69.20%	1.00



4.6.3 四種代表性風險情境之十年疾病進展模擬

為進一步了解模型所辨識之高、低風險族群之臨床特徵，本研究選取各模型預測風險最低十分位（第 1 層）與最高十分位（第 10 層）作為代表，進行描述性分析，比較不同風險層之臨床特徵差異。

表 4.18 「分層 q_{12} 」與「分層 q_{34} 」模型低/高風險十分位族群之臨床特徵

變項	分層 q_{12}	分層 q_{12}	分層 q_{34}	分層 q_{34}
	低風險組	高風險組	低風險組	高風險組
	(n=158)	(n=158)	(n=10)	(n=10)
連續變項 中位數 [IQR]				
肌酸酐(mg/dL)	0.70	0.90	0.89	0.69
	[0.60-0.80]	[0.71-1.18]	[0.72-1.00]	[0.53-0.85]
血尿素氮	12.00	14.00	15.50	13.50
	[11.00-15.17]	[11.25-17.00]	[12.00-18.32]	[10.25-14.75]
尿酸(mg/dL)	4.10	6.00	5.95	4.55
	[3.60-4.60]	[5.10-6.90]	[5.15-6.52]	[4.12-5.22]
GPT(U/L)	18.00	26.50	27.50	20.00
	[15.00-21.00]	[20.00-38.00]	[22.00-30.50]	[17.25-24.50]
GOT(U/L)	21.00	25.50	23.00	21.00

變項	分層 q_{12}	分層 q_{12}	分層 q_{34}	分層 q_{34}
	低風險組	高風險組	低風險組	高風險組
	(n=158)	(n=158)	(n=10)	(n=10)
	[18.00-24.75]	[20.25-33.00]	[21.00-25.00]	[17.50-22.75]
高密度脂蛋白膽固醇	62.00	49.00	53.20	62.65
	[48.83-76.75]	[41.10-56.00]	[44.50-55.77]	[52.15-75.30]
三酸甘油脂(mg/dL)	60.00	134.50	237.50	57.50
	[50.00-76.00]	[93.25-202.00]	[117.50-262.50]	[43.25-111.50]
空腹血糖(mg/dL)	92.00	100.00	97.50	94.00
	[88.00-97.75]	[91.00-115.75]	[87.00-104.75]	[90.25-113.75]
舒張壓(mmHg)	72.00	90.50	89.50	73.00
	[68.00-78.00]	[82.00-99.00]	[80.25-94.75]	[68.00-79.00]
收縮壓(mmHg)	113.00	140.00	141.50	119.00
	[105.25-126.00]	[128.00-152.00]	[124.75-154.75]	[113.25-127.00]
腰圍	73.00	89.25	86.50	74.50
	[69.00-78.00]	[84.00-96.00]	[82.75-87.75]	[69.25-87.00]
誘導時間	13.87	11.49	10.65	0.08

變項	分層 q_{12}	分層 q_{12}	分層 q_{34}	分層 q_{34}
	低風險組	高風險組	低風險組	高風險組
	(n=158)	(n=158)	(n=10)	(n=10)
	[12.36-14.48]	[8.94-13.49]	[9.57-12.25]	[0.05-2.39]
運動習慣				
無	76 (48.1%)	69 (43.7%)	7 (70.0%)	3 (30.0%)
有	82 (51.9%)	89 (56.3%)	3 (30.0%)	7 (70.0%)
大腸癌家族病史				
無	150 (94.9%)	150 (94.9%)	9 (90.0%)	7 (70.0%)
有	8 (5.1%)	8 (5.1%)	1 (10.0%)	3 (30.0%)
性別				
男性	5 (3.2%)	89 (56.3%)	3 (30.0%)	3 (30.0%)
女性	153 (96.8%)	69 (43.7%)	7 (70.0%)	7 (70.0%)
有序類別變項 人數 (%)				
咖啡攝取				
低	129 (81.6%)	125 (79.1%)	9 (90.0%)	8 (80.0%)

變項	分層 q_{12}	分層 q_{12}	分層 q_{34}	分層 q_{34}
	低風險組	高風險組	低風險組	高風險組
	(n=158)	(n=158)	(n=10)	(n=10)
中	17 (10.8%)	17 (10.8%)	1 (10.0%)	2 (20.0%)
高	12 (7.6%)	16 (10.1%)	0(0%)	0(0%)
水果攝取				
低	45 (28.5%)	33 (20.9%)	2 (20.0%)	1 (10.0%)
中	61 (38.6%)	62 (39.2%)	4 (40.0%)	5 (50.0%)
高	52 (32.9%)	63 (39.9%)	4 (40.0%)	4 (40.0%)
肉類攝取				
低	96 (60.8%)	49 (31.0%)	3 (30.0%)	3 (30.0%)
中	52 (32.9%)	77 (48.7%)	6 (60.0%)	7 (70.0%)
高	10 (6.3%)	32 (20.3%)	1 (10.0%)	0(0%)
蔬菜攝取				
低	32 (20.3%)	35 (22.2%)	2 (20.0%)	2 (20.0%)
中	55 (34.8%)	63 (39.9%)	5 (50.0%)	2 (20.0%)

