

中華民國第 63 屆中小學科學展覽會 作品說明書

高中組 電腦與資訊學科

佳作;探究精神獎

052513

運用疾病軌跡搭配深度學習偵測胰臟癌風險

學校名稱：臺北市立建國高級中學

作者： 高二 黃楷翔	指導老師： 王鼎中
---------------	--------------

關鍵詞：胰臟癌風險偵測、疾病軌跡、深度學習

摘要

目的：在台灣，胰臟癌雖非癌症發生數最高，但 85%為末期，五年存活率小於 5%，有癌王之稱。如何早期偵測胰臟癌風險，一直是醫療上的重大議題。本研究運用時間矩陣搭配深度學習進行大量變數之胰臟癌風險偵測。

方法：健保資料庫收集國人疾病、用藥等結構化資料，可藉此反映每人健康狀態。**將三年內診斷與藥物碼轉為時間矩陣**，以卷積神經網路訓練，訓練組與測試組比例為 9:1。

結果：共計案例組 1095 名及對照組 10950 名，訓練後測試組表現之 AUROC(area under the receiver operating characteristic curve)為 0.937，六十歲以上及以下 AUROC 分別為 0.846 及 0.897。

結論：結論：能將診斷、藥物、時間轉成矩陣，是**以疾病軌跡預測胰臟癌風險**，且能找到新特徵，未來搭配健康存摺，為**低成本、快速、大量的胰臟癌數位快篩**。

壹、前言

一、研究動機

癌症自 1982 至 2021 年皆為國人十大死因排行首位，以 2021 年為例，死亡率為每十萬人口 220.1 人，遠高於第二名心臟疾病的 93.1 人。胰臟癌 2021 年造成 2659 人死亡，且年增率為十大癌症中次高的 8.5% [1]。

胰臟癌佔台灣全部癌症個案數的 1.55%，而死亡人數佔全部癌症死亡個案數的 2.93%，男性發生率為每十萬人口 4.64 位，女性為每十萬人口 3.21 位 [2]。**胰臟癌雖非癌症的發生數的第一名，但 85%為末期癌症，同時末期的五年存活率小於 5%。**

目前偵測方法包括腫瘤標記 CA19-9 和 CEA，並以 X 射線、電腦斷層、超聲波等放射檢查作為主要診斷方法[2]。然而，血液中腫瘤標記對胰臟癌不具特異性，仍有待改善 [3]，且胰臟位置導致放射診斷難以早期發現腫瘤，發現時已擴散到其他區域，導致低存活率。若使用基因定序等更精準之檢測，會造成更高的醫療成本。

因此胰臟癌現階段的預後及治療效果不佳，最好的應對方式為**早期預防及治療**，對於高風險族群的篩檢相當重要，以便早期發現進行後續處置。

台灣的全民健保涵蓋了大多數人口，在 2021 年的西醫門診達 342,408,162 次（平均每人 14.9 次），其中包括胰臟癌的就診次數為 113,147 次 [4]。因此，健保資料庫是一個能夠反映健康的工具，並提供了包括診斷、藥物、開立天數及時間點等資訊。

表一為胰臟癌在各個分期的存活月數中位數[5]，可見**如果能將發現腫瘤的時間提早，能提高存活率**，並能省下醫療成本。如果再考慮病人罹癌時的工作、看護照顧成本，以及手術、復健等費用，若能有一個低成本的模型能做早期預測，對胰臟癌治療會有很大的幫助。此外，**預測模型也能進一步偵測高風險因子，包括目前醫療界尚未視為與胰臟癌相關之隱藏疾病**，進而早期介入治療，以更積極的角度降低胰臟癌的死亡率。

表一：胰臟癌分期、症狀與存活月數中位數

分期	狀況	存活月數中位數(月)
1A	癌腫瘤小於或等於 2 公分	24.1
1B	癌腫瘤 2~4 公分	20.6
2A	癌腫瘤大於四公分	15.4
2B	轉移至 1~3 個局部淋巴結	12.7
3	癌腫瘤侵犯到腹腔動脈幹、上腸繫膜動脈或肝總動脈，或轉移至超過 4 個局部淋巴結	10.6
4	有遠處轉移	4.5

二、研究目的

由於健保資料庫收集國人長期的疾病、用藥等項目，可以視為反映每人的健康狀態，本研究希望使用豐富的健保資料庫資料，將健保資料庫的診斷、藥物及時間資料，轉化成疾病軌跡，並運用深度學習的卷積神經網路(Convolutional Neural Network, CNN)為模組，**提供臨床一個低成本、非侵入式、高準確度的胰臟癌風險預測。**

針對高風險胰臟癌族群，建議進行使民眾盡早進行適當醫療檢驗，進行胰臟癌的診斷。協助未來醫療院所或健檢中心，透過民眾的健康存摺，進行胰臟癌數位快篩。

此外，針對胰臟癌在高年齡層發生率較高的特性，**本研究亦針對年齡層分開製作模型**，觀察在 60 歲以上（包含 60）及小於 60 歲（不包含 60）分別的預測效果。為了確認同時使用診斷及藥物代碼為適當的做法，本研究將診斷碼及藥物碼分別單獨使用進行預測並進行比較，開發預測 180 天及 360 天的癌症風險，讓模型的應用層面更廣。最後，**透過計算勝算比並 leave one out 觀察預測效能的下降程度**，了解重要特徵的風險，進而從更積極的角度銜接至臨床應用，觀察到此特徵就能介入胰臟癌偵測及治療，以科技輔助醫師，為智慧醫療的發展盡一份心力。

本研究的預期結果包含如下：

- （一）使用**診斷碼和藥物碼的時序資料**透過深度學習建立可預測胰臟癌之模型
- （二）比較**不同年齡層**預測效果之差異
- （三）比較僅使用診斷碼或藥物碼與同時使用之差異
- （四）比較預測**未來 180 天及 360 天**後效果之差異
- （五）透過**計算勝算比與 leave one out 驗證**選取胰臟癌之高風險因子

三、文獻回顧

本研究針對胰臟癌、類神經網路、病例對照研究、資料標準、胰臟癌預測進行文獻分析回顧。

（一）胰臟癌 (Pancreatic ductal adenocarcinoma)

本研究使用的是關於胰臟癌的診斷、治療和藥物代碼，因此需預先了解主要的症狀、診斷方式和治療方式。此外透過了解目前醫療界認定的風險因素，能了解最後透過模型中採用特徵與參數的合理性。

1. 跡象和症狀

胰臟癌是由胰臟細胞生長出來的惡性腫瘤。約 92%的胰臟癌為胰腺癌，源自胰臟的外分泌腺體，而胰臟癌發生在胰臟頭部約 70%，體部約 20%，尾部約 10%。

由於胰臟位置隱密，當腫瘤開始生長時，可能**早期症狀不明顯或出現與上腸胃道疾病類似症狀，極可能為人們所疏忽**，常在癌細胞擴散導致上腹痛、背痛、黃疸、體重減輕和消化不良等症狀時，才就醫診斷出來。

若癌症發生於胰臟頭部，則癌細胞有可能會侵犯到膽管而導致膽汁無法流至十二指腸，導致皮膚及眼角膜變黃、尿液呈現深茶色或是皮膚發癢，嚴重者甚至會有發燒等症狀[2]。腫瘤指數又稱血清檢查，包括 CA-199 和 CEA。然而，腫瘤指數不具特異性，除了胰臟癌，胃癌、肝癌也能檢測到 CA-199 升高，臨床上無法使用 CA-199 為胰臟癌的診斷，只能作為初步篩檢。

2. 風險因素

胰臟癌的風險因子包括吸菸、異常代謝、肥胖和幽門螺旋桿菌。一般以 60 歲以上為高風險族群，以及吸菸、喝酒等習慣。糖尿病的病人因為接觸針對胰臟的化學藥品，也有胰臟癌的風險[2]。

3. 診斷方法

接受上消化道檢查排除食道和胃腸問題後，可以接受腹部超音波或電腦斷層等放射診斷[2]。

4. 治療

最常見的治療方式是胰十二指腸切除術，又稱 Whipple procedure。此外，膽道繞道手術及腹腔神經叢阻斷也是能改善胰臟癌的治療，大約有 5%到 22%的病人腫瘤是可切除的。化療、放射治療和標靶治療皆為目前發展的治療方向，**若能提早發現，便能提升治療的效果**[2]。

(二) 病例對照研究 (Case-control study)

病例對照研究是研究健康問題發生原因之研究方法。透過罹患特定疾病的病人作為病例組 (cases)，另一組沒有罹患其疾病的民眾為對照組(control)，兩者進行匹配(matching)，此研究方法的優點為適用發生率低的事件，例如胰臟癌。研究中可藉由兩者之間變數差異（如年齡、性別、疾病史及藥物開立史），探索與疾病發生的關聯[6]。

(三) 類神經網路

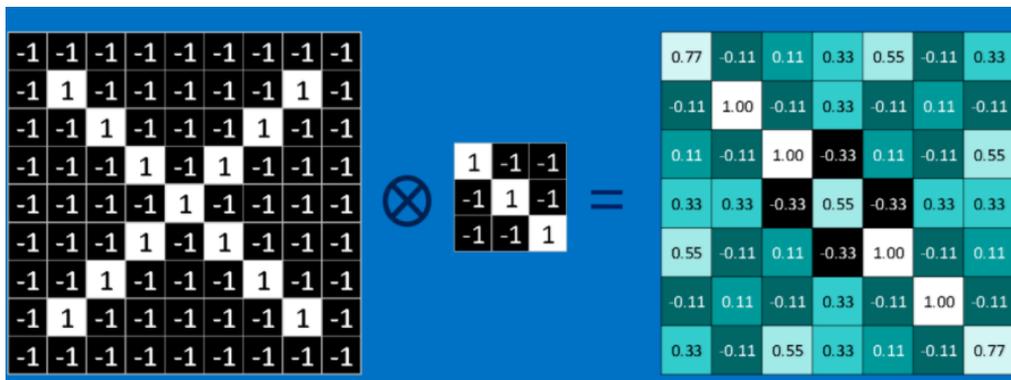
類神經網路的基本元素為神經元，連線在一起為神經網路。

$$t = f(\overline{W}'\vec{A} + b)$$

其中 \overline{W}' 為權重向量的轉置， \vec{A} 為輸入向量， b 為偏置值(bias)， f 為激勵函式(activation function)。將輸入與權重兩向量內積後，經過一個非線性函式得到結果。若無激勵函式，輸出會呈現 0 和 1 的 step function，透過 sigmoid、tanh、或 ReLU 可呈現較和緩的變化。

卷積神經網路(CNN)將輸入資料(特徵)以特徵過濾器(filter)做卷積運算，如下圖一，如果和特徵過濾器(filter)有相符的特徵，運算後會有較大的結果值，若無則結果值較小[7]。

特徵過濾器可運算出特徵圖(feature map)，經過多個 epoch 訓練並修正，使 loss 達到最小，即能獲得理想模型所使用的特徵過濾器。



圖一、特徵過濾器做卷積運算 [7]

卷積神經網路通常使用在二維影像處理上，本研究為了同時考量時間和疾病、藥物的二維特徵，將診斷、藥物碼和時間軸處理成二維的疾病軌跡，再利用 CNN 運算，成功達到預測效果。

(四) 國際疾病分類 ICD-9-CM (International Classification of Diseases, ICD)

