利用類神經網路預測糖尿病發生情形

Using Automated Neural Network to forecast the occurrence of diabetes cases

饒孝先[a], 蔡昆原[a], 邱泓文[a], 徐建業[a*]
Hsiao-Hsien Rau[a], Kun-Yuan Tsai[a], Hung-Wen Chiu[a], Chien-Yeh Hsu[a*]
[a] *臺北醫學大學醫學資訊研究所*
[*] *通訊作者: 徐建業*, cyhsu@tmu.edu.tw

摘要

近年來我國罹患糖尿病的人口持續上昇。專家預測，2010 年的糖尿病患者可能增加到 1994 年的 2 倍。另外依據國民健康保險局的資料，台灣糖尿病之成人盛行率已達 5%，而且逐年上升中。糖尿病會造成急性與慢性的併發症，急性的併發症可以造成急性的症狀，甚至致死。而慢性的併發症，會造成種種的病患失能，或器官衰退，雖然未必立即造成死亡，但其將造成整個社會的負擔。如果不給予完善的治療與預防，將耗用更多的醫療資源，造成整體社會的生活品質急速下降，實在相當的重要。

本研究以台北市某醫學中心之門診檢驗檢查資料為母體，抽取 500 位大於 20 歲沒有糖尿病診斷且在 93-95 年與 96-97 年間皆有到該醫院就診的病患。整理後去除檢驗報告有欠缺個案後，完整個案有 136 筆個案資料，並以 STATISTICA 工具訓練多種類神經網路的模型，並以預測結果來比較各種參數對於糖尿病之預測能力。

最後本研究發現，使用三酸甘油脂，總膽固醇，與高密度膽固醇，麩氨酸草醋酸轉氨素與麩氨酸丙酮酸轉氨酵素的平均值的組合具有最佳的預測力，所得的五個模型平均預測的正確率為 79%，其中最佳的正確預測率為 86%。

關鍵字： 糖尿病、類神經網路

1、前言

近年來由於台灣地區人口結構、飲食西化及生活型態的改變，糖尿病已經成為國內重要的慢性疾病之一。根據世界衛生組織的資料顯示，全球每年有 320 萬人死於糖尿病，而且人數還持續在增加。根據國民健康保險局的資料，台灣糖尿病之成人盛行率已達 5%，而且逐年上升中。

典型糖尿病的症狀定義為多喝，多尿，不明原因的體重減輕[1]。然而大部分的糖尿病患者在糖尿病的初期卻大部分沒有症狀，在這段時期，糖尿病的診斷只能倚靠抽血追蹤檢查，沒有辦法依據臨床症狀來引導病患發現自己已經患病[2]。然而這樣沒有症狀的高血糖就會開始造成各種糖尿病的慢性併發症，如眼睛病變，腎臟病變，神經病變，或血管病變[2-4]。

美國糖尿病醫學會(American Diabetes Association)就建議在較肥胖或年紀較大的病患，或是有強烈家族病史，再者具有三酸甘油脂過高，高密度膽固醇過低，血糖耐受不良，空腹血糖偏高等代謝症候群相關因子時，臨床醫師可檢查空腹血糖來篩檢病患有無糖尿病，以提高糖尿病的診斷率[1]，但即使是這樣還是有超過一半的糖尿病人因為沒有症狀而延遲治療[5]。

因此本研究想利用醫院檢驗資訊系統內容易獲得的電腦資訊，建立預測的模型，利用類神經網路來訓練該模型並試驗模型的效果。

2. 文獻探討

(1)糖尿病(Diabetes)

糖尿病是一群表現出高血糖症的統稱，屬於一種慢性代謝性障礙疾病，病患可能會有胰島素分泌不足或胰島阻抗的問題，或者兩者皆有。其主要的臨床症狀有尿多、口渴、飢餓、疲勞、視力模糊、體重減輕或傷口不易癒合等症狀。美國糖尿病協會將糖尿病分類為：

