國立臺灣大學公共衛生學院流行病學與預防醫學研究所

碩士論文

Graduate Institute of Epidemiology and Prevention Medicine

College of Public Health

National Taiwan University

Master Thesis

在貝氏統計方法下且考量疾病盛行率的診斷工具網絡統合分析

Diagnostic Test Performance Network Meta-Analysis in Bayesian Approach and Prevalence

藍正翔

Cheng-Hsiang Lan

指導教授：杜裕康　博士

Advisor: Yu-Kang Tu, DDS, PhD

中華民國 112 年 07 月

July, 2023

doi:10.6342/NTU202302826

國立臺灣大學碩士學位論文
口試委員會審定書

在貝氏統計方法下且考量疾病盛行率的診斷工具網絡統合分析

Diagnostic Test Performance Network Meta-Analysis in Bayesian Approach and Prevalence

本論文係 藍正翔 君（學號 R09849005）在國立臺灣大學流行病學與預防醫學研究所完成之碩士學位論文，於民國 112 年 07 月 24 日承下列考試委員審查通過及口試及格，特此證明

口試委員：

（簽名）

（指導教授）

doi:10.6342/NTU202302826

誌謝

能夠完成這篇論文，首先要好好感謝我的指導教授杜裕康博士。很幸運能在這段碩班的生活中，跟著老師好好學習統合分析的相關知識，在這之前我對統合分析是完全不了解的，但在老師的悉心教導與耐心指導下，讓我能夠完成這篇論文。老師總是很有耐心，花費許多時間解決我在這篇論文中的許多疑難雜症，以及協助我處理無法應付的各種問題。另外也要感謝老師在論文與學業之外，也給了我許多未來的人生方向，讓我對未來規劃能夠有許多想法，不單單只是侷限於眼前，很高興能在這些日子跟隨杜老師學習。

要謝謝實驗室眾多的學長、學姐，以及一同學習的同學們，能與大家一起度過這段日子十分令人開心，從實驗室的大家身上，我學到許多知識，不只是統計，同時也有許多醫學相關的內容，讓我的眼界不再僅僅侷限於統計，知道統計也可以在醫療領域發揮強大的力量，因為有你們，讓每次的 meeting 都是非常快樂的，不是嚴肅沉悶的場合。也要謝謝身邊的其他碩士班生統組的同學，從你們身上也讓我看到了許多各個不同統計領域的發展，因為有你們的陪伴，也讓我碩班這幾年的日子過得多彩多姿。

時光飛逝，不知不覺也來到碩班的尾聲，在臺大的這段時間學到了無窮的知識，這是這輩子珍貴的寶藏，不僅僅是書本與領域的知識，更多的是面對與解決問題的方法。未來我也將以此為圭臬，繼續在知識中努力追求真理，面對人生中的各種問題。

最後要感謝我身邊的所有家人、朋友，謝謝你們在這段日子裡的陪伴，也協助我解決許多論文之外的其他煩惱，讓我可以專心完成論文。要感謝的人真的太多了，也祝福所有幫助我的人在未來人生的道路上，也能事事順利。

藍正翔

2023/06/30

doi:10.6342/NTU202302826

摘要

背景

有別於過往介入方式的統合分析，診斷工具的統合分析因為需要同時考慮敏感度與特異度兩個變項，但因為在不同研究之間此二變項並不獨立，因此除了同時考量二者外，估計上亦須考量其相關性。目前常見的診斷工具統合分析的方法為二元模型，而若要進行診斷工具的網絡統合分析，則以 ANOVA 模型為主。此外因診斷工具的網絡統合分析需要估計許多參數，因此貝氏統計透過資料更新先驗分配，以進行參數估計的方式也較為常用的方法。

目標

在本研究中，將採取貝氏統計的做法，進行診斷工具的網絡統合分析，並以一般診斷工具統合分析的二元模型與網絡統合分析的 ANOVA 模型為主，且同時將疾病盛行率透過潛在類別分析的方式納入參數考量，以使其能夠更為準確地同時比較多種不同的診斷工具，解決黃金標準可能並不完美的問題。並以 SUCRA 值、Superiority Index 與 Youden Index 三項指標，評估不同診斷工具之優劣排名。