國際疾病分類碼是關於疾病、症狀與異常等分類的代碼，是健保資料採用之**疾病標準代碼**。台灣於 1992 年始用 ICD-9-CM，而 2015 後逐步轉為 ICD-10-CM，兩個版本間可互相對照。ICD-9-CM 共分為 18 大類，如下表二所示，而本研究採用的前三碼可大致分類診斷，後面的一碼或二碼可更詳細描述疾病。

表二、ICD 編碼分組與疾病對照表

組別	編碼	疾病
1	001-139	傳染病和寄生蟲疾病
2	140-239	贅生物 55(腫瘤)
3	240-279	內分泌、營養、新陳代謝及免疫系統疾病
4	280-289	血液及造血器官疾病
5	290-319	精神失常
6	320-359	神經系統疾病
7	360-389	感覺器官疾病
8	390-459	循環系統疾病
9	460-519	呼吸系統疾病
10	520-579	消化系統疾病
11	580-629	泌尿生殖系統疾病
12	630-676	妊娠、分娩和產後合併症
13	680-709	皮膚及皮下組織疾病
14	710-739	肌肉骨骼系統及結締組織疾病
15	740-759	先天畸形
16	760-779	產期引起的特定情況
17	780-799	症候、徵候及不明情況
18	800-999	受傷及中毒
19	E	外傷
20	V	補充分類

(五) 癌症風險預測或偵測

1. 癌症預測系統（以 QCancer 為例）

QCancer 使用英國的家醫科聯合會的資料庫，以性別、年齡、身高體重、生活型態及疾病史等變因，計算 11 種癌症的風險高低。針對胰臟癌的 AUROC 約為 0.855 至 0.875，不過風險為絕對風險值，不能比較高低風險。

Cancer	Type	Risk
No cancer		48.46%
Any cancer		51.54%
	colorectal	18.68%
	other	13.06%
	pancreatic	10.62%
	blood	3.57%
	gastro-oesophageal	2.38%
	lung	1.63%
	prostate	1.42%
	renal tract	0.16%
	testicular	0%

圖二、QCancer 使用介面

2. 以電子病歷預測胰臟癌之實例 [8]

2023 年 5 月 Davide Placido 及團隊使用美國及丹麥電子病歷資料預測胰臟癌風險，並使用丹麥 (Danish National Patient Registry, DNPR) 六百萬位病人資料，其中約 2.4 萬個病人有胰臟癌的紀錄。他們嘗試了不同的機器學習方法，如 GRU、Transformer 等，並測試 3 個月到 60 個月後預測胰臟癌的準確程度。在 36 個月後的風險預測結果模型達到了 AUROC=0.88。最後將模型應用在陌生的資料集，並得到 AUROC 略為下降的表現，經重新訓練後效果更好。

3. 運用深度學習於時序資料進行預測疾病之實例

利用深度學習以診斷及用藥歷史預測罹癌風險-以肝癌為例 [9]

此文獻使用健保資料庫中疾病診斷結果及長期藥物開立時序資料以 CNN 和 MLP 類神經網路訓練，並以逐步抽取變數觀察準確性損失法、隨機森林法及風險係數得知各變數之重要性，對於提早 0.5 至 3 年預測的 AUROC 可達 0.880~0.917。此方法可使用疾病或藥物有或無的資訊以及時間資訊，考量遠近和嚴重程度，進行肝癌預測。本研究參考其二維表作法，可考慮時序性。

藉由上述相關文獻研究，得知以電子病歷及機器學習預測胰臟癌等癌症為可行的做法，若能加入更多變因測試模型的不同效果，並嘗試將時間當作特徵，並採用 CNN 訓練，可能會有不同的結果。本研究參考此相關文獻將時序性資料製成疾病軌跡，並加以探討僅用診斷、藥物碼，另外針對胰臟癌年齡層普遍較高的特性，探討不同的年齡層預測的效果是否有提升。

貳、研究設備及器材

一、程式軟體環境：

前處理：SQL, Microsoft SQL Server Management Studio (SSMS)

機器學習：Python 3.9.12, Anaconda 4.12.0, Tensorflow 2.10.0

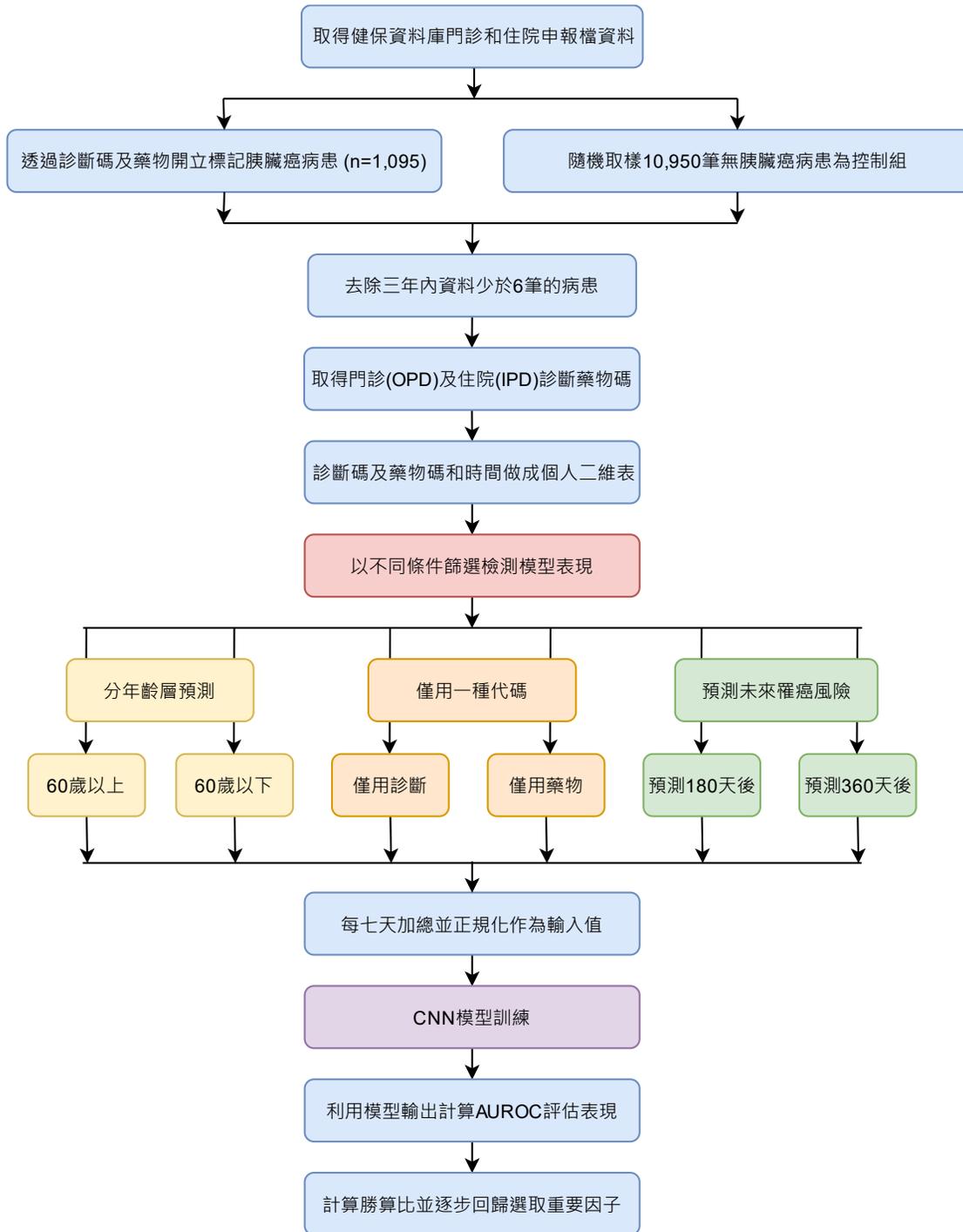
二、硬體配備：Intel i7-1185G7 @ 3.00GHz CPU、16GB RAM、GeForce GTX 1650 Ti

三、研究資料庫：全民健康保險研究資料庫 (NHIRD)，採用 1999 至 2013 間的資料，包含門診處方及治療明細檔 (CD)、門診處方醫令明細檔 (OO)

參、研究過程或方法

一、 研究架構

下圖三為本研究流程圖，經資料清洗、病人選取後，進行預測分析。



圖三、研究架構圖

本研究使用全民健康保險研究資料庫（NHIRD）中的門診處方及治療明細檔（CD）及門診處方醫令明細檔（OO），在門急診和住院的 200 萬人抽樣檔中取得 1999 到 2013 的資料，並使用去識化 ID、出生日期、性別、就診日期、ICD-9 代碼、藥物代碼、給藥天數做資料前處理。本研究使用 Microsoft SQL Server Management Studio (SSMS) 進行資料庫檢索，以 SQL 進行資料前處理。

```

SELECT TOP (1000) [id], [FUNC_DATE], [ID_BIRTHDAY], [ACODE_ICD9_1]
, [ACODE_ICD9_2], [ACODE_ICD9_3], [ID_SEX]
FROM [DB_2M].[dbo].[cd_2002]

```

	id	FUNC_DATE	ID_BIRTHDAY	ACODE_ICD9_1	ACODE_ICD9_2	ACODE_ICD9_3	ID_SEX
1	1278557	2002	194	4659			F
2	1150005	2002	196	37021			F
3	1509770	2002	196	4659			F
4	1350112	2002	197	4659			M
5	1394064	2002	196	460			M
6	1052426	2002	198	64003			F
7	1854871	2002	196	460			M
8	1200180	2002	194	460			M
9	1439662	2002	199	4661			M
10	1278605	2002	198	4659			F
11	1092383	2002	196	6280			F
12	1537488	2002	195	8489			F
13	1299305	2002	196	7242			F
14	1742956	2002	193	8489			F

圖四、SSMS 檢索畫面（選取 2002 年 CD 檔中 ID、就診日期、出生日期、ICD9 代碼、性別）

二、 標記病人(Label)

為避免錯誤診斷的申報，本研究判斷胰臟癌病人的標準包括：

- （一）登記於全民健保**重大傷病檔**
- （二）**住院的主診斷**為胰臟癌
- （三）**有癌症診斷**且接受**癌症相關藥物**，或接受**癌症相關處置**，或接受**癌症相關治療**。

納入條件：年齡大於 20 歲、胰臟癌前 3 年內有六次診斷以上。

排除條件：年齡小於 20 歲

下表為本研究使用的胰臟癌病例組及對照組病人的基本資料及平均每年每人出現各個診斷碼及藥物碼的次數，並以比值向下排列，可得知資料分布。經條件篩選得到 1095 筆罹患胰臟癌之病人，與 10950 筆（採 1:10）對照組病人，共 12,045 筆。下表僅採用出現多於 20 次之資料，避免某些代碼出現過少導致比值過大的狀況。