- 第一型糖尿病(Type I diabetes)：之前稱為胰島素依賴型糖尿病(Insulin Dependent Diabetes Mellitus, IDDM)，常見發生於幼年的糖尿病，主要因為來自胰島細胞遭到免疫反應破壞造成。可能是因為病人的遺傳、生活環境、或是病毒感染引發自體免疫的

反應，該免疫反應攻擊胰臟內的 β 細胞，造成破壞。以致病患身體無法製造足夠胰島素。

- 第二型糖尿病(Type II diabetes)：之前稱非胰島素依賴型糖尿病(Noninsulin Dependent Diabetes Mellitus, NIDDM)。較常發生在成人，一般多發病在 40 歲左右。通常同時存在胰島素分泌不足與胰島素阻抗的問題。發病的原因是多因性的。一般認為跟遺傳、肥胖或缺乏運動有關。
- 其他特殊型式的糖尿病：包含如基因缺陷造成的糖尿病、胰臟外分泌破壞胰臟造成之糖尿病、由藥物或化學物質所引起的糖尿病等。
- 妊娠糖尿病(gestational diabetes mellitus, GDM)：懷孕時可能因荷爾蒙或代謝改變，造成胰島素的抵抗作用。如胰島素代償性分泌不足造成的血糖上升。

美國糖尿病協會另外定義了早期糖尿病(Pre-diabetes)，其中包含了：

- 空腹血糖異常(Impaired Fasting Glucose, IFG)：在不進食有卡路里食物超過 8 小時後，血漿血糖值大於等於 100 mg/dL 而且小於 126mg/dL 者。
- 葡萄糖耐受異常(Impaired Glucose Tolerance, IGT)：受測者前一天正常進食，受測當天不進食有卡路里食物超過 8 小時後，服用含無水葡萄糖 75g 後，兩小時之血漿血糖濃度大於等於 140mg/dL 而且小於 200mg/dL 者。

「第二型糖尿病」的發生率是相當驚人。世界衛生組織的資料中顯示，全球有超過 1 億 7 千以上的糖尿病患者。而其中九成以上是第二型的糖尿病。糖尿病會造成的病患包括眼睛、腎臟、神經、與大血管的傷害。相較於非糖尿病者，糖尿病患的壽命少了十歲。而且很多大型的糖尿病研究都證實早期診斷並給予積極的治療，可以降低併發症之發生及進展。

流行病學將疾病預防分三級，即是初級、次級、三級等。早期診斷糖尿病，並給予治療是屬於次級預防。而次級預防的主要目標是早期診斷早期治療。因此發展好的篩檢方法，應是糖尿病的次級預防裡相當重要的方式。利用好的預防方法，才能早期發現病患，給予適當的衛教與適當的治療。

(2)各檢驗檢查值與糖尿病之關聯性

BMI(身體質量指數)的數值在電子病歷尚未標準化之前，並不容易直接獲得，但肥胖或腰圍是胰島素阻抗的重要因子。在近幾年的研究中，非酒精性脂肪肝(Non-alcoholic fatty liver disease) 與代謝症候群的關係是常常被談到的[14, 15]，跟 BMI，腰圍還有肥胖也都有相關性。而且因為非酒精性脂肪肝會造成特異性的肝功能異常。[16] 因此研究中特別選取麩氨酸草醋酸轉氨素(GOT)與麩氨酸丙酮酸轉氨酵素(GPT)作為相關的輸入因子。

在併發症的部份，血糖上升會影響腎絲球的組織灌流量，對腎臟功能會有影響。血清肌酸酐(Creatinine)是評估腎臟功能的因子，因此使用血清肌酸酐來做為血糖的預測因子[17]。鉀離子(K)本身在腎臟功能不良的時候會造成上升[18]。

(3)類神經網路(Automated Neural Network, ANN)