方法

我們以 ANOVA 模型為基礎，並在參數中考量疾病盛行率，使模型之結果更加穩定精確。資料驗證上我們以 Hoyer and Kuss 於 2018 年研究中比較第二型糖尿病的兩種不同診斷工具之資料，以及 Veroniki et al.於 2021 年比較 CIN2+型的子宮頸癌的三種不同診斷工具的研究作為驗證的資料，以確認模型所估計之參數結果之正確性，並嘗試透過不同指標將不同診斷工具之效果進行優劣排名。

doi:10.6342/NTU202302826

結果

本研究中所提出之模型，在 Hoyer and Kuss 於 2018 年研究與 Veroniki et al. 於 2021 年的研究中皆可求得與論文所發表之研究數據相近之結果。此外對於診斷工具的排名，可以發現 SUCRA 值因將敏感度與特異度分開考量，可能會導致不同的解讀結果；而 Superiority Index 則是將敏感度與特異度同時考量，以求出一個理想的排名結果，但沒有絕對範圍；Youden Index 則可同時考量敏感度與特異度，並將指標之數值限制於 0 至 1 之間，提供一個絕對範圍。

結論

二元模型與 ANOVA 模型為診斷工具的統合分析提供了優良的架構，以評估多種不同的診斷工具，而將疾病盛行率透過潛在類別分析的方式納入模型參數中，可以為黃金標準並非完全準確時，提供一個解決此問題的方法，在統計估計上可能出現偏差的情況。搭配上貝氏統計的方式，亦能協助我們更有效率地估算參數，以比較不同診斷工具的效果，也提供一個更有彈性的方式，讓使用者可自行調整對參數的分配假設。

關鍵字：網絡統合分析、診斷工具、貝氏統計、二元模型、ANOVA 模型、疾病盛行率

doi:10.6342/NTU202302826

Abstract

Background

Unlike the intervention meta-analysis, we need to consider sensitivity and specificity at the same time when we do the diagnostic test performance meta-analysis. However, we not only need to estimate both sensitivity and specificity but also need to estimate their correlation, because they may not be independent across different studies. The general method for diagnostic test performance meta-analysis is the bivariate model. If we want to undertake a diagnostic test performance network meta-analysis to compare several diagnostic tests, the ANOVA model is common method. Besides, the Bayesian approach is also often used for diagnostic test performance network meta-analysis because we can update the prior distribution by the data and estimate the large number of parameters in the network meta-analysis.

Objectives

We use the Bayesian approach to undertake a diagnostic test performance network meta-analysis in this study. We use the bivariate model and ANOVA model as the basement, and we consider the prevalence by latent class analysis as a parameter in this method to compare several diagnostic tests at the same time to deal with the problem of the imperfect gold standard. We also use three different indexes such as SUCRA, superiority index, and Youden index to rank the different diagnostic tests.

doi:10.6342/NTU202302826

Methods

We use the bivariate model and ANOVA model as the basis, and add the prevalence as a parameter into these models to obtain more precise estimates. For the demonstration of our models, we use two datasets: one, reported by Hoyer and Kuss, compared two different diagnostic tests for the diagnosis of type 2 diabetes mellitus; the other, reported by Veroniki et al. in 2021, compared three different diagnostic tests for the diagnosis of invasive cervical cancer (CIN2+). We also use different indexes to rank the different diagnostic tests in these researches.

Results

The results obtained by our model were similar to those given by Hoyer and Kuss or Veroniki et al. In terms of the ranking of diagnostic tests, SUCRA may lead to different decisions because it considers sensitivity and specificity separately. The superiority index considers both sensitivity and specificity together to calculate the ranking, but the range of its value has no limit. Youden index not only considers both sensitivity and specificity simultaneously but also ranges from 0 to 1.

doi:10.6342/NTU202302826

Conclusions

Both bivariate model and ANOVA model provides a useful statistical framework for assessing the performance of several diagnostic tests. Including the prevalence by latent class analysis as a parameter in the model may provide a solution to the problem of an imperfect gold standard. The Bayesian approach help we estimate the parameters more efficiently and flexibly, and it also allows the users to adjust the prior distributions of model parameters.