變項	分層 q_{12}	分層 q_{12}	分層 q_{34}	分層 q_{34}
	低風險組	高風險組	低風險組	高風險組
	(n=158)	(n=158)	(n=10)	(n=10)
高	71 (44.9%)	60 (38.0%)	3 (30.0%)	6 (60.0%)
嚼檳榔				
無	157 (99.4%)	129 (81.6%)	8 (80.0%)	10 (100.0%)
已戒	0(0%)	16 (10.1%)	0(0%)	0(0%)
仍在嚼檳榔中	1 (0.6%)	13 (8.2%)	2 (20.0%)	0(0%)
喝酒				
無	157 (99.4%)	128 (81.0%)	7 (70.0%)	9 (90.0%)
已戒	0(0%)	3 (1.9%)	0(0%)	0(0%)
仍在喝酒中	1 (0.6%)	27 (17.1%)	3 (30.0%)	1 (10.0%)
抽菸				
無	157 (99.4%)	110 (69.6%)	8 (80.0%)	8 (80.0%)
已戒	0(0%)	11 (7.0%)	0(0%)	0(0%)
仍在抽菸中	1 (0.6%)	37 (23.4%)	2 (20.0%)	2 (20.0%)

變項	分層 q_{12}	分層 q_{12}	分層 q_{34}	分層 q_{34}
	低風險組	高風險組	低風險組	高風險組
	(n=158)	(n=158)	(n=10)	(n=10)
教育程度				
低	120 (75.9%)	114 (72.2%)	9 (90.0%)	8 (80.0%)
中	28 (17.7%)	28 (17.7%)	1 (10.0%)	1 (10.0%)
高	10 (6.3%)	16 (10.1%)	0(0%)	1 (10.0%)
年齡分組				
50-54	48 (30.4%)	46 (29.1%)	0(0%)	5 (50.0%)
55-59	28 (17.7%)	43 (27.2%)	1 (10.0%)	4 (40.0%)
60-64	28 (17.7%)	45 (28.5%)	7 (70.0%)	1 (10.0%)
65-69	54 (34.2%)	23 (14.6%)	2 (20.0%)	0(0%)
70-74	0(0%)	1 (0.6%)	0(0%)	0(0%)
FIT 濃度分組 (ng/mL)				
0	115 (72.8%)	45 (28.5%)	2 (20.0%)	1 (10.0%)
1-9	24 (15.2%)	13 (8.2%)	6 (60.0%)	1 (10.0%)

變項	分層 q_{12}	分層 q_{12}	分層 q_{34}	分層 q_{34}
	低風險組	高風險組	低風險組	高風險組
	(n=158)	(n=158)	(n=10)	(n=10)
10-19	14 (8.9%)	9 (5.7%)	1 (10.0%)	1 (10.0%)
20-29	1 (0.6%)	18 (11.4%)	1 (10.0%)	0(0%)
30-39	1 (0.6%)	8 (5.1%)	0(0%)	3 (30.0%)
40-49	3 (1.9%)	5 (3.2%)	0(0%)	0(0%)
50+	0(0%)	60 (38.0%)	0(0%)	4 (40.0%)
預測機率 中位數 [IQR]				
預測目標機率	0.0006	0.0505	0.0414	0.6657
	[0.0006-0.0013]	[0.0411-1.0000]	[0.0411-0.0629]	[0.5857-0.7993]

表 4.18 顯示分層 q_{12} 正常至腺瘤 (Normal→Adenoma) 及分層 q_{34} 腺瘤至大腸癌 (Adenoma→CRC) 兩階段模型中，最低十分位與最高十分位風險群體之臨床特徵。連續變項以中位數與四分位距 (median [IQR]) 表示，類別變項則以人數及百分比呈現。前述 SHAP 分析已辨識各階段模型中影響預測結果之重要變項，本節則進一步比較高、低風險群體之臨床特徵，檢視 SHAP 所辨識之重要變項是否反映於不同風險族群。

在正常至腺瘤 (Normal→Adenoma) 階段，高風險組相較於低風險組呈現較高之



尿酸 (6.00 vs. 4.10 mg/dL)、三酸甘油酯 (134.50 vs. 60.00 mg/dL)、收縮壓 (140.00 vs. 113.00 mmHg)、舒張壓 (90.50 vs. 72.00 mmHg) 及腰圍 (89.25 vs. 73.00 cm)，而高密度脂蛋白膽固醇則較低 (49.00 vs. 62.00 mg/dL)，誘導時間亦較短 (11.49 vs. 13.87 年)。此外，高風險組之 FIT 濃度明顯偏向高濃度組 (50+ ng/mL: 38.0% vs. 0%)，吸菸、飲酒及嚼檳榔比例亦較高。上述結果與前述 SHAP 分析所辨識之重要特徵，包括 FIT 濃度、誘導時間、腰圍、舒張壓、尿酸及三酸甘油酯等，大致呈現一致趨勢，顯示模型所辨識之重要變項亦反映於高風險族群之臨床特徵。