「類神經網路是一種計算系統，包括軟體與硬體，它使用大量簡單的相連人工神經元來模仿生物神經網路的能力。人工神經元是生物神經元的簡單模擬，它從外界環境或者其它人工神經元取得資訊，並加以非常簡單的運算，並輸出其結果到外界環境或者其它人工神經元。」

人工類神經網路使用數個微處理器，用來當做人腦之中的神經元，將它們組合成的神經網路結構型態，然後選定一個數學推論出來的演算法，將這演算法置入這個類神經網路中。

要使得類神經網路能正確的運作必須透過訓練（Training）的方式，讓類神經網路反覆的學習，直到對於每個輸入都能正確對應到所需要的輸出，因此在類神經網路學習前，必須建立出一個訓練樣本（Training Pattern）使類神經網路在學習的過程中有一個參考，訓練樣本的建立來自於實際系統輸入與輸出或是以往的經驗。

一般常用來總結 ROC 曲線訊息的指標為 ROC 曲線下面積與 ROC 曲線下部分面積。ROC 曲線下面積為特異度(specificity)全域內的平均可靠度(sensitivity)；ROC 曲線下部分面積則是特異度被限制在臨床上有意義的範圍內之平均可靠度。在診斷試驗中，比較新診斷工具與現行標準診斷工具的準確性(accuracy)是一項重要的課題。

(3).ANN 於糖尿病之預測研究

外國發展了很多不同的預測模型，例如 San Antonia model 與 Framingham mode 等等。其中預測效果較好的是 San Antonia model，它的 receiver-operating characteristic (ROC) curve 下的面積有 84.3% [7, 19]。San Antonia model 與 Framingham mode 這兩個預測模型都需要空腹血糖，收縮壓，高密度膽固醇，身體質量指數，與家族病史。而 Framinham model 多了三酸甘油脂，而 San Antonia model 中則多了年齡，性別，種族來做為預測因子。[6-8]。 但是這兩個模型需要的預測因子需要病患的身體質量指數，與家族病史，這些資料在現行的醫院電腦系統裡並沒有完整的資訊化，並不容易獲得。而很多在醫院長期追蹤的病患，雖然不同科的醫師提供抽血的檢查，卻不一定有頻繁的檢驗血糖的數字。

3. **研究步驟與方法**

(1).研究對象

本研究以台北市某醫學中心的病患為抽樣母體，隨機挑選 500 位大於 20 歲沒有糖尿病診斷且在 93-95 年與 96-97 年間皆有就診的病患。

篩選在 93-95 年間同時具備空腹血糖，三酸甘油脂，總膽固醇，高密度膽固醇，血清尿酸，麩氨酸草醋酸轉氨素，麩氨酸丙酮酸轉氨酵素，血清肌酸酐，與鉀離子檢驗資料的病患共 408 筆實驗室檢驗檢查資料，個案數則為 136 人。

(2)資料前處理

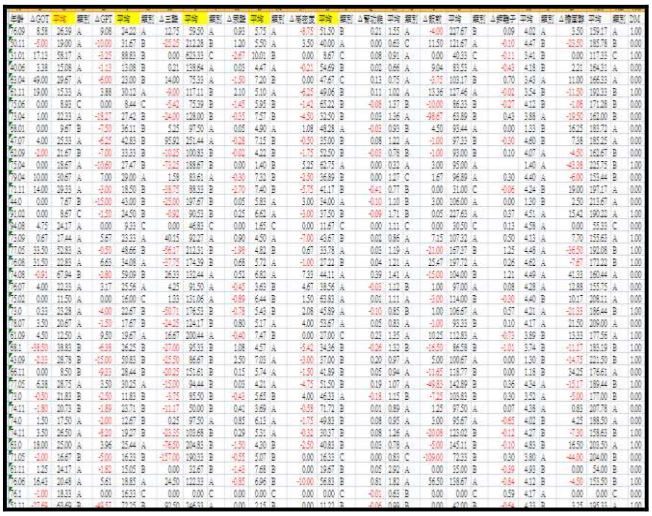
本研究利用前述所蒐集到的資料計算出每一個案在過去兩年內各項檢查項目的平均值以及其平均變化量(Δ)，當做 ANN 模型的預測因子，本研究所應用之資料整理如下：