Keywords: network meta-analysis, diagnostic tests, Bayesian statistics, binomial model, ANOVA model, prevalence

doi:10.6342/NTU202302826

目錄

doi:10.6342/NTU202302826

doi:10.6342/NTU202302826

圖目錄

doi:10.6342/NTU202302826

表目錄

doi:10.6342/NTU202302826

第一章　前言

1.1 研究背景

常見比較不同介入方式的統合分析，若是連續型資料，效果量通常是用平均差異（Mean Difference, MD），而類別型資料則是考慮使用風險比（Risk Ratio, RR）或勝算比（Odds Ratio, OR）。然而還有一種較為特別的診斷工具（Diagnostic Test）統合分析，診斷工具一般是用以檢測病患是否罹患疾病的某種治療或標準，針對同一疾病，不同的診斷工具判斷出的結果也可能有所差異，因此如何評估一個診斷工具的優劣就會是一大重點。想要透過統合分析來進行診斷工具的評估，我們會同時需要考慮兩個變項：敏感度（Sensitivity）與特異度（Specificity），由於當在不同研究之間時，此兩個變項並不獨立，因此我們應同時估計，此時如何對其個別的參數及兩者的相關性進行假設，便是其分析上的重點考量。

若是僅分析一種診斷工具，這在傳統統合分析上仍然相對簡單，只需要計算該診斷工具的敏感度、特異度、敏感度的變異程度、特異度的變異程度，以及敏感度與特異度的相關係數，一共 5 個參數，即可得到結果。但若要比較兩種以上的診斷工具的話，除了要找到剛好符合我們想要比較的兩種診斷工具的資料並不容易，同時大多情況其會是某一種診斷工具與黃金標準（Gold Standard）的診斷工具進行比較。此處我們可以將黃金標準亦視為一種診斷工具，因此現在我們就需要比較超過兩種的診斷工具，這時就需要借助網絡統合分析（Network Meta-Analysis, NMA），將不同的診斷工具彼此串聯，即可以間接比較的方式來比較不同診斷工具的結果，即可解決我們要比較原先我們想要知道的那些不同診斷工具之間差異的這個問題。但此時因為資料會變成提供兩個或以上的2×2列聯表，要建立的模型就會變得十分複雜，需要估計的參數會變得很多，此時如何妥善估計便是一大考驗。

doi:10.6342/NTU202302826

除此之外，在診斷工具統合分析的模型假設中，我們經常以黃金標準作為完美標準，即其 100%準確。但若是黃金標準並不完美，意即其並非完全準確，就可能會使我們在估計診斷工具的敏感度與特異度時出現偏差，造成統計結果估計上出現錯誤。由於一般而言假設疾病盛行率與敏感度、特異度是獨立的，彼此不互相影響，此時透過疾病盛行率並搭配研究中的各個診斷工具所計算而出的敏感度、特異度，重新對每個診斷工具的敏感度與特異度進行估計，便是可以考慮的作法，這樣的作法能夠避免因為黃金標準不準確所造成的估計偏差。因此如何將疾病盛行率作為參數加入模型之中，使模型得以準確估計各個診斷工具的敏感度、特異度，便是可以思考的課題。

在貝氏統計之下，我們能夠給予每個參數不同的先驗分配（Prior Distribution），並透過我們在系統性回顧下所取得的資料，更新此先驗分配的資訊，以求得其後驗分配（Posterior Distribution）。這樣的方法最大的好處即便不具有大樣本，我們仍然可以透過先驗分佈的資訊及其系統性回顧所取得的樣本，對參數進行估計。此外，若是要透過網絡統合分析比較多種診斷工具時，我們要估計的參數眾多，若能針對每個參數給予其分佈，我們可以直接針對每個不同參數的先驗分佈進行調整，亦能在此處針對每個參數進行分佈的假設與加入我們所考量的資訊，讓參數不會僅受限於樣本資料，也能增加模型的彈性。

因此，在本論文中我們希望透過貝氏統計的方式，透過給定每個參數不同的先驗分佈，並納入疾病盛行率作為參數之一，建立貝氏統計模型，以更有效的方式估計出其平均的敏感度與特異度。