相較之下，腺瘤至大腸癌 (Adenoma→CRC) 階段之高、低風險組樣本數皆僅 10 人，因此相關結果應以探索性方式解讀。高風險組之預測機率中位數高於低風險組 (0.6657 vs. 0.0414)，且誘導時間明顯較短 (0.08 vs. 10.65 年)，與前述 SHAP 分析所辨識之重要特徵相符。然而，部分代謝指標如三酸甘油酯、腰圍及血壓並未呈現與第一階段一致之變化方向，顯示腺瘤形成後之疾病進展可能涉及不同之風險特徵組合，而非單純反映代謝症候群相關危險因子；此外，亦可能受到第二階段樣本數較少之影響，因此相關結果仍須審慎解讀。

第一階段模型所辨識之重要特徵與高風險族群之臨床特徵具有較高一致性；第二階段模型則反映不同於第一階段之風險特徵模式，顯示兩階段疾病進展可能具有不同之影響因子。基於此，本研究進一步依兩階段模型之最低與最高風險十分位建立四種代表性風險情境 (Case A–D)。由於兩階段模型係分別建立於不同資料集，各情境係由兩資料集中臨床特徵最相近之代表個案配對形成，並非同一受試者之縱向追蹤資料。Case A、B、C 及 D 分別代表 q_{12} 低/ q_{34} 低、 q_{12} 高/ q_{34} 低、 q_{12} 低/ q_{34} 高及 q_{12} 高/ q_{34} 高四種疾病進展情境。各情境係分別選取屬於相同風險層級且臨床特徵相近之代表個案進行配對，以模擬不同風險組合下之疾病進展；因此 Case 中之兩位代表個案並非同一人，而是同一類型風險個案之配對。

表 4. 19 四種代表性風險情境之配對代表個案臨床特徵

臨床特徵	Case A	Case A	Case B	Case B	Case C	Case C	Case D	Case D
	分層 q_{12}	分層 q_{34}	分層 q_{12}	分層 q_{34}	分層 q_{12}	分層 q_{34}	分層 q_{12}	分層 q_{34}
風險組合	低	低	高	低	低	高	高	高
預測機率	0.0006	0.0700	1.0000	0.0414	0.0013	0.6806	1.0000	0.5845
年齡分組	65–69	60–64	60–64	60–64	60–64	55–59	55–59	55–59
FIT 分組	0	1–9	0	0	0	10–19	≥50	30–39
咖啡攝取	中	中	中	中	高	高	高	高
水果攝取	中	中	高	高	高	高	高	高
肉類攝取	中	中	中	中	低	低	中	中
蔬菜攝取	高	高	低	低	高	高	低	低

臨床特徵	Case A	Case A	Case B	Case B	Case C	Case C	Case D	Case D
	分層 q_{12}	分層 q_{34}	分層 q_{12}	分層 q_{34}	分層 q_{12}	分層 q_{34}	分層 q_{12}	分層 q_{34}
規律運動	無	無	有	有	有	有	無	無
吸菸	無	無	無	無	無	無	有	有
飲酒	無	無	無	無	無	無	無	無
嚼檳榔	無	無	無	無	無	無	無	無
大腸癌家族史	無	無	有	有	無	無	有	有
教育程度	低	低	中	中	中	低	中	中
肌酸酐(mg/dL)	0.70	0.50	1.20	1.20	0.67	0.51	1.10	1.10
血尿素氮(mg/dL)	12	12	19	19	10	11	21	21

臨床特徵	Case A	Case A	Case B	Case B	Case C	Case C	Case D	Case D
	分層 q_{12}	分層 q_{34}	分層 q_{12}	分層 q_{34}	分層 q_{12}	分層 q_{34}	分層 q_{12}	分層 q_{34}
尿酸 (mg/dL)	5.4	5.8	6.6	6.6	3.5	4.1	5.3	5.3
GPT (U/L)	14	28	21	21	25	16	29	29
GOT (U/L)	21	27	21	21	30	17	19	19
HDL (mg/dL)	41.9	55.0	27.1	27.1	49.8	77.0	51.0	51.0
三酸甘油脂(mg/dL)	66	104	249	249	51	43	179	179
空腹血糖 (mg/dL)	87	106	93	93	91	87	128	128
舒張壓 (mmHg)	82	84	85	85	68	62	79	79
收縮壓 (mmHg)	125	145	122	122	114	104	120	120



臨床特徵	Case A	Case A	Case B	Case B	Case C	Case C	Case D	Case D
	分層 q_{12}	分層 q_{34}	分層 q_{12}	分層 q_{34}	分層 q_{12}	分層 q_{34}	分層 q_{12}	分層 q_{34}
腰圍 (cm)	84	88	86	86	70	62	91	91
誘導時間 (年)	13.41	12.33	10.11	10.11	10.63	6.18	0.05	0.05