- 輸入變項：

  皆為連續型資料(Continuous)，包含年齡、GOT 平均變化量、GOT 平均值、GPT 平均變化量、GPT 平均值、鉀離子平均變化量、鉀離子平均值、尿酸平均變化量、尿酸平均值、腎功能指數平均變化量、腎功能指數平均值、膽固醇平均變化量、膽固醇平均值、三酸甘油酯平均變化量、三酸甘油酯平均值、高密度脂蛋白平均變化量、高密度脂蛋白平均值、飯前血糖平均變化量、飯前血糖平均值

- 輸出變項：

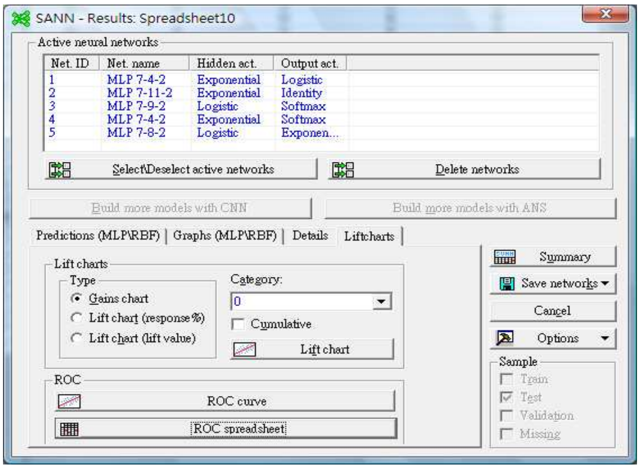
  為類別型資料(Categorical)，其值為兩年後是否被診斷為糖尿病，以 1 表示是糖尿病病患，0 則表是非為糖尿病病患。



圖：本研究之原始之資料

(3).使用工具

本研究使用 STATISTICA 7.0 為 ANN 的分析工具，STATISTICA 是一套統計資料分析、圖表、資料管理、應用程式開發的系統，也提供了對其他技術、包含提供類神經網路模型的訓練與測試等功能、資料挖掘應用的功能模組。對使用者而言，提供完整且可選擇性的使用介面；亦可廣泛使用程式語言嚮導建立模型或整合 Statistica 與其他應用程序進行計算。



SANN - Results: Spreadsheet10

Active neural networks

Net. ID	Net. name	Hidden act.	Output act.
1	MLP 7-4-2	Exponential	Logistic
2	MLP 7-11-2	Exponential	Identity
3	MLP 7-9-2	Logistic	Softmax
4	MLP 7-4-2	Exponential	Softmax
5	MLP 7-8-2	Logistic	Exponen...

Lift charts — Type: ● Gains chart

Category: 0

☐ Cumulative

Sample: ☐ Train ☑ Test ☐ Validation ☐ Missing

圖：STATISTICA 操作介面

(4)類神經網路訓練

本研究首先將 136 筆資料分為訓練資料與測試資料兩部分，其中 80%(108 筆)做為訓練資料，20%(28 筆)做為測試資料。

類神經網路的訓練步驟如下

- 讀入訓練資料進行訓練

  在 STATISTICA 中新加入資料陣列，本研究的訓練資料共有 29 個參數，108 筆資料。

  ■ 定義輸入、出變項(連續型/離散型)

  本研究是要依據各檢驗檢查值來預測是否會被診對於糖尿病，因此要訓練一個新的分類 ANN，依據資料的型態，將各遍項定義為輸入變項或輸出變項，其中輸入變項包含連續型與離散型兩類。