1.2 研究目的

假設僅納入 1 個研究以比較一個診斷工具與黃金標準，此時在資料僅有真陽性、真陰性、偽陽性、偽陰性的情況下，受限於總樣本數，其僅能提供 3 個

doi:10.6342/NTU202302826

自由度，此處卻要估計診斷工具的敏感度、特異度、敏感度的變異程度、特異度的變異程度、敏感度與特異度的相關係數及疾病盛行率，一共 5 個參數，會產生無法估計的情況，此時則得多納入更多研究。但在比較多種診斷工具的網絡統合分析時，由於同時納入多種診斷工具，此時樣本變得以提供足夠的自由度對於個診斷工具之敏感度、特異度及其變異程度與相關係數，以及疾病盛行率進行估計。

本論文將著重於比較多種診斷工具時的網絡統合分析，並以貝氏統計的方式對模型參數進行估計。主要目標將著重於以傳統診斷工具統合分析之二元模型、診斷工具網絡統合分析之 ANOVA 模型為基礎，並以潛在類別分析（Latent Class Analysis, LCA）的方式，同時加入疾病盛行率（Prevalence）作為參數，以避免估計結果因黃金標準不準確而使估計結果出現偏誤，並調整其中各參數之先驗分配設定以作為模型延伸與改善之基礎。同時亦可計算其累積排行曲線下的面積（Surface Under The Cumulative Ranking Curve, SUCRA）、優勢指標（Superior Index）以及約登指標(Youden Index)，做為比較不同診斷工具優劣之指標。

doi:10.6342/NTU202302826

第二章　文獻回顧

2.1 診斷工具

診斷工具指的是一種檢測病患是否罹患疾病的方法[1, 2]。我們之所以要探討診斷工具的好壞，是因為診斷工具最主要的功能，就是在判定病患是否罹病[2]。但既然是判定其是否罹病，就有誤判的可能性發生，因此我們需要衡量診斷工具判斷的準確性（Accuracy），以確保在一個新的診斷工具出現時，其是具有較高的準確性能斷定病患是否罹病[2]。

診斷工具的網絡統合分析，較早的研究出現於 1993 年由 Midgette et al.所提出的研究[3]，在該文獻中其透過真陽性率（True Positive Rate, TPR）與真陰性率（True Negative Rate, TNR）做為分析診斷工具好壞的指標，而這兩個指標也就對應到我們日後的敏感度與特異度。在後續 Irwig et al.於 1995 年的研究中[4]，則點出在診斷工具的統合分析之中，通常會分別透過將敏感度與特異度兩者分別合併（Pooled）的方式進行分析，而後多數的統合分析與網絡統合分析對於診斷工具的研究，也多以敏感度與特異度呈現，因此本文後續也將以敏感度與特異度對診斷工具進行分析與評估。

2.2 敏感度與特異度

通常我們會使用一張2 × 2的列聯表來說明何謂敏感度與特異度。如表 1 所示[5, 6]，表格的列變數為其是否真的有病，而欄變數則是檢測時是否有病，若其檢測有病且真實情況有病，我們稱其為真陽性（True Positive, TP），若其檢測有病但真實情況並未得病，則我們稱其為偽陽性（False Positive, FP），以此類推，若其檢測無病但真實情況有病，我們稱其為偽陰性（False Negative, FN），若其檢測無病且真實情況也並未得病，則我們稱其為真陰性（True Negative, FP）。

doi:10.6342/NTU202302826

依據表 1[5, 6]，若我們將其橫向計算比例，即可求得陽性預測值（Positive Predictive Value, PPV）為真陽性人數除以所有檢測有病之人數；陰性預測值（Negative Predictive Value, NPV）則為真陰性人數數除以所有檢測無病之人數。若採用縱向計算比例，即可求得敏感度為真陽性人數除以真實有病人數之比例；特異度則為真陰性人數除以真實無病人數之比例。