表 4. 20 呈現四個原型個案追蹤十年之疾病進展機率。Case A 兩段風險均低，十年內進展至腺瘤以上之機率為 0.0190%，至臨床期癌症僅 0.0056%，屬低風險族群。Case D 兩段風險皆高，進展至腺瘤以上達 8.2606%，至臨床期癌症達 6.4128%，因此需積極監測之高風險個體。

表 4. 20 四種代表性特徵自正常狀態出發追蹤十年之疾病進展機率

個案	風險象限	q_{12}	q_{34}	腺瘤風險	癌症風險
Case A	低，低	0.000019	0.1101	0.0190%	0.0056%
Case B	高，低	0.008645	0.1101	8.2606%	2.4642%
Case C	低，高	0.000019	1.6270	0.0190%	0.0146%
Case D	高，高	0.008645	1.6270	8.2606%	6.4128%

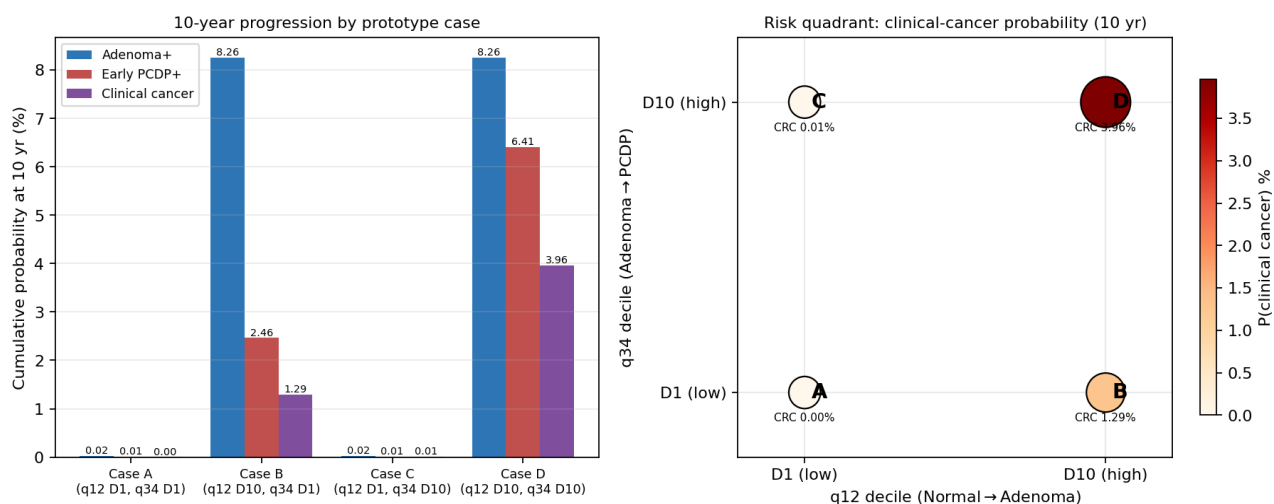


圖 4.14 四個代表性特徵之十年進展機率(左)及風險象限圖(右，圓圈大小與顏色對應臨床期癌症機率)

特別值得注意者為 Case B 與 Case C 之對比。兩者分別僅於單一階段呈現高風險，惟 Case B(q_{12} 高、 q_{34} 低)進展至臨床期癌症之機率為 2.4642%，而 Case C(q_{12} 低、 q_{34} 高) 僅為 0.0146%，前者約為後者之 169 倍。此一顯著差異反映於本研究



之疾病自然史架構下，個體須先跨越「形成腺瘤」之階段方能進入後續惡化路徑，故 q_{12} (是否容易形成腺瘤)對整體癌症風險之影響，遠大於 q_{34} (腺瘤形成後之惡化速度)。換言之，在腺瘤－癌症序列之循序進展假設下， q_{12} 構成整體風險之上游主閘門。

本節結果顯示，將個體層級之機器學習預測機率轉換為相對風險倍數並回饋調節疾病自然史模型之轉移強度，能於維持母體平均效應一致之前提下，有效刻劃不同風險族群之疾病進展差異。其中，Normal 至 Adenoma 階段之風險分層具有最高之區辨度，而辨識「容易形成腺瘤」之高風險個體，較辨識「腺瘤惡化速度快」之個體更能於上游有效攔截大腸直腸癌之發生風險。此結果可作為未來發展風險導向之差異化篩檢策略之實證依據。



第五章 討論與結論

5.1 主要研究發現

本研究最大的貢獻，不僅在於建立一套能描述大腸直腸癌疾病自然史之九狀態連續時間隱藏狀態模型，更重要的是首次整合群體層級疾病自然史推論與個人化風險預測，提出一套由群體疾病自然史 (Population-level disease progression) 延伸至個人化疾病進展 (Personalized disease progression) 之整合分析架構。相較於傳統疾病自然史研究多著重於估計整體族群平均疾病進展速度，或機器學習研究僅專注於疾病分類及風險預測，本研究將兩者加以整合，使群體層級疾病轉移資訊得以進一步轉化為個體層級疾病進展歷程，提供疾病自然史研究一項新的分析模式。