表三、常見 ICD-9 診斷碼、常見 ATC Code 藥物碼

	胰臟癌 (n=1,095)	Control (n=10,950)	All (n=12,045)
平均年齡 (標準差)	63.3 (14.7)	46.1 (16.6)	47.6 (17.2)
60 歲以上人數 (%)	654 (59.7%)	2376 (21.7%)	3030 (25.2%)
男性 (%)	585 (53.4%)	5046 (46.1%)	5631 (46.7%)
總診斷次數	141,176	775,557	916,733
常見 ICD-9 診斷 次/年			比值
577 (胰臟疾病)	0.2892	0.0055	52.6
235 (性態未明之腫瘤)	0.0381	0.0016	23.8
576 (膽道之其他疾患)	0.0871	0.0047	18.5
251 (胰臟內分泌疾患)	0.0201	0.0016	12.6
239 (性質未明之腫瘤)	0.0886	0.0086	10.3
560 (腸及腹膜之疾病)	0.0496	0.0057	8.7
常見 ATC Code 藥物碼 次/年			
A10BA02 (Metformin)	0.9909	0.2887	3.4
B01AC06 (Aspirin)	1.2977	0.4067	3.2

三、資料前處理

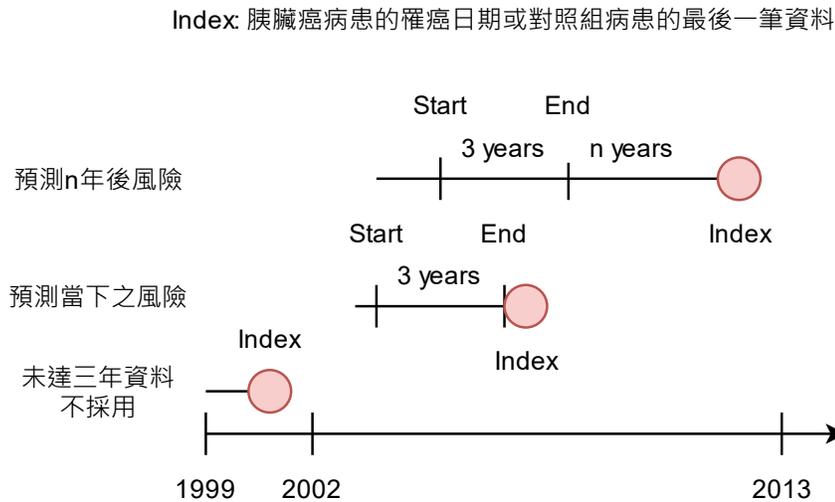
首先將取得的病例組病人和對照組病人去識化 ID 和資料庫的診斷碼及藥物碼比對，取得資料。

因為**病例組病人的平均年齡大於對照組病人**，本研究加入大於等於六十歲或小於六十歲的條件，**針對不同族群做更精準的預測**。本研究也嘗試僅使用診斷碼或藥物碼進行預測觀察效果。

胰臟癌病人確診時間為**標示日**及**觀察終止日**，對照組則以最後一筆資料為**標示日**及**觀察終止日**，而三年前即為**觀察起始日**。本研究採取三年為觀察期(Observation time)，而本研究使用的是**起始日**和**終止日**組成觀察窗(Observation window)中的資料。

若要預測 180 天或 360 天後的癌症風險，則改成使用 180 天或 360 天前的數據為**觀察終止日**，此時有些病人資料量不足，不採用。**觀察起始日**為**觀察終止日**前三年。

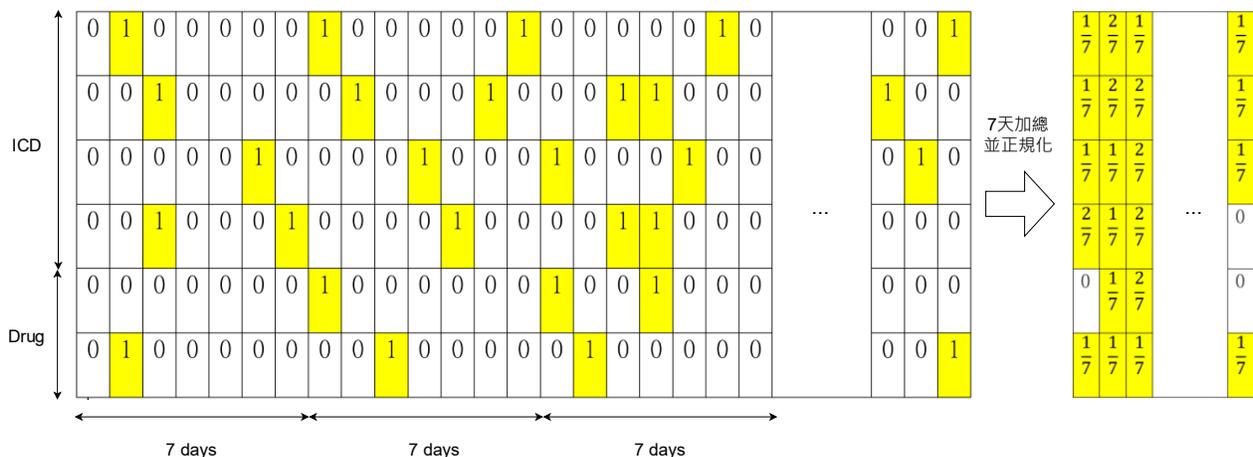
觀察期需要有夠多的病人可供預測，也能讓每個人分別有夠多資料。觀察期較長則資料長度符合觀察期的病人越少，但每個人每次能選到較多資料。健康存摺一次能提供三年資料，因此模型能配合健康存摺使用，可以直接將資料輸入進模型使用。



圖五、資料前處理示意圖

本研究以 ICD-9-CM 的前三碼代表診斷，再加上藥物碼作為「**疾病藥物特徵值軸**」，並將每筆診斷發生的時間（天數）除以七，作為時間軸的數值。在「疾病藥物-時間」二維表中，每當某病人在某週有一次某個診斷或藥物開立，即在該欄位設為一分，為期三年。

訓練模型前，因為每天最多出現一次同樣的診斷碼，所以須將每個數值除以 7 做**正規化**，轉化成 0 到 1 之間的值，最大為 1，最小為 0，讓**每項變因的輸入值為所有人對於同個變因的頻率**。



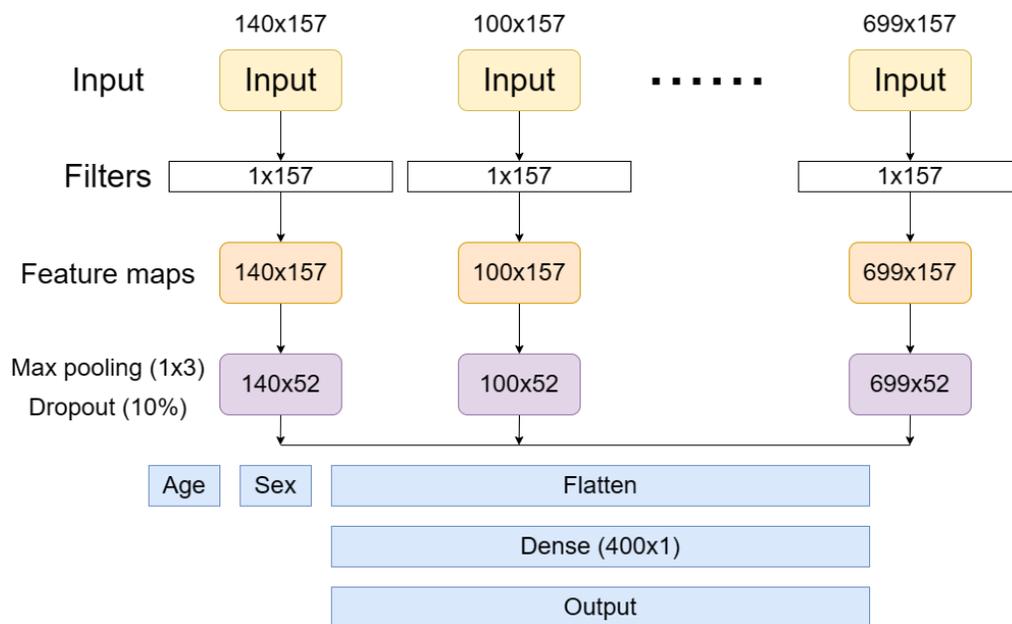
圖六、個人二維表及正規化示意圖

最後，將選取出來的資料（包含病例組對照組）分別以 9:1 的比例分為訓練和測試資料。

四、類神經網路訓練

將前處理後的資料，**加上年齡和性別做為特徵**，輸入至以 Tensorflow 為框架的**卷積神經網路 (CNN)**訓練。每個病人的二維表因為資料量較大，以診斷碼分組輸入模型中，採 1x157 大小的特徵過濾器過濾出特徵圖，採取 ReLU 激活函數，此為第一層（卷積層）。第二層採用 1x3 大小的 Max pooling 層，取得顯著特徵。第三層為 Dropout 層，不納入 10%的數據，避免模型的過度擬合。第四層攤平上一層的輸出，並將年齡和性別等資料納入。第五層為全連接層(Fully Connected Layer, Dense Layer)，最後以 Softmax 函數輸出。

選擇 Epoch=20、Batch-size=16 為參數訓練，每 16 筆資料更新一次權重，並將所有資料訓練 20 次。編譯器(Compiler)：優化器(Optimizer)為 Adam、損失函數(Loss function)為二值交叉熵(Binary cross-entropy)、Metrics(成效衡量方式)為 Accuracy。



圖七、CNN 深度學習模型架構

本研究採取 **5-fold cross validation**，將資料亂數排列後分成五部分，選擇其中四個部分訓練，使用最後一部份為驗證，並重複五次，可避免依賴固定的訓練及測試資料產生偏差，也能防止過度擬合。最後再以**測試資料輸入**得到 score，可以理解為得到癌症的相對風險。取 5-fold 以後五個 score 的平均值再以 AUROC 及敏感度、特異度、PPV、NPV 等數值為判斷效能指標。

五、 重要因子選取

為了選取出對於模型預測貢獻較大的重要因子，進而選取出胰臟癌的風險因子，並進一步進行早期介入治療，本研究透過**勝算比計算及 leave one out 驗證進行重要因子選取**。

勝算比的計算是由「出現此特徵的病例組病人數」乘上「沒有出現此特徵的對照組病人數」，除以「沒有出現此特徵的病例組病人數」乘上「有出現此特徵的對照組病人數」的結果，相當於觀察胰臟癌和該特徵同時出現或同時不出現的頻率，可衡量因子的重要性，如下表所示。

表四、勝算比(odds ratio)

	病例組病人	對照組病人
有此特徵	a 人	b 人
沒有此特徵	c 人	d 人

勝算比(odds ratio)即為 $\frac{ad}{bc}$ ，可衡量特徵的重要性

採用勝算比最高的十個特徵和最低的十個特徵，每次抽取掉一個特徵不採用並訓練模型及驗證，再以原本使用完整資料集的模型比較效果，能衡量特徵的重要性，也就是**移除該特徵後損失的預測能力**，此方法為 **leave one out 驗證**。