由其計算公式我們可以得知，敏感度代表的是真實有病下檢測出有病的機率，特異度則是代表真實無病下檢測出無病的機率。若是在單一研究之中，敏感度與特異度兩者獨立，因為其人數計算分別來自於真實有病的人與真實無病的人，兩者並不關聯；但若是針對不同研究之間，敏感度與特異度兩者可能不獨立，因為敏感度與特異度的結果會受臨界值（Threshold）的影響，不同研究的臨界值不同，從而影響其敏感度與特異度。當臨界值設定較大時，結果會傾向於將患者判定為無病，就容易呈現出敏感度較小但特異度較大的結果，反之亦然，因此敏感度與特異度會呈現反向變動[5, 7]。由於在進行統合分析時，我們會考量不同研究的敏感度與特異度，而兩者並不獨立，因此必須考量其相關性，無法像介入方式的統合分析一樣，僅單純考量單一之應變數。

表 1 透過列聯表呈現敏感度與特異度計算與關聯

	真實有病	真實無病	
檢測有病	真陽性人數 (True Positive, TP)	偽陽性人數 (False Positive, FP)	陽性預測值$=\frac{TP}{TP+FP}$
檢測無病	偽陰性人數 (False Negative, FN)	真陰性人數 (True Negative, TN)	陰性預測值$=\frac{TN}{FN+TN}$
	敏感度$=\frac{TP}{TP+FN}$	特異度$=\frac{TN}{FP+TN}$	

doi:10.6342/NTU202302826

2.3 黃金標準

因為我們無法得知一名病患在真實情況下是否真的有罹患疾病，因此我們仍需要一個診斷工具來作為評斷標準。通常我們會使用「黃金標準」來稱作現階段世界上認為最為準確能否判斷病患是否真實有罹患疾病的診斷工具，而其他的診斷工具則通常會拿來與黃金標準進行比較。然而，黃金標準並非永恆不變，其仍可能隨著不同的時代而產生出更優良的診斷工具而使其被替換[5]。

由於我們會透過黃金標準判斷是否真實有病，因此原先的表 1，可重新修改成表 2，其餘計算公式則不變。

表 2 以黃金標準呈現真實有病及敏感度與特異度計算與關聯

	黃金標準之下		
	有病	無病	
檢測有病	真陽性人數 (True Positive, TP)	偽陽性人數 (False Positive, FP)	陽性預測值$=\frac{TP}{TP+FP}$
檢測無病	偽陰性人數 (False Negative, FN)	真陰性人數 (True Negative, TN)	陰性預測值$=\frac{TN}{FN+TN}$
	敏感度$=\frac{TP}{TP+FN}$	特異度$=\frac{TN}{FP+TN}$	

從上述內容中我們可以得知，因為大多數的診斷工具會跟黃金標準比較，若此時若我們要比較兩種不同診斷工具，即便其呈現於不同的研究之中，我們還是能夠透過網絡統合分析以及不同診斷工具與黃金標準的直接比較，將不同的診斷工具進行串聯，以間接的方式比較其差異，進而推算出不同診斷工具之優劣。

若此時有兩種診斷工具，則表 2 則可繼續延伸。此處為便於說明，我們改

doi:10.6342/NTU202302826

將黃金標準之結果呈現於列變數，而兩種不同診斷工具之結果則改為置放於欄變數的位置，此時相較於先前之表格，此處將變為有 8 個細格，其中真陽性、偽陽性、偽陰性、真陰性對應到兩種診斷工具與黃金標準之比較，將各產生兩個項目，其結果將如表 3 所呈現。

表 3 黃金標準與兩種診斷工具呈現之列聯表

		診斷工具 I		診斷工具 II	
		有病	無病	有病	無病
黃金標準之下	有病	TP_1	FP_1	TP_2	FP_2
	無病	FN_1	TN_1	FN_2	TN_2
		敏感度$_1$ $= \frac{TP_1}{TP_1+FN_1}$	特異度$_1$ $= \frac{TN_1}{FP_1+TN_1}$	敏感度$_2$ $= \frac{TP_2}{TP_2+FN_2}$	特異度$_2$ $= \frac{TN_2}{FP_2+TN_2}$

從表 3 中，我們亦可算出兩種診斷工具其各自的敏感度與特異度，但此時若針對每個敏感度與特異度都考量其隨機效果，此時需要考量所有診斷工具在不同研究之間敏感度與特異度彼此的相關性，而並非只考量同一診斷工具之中的敏感度與特異度的相關程度[8, 9]。