在群體層級方面，本研究利用全國性糞便免疫化學檢查 (FIT) 篩檢資料，建立九狀態連續時間隱藏狀態馬可夫鏈 (Continuous-time hidden-state Markov chain) 疾病進展模型，並透過發射矩陣 (Emission matrix) 建立潛在疾病狀態與實際篩檢結果之連結，進一步估計生成矩陣 (Generator matrix) 及各疾病狀態間之轉移強度。此模型不僅重建大腸直腸癌自然史，也提供疾病停留時間、疾病進展速度及疾病轉移方向之完整資訊，使疾病自然史由傳統概念性模型提升為可量化之數學模型。

另一方面，本研究利用人工神經網路及隨機森林建立正常至腺瘤及腺瘤至篩檢癌之個人化風險預測模型。研究結果顯示，Random Forest 於正常與腺瘤辨識具有較佳預測能力，而 SHAP 分析亦證實年齡、誘導時間及糞便血紅素濃度為疾病進展最重要之影響因子。更重要的是，本研究並未止於疾病分類，而是進一步將 Random Forest 所得之個人化風險倍數導入群體生成矩陣，建立個人化生成矩陣 (Personalized generator matrix)，推估不同個體之疾病進展速度及疾病進展光譜

(personalized disease progression spectrum)。此一架構代表疾病自然史研究已由描述「平均疾病進展」提升至推估「個人疾病進展」，亦為本研究最具創新性之方法學貢獻。



5.2 群體層級疾病自然史模型之方法學意義

疾病自然史模型長期被應用於癌症篩檢政策評估，其核心目的在於重建無法直接觀察之疾病進展歷程。然而，傳統疾病自然史模型多以三至五狀態描述疾病演變，且多假設觀測結果即可代表真實疾病狀態，較少同時考慮潛在疾病狀態、篩檢敏感度及觀測誤差。

本研究生成矩陣之估計，不僅提供各疾病狀態間之轉移強度，更代表整體族群疾病自然史之核心參數。生成矩陣可視為疾病演化之「引擎」，所有疾病進展方向、疾病停留時間及長期疾病分布皆由生成矩陣所決定。因此，本研究第四章估計之生成矩陣，不應僅解讀為數值估計結果，而應視為描述臺灣 FIT 篩檢族群疾病自然史的重要數學模型。相較於僅估計單一轉移率，本研究生成矩陣同時整合多個疾病狀態之動態資訊，使疾病自然史由靜態描述提升為完整之連續時間動態模型。

5.3 個人化風險預測與機器學習

近年人工智慧已廣泛應用於癌症風險預測，然而多數研究仍停留於疾病分類或事件預測，較少探討如何與疾病自然史模型整合。本研究比較 ANN 與 Random Forest 兩種監督式機器學習模型，結果顯示 Random Forest 於正常與腺瘤分類具有較佳預測能力，顯示樹狀集成模型較能捕捉人口學變數、生化指標及生活型態間之非線性交互作用。相較之下，ANN 於樣本數有限及類別不平衡情況下，其泛化能力較受限制，顯示模型效能仍受到資料結構影響。

值得注意的是，腺瘤與篩檢癌之分類能力明顯下降，此結果反映腺瘤至癌症

之疾病進展並非單一事件，而是一連串累積之生物學過程。單一時間點之臨床資料不足以完全描述癌化過程，因此亦支持將疾病自然史資訊納入個人風險預測的重要性。



SHAP 分析結果進一步顯示，年齡、誘導時間及糞便血紅素濃度在兩項分類模型中均具有高度重要性。其中，年齡代表長期暴露及生物老化之累積效應；誘導時間反映疾病演化速度；而糞便血紅素則代表腫瘤微量出血及病灶活動程度。此結果不僅與既有流行病學研究一致，也說明時間因素與篩檢指標在疾病進展中具有持續且重要之角色。更重要的是，SHAP 提供各變數對模型預測結果之個別貢獻，使黑盒子模型具有可解釋性，提升人工智慧於臨床決策之可信度。

5.4 群體疾病自然史與個人化疾病進展之整合

本研究最重要的方法學創新，在於首次將群體層級疾病自然史模型與個人化風險預測模型加以整合，建立一套由群體疾病自然史 (Population-level disease progression) 延伸至個人化疾病進展 (Personalized disease progression) 之分析架構。相較於過去疾病自然史模型著重於估計整體族群平均疾病進展速度，而機器學習模型則主要著重於疾病分類或個人風險預測，本研究進一步建立兩者之橋樑，使群體層級疾病自然史所估計之生成矩陣 (Generator matrix) 能夠依據個體風險特徵進行調整，形成個人化生成矩陣 (Personalized generator matrix)，進而推估不同個體之疾病進展歷程與累積疾病風險。