  ■ 定義取樣率

  本研究將資料分為訓練資料與測試資料，在訓練時以所有的訓練資料為對象，因此將 Sampling Rate 設為 100%。

  ■ 完成訓練並儲存模型

  將訓練完成的模型以 PMML（Predictive Model Markup Language, 預測模型標記語言）儲存演算法(.xml) 。

- 讀入測試資料進行測試

  在 STATISTICA 中新加入資料陣列，本研究的訓練資料共有 29 個參數，28 筆資料。定義完成後便開啟 ANN 的工具，選擇之前儲存的模型進行測，最後得到 ROC Threashold、ROC Area、ROC Curve 以及 Prediction 等結果

(5)分組

為了瞭解並比較檢驗檢查的平均值與變異量對於糖尿病診斷之預測能力本研究在做 ANN 的訓練與測試時均將資料分為下列幾類來計算期預測準確度：

- 使用所有參數(即年齡與每項檢驗檢查值的平均值與平均變化量)做預測。
- 僅使用年齡及每項檢驗檢查項目的平均值做預測
- 僅使用年齡及每項檢驗檢查項目的平均變化量做預測。

另一方面，為了瞭解每一種檢驗檢查值對於糖尿病診斷之預測能力，本研究亦將其分為下列三組分別訓練與測試模型：

- 使用年齡、膽固醇、高密度脂蛋白、三酸甘油酯、飯前血糖等五個項目做預測。
- 使用鉀離子、GOT、GPT、腎功能指數、尿酸值等五個項目做預測。
- 使用年齡、膽固醇、高密度脂蛋白、三酸甘油酯、飯前血糖、GOT、GPT 等七個項目做預測。

4. 研究結果

經由上述的 Input、訓練 ANN、儲存結果模型、測試模型等等動作之後本研究得到下列六組模型，每一組皆得到五個模型，且皆為多層感知神經網路模型(Multi-layer Perception, MPL) ，以下分別說明每類模型的預測力：

(1) 類別一：使用所有參數做為輸入變項

如下表所示，若將所有參數均設為輸入變項，所得到的五個模型中最好的有 86%，僅有一個預測率較差(61%)，平均預測的正確率達 72%

表一：類別一的預測力

	ROC Area	ROC Threashold	Correct	Wrong	C%
模型 1	0.80	0.99	24	4	0.86
模型 2	0.66	1.00	20	8	0.71
模型 3	0.57	0.75	19	9	0.68
模型 4	0.58	0.58	17	11	0.61
模型 5	0.79	0.99	21	7	0.75

(2) 類別二：使用各檢察項目平均值做為輸入變項

如下表所示，若將所有各項檢驗檢查項目的平均值設為輸入變項，所得到的五個模型中最好的僅有 75%，最差的正確率則為 50%，平均預測的正確率為 65%，顯示這個單用平均值的預測能力較以所有參數做為預測變項的預測力略差。

表二：類別二的預測力

	ROC Area	ROC Threashold	Correct	Wrong	C%
模型 1	0.46	0.93	18	10	0.64
模型 2	0.58	1.00	19	9	0.68
模型 3	0.67	0.99	19	9	0.68
模型 4	0.66	1.00	21	7	0.75
模型 5	0.44	1.00	14	14	0.50

(3) 類別三：使用各檢察項目平均變化值做為輸入變項

如表三所示，若將所有各項檢驗檢查項目的平均變化值設為輸入變項，所得到的五個模型中最好的為 79%，最差的正確率則為 57%，平均預測的正確率為 68%，顯示這個單用平均變化量的預測能力與擔用平均值得相當接近，皆比使用所有參數做為預測變項的預測力略差。

表三：類別三的預測力

	ROC Area	ROC Threashold	Correct	Wrong	C%
模型 1	0.49	0.57	16	12	0.57
模型 2	0.83	0.54	20	8	0.71
模型 3	0.67	0.43	16	12	0.57
模型 4	0.81	0.58	21	7	0.75
模型 5	0.57	0.33	22	6	0.79