一般而言在診斷工具的統合分析之中，我們會假設黃金標準為完美黃金標準，即其 100%正確，然而當黃金標準可能有偏差，或是對於某種疾病的檢測並不存在黃金標準時，這樣的估計就可能使得模型不準確。此時就需要調整對於診斷工具統合分析的模型假設，以使模型能夠在黃金標準不準確時仍得以運用[10]。

doi:10.6342/NTU202302826

2.4 貝氏統計與網絡統合分析

在 Jansen et al.於 2014 的研究中點出[11]，在網絡統合分析之中，可以透過頻率學派或貝氏學派的作法來進行分析。通常在頻率學派之下，結果會直接透過 95%信賴區間（Confidence interval, CI）呈現，並同時也會呈現出 *p* 值（*p*-value），以藉此判定其結果在統計上是否顯著；然而由於貝氏學派認為在考慮資料前，應先考量資料的先驗分配，再透過系統性回顧的資料更新分配，以求取後驗分配。也因此在貝氏學派的架構下，我們最後會呈現出後驗分配之點估計（point estimate）與 95%可信區間（Credible Interval, CrI）。由於貝氏統計假設參數服從一個分配，因此其結果亦會針對母體參數估計出一個分配，並非像頻率學派直接以樣本統計量對母體參數進行估計，因此不能直接像頻率學派的架構下，將 95%可信區間採取與 95%信賴區間相同的解釋，而誤將其解釋為「有 95%的信心其診斷工具之效果包含真正的母體參數之估計」。在貝氏學派的架構之中，其會針對每個不同的診斷工具，給與其一個可能的機率，因此我們可以知道每個不同的診斷工具會有多大的機率其能夠有最高的準確率，以藉此比較不同診斷工具的好壞。

若想要求取貝氏統計的點估計與 95%置信區間，通常我們會透過馬可夫鍊蒙地卡羅法（Markov Chain Monte Carlo Method, MCMC Method），透過其演算法分析結果的收斂，以求取其分析結果[12]。因為其實透過演算法求取收斂，儘管每次分析上數值可能會有些微差異，但通常相差無幾，因此我們仍可以求取相似的結果。

由於貝氏統計的估計結果是給予其一個後驗分配，並非像頻率學派直接透過樣本求取母體發生的可能性，因此不能直接像頻率學派的架構下，將 95%可信區間採取與 95%信賴區間一樣解釋為「有 95%的信心其診斷工具之效果包含真正的母體參數之估計」。在貝氏學派的架構之中，其會針對每個不同的診斷工

doi:10.6342/NTU202302826

具，給與其一個可能的機率，因此我們可以知道每個不同的診斷工具會有多大的機率其能夠有最高的準確率，以藉此比較不同診斷工具的好壞[11]。因為貝氏統計透過機率的方式呈現其結果，也給予了其更好的預測能力，而透過機率的方式呈現不確定性（Uncertainty），也使得貝氏統計之模型相較於頻率學派之模型更加具有彈性[12]。

2.5 診斷工具網絡統合分析

2.5.1 二元模型（Bivariate Model）

此處我們退回到只有一種診斷工具時的狀況，若在只有一種診斷工具時，此時僅有一個敏感度與一個特異度，我們須考量的參數只有 5 個：敏感度的平均值、特異度的平均值、敏感度的變異數、特異度的變異數，以及敏感度與特異度的相關係數[10, 13]。該模型中我們假設敏感度為p_{Ai}、特異感度為p_{Bi}，其中下標 i 表示其為第 i 個研究，透過 logit 轉換，我們可以得到下列結果：

$$logit(p_{Ai}) = \mu_{Ai},\ i = 1, \dots, I \tag{1}$$

$$logit(p_{Bi}) = \mu_{Bi},\ i = 1, \dots, I \tag{2}$$

式(1)與式(2)中的μ_{Ai}與μ_{Bi}則服從一個多變量常態分配（Multi-Variate Normal Distribution），如式(3)所示。

$$\begin{pmatrix}\mu_{Ai}\\ \mu_{Bi}\end{pmatrix} \sim MVN\left(\begin{pmatrix}\mu_A\\ \mu_B\end{pmatrix}, \begin{pmatrix}\sigma_A^2 & \rho\sigma_A\sigma_B\\ \rho\sigma_A\sigma_B & \sigma_B^2\end{pmatrix}\right) \tag{3}$$