傳統疾病自然史模型通常假設所有受檢者共享相同之疾病轉移強度，因此生成矩陣所代表者為整體族群之平均疾病自然史。然而，實際臨床觀察顯示，即使具有相同年齡或相同 FIT 結果，不同個體仍可能因生活型態、生理特徵、代謝狀況及其他危險因子不同，而具有截然不同之疾病進展速度。因此，若僅利用群體平均生成矩陣描述疾病進展，將無法充分反映疾病異質性 (Disease

heterogeneity)，亦限制疾病自然史模型於個人化醫療之應用。

為克服上述限制，本研究首先利用 Random Forest 模型估計每位受檢者由正常狀態進展至腺瘤，以及由腺瘤進展至大腸直腸癌之個人化疾病風險。不同於過去僅將機器學習模型視為疾病分類工具，本研究進一步將模型所估計之預測機率轉換為相對風險倍數 (Relative risk multiplier)，並以此作為修正群體生成矩陣各疾病轉移強度之依據。換言之，群體生成矩陣提供疾病自然史之基本結構，而機器學習所估計之個人風險則決定疾病轉移速度，使每位受檢者均可建立專屬之生成矩陣。此一概念使疾病自然史由「平均疾病進展」正式發展為「個人化疾病進展」，亦代表疾病自然史研究由群體推論邁向個體推論的重要突破。

本研究所建立之個人化生成矩陣，不僅保留疾病自然史模型之理論基礎，同時納入個體危險因子對疾病演化速度之影響，使不同個體即使具有相同初始疾病狀態，仍可因不同風險特徵而產生不同疾病進展軌跡 (Disease trajectory)。因此，本研究第四章所提出之個人化疾病進展光譜 (Personalized disease progression spectrum)，並非單純之風險分層結果，而是建立於疾病自然史模型基礎上之動態疾病演化模型，可同時呈現不同風險族群於未來各時間點疾病進展之累積機率與疾病狀態分布。

另一方面，本研究亦利用 SHAP 分析探討不同危險因子對疾病進展之影響機制。研究結果顯示，年齡、誘導時間及糞便血紅素濃度均為影響個人疾病進展最重要之預測變數。其中，年齡反映長期暴露及生物老化效應；誘導時間代表疾病進展歷程之時間維度；糞便血紅素則反映病灶出血程度及腫瘤生物活性。此三項變數不僅於 Random Forest 模型中具有高度重要性，同時亦與群體疾病自然史模型所估計之疾病轉移方向相互呼應，顯示疾病進展同時受到時間累積效應與生物學特徵共同影響。



值得注意的是，本研究並未將機器學習模型視為疾病自然史模型之替代方法，而是視其為補充群體模型不足之工具。生成矩陣仍然提供疾病演化之主要機制，而 Random Forest 則負責量化個體疾病風險，兩者共同構成完整之疾病進展架構。此種 Hybrid Disease Progression Framework 兼具機制解釋能力 (Mechanistic interpretability) 與個人化預測能力 (Personalized prediction)，突破過去疾病自然史模型缺乏個人化資訊，以及機器學習缺乏疾病機制解釋之限制。

更重要的是，本研究所提出之方法亦可視為未來 AI 數位雙生 (AI-enabled Digital Twin) 之基礎架構。數位雙生之核心概念，在於利用真實世界資料建立可隨時間持續更新之虛擬個體模型，而本研究建立之個人化生成矩陣正可視為每位受檢者之疾病進展數位模型。未來若能持續整合歷次 FIT 檢測結果、大腸鏡發現、生化指標、生活型態及其他臨床資訊，個人化生成矩陣將可隨資料更新而動態調整，使疾病進展預測由靜態模型發展為動態自我更新模型 (Dynamic self-updating disease progression model)，提供更符合精準醫療需求之風險評估架構。

因此，本研究所提出之整合架構，其貢獻已不限於大腸直腸癌疾病自然史研究，而是建立一套兼具疾病自然史推論、人工智慧風險預測及個人化疾病進展分析之整合方法，可望作為其他慢性疾病及癌症篩檢研究之重要分析框架。

5.5 臨床與公共衛生意涵

本研究所建立之整合模型，不僅具有方法學上的創新，更具有重要之臨床與公共衛生應用價值。現行大腸直腸癌篩檢策略多採固定篩檢間隔，主要依據年齡及篩檢結果決定後續追蹤方式。然而，大量研究已指出，不同受檢者即使具有相同 FIT 結果，其未來疾病進展風險仍存在顯著差異。因此，固定篩檢策略雖易於執行，但未必能兼顧篩檢效益與醫療資源配置效率。

本研究提出之個人化疾病進展模型，可依據每位受檢者之疾病進展速度建立



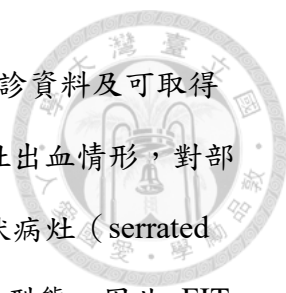
風險分層，使篩檢策略由傳統年齡導向 (age-based screening) 逐步發展為風險導向 (risk-adaptive screening)。對於低風險族群，可適度延長篩檢間隔，以降低不必要之醫療支出及檢查負擔；而高風險族群則可縮短追蹤時間，提早安排大腸鏡檢查或其他診斷措施，以提升臨床前可偵測癌症階段及早期癌症之發現率。