為了避免較少出現的診斷對勝算比產生很大的影響，本研究**不採用過去三年在少於十人**的資料中出現的特徵。

透過比較勝算比最高的十個特徵，可了解與胰臟癌呈**正相關**的十個特徵。

透過比較勝算比最低的十個特徵，可了解與胰臟癌呈**負相關**的特徵，也就是擁有這些特徵的病人較不易出現胰臟癌的診斷。

以兩種方式衡量重要因子，可互相參考，並選擇出與胰臟癌相關程度最高的風險因子。

肆、研究結果

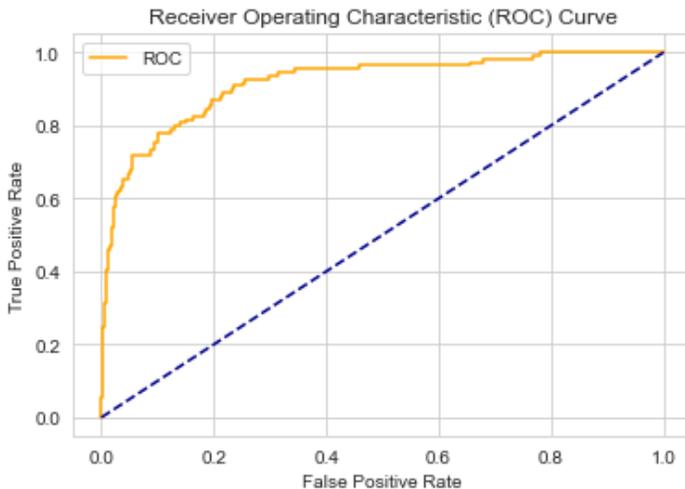
一、以診斷碼及藥物碼預測胰臟癌的結果

以下為部分**測試資料**輸入後的測試結果。從左邊開始是去識化 id、分類 grp（其中有胰臟癌是 1、對照組為 0）、年齡 age、性別 sex、及將測試資料輸入模型得到的 Score 與平均。

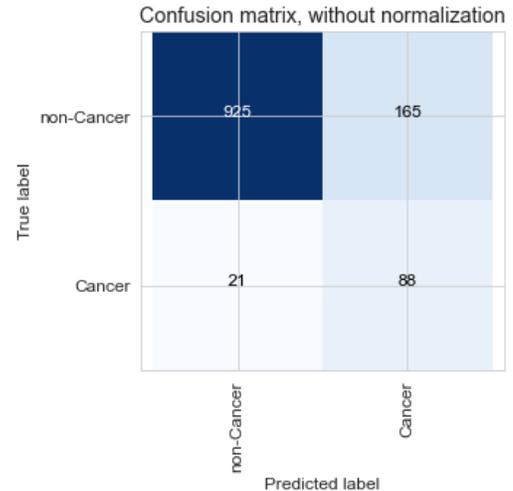
表五、部分測試資料之測試結果

id	grp	age	sex	average	Score	Score	Score	Score	Score
3246	1	52	1	0.695266	0.968334	0.575949	0.516409	0.832838	0.582802
3494	0	37	1	0.105275	0.005158	0.134026	0.292314	0.026124	0.068753
4862	0	63	0	0.159318	0.002464	0.327099	0.157202	0.002763	0.307061
4881	1	84	1	0.311423	0.013523	0.197405	0.508503	0.271891	0.565794
4920	0	33	1	0.043766	0.002202	0.070322	0.125161	0.00717	0.013974
4976	0	26	0	0.000683	3.41E-05	0.000401	0.000308	4.98E-05	0.002621
5440	1	41	0	0.005977	7.14E-08	0.00034	0.001009	1.85E-05	0.028517
6795	0	55	1	0.086431	0.000654	0.007848	0.031472	0.012519	0.379661
7763	0	47	0	0.020814	0.044147	0.006821	0.026212	0.006056	0.020833
9373	0	40	0	0.147336	0.003622	0.383822	0.022854	0.115027	0.211356
10056	0	73	0	0.734358	0.992896	0.600473	0.626906	0.842158	0.609356
10628	0	25	0	0.034957	0.009995	0.031359	0.035806	0.017087	0.080536
11329	0	39	0	0.023401	0.006838	0.041854	0.038851	0.003179	0.026285

Score 可以理解為得到癌症的風險，因為 Score 通常較接近 0，因此須觀察不同閾值(threshold)下的模型表現，大於閾值為 1（有胰臟癌）否則為 0。最後**計算不同閾值下的真陽率、偽陽率繪製 ROC 曲線**，將**實際結果 (ground truth) 和預測出的結果比較製成混淆矩陣**。



圖八、使用診斷碼及藥物碼預測之 ROC 曲線



圖九、用診斷及藥物碼預測之混淆矩陣

若預測為陽性而實際上為陽性，即為**真陽性** (true positive, tp)。

若預測為陽性而實際上為陰性，即為**偽陽性** (false positive, fp)。

若預測為陰性而實際上為陽性，即為**偽陰性** (false negative, fn)。

若預測為陰性而實際上為陰性，即為**真陰性** (true negative, tn)。

以上四種數據構成混淆矩陣。採取真陽率除以偽陽率最大的點為效果最好的閾值，並計算敏感度、特異度、PPV、NPV、accuracy 及 f1-score。

其中，**PPV (陽性預測值)** 為 $tp/(tp+fp)$ (precision 精準度)，預測陽性中實際上陽性的比率。

NPV (陰性預測值) 為 $tn/(tn+fn)$ ，預測陰性中實際上陰性的比率。

Sensitivity(敏感度) 為 $tp/(tp+fn)$ (recall 召回率)，實際上為陽性中預測正確的比率。

Specificity(特異度) 為 $tn/(tn+fp)$ ，實際上為陰性中預測正確的比率。

F1-score 為 2 倍的 precision 乘以 recall 再除以 precision+recall

Accuracy 為預測正確的結果，即為 $(tp+tn)/$ 全部樣本數。

support 為各樣本數，即顯示測試資料中有 109 筆癌症病人，1090 筆對照組病人，共 1199 筆。

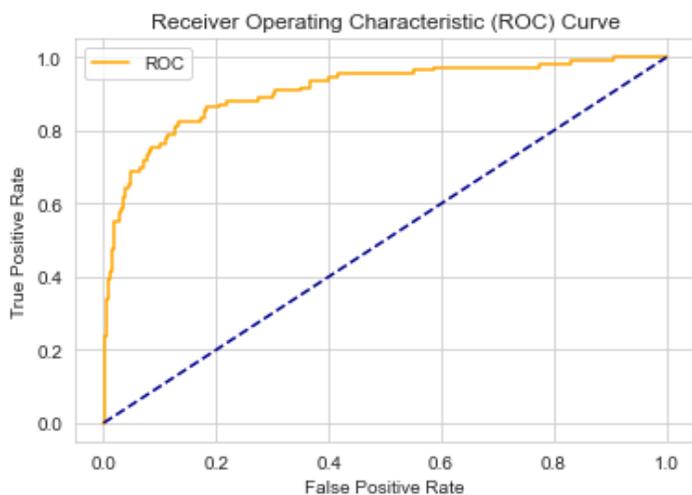
AUROC 即為 ROC 曲線下的面積，能客觀的觀察各種閾值下的模型效能，面積大則真陽率大於偽陽率，模型表現佳。

表六、使用診斷碼及藥物碼預測之各項效能

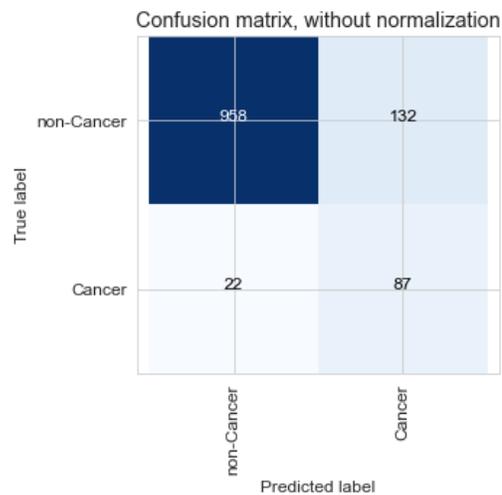
AUROC	Sensitivity	Specificity	F1-Score	Accuracy	PPV	NPV
0.937	0.825	0.925	0.641	0.916	0.523	0.982

二、僅用診斷碼或僅用藥物碼預測胰臟癌的結果

(一) 僅用診斷碼為輸入：（測試資料為 109 筆癌症病人、1090 筆對照組病人）

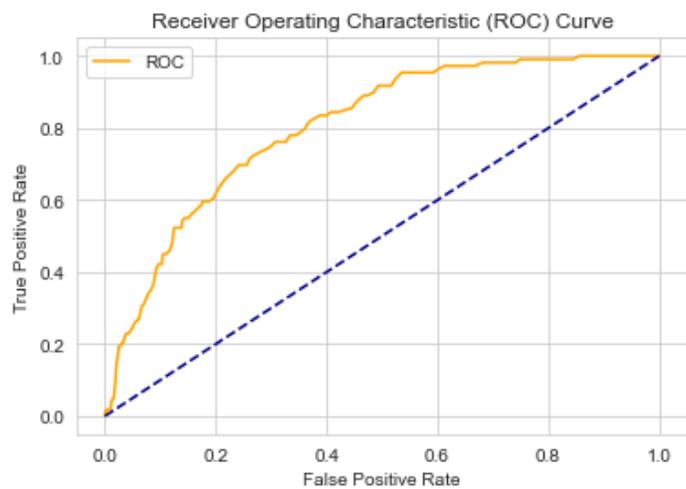


圖十、使用診斷碼預測之 ROC 曲線

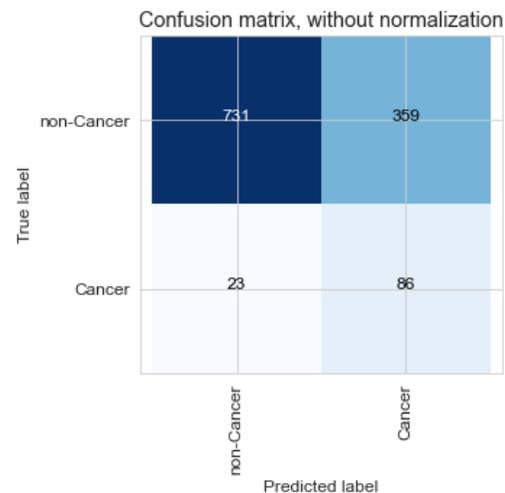


圖十一、使用診斷碼預測之混淆矩陣

(二) 僅用藥物碼為輸入：（測試資料為 109 筆癌症病人、1090 筆對照組病人）



圖十二、使用藥物碼預測之 ROC 曲線



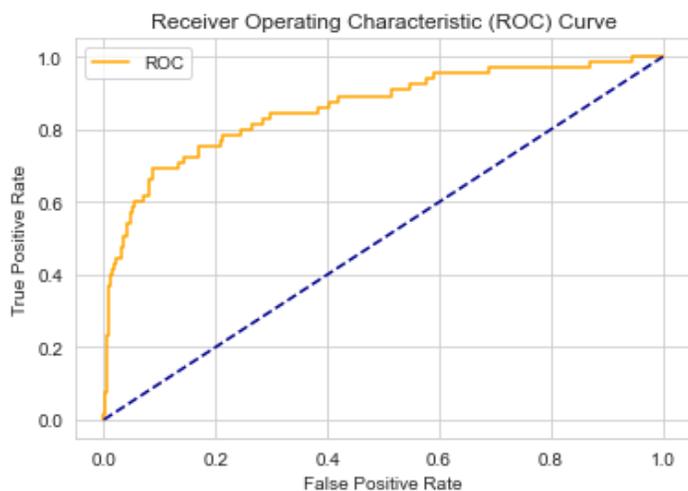
圖十三、使用藥物碼預測之混淆矩陣

表七、僅使用診斷碼或藥物碼預測之各項效能

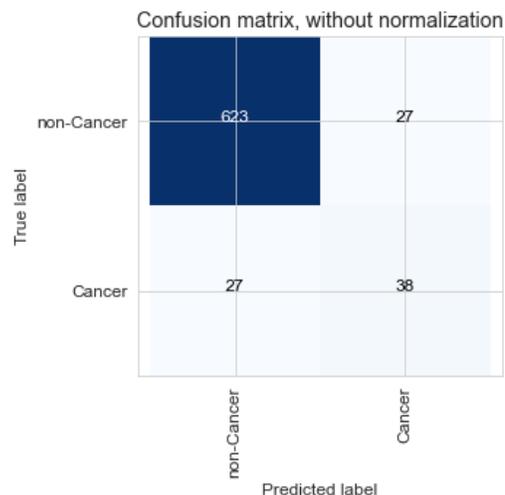
組別	AUROC	Sensitivity	Specificity	F1-Score	Accuracy	PPV	NPV
診斷	0.928	0.817	0.881	0.543	0.875	0.406	0.980
藥物	0.901	0.725	0.933	0.605	0.914	0.520	0.971