其中，參考 Nishiumra 的研究，給予了式(3)當中的μ_A服從一平均數為 0、變異數為 4 的常態分配作為先驗分配，μ_B也服從另一平均數為 0、變異數為 4 的常態分配為先驗分配。而σ_A與σ_B則在其中給定各自的θ_A與θ_B，並假設θ_A與θ_B

doi:10.6342/NTU202302826

的先驗分配各自服從一個α為 2、λ為 5 的伽瑪分配（Gamma Distribution），ρ因為其應界於-1 至 1 之間，因此則給予其一個範圍為-1 至 1 的均勻分配（Uniform Distribution）作為先驗分配。若將其整理成數學式，將式(4)至式(7)所示[14]。

$$\mu_A \sim N(0,4),\ \mu_B \sim N(0,4) \tag{4}$$

$$\sigma_A = \sqrt{\theta_A},\ \theta_A \sim Gamma(2,0.5) \tag{5}$$

$$\sigma_B = \sqrt{\theta_B},\ \theta_B \sim Gamma(2,0.5) \tag{6}$$

$$\rho \sim Unif(-1,1) \tag{7}$$

此處我們則假設真陽性、真陰性分別服從一個二項分配（Binomial Distribution），如式(8)與式(9)所示。

$$TP_i \sim Bin(TP_i + FN_i, p_A) \tag{8}$$

$$TN_i \sim Bin(TN_i + FP_i, p_B) \tag{9}$$

以上在經過馬可夫鍊蒙地卡羅法迭代後，計算出之μ_A與μ_B，再經過反函數計算，如式(10)與式(11)，即可求出整體之敏感度p_A與特異度p_B。

$$p_A = antilogit(\mu_A) = \frac{e^{\mu_A}}{1 + e^{\mu_A}} \tag{10}$$

$$p_B = antilogit(\mu_B) = \frac{e^{\mu_B}}{1 + e^{\mu_B}} \tag{11}$$

但在二元模型中，有一假設為其所比較之黃金標準為完美黃金標準，意即該標準為 100%準確，若黃金標準並不完美時，其估計結果就可能產生偏誤，因

doi:10.6342/NTU202302826

此當黃金標準不完美時，該如何去對觀察到之資料進行調整，便是一個延伸的課題。

2.5.2 潛在類別分析（Latent Class Analysis, LCA）

由於在潛在類別分析中，其將未觀察到的潛在類別視為 K 個潛在類別變數[15]，而在診斷工具的統合分析中，我們將患者是否真實罹病視為一未觀察到的變數，因此此處僅包含兩個類別：罹病與未罹病。我們將透過這樣的方式，以針對黃金標準不完美時的觀察結果進行調整，以使觀察結果準確。

在潛在類別分析中[10, 15]，為了同時納入對於是否罹病的考量，我們添加了隨機變數π_i作為一潛在變數，其代表的涵義為發病率（Disease Rate），或疾病盛行率（Prevalence），主要目的是為了增加求算真陽性、偽陽性、偽陰性、真陰性四者時得以同時考到其他研究的資料，以避免各研究所直接給予的數值，可能因黃金標準不準確而有偏差，如式(12)至式(16)所示，其中Se_i與Sp_i分別表我們感興趣之診斷工具的敏感度與特異度，Se_{Ri}與Sp_{Ri}則分別代表作為不完美黃金標準的診斷工具，此處假設感興趣之診斷工具與作為不完美黃金標準的診斷工具兩者的敏感度互為獨立、兩者的特異度亦互為獨立，方能透過式(12)至式(15)計算其實際之真陽性、偽陽性、偽陰性、真陰性之比例。由於疾病的發病率應界於 0 與 1 之間，故此處給予其一個α與β均為 1 的 Beta 分配（Beta Distribution）作為先驗分配，如式(16)所示。

$$Prob_i^{TP} = \pi_i \times Se_i \times Se_{Ri} + (1-\pi_i) \times (1-Sp_i) \times (1-Sp_{Ri}) \tag{12}$$

$$Prob_i^{FP} = \pi_i \times Se_i \times (1-Se_{Ri}) + (1-\pi_i) \times (1-Sp_i) \times Sp_{Ri} \tag{13}$$