此外，個人化疾病進展光譜亦可提供臨床醫師與受檢者進行共享決策 (shared decision making) 之重要依據。相較於僅提供單一風險分數，本研究可呈現不同時間點之疾病進展機率，使受檢者更容易理解自身疾病演化趨勢，有助於提升篩檢遵從性及健康促進效果。

從公共衛生觀點而言，本研究亦提供未來國家癌症篩檢政策新的發展方向。隨著人工智慧與精準醫療之快速發展，未來癌症篩檢已不應僅追求提高篩檢率，更應依據不同個體疾病風險配置醫療資源。本研究建立之方法學，可望協助國家篩檢計畫建立風險分層管理系統，提高篩檢效率，同時兼顧成本效益與健康公平，作為精準公共衛生 (precision public health) 之重要基礎。

5.6 研究限制與未來發展

雖然本研究建立完整之群體疾病自然史與個人化疾病進展整合架構，但仍存在若干限制。首先，本研究之個人化風險模型主要利用彰化地區社區篩檢資料建立，且研究樣本係依疾病模式 (Disease mode) 進行抽樣，而非直接反映一般社區篩檢族群之疾病分布。因此，各疾病類別之樣本比例係依研究設計設定，樣本中各疾病狀態之分布及其平均比例並不同於實際族群之疾病盛行率 (Prevalence)。雖然本研究於模型訓練過程採用交叉驗證 (cross-validation) 以降低模型過度擬合之風險，但模型之泛化能力仍有待利用其他地區、不同族群及不同國家之資料進行外部驗證 (External validation)，以進一步評估模型之可攜性、穩健性及臨床應用價值。



此外，本研究之疾病進展模型主要依據 FIT 篩檢結果、確診資料及可取得之臨床變項建立個人化風險分層。然而，FIT 篩檢主要反映病灶出血情形，對部分出血量較少之病灶，其偵測能力可能受到限制。例如，鋸齒狀病灶（serrated lesions）通常具有較低之出血傾向，且多位於右側大腸或呈扁平型態，因此 FIT 濃度未必明顯升高，可能增加篩檢漏失之機率。若研究資料未能完整納入病理亞型、病灶位置或相關分子特徵，模型對此類病灶之疾病進展風險仍可能有所低估。未來若能進一步整合病理分類、病灶位置及分子生物標記等資訊，將有助於提升不同腺瘤亞型之風險辨識能力，並進一步改善個人化疾病進展模型之預測效能。

本研究所建立之個人化生成矩陣仍屬靜態估計，主要利用單一時間點之危險因子修正疾病轉移強度。未來若能結合重複 FIT 檢測、生化指標、生活型態變化及長期追蹤資料，建立縱向動態更新 (longitudinal dynamic updating) 模型，將可更真實反映疾病風險隨時間之變化。

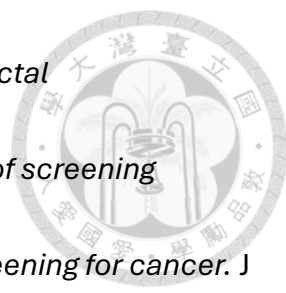
此外，近年精準醫療快速發展，未來可進一步整合多體學 (multi-omics) 資訊，包括全基因體資料、Polygenic Risk Score、腸道微生物相、代謝體學及蛋白質體學等生物標誌，建立更完整之個人化疾病進展模型。同時，亦可結合影像人工智慧、大腸鏡影像特徵及病理影像資訊，進一步提升疾病進展推估之精確性。

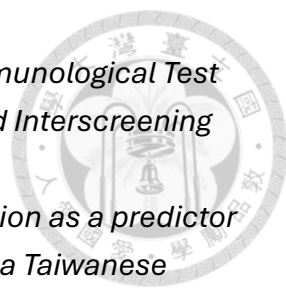
另一方面，本研究所提出之架構亦可進一步發展為 AI Digital Twin 系統。透過持續接收臨床資料及篩檢資訊，建立可即時更新之虛擬疾病模型，再結合強化學習 (Reinforcement learning)，動態模擬不同篩檢策略、追蹤間隔及介入措施對疾病進展之影響，進而提供最佳化之個人化篩檢建議。此種自我學習之智慧化系統，將使疾病自然史模型由傳統描述性分析提升為主動式臨床決策支援工具。

5.7 結論

本研究首次整合九狀態連續時間隱藏狀態疾病自然史模型與人工智慧個人化風險預測模型，建立由群體疾病自然史推論延伸至個人化疾病進展之整合分析架構。透過 Observation Mapping、生成矩陣及個人化生成矩陣之建立，本研究不僅重建大腸直腸癌疾病自然史，更進一步描述不同個體疾病進展速度與疾病進展光譜，突破傳統疾病自然史模型僅能描述群體平均疾病進展之限制。