三、分成不同年齡層預測胰臟癌的結果

(一) 60 歲以上 (包含 60) : (測試資料為 65 筆癌症病人、650 筆對照組病人)

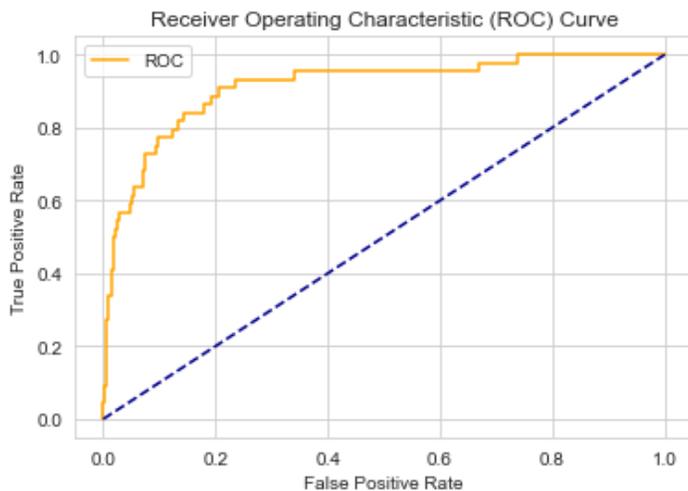


圖十四、60 歲以上預測之 ROC 曲線

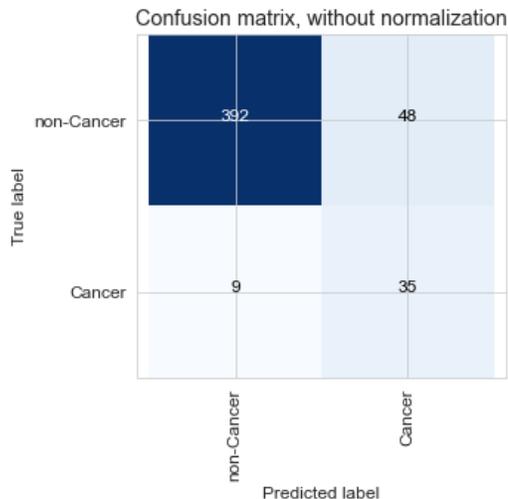


圖十五、60 歲以上預測之混淆矩陣

(二) 60 歲以下 (不包含 60) : (測試資料為 44 筆癌症病人、440 筆對照組病人)



圖十六、60 歲以下預測之 ROC 曲線



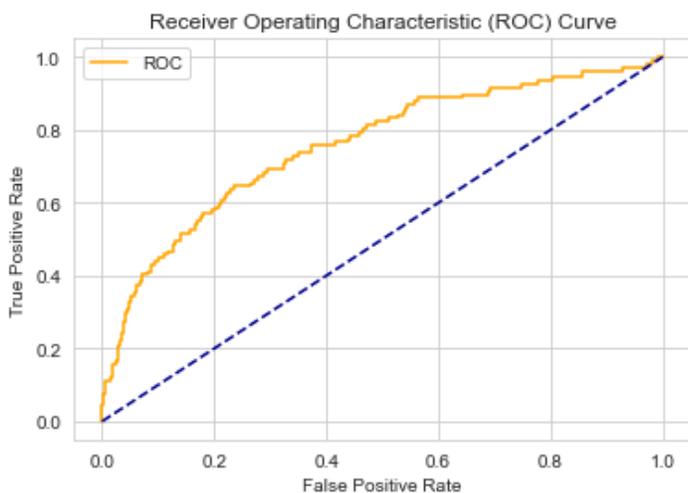
圖十七、60 歲以下預測之混淆矩陣

表八、分年齡層預測之各項效能

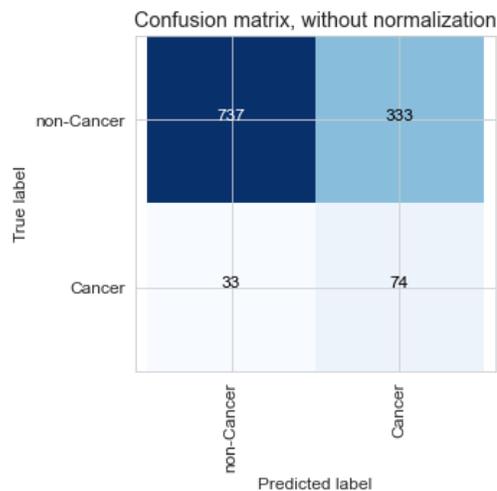
組別	AUROC	Sensitivity	Specificity	F1-Score	Accuracy	PPV	NPV
60 以上	0.846	0.600	0.952	0.578	0.920	0.557	0.960
60 以下	0.897	0.750	0.921	0.589	0.905	0.485	0.974

四、預測未來不同時間點胰臟癌風險的結果（部分病人資料量不足未採用）

（一）180 天（測試資料為 107 筆癌症病人、1070 筆對照組病人）

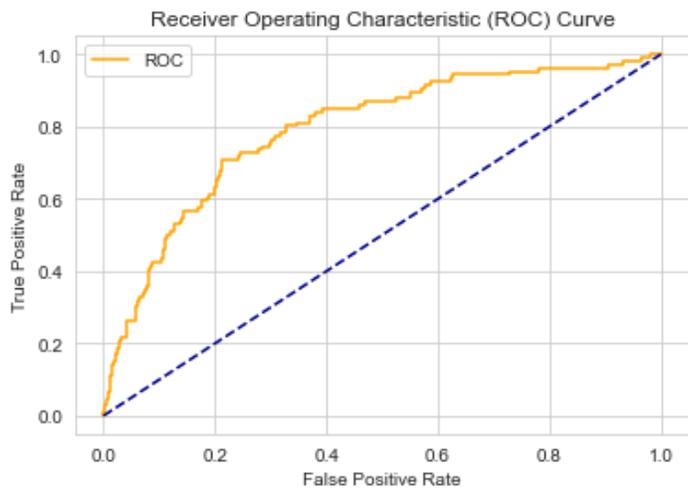


圖十八、180 天後預測之 ROC 曲線

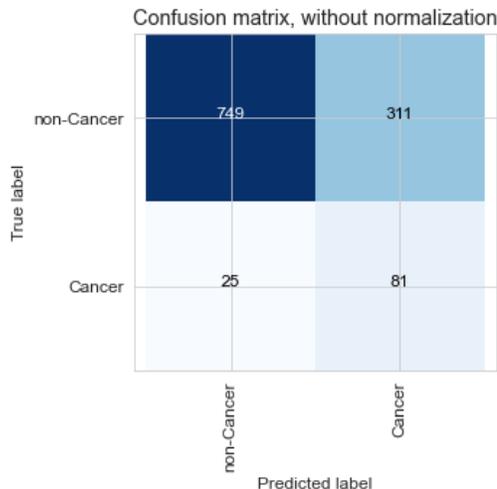


圖十九、180 天後預測之混淆矩陣

（二）360 天（測試資料為 106 筆癌症病人、1060 筆對照組病人）



圖二十、360 天後預測之 ROC 曲線



圖二十一、360 天後預測之混淆矩陣

表九、預測 180 天及 360 天後之各項效能

組別	AUROC	Sensitivity	Specificity	F1-Score	Accuracy	PPV	NPV
180 天	0.830	0.701	0.748	0.332	0.743	0.217	0.962
360 天	0.837	0.528	0.916	0.446	0.881	0.386	0.951

五、重要因子選取 (leave one out)

採用勝算比最高和最低的特徵，移除該特徵並使用剩下資料訓練，觀察 AUROC 下降。

(一) 勝算比最高之診斷碼 (標紅字為**消化疾病**)：

表十、勝算比最高之診斷碼、代表疾病、勝算比及 AUROC 下降量

診斷碼	代表診斷 [12]	勝算比	AUROC 下降 (%)
577	胰臟疾病	47.036	0.0156 (1.71%)
576	膽道之其他疾患	22.828	0.0155 (1.70%)
235	性態未明之腫瘤	16.500	0.0044 (0.48%)
239	性質未明之腫瘤	12.697	0.0136 (1.48%)
570	急性及亞急性肝壞死	10.091	0.0075 (0.81%)
574	膽囊結石	9.036	0.0133 (1.46%)
572	肝膿瘍及慢性肝病之後遺症	8.537	0.0033 (0.36%)
575	膽囊之其他疾患	7.935	0.0090 (0.98%)
532	急性十二指腸潰瘍	7.532	0.0067 (0.74%)
251	胰臟內分泌之其他疾患	6.794	0.0153 (1.67%)

(二) 勝算比最低之診斷碼 (標紅字為**60歲以下或懷孕**較容易出現之診斷)：

表十一、勝算比最低之診斷碼、代表疾病、勝算比及 AUROC 下降量

診斷碼	代表診斷 [12]	勝算比	AUROC 下降 (%)
V22	正常妊娠	0.242	0.0591 (6.45%)
626	月經及其他女性生殖道異常出血	0.298	0.0172 (1.89%)
520	牙齒發育及萌發之疾患	0.372	0.0127 (1.38%)
625	與女性生殖器官有關之疼痛及其他	0.376	0.0176 (1.92%)
706	皮脂腺疾病	0.437	0.0108 (1.18%)
524	齒及面部異常，包括咬合不正	0.481	0.0110 (1.20%)
220	卵巢良性腫瘤	0.512	0.0233 (2.54%)
463	急性扁桃腺炎	0.576	0.0152 (1.66%)
705	汗腺疾患	0.588	0.0185 (2.02%)
461	急性鼻竇炎	0.592	0.0192 (2.10%)

(三) 勝算比最高之藥物碼 (標紅字為**消化或胰臟相關疾病**可能出現之藥物)

表十二、勝算比最高之藥物碼、適應症、勝算比及 AUROC 下降量

藥物碼	適應症 [13]	勝算比	AUROC 下降 (%)
N02AC	輕度到中度疼痛之緩解	13.657	0.0126 (1.38%)
A03AX95	緩解膽道疾病、胰臟炎引起的痙攣性症狀	12.268	0.0180 (1.96%)
A05AA	膽結石之溶解、原發性膽道肝硬化	10.642	0.0062 (0.67%)
A11BA	多種維生素，單方	10.091	0.0127 (1.38%)
C04AX02	周邊循環障礙、糖尿病性壞疽等	10.091	0.0171 (1.87%)
A11AA	多種維生素與礦物質	8.363	0.0069 (0.75%)
A11JC	維生素，其他複方	7.886	0.0026 (0.29%)
A07XA01	止瀉藥、腸道抗炎、抗感染	7.098	0.0275 (3.00%)
H05BA	高血鈣、骨骼的帕哲特氏病	6.935	0.0156 (1.71%)
A05BA	維護肝臟正常功能	6.895	-0.0016 (-0.18%)