除各種平均值之預測能力外，下面類別四~六將預測各種代謝症候群相關因子的預測能力

(4) 類別四：使用年齡、膽固醇、高密度脂蛋白、三酸甘油酯、飯前血糖等五個項目的平均值做預測

如下表四所示，若使用上述五種因子的平均值做預測，所得到的五個模型中最好的為 86%，最差的正確率則為 68%，平均預測的正確率為 74%。

表四：類別四的預測力

	ROC Area	ROC Threashold	Correct	Wrong	C%
模型 1	0.77	0.76	22	6	0.79
模型 2	0.80	0.74	24	4	0.86
模型 3	0.76	1.00	19	9	0.68
模型 4	0.61	0.76	19	9	0.68
模型 5	0.70	0.97	20	8	0.71

(5) 類別五：使用鉀離子、GOT、GPT、腎功能指數、尿酸值等五個項目做預測

如下表五所示，若使用上述五種代謝症候群相關因子的平均值做預測，所得到的五個模型期預測能力均大幅降低，其中最好的僅為 39%，最差的甚至僅有 25%的預測能力，平均預測的正確率為 34.8%，與使用前述四種因此的預測能力有相當大的差異。

表五：類別五的預測力

	ROC Area	ROC Threashold	Correct	Wrong	C%
模型 1	0.46	0.73	9	19	0.32
模型 2	0.43	0.60	7	21	0.25
模型 3	0.47	0.66	11	17	0.39
模型 4	0.41	0.57	11	17	0.39
模型 5	0.44	0.67	11	17	0.39

(6) 類別六：使用年齡、膽固醇、高密度脂蛋白、三酸甘油酯、飯前血糖、GOT、GPT 等七個項目的平均值做預測

由於有相關研究顯示，肥胖所造成的非酒精性脂肪肝會造成特異性的肝功能異常，因此我們將 GOT、GPT 這個檢驗值的平均值加入到類別四的項目之中。

如下表六所示，若使用上述七種因子的平均值做預測，所得到的五個模型中最好的為 86%，最差的正確率則為 68%，平均預測的正確率為 78.8%。此結果亦顯示，加入 GOT、GPT 後，使得這些因此對於糖尿病的預測力更為提高(平均增加 5%)，甚至高於使用所有參數最為輸入變項的預測方式(平均增加 7%)，是所有模型中預測準確率最高的

表六：類別六加入 GOT、GPT 的預測力

	ROC Area	ROC Threashold	Correct	Wrong	C%
模型 1	0.77	0.98	23	5	0.82
模型 2	0.72	0.67	22	6	0.79
模型 3	0.84	0.64	22	6	0.79
模型 4	0.62	0.96	19	9	0.68
模型 5	0.73	0.64	24	4	0.86

5.討論

本研究使用代謝症候群相關因子(不包含血壓與 BMI)，可以得到 70%以上的疾病預測力。若再加入麩氨酸草醋酸轉氨素(GOT)與麩氨酸丙酮酸轉氨酵素(GPT)來訓練類神經網路模型預測糖尿病，可以得到更好的疾病預測力。

糖尿病的發生是胰島細胞逐漸衰退不足以與胰島素阻抗相抗衡的結果。因此影響到糖尿病發生的因子，必是與胰島細胞衰退跟胰島素阻抗相關的因子。三酸甘油脂，總膽固醇，與高密度膽固醇的這三個代謝症侯

群相關因子，與病患的胰島素阻抗有相關性。而年齡，與空腹血糖則與胰島細胞衰退情況有相關。在分析中也發現扣除血糖輸入因子，的確會大幅惡化該模型的預測力。

本研究過程中亦碰到些許限制，因為完整收集到所有輸入因子的病患數目只有 136 位，而且都侷限在台北市某醫學中心的資料。所以其實需要再用更大量的資料來做訓練與測試，才能知道該模型是否適合於不同的醫院的病患族群，或是所有的東方族群。