本研究所提出之 Hybrid Disease Progression Framework 兼具疾病機制解釋能力與人工智慧預測能力，不僅可作為未來風險分層篩檢、個人化篩檢間隔設計及臨床決策之理論基礎，更可望成為 AI Digital Twin 及智慧化精準公共衛生發展的重要方法學平台。未來隨著多體學資料、人工智慧及真實世界資料之持續整合，本研究所建立之架構將具有更廣泛之應用潛力，並可延伸至其他癌症及慢性疾病之疾病自然史研究，促進精準醫療與精準公共衛生之發展。

- 
1. Fearon, E.R. and B. Vogelstein, *A genetic model for colorectal tumorigenesis*. Cell, 1990. **61**(5): p. 759–67.
 2. Feinleib, M. and M. Zelen, *Some pitfalls in the evaluation of screening programs*. Arch Environ Health, 1969. **19**(3): p. 412–5.
 3. Cole, P. and A.S. Morrison, *Basic issues in population screening for cancer*. J Natl Cancer Inst, 1980. **64**(5): p. 1263–72.
 4. Chen, T.H.-H., et al., *Evaluation of a selective screening for colorectal carcinoma*. Cancer, 1999. **86**(7): p. 1116–1128.
 5. Chiu, Y.-H., et al., *Health information system for community-based multiple screening in Keelung, Taiwan (Keelung Community-based Integrated Screening No. 3)*. International Journal of Medical Informatics, 2006. **75**(5): p. 369–383.
 6. Yang, K.C., et al., *Colorectal cancer screening with faecal occult blood test within a multiple disease screening programme: an experience from Keelung, Taiwan*. J Med Screen, 2006. **13 Suppl 1**: p. S8–13.
 7. Chiu, H.-M., et al., *Long-Term Effectiveness Associated With Fecal Immunochemical Testing for Early-Age Screening*. JAMA Oncology, 2025. **11**(8): p. 846–854.
 8. Hsu, W.-F., et al., *Classifying interval cancers as false negatives or newly occurring in fecal immunochemical testing*. Journal of Medical Screening, 2021. **28**(3): p. 286–294.
 9. Chiu, H.M., et al., *Long-term effectiveness of faecal immunochemical test screening for proximal and distal colorectal cancers*. Gut, 2021. **70**(12): p. 2321–2329.
 10. Yen, A.M.-F. and H.-H. Chen, *Modeling the overdetected of screen-identified cancers in population-based cancer screening with the Coxian phase-type Markov process*. Statistics in Medicine, 2020. **39**(5): p. 660–673.
 11. Lin, T.-Y., et al., *Assessing overdiagnosis of fecal immunological test screening for colorectal cancer with a digital twin approach*. npj Digital Medicine, 2023. **6**(1): p. 24.
 12. Yen, A.M.-F. and H.-H. Chen, *Bayesian measurement-error-driven hidden Markov regression model for calibrating the effect of covariates on multistate outcomes: Application to androgenetic alopecia*. Statistics in Medicine, 2018. **37**(21): p. 3125–3146.
 13. Johnson, C.M., et al., *Meta-analyses of colorectal cancer risk factors*. Cancer Causes Control, 2013. **24**(6): p. 1207–22.

- 
14. Yen, A.M.-F., et al., *Precision Colorectal Cancer Fecal Immunological Test Screening With Fecal-Hemoglobin-Concentration-Guided Interscreening Intervals*. JAMA Oncology, 2024. **10**(6): p. 765–772.
 15. Chen, L.S., et al., *Baseline faecal occult blood concentration as a predictor of incident colorectal neoplasia: longitudinal follow-up of a Taiwanese population-based colorectal cancer screening cohort*. Lancet Oncol, 2011. **12**(6): p. 551–8.
 16. Rajkomar, A., J. Dean, and I. Kohane, *Machine Learning in Medicine*. N Engl J Med, 2019. **380**(14): p. 1347–1358.
 17. Nartowt, B.J., et al., *Robust Machine Learning for Colorectal Cancer Risk Prediction and Stratification*. Frontiers in Big Data, 2020. **Volume 3 - 2020**.
 18. Burnett, B., et al., *Machine Learning in Colorectal Cancer Risk Prediction from Routinely Collected Data: A Review*. Diagnostics (Basel), 2023. **13**(2).
 19. Lundberg, S.M. and S.-I. Lee, *A unified approach to interpreting model predictions*, in *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017, Curran Associates Inc.: Long Beach, California, USA. p. 4768–4777.
 20. Allman, E.S., C. Matias, and J.A. Rhodes, *Identifiability of parameters in latent structure models with many observed variables*. The Annals of Statistics, 2009. **37**(6A): p. 3099–3132, 34.
 21. Andersen, P.K., et al., *Introduction*, in *Statistical Models Based on Counting Processes*, P.K. Andersen, et al., Editors. 1993, Springer US: New York, NY. p. 1–44.