(四) 勝算比最低之藥物碼 (標紅字為**60歲以下或懷孕**較容易出現之藥物)：

表十三、勝算比最低之藥物碼、適應症、勝算比及 AUROC 下降量

藥物碼	適應症 [13]	勝算比	AUROC 下降 (%)
G02AB	分娩時的子宮弛緩性出血或產後出血	0.281	0.0026 (0.28%)
D10AF	尋常性痤瘡	0.382	0.0001 (0.01%)
G03DC	需黃體荷爾蒙治療的婦科疾患、子宮出血	0.412	0.0106 (1.16%)
D10AE	尋常性痤瘡	0.514	0.0038 (0.42%)
G03DA	停經婦女雌激素不足所引起之症狀	0.534	0.0158 (1.72%)
R06AX26	季節性過敏性鼻炎及慢性自發性蕁麻疹	0.592	0.0092 (1.00%)
D07XC01	皮膚炎、皮膚過敏	0.595	0.0133 (1.45%)
G01AX03	子宮頸糜爛、白帶、陰道炎、子宮頸炎	0.640	0.0132 (1.44%)
R03CA	支氣管性氣喘、過敏性鼻炎	0.640	0.0068 (0.74%)
D10AD	尋常性痤瘡	0.646	0.0104 (1.14%)

伍、研究討論

上述研究成果彙整為表如下：

表十四、研究結果彙整

	AUROC	Sensitivity	Specificity	F1-Score	Accuracy	PPV	NPV
診斷、藥物	0.937	0.825	0.925	0.641	0.916	0.523	0.982
僅用診斷	0.928	0.817	0.881	0.543	0.875	0.406	0.980
僅用藥物	0.901	0.725	0.933	0.605	0.914	0.520	0.971
60 歲以上	0.846	0.600	0.952	0.578	0.920	0.557	0.960
60 歲以下	0.897	0.750	0.921	0.589	0.905	0.485	0.974
180 天以後	0.830	0.701	0.748	0.332	0.743	0.217	0.962
360 天以後	0.837	0.528	0.916	0.446	0.881	0.386	0.951

- 一、 使用診斷碼及藥物碼預測胰臟癌是否發生，可得 AUROC 約為 0.937。
- 二、 僅使用診斷碼或藥物碼預測疾病時，可觀察到兩者的 AUROC 皆低於同時使用兩者來預測疾病的效果。藥物碼預測效果較差，AUROC 約為 0.901，推測是因為不同健康狀況可對應到同一種藥物，因此藥物僅能部分表現當下健康狀況，而診斷碼較能對應到開立當下的健康狀態，預測效果較好，AUROC 為 0.928。
- 三、 依照年齡層預測時，可觀察到 60 歲以上（含 60）的分組中，胰臟癌預測效果較差，AUROC 約為 0.859。推測是因為在同樣的年齡層當中，得到長期疾病的種類較為類似，因此增加分類的難度。另外，年齡也是預測疾病的特徵之一，是在分年齡層後預測能力下降的原因。60 歲以上的 PPV 較高，是因為盛行率提高會使 PPV 上升。
- 四、 在 60 歲以下（不含 60）的分組中，可觀察到 AUROC 較 60 歲以上效果好，可達約 0.897，推測是因 60 歲以下得到與胰臟癌相關疾病的風險較低，因此較易進行分類。相對的，此年齡層沒有固定做篩檢的意識，凸顯此模型在早期發現疾病之重要性，增加應用價值。

- 五、 預測未來的疾病風險時，可發現在 180 天與 360 天後預測效果皆有顯著下降，推測是因為模型對於**觀察終止日前的診斷與藥物有調整權重**，因此如果選用觀察終止日後一段時間的真實結果(ground truth)做比較，預測效果會下降。針對 360 天後預測效果較 180 天效果好，不符合預期之經過時間越長預測效果越差的結果，是因為**在觀察終止日和標示日之間的健康狀態為未知**，不能比較最後效果好壞。
- 六、 在勝算比最高的診斷方面，可以發現**多數為良性腫瘤相關或是消化道相關之疾病**，符合胰臟癌之特性。此外，勝算比較高的 577（胰臟疾病）、576（膽道之其他疾患）診斷，符合醫學原理推斷，胰臟和膽道在位置上及功能上皆有相關性。但因為**彼此可以相互解釋，因此在 AUC loss 會互相影響**。
- 七、 在勝算比最低的診斷方面，可以發現部分診斷可以反映特定情況（如懷孕）或年齡，包括 V22（正常妊娠）、626（月經及其他女性生殖道異常出血）、520（牙齒發育及萌發之疾患）、625（與女性生殖器官有關之疼痛及其他）等皆是**較低年齡層較容易出現的診斷**，符合胰臟癌分布年齡高的特性。本研究納入良性腫瘤的資料，如 220（卵巢良性腫瘤）會定期追蹤，推測可能**提高自主健康管理意識**，因此對胰臟癌預測也有幫助。
- 八、 在勝算比最高的藥物方面，可以發現與診斷碼相似，**多數為和消化或胰臟疾病相關的藥物**。其中藥品 ATC code 為 A03AX95，如 FLOPROPIONE，其適應症為緩解膽道疾病、胰臟炎引起的痙攣性症狀。A07XA01，如 ALBUMIN TANNATE，其適應症為緩解輕微或中度急性腹瀉，符合胰臟癌之特性。A05AA，如 URSODEOXYCHOLIC ACID，其適應症為膽固醇系膽結石之溶解、原發性膽道肝硬化(primary biliary cirrhosis, PBC)之肝功能改善，可能因為膽結石造成膽道阻塞，進而引發胰臟癌的結果。C04AX02，如 PHENOXYBENZAMINE HCL，其適應症為周邊循環障礙、糖尿病性壞疽等，推測與糖尿病有關，**藉由藥物-疾病方式，能進行胰臟癌預測**。
- 九、 在勝算比最低的藥物方面，可以發現和診斷碼相似，**多為較低年齡層或懷孕易開立的藥物**，包括 G02AB，如 ERGONOVINE MALEATE，其適應症為分娩時子宮弛緩性出血或產後出血等，**符合胰臟癌分布年齡普遍較高**之特性，故此特徵有助於非胰臟癌的病人預測。
- 十、 上述重要因子，仍須和臨床醫師或藥師進行深入討論，期望找到新的預測因子，同時目前尚未能全面解釋不同特徵與 AUROC 的關係，就特徵的影響，仍需要其他觀察性研究。

研究限制如下：

- 一、 本研究使用的是健保資料庫，因此缺乏生活型態、身體特徵、檢驗數據、遺傳病史及基因資料，若能再加上這些資料變成更加精準的預測癌症風險。不過，**使用診斷資料能彌補部分未統計到的資料**，為本研究嘗試克服此限制的方法。
- 二、 本研究亦包含可能的取樣偏差、錯誤標示、受試者環境改變、數據需求龐大等缺點，且部分健康狀態無法透過診斷碼表達，也會有醫師誤診的問題，可能影響預測效果。

與前人研究比較：

表十五、前人研究比較

	前人研究 [8]	本組創新
研究資料庫	Danish National Patient Registry	使用 全民健康保險研究資料庫 (NHIRD) ，具完整疾病、用藥， 更符合國人背景 。
研究材料	僅使用疾病代碼	使用 疾病代碼及用藥代碼 ，使用更多特徵搭配時間軸進行 更全面的預測
分年齡層	無	針對胰臟癌 普遍較高齡的特性 ，分年齡層進行訓練， 提早發現年輕病人胰臟癌風險
配合系統	無	採取過去三年資料可 配合健康存摺 ，格式相通可直接連結資料進行 數位快篩
研究結果	AUROC=0.88	AUROC=0.937 ，預測模型有更好的表現
重要因子	僅使用預測效果下降萃取因子	使用勝算比配合預測效果下降， 以兩種方式衡量模型的重要因子 ，更為完整
訓練資料	採 80%/10%/10%為訓練/驗證/測試	採用 5-fold cross validation ，充分運用資料避免過度擬合
訓練模型	採 GRU、Transformer 等	使用 CNN 架構 ，針對有時序性的資料訓練

陸、結論與應用

一、結論

(一) 本研究使用健保資料庫進行機器學習訓練，使用疾病、用藥及時間等特徵，**可以找出隱藏的風險因子，預測病患得到癌症的風險**，AUROC 達 0.937，提供**低成本且快速的初步檢驗**。

(二) 證明同時使用診斷碼及藥物碼能達到更好的預測效果，**並能針對年齡層選擇較為適合的預測模型**，在 60 歲以上（含 60）能達到 AUROC 約 0.846，在 60 歲以下（不含 60）可達到約 0.897，**提早發現疾病發生的時間，增加癌症篩檢的早期意識**。

(三) **能得知未來不同時間點得到癌症的機率**，180 天 AUROC 約 0.830，360 天約 0.837。

(四) 透過模型篩選出重要因子，能從**積極的角度進行早期偵測**，如果發現該特徵就能提早進行模型篩檢，增加模型的應用價值。

透過個人的健保資料能預測未來得到癌症的風險，也能在沒有懷疑胰臟癌發生時提早檢驗並針對高風險病患做後續檢查，**減少在早期沒有發現癌症導致癌細胞迅速擴散造成更大傷害的可能，可以提升存活率**，並節省不必要的檢查所造成的醫療成本。

本研究提出之研究方法可運用於其他癌症或疾病，也能使用其他國家的資料訓練，更符合各國資料的差異，進行精準偵測。

本研究成功將人工智慧的技術應用在醫療界，透過時間矩陣搭配深度學習技術，實現**快速、大量、高準確度、低成本、個人化的胰臟癌數位快篩**。

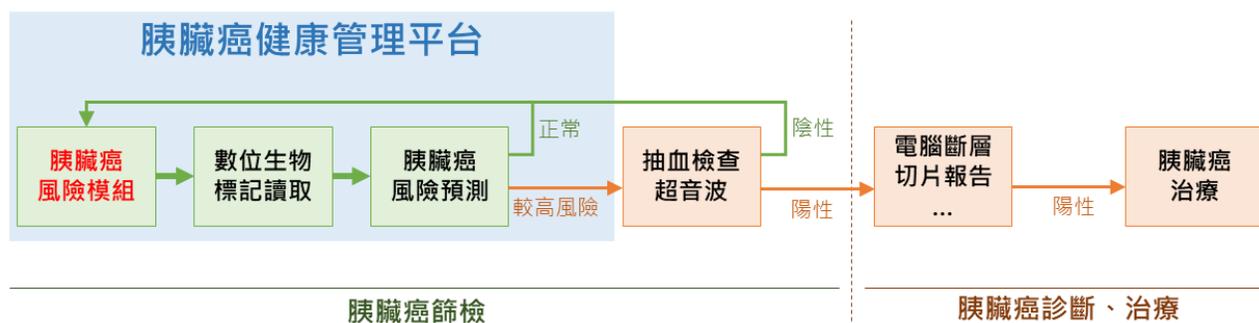
二、未來展望

(一) 本研究所提出的研究方法未來**能進一步推展至其他癌症或疾病，並和健康存摺配合**，輸入一次資料便可預測獲得其他癌症或疾病之風險。

(二) 亦可納入更多資料，包括**實驗室檢驗結果、個人習慣、基因**等，或與其他胰臟癌的風險模組結合，進行胰臟癌的 fusion screening，來進行更全面的預測。透過此方法能配合進一步的治療，鎖定高風險病患提前介入改善生活型態。