6.結論

本研究期望在糖尿病的預測上提供臨床照護團隊容易進行判斷病患之風險值，方便決定追蹤血糖的時間。

本研究建立數種模型比較期預測力後發現，三酸甘油脂、總膽固醇、與高密度膽固醇、麩氨酸草醋酸轉氨素(GOT)與麩氨酸丙酮酸轉氨酵素(GPT)可達成較好的預測能力。

參考文獻

1. Standards of medical care in diabetes--2009. Diabetes Care, 2009. 32 Suppl 1: p. S13-61.
2. Colagiuri, S. and D. Davies, The value of early detection of type 2 diabetes. Curr Opin Endocrinol Diabetes Obes, 2009. 16(2): p. 95-9.
3. Tapp, R.J., et al., Foot complications in Type 2 diabetes: an Australian population-based study. Diabet Med, 2003. 20(2): p. 105-13.
4. Tapp, R.J., et al., Albuminuria is evident in the early stages of diabetes onset: results from the Australian Diabetes, Obesity, and Lifestyle Study (AusDiab). Am J Kidney Dis, 2004. 44(5): p. 792-8.
5. IDF 2006, 19th World Diabetes Congress, 3-7 December 2006, Cape Town, South Africa. Abstracts. Diabet Med, 2006. 23 Suppl 4: p. 1-788.
6. Stern, M.P., K. Williams, and S.M. Haffner, Identification of persons at high risk for type 2 diabetes mellitus: do we need the oral glucose tolerance test? Ann Intern Med, 2002. 136(8): p. 575-81.
7. McNeely, M.J., et al., Comparison of a clinical model, the oral glucose tolerance test, and fasting glucose for prediction of type 2 diabetes risk in Japanese Americans. Diabetes Care, 2003. 26(3): p. 758-63.
8. Lorenzo, C., et al., The metabolic syndrome as predictor of type 2 diabetes: the San Antonio heart study. Diabetes Care, 2003. 26(11): p. 3153-9.
9. Niemeijer-Kanters, S.D., J.D. Banga, and D.W. Erkelens, [Dyslipidemia in diabetes mellitus]. Ned Tijdschr Geneeskd, 2001. 145(16): p. 769-74.
10. Chan, J.C., et al., Diabetes in Asia: epidemiology, risk factors, and pathophysiology. JAMA, 2009. 301(20): p. 2129-40.
11. Bradna, P., [Gout and diabetes]. Vnitr Lek, 2006. 52(5): p. 488-92.
12. Dehghan, A., et al., High serum uric acid as a novel risk factor for type 2 diabetes. Diabetes Care, 2008. 31(2): p. 361-2.
13. 查顯友;周燕, 老年高尿酸血症與高血糖、高血脂、高血壓關係的研究. 實用老年醫學, 2007/10. 21 卷 5 期: p. 355-356.
14. Watanabe, S., et al., Liver diseases and metabolic syndrome. J Gastroenterol, 2008. 43(7): p. 509-18.
15. Gronbaek, H., et al., Role of nonalcoholic fatty liver disease in the development of insulin resistance and diabetes. Expert Rev Gastroenterol Hepatol, 2008. 2(5): p. 705-11.
16. Wieckowska, A. and A.E. Feldstein, Diagnosis of nonalcoholic fatty liver disease: invasive versus noninvasive. Semin Liver Dis, 2008. 28(4): p. 386-95.
17. Lamb, E.J., C.R. Tomson, and P.J. Roderick, Estimating kidney function in adults using formulae. Ann Clin Biochem, 2005. 42(Pt 5): p. 321-45.
18. Nichols, C.G. and J.C. Koster, Diabetes and insulin secretion: whither KATP? Am J Physiol Endocrinol Metab, 2002. 283(3): p. E403-12.
19. Chien, K., et al., A prediction model for type 2 diabetes risk among Chinese people. Diabetologia, 2009. 52(3): p. 443-50.

20.張俊郎、陳啟浩、曾輝鈺，結合類神經網路與決策樹於糖尿病前期診斷之研究