本研究未來擬提出「**胰臟癌健康管理平台**」，使用訓練好的胰臟癌風險模組讀取胰臟癌風險，若風險較高，則採用目前醫療體系使用的篩檢方式（如抽血檢查血液腫瘤標記、超音波等）。本研究並非取代目前存在的篩檢方式，而是能提高現有篩檢方式的成效，同時亦能針對無傳統胰臟癌風險因子的民眾，也能早期偵測胰臟癌風險，藉由上述運用，能**早期發現胰臟癌，提高存活率，並避免過晚發現所造成的龐大家庭負擔與醫療成本**。若檢測方式的敏感度和特異度固定，則提高盛行率能提高檢測方式的陽性預測值(PPV)，本研究亦可提高做抽血、超音波前的盛行率，進而提高篩檢的陽性預測值，使篩檢出陽性的病人中實際上為陽性的比率更大。

最後，再以更精確的電腦斷層或切片報告進行胰臟癌的診斷和後續治療。



圖二十二、胰臟癌健康管理平台

柒、參考文獻資料

1. 110 年國人死因統計結果（2022 年 6 月 30 日）。衛生福利部：統計處。取自 <https://www.mohw.gov.tw/cp-16-70314-1.html>
2. 戴鋒泉（2022 年 10 月 13 日）。胰臟癌。取自 https://www.cgh.org.tw/ec99/rwd1320/category.asp?category_id=934
3. 胰臟癌（無日期）。三軍總醫院。取自 <https://www.tsgh.ndmctsgh.edu.tw/unit/10021/20000>
4. 110 年度全民健康保險醫療統計年報（2022 年 12 月 12 日）。衛生福利部統計處。取自 <https://dep.mohw.gov.tw/dos/lp-5103-113-xCat-y110.html>
5. 胰臟癌分期（無日期）。嘉義長庚一般外科。取自 <https://www1.cgmh.org.tw/intr/intr5/c6210/pancreas%20cancer%20stage.html>
6. 病例對照研究 (Case-control study)（2015 年 11 月 24 日）。國家醫療科技評估中心。取自 https://nihta.cde.org.tw/Knowledge/cyclopedia_more?id=23
7. Rohrer, B. (2016, August 18). How do Convolutional Neural Networks work? Retrieved January 25, 2023, from https://e2eml.school/how_convolutional_neural_networks_work.html
8. Placido, D., Yuan, B., Hjaltelin, J.X. et al. A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories. Nat Med 29, 1113 – 1122 (2023). <https://doi.org/10.1038/s41591-023-02332-5>
9. 梁家維（2018）。利用深度學習以診斷及用藥歷史預測罹癌風險-以肝癌為例。臺北醫學大學醫學資訊研究所
10. 朱昱瑾（2022）。用於辨識蓄意自傷之深度學習預測模型。臺北醫學大學醫學資訊研究所。
11. 葛婷婷（2022）。利用健保資料庫時間區段之預測中風發生事件。臺北醫學大學醫學資訊研究所。
12. ICD-9-CM 代碼（無日期）。衛生福利部中央健康保險署。取自 https://www.nhi.gov.tw/Resource/webdata/23348_2_ICD-10-CM.pdf
13. 國軍台中總醫院藥品資料庫暨外觀辨識系統（2014）。國軍台中總醫院藥劑科。取自 <http://web-reg-server.803.org.tw:8090/index.asp>

【評語】 052513

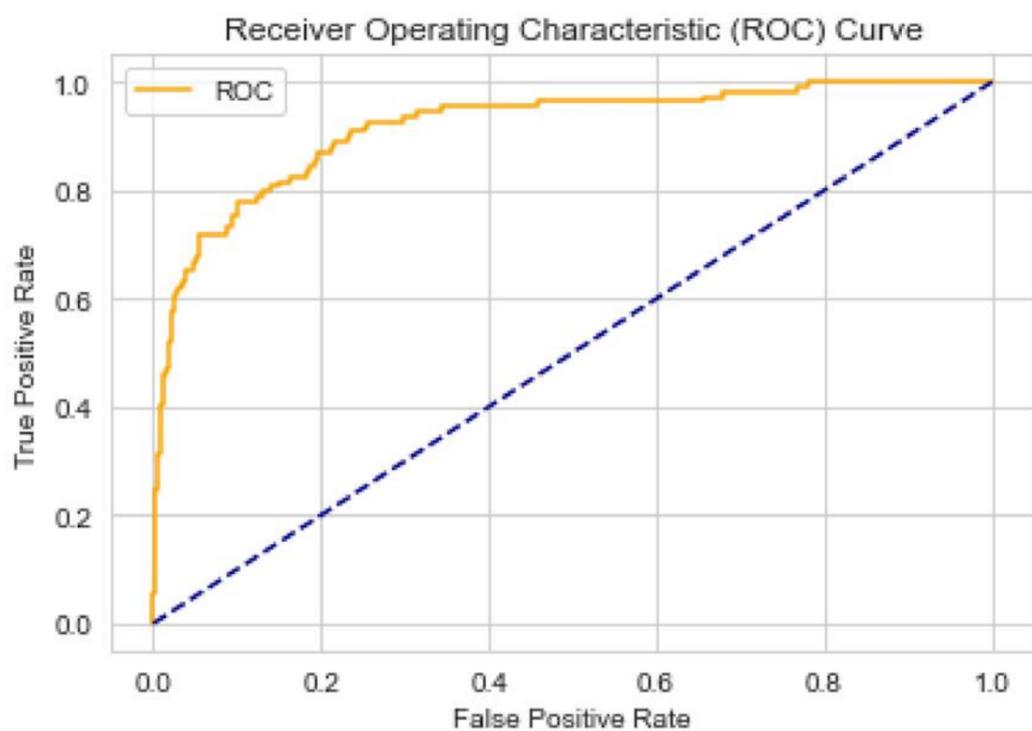
1. 本作品透過健保資料庫，以卷積神經網路訓練後用以偵測胰臟癌。研究頗具重要性及實用性。實驗結果顯示可以找出隱藏的風險因子來預測病患得到癌症的風險，AUROC 達 0.937，已有不錯的偵測效果。
2. 本作品參加電腦與資訊學科組的競賽，應可增加一些資訊技術研究與測試分析的討論，而非僅進行偵測效果的討論。

作品海報

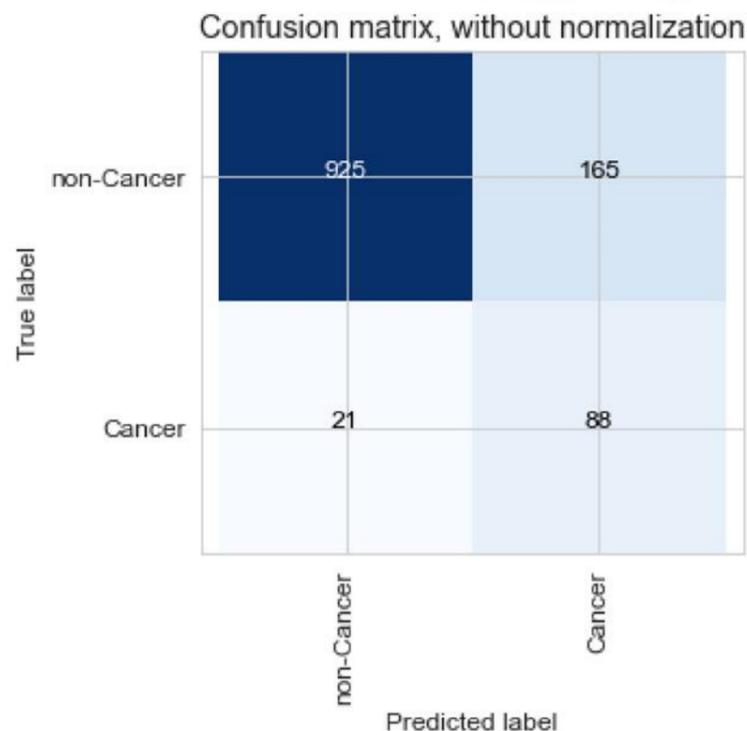
運用疾病軌跡搭配深度學習
偵測胰臟癌風險

研究結果與討論

模型輸出的Score和閾值(threshold)比較大小後，與病人的實際結果核對，可呈現於混淆矩陣中。不同閾值下真陽率和偽陽率的比較可以ROC曲線表示，為客觀衡量機器學習效能的方式。模型表現好則曲線較高，模型下面積(AUROC)較大。本研究AUROC達0.937。

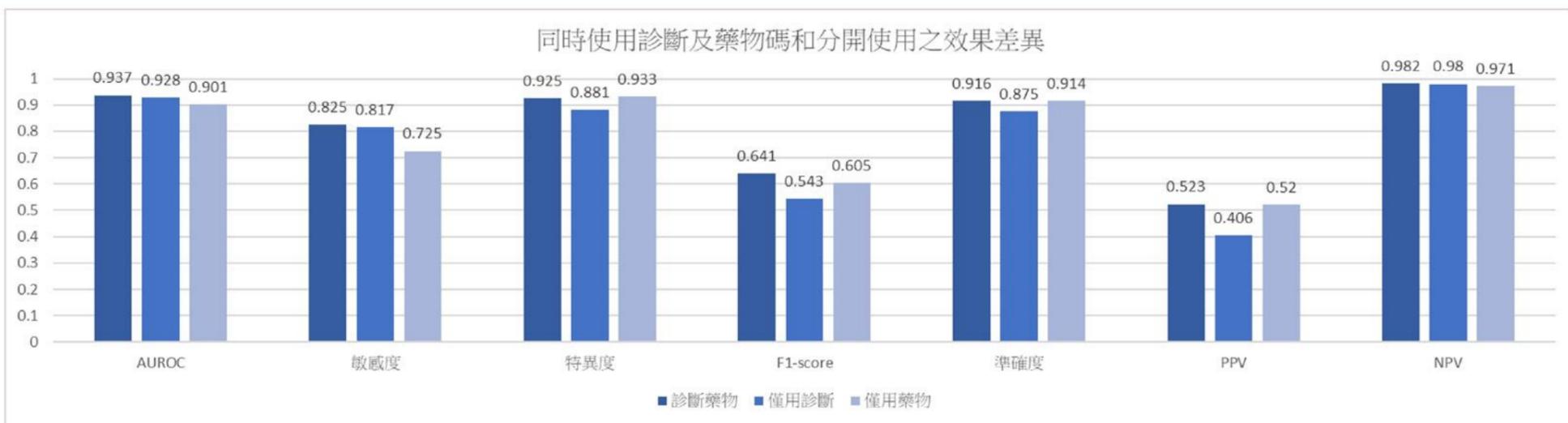


圖四、ROC曲線

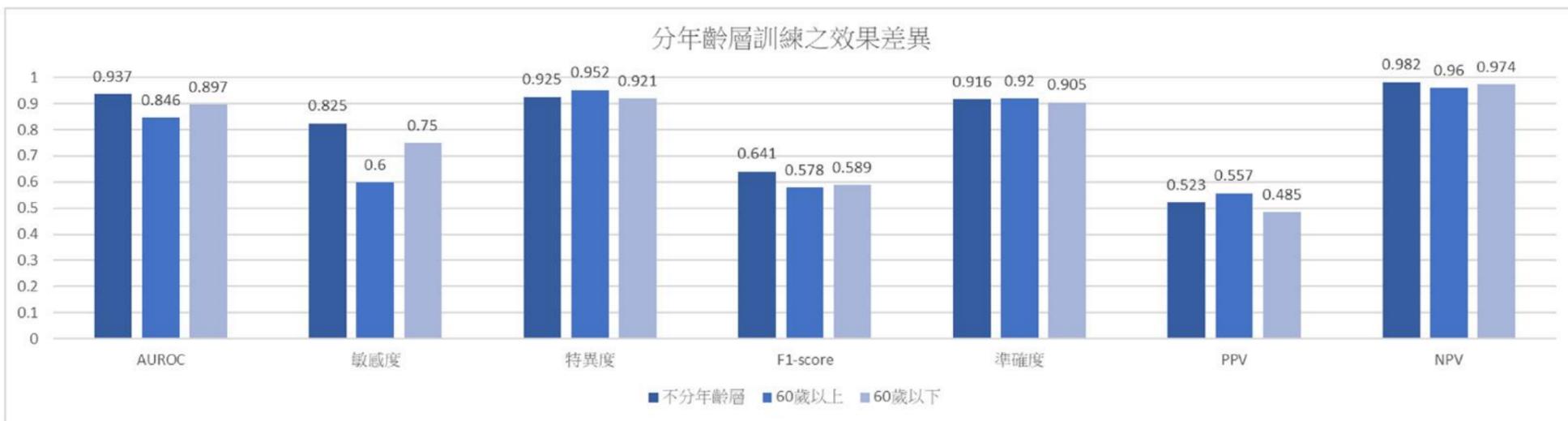


圖五、混淆矩陣

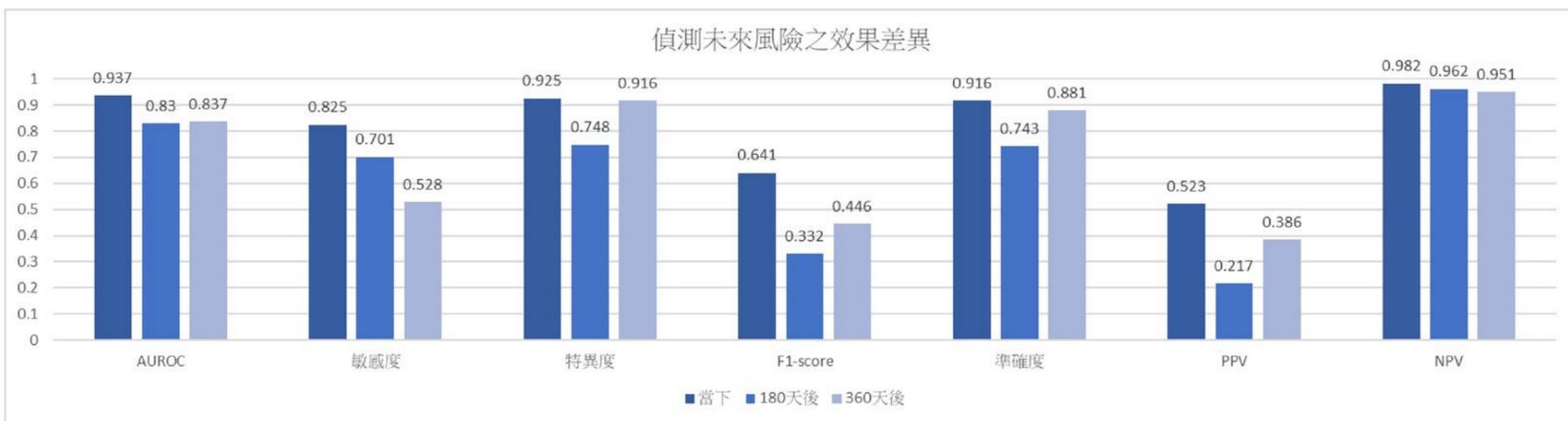
次群組分析



圖六、同時使用診斷及藥物碼和分開使用之效果差異



圖七、分年齡層訓練之效果差異



圖八、偵測未來風險之效果差異

討論：

- 同時使用診斷碼及藥物碼偵測癌症效果較分開使用好，因同時使用較能描述病人的整體健康狀況，同一藥物碼可能對應不同健康狀態（如止痛藥），故效果較診斷碼略差
- 分開年齡層偵測因背景、慢性疾病相似，較難進行分類，而低年齡層偵測效果較好可提醒無篩檢意識的淺在風險族群
- 偵測未來風險因病人的觀察終止日至標示日有一段時間無資料，偵測效果稍微下降，仍維持AUROC=0.830及0.837

重要因子選取

本研究採用勝算比和leave one out進行重要因子的衡量，leave one out為一次抽出一個特徵重新訓練剩下資料，並觀察偵測效果下降量，下降多則該特徵重要性高。勝算比為右表中 ad/bc ，勝算比較大則特徵重要性高。

	有此特徵	無此特徵
有胰臟癌	a人	b人
無胰臟癌	c人	d人

以下為參考兩者後評斷正向及負向關聯強之特徵：

表一、正向及負向關聯強之特徵

排序	正向關聯強診斷	正向關聯強藥物	負向關聯強診斷	負向關聯強藥物
1	胰臟疾病	輕度到中度疼痛之緩解	正常妊娠	子宮弛緩性或產後出血
2	膽道之其他疾患	緩解膽道疾病、胰臟炎	月經、生殖道異常出血	尋常性痤瘡
3	性態未明之腫瘤	膽結石、膽道肝硬化	牙齒發育及萌發之疾患	婦科疾患、子宮出血
4	性質未明之腫瘤	多種維生素，單方	女性生殖器官疼痛	尋常性痤瘡
5	急性及亞急性肝壞死	周邊循環障礙、糖尿病	皮脂腺疾病	停經婦女雌激素不足
6	膽囊結石	多種維生素與礦物質	齒及面部異常	鼻炎及自發性蕁麻疹
7	肝膿瘍及慢性肝病	維生素，其他複方	卵巢良性腫瘤	皮膚炎、皮膚過敏
8	膽囊之其他疾患	止瀉藥、腸道抗炎	急性扁桃腺炎	子宮頸糜爛、子宮頸炎
9	急性十二指腸潰瘍	高血鈣、帕哲特氏病	汗腺疾患	氣喘、過敏性鼻炎
10	胰臟內分泌之其他疾患	維護肝臟正常功能	急性鼻竇炎	尋常性痤瘡

討論：

- 與胰臟癌關聯性高之特徵多與消化道或胰臟功能相關（如上表紅色標示之特徵），符合醫學原理。
- 與胰臟癌關聯性低之特徵多分布於年齡較低的族群（如上表藍色標示之特徵），符合胰臟癌發生年齡較高。

結論與未來展望

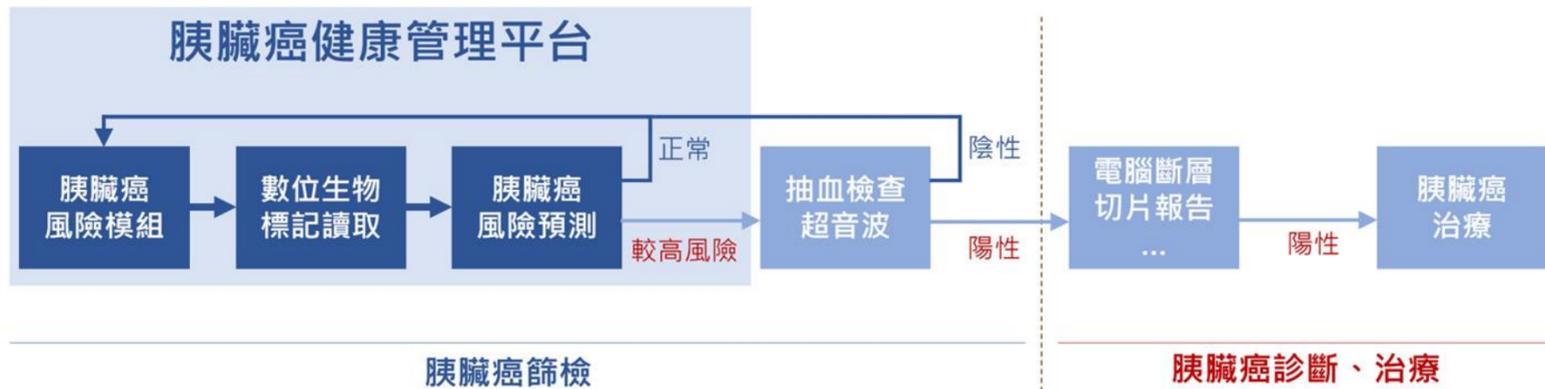
- 透過個人健保資料偵測未來得到癌症的風險，能在沒有懷疑胰臟癌時提早檢驗，針對高風險病患做後續檢查，避免早期沒發現癌症而導致癌細胞迅速擴散，造成更大傷害。可提升存活率，節省晚期發現的龐大醫療支出。
- 本研究提出之研究設計可運用於其他癌症或疾病，能使用、訓練其他國家的資料，符合各國資料的差異。
- 將人工智慧的技術應用在醫療，實現快速、大量、高準確度、低成本、個人化、非侵入式的胰臟癌數位快篩。

表二、與前人研究[1]比較

	研究資料庫	研究材料	配合系統	訓練模型	效果驗證	重要因子
前人研究	Danish National Patient Registry	診斷	無	Transformer GRU等	訓練：驗證：測試 8:1:1	衡量對模型貢獻
本組創新	全民健保資料庫(NHIRD) 完整、更符合國人背景	診斷+藥物	健康存摺	卷積神經網路 適合二維資料	5-fold交叉驗證	先以勝算比篩選 再逐步抽取觀察

未來展望：

- 可納入更多資料，包括實驗室檢驗結果、個人習慣、基因等，或與其他胰臟癌風險模組結合，進行胰臟癌 fusion screening。透過此方法能配合進一步的治療，鎖定高風險病患提前介入改善生活型態。
- 未來擬提出「胰臟癌健康管理平台」，使用已訓練的胰臟癌風險模組讀取胰臟癌風險，再建議採用目前醫療體系使用的篩檢方式（如抽血檢查、超音波等）。本研究並非取代目前存在的篩檢方式，而是能提高現有篩檢方式的成效，同時亦能針對無傳統胰臟癌風險因子的民眾，也能早期偵測胰臟癌風險。若檢測方式的敏感度和特異度固定，則提高盛行率能提高檢測方式的陽性預測值(PPV)，本研究可提高做抽血、超音波前的盛行率，進而提高篩檢的陽性預測值，使篩檢出陽性的病人中實際上為陽性的比率更大。



圖九、胰臟癌健康管理平台

參考資料

- Placido, D., Yuan, B., Hjaltelin, J.X. et al. A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories. Nat Med 29, 1113-1122 (2023).
- 梁家維 (2018)。利用深度學習以診斷及用藥歷史預測罹癌風險-以肝癌為例。臺北醫學大學醫學資訊研究所
- 陳冠文 (2021) 建立胰臟癌風險識別預測模型—以健保資料庫為例。國立臺北護理健康大學資訊管理研究所