$$Prob_i^{FN} = \pi_i \times (1-Se_i) \times Se_{Ri} + (1-\pi_i) \times Sp_i \times (1-Sp_{Ri}) \tag{14}$$

doi:10.6342/NTU202302826

$$Prob_i^{TN} = \pi_i \times (1 - Se_i) \times (1 - Se_{Ri}) + (1 - \pi_i) \times Sp_i \times Sp_{Ri} \quad (15)$$

$$\pi_i \sim Beta(1,1) \quad (16)$$

資料裡中的概似函數，我們則設定其資料服從一多項分佈，其參數為先前估計出真陽性、偽陽性、偽陰性、真陰性以及其總人數，其呈現如式(17)。

$$\boldsymbol{r_i} \sim Multi\left(N_i, (Prob_i^{TP}, Prob_i^{FP}, Prob_i^{FN}, Prob_i^{TN})\right) \quad (17)$$

站在流行病學的角度上，通常我們會視疾病盛行率與敏感度及特異度為獨立，因此不會於考量診斷工具時，將疾病盛行率納入考量[16, 17]。但若是黃金標準並非最準確的診斷工具，或是黃金標準不存在時，此時要評估診斷工具之準確性，就會變得很困難，此時若加入疾病盛行率作為參數，同時考量不同研究所提供之數值，可以使模型估計出之診斷工具的敏感度與特異度數值準確，而不會使其可能受到黃金標準不準確而出現偏差。

2.5.3 ANOVA 模型（ANOVA Model）

在 Nyaga et al.於 2018 年發表的 ANOVA 模型論文中[9]，引述了 Jing Zhang et al.於 2014 年的研究之中[18]，其假設第 i 個事件y_i為二項分配，如式(18)所示。而p_{ik}則可透過平均加上其隨機效果，以式(19)進行估計，並且進一步透過式(20)估計各治療的效果π_k。

$$P(Y_i = y_i) \propto \prod_{k \in S_i} (p_{ik})^{y_{ik}} (1 - p_{ik})^{n_{ik} - y_{ik}},\ i = 1, \ldots, I \quad (18)$$

$$\Phi^{-1}(p_{ik}) = \mu_k + \sigma_k \nu_{ik},\ i = 1, \ldots, I,\ k = 1, \ldots, K \quad (19)$$

doi:10.6342/NTU202302826

$$\pi_k = E(p_{ik}|\mu_k, \sigma_k) = \int \Phi(\mu_k + \sigma_k z)\phi(z)dz = \Phi\left(\mu_k/\sqrt{1+\sigma_k^2}\right), \quad k = 1, \ldots, K \tag{20}$$

由於這樣的積分程序並不容易執行，因此多數時間會直接透過貝氏統計軟體，並以馬可夫鍊蒙地卡羅法直接進行模擬，以求出各治療的效果。

回到 ANOVA 模型本身[9]，由於我們想要同時估計與比較兩種或兩種以上診斷工具的敏感度與特異度，但在比較多種診斷工具時，可能並非所有研究皆包含所有診斷工具的資料，在該模型中假設這些研究若未包含該診斷工具之資料時，其為隨機缺失(Missing at Random, MAR)，而在該模型中即可透過多種診斷工具所建構之網絡結構，並以此去插補遺失資料，以估計與比較所有的診斷工具之效果。基於 Arm-Based Model，假定有 I 個研究項目與 K 個診斷工具，假設參數為真陽性人數（Y_{i1k}）與真陰性人數（Y_{i2k}），這兩個參數服從二項分配，其對應之機率分別為敏感度（p_{i1k}）與特異度（p_{i2k}），樣本數則為有病人數（N_{i1k}）與無病人數（N_{i2k}），如式(21)與式(22)所示。

$$Y_{i1k}|p_{i1k} \sim Bin(N_{i1k}, p_{i1k}),\ i = 1, \ldots, I,\ k = 1, \ldots, K \tag{21}$$

$$Y_{i2k}|p_{i2k} \sim Bin(N_{i2k}, p_{i2k}),\ i = 1, \ldots, I,\ k = 1, \ldots, K \tag{22}$$

而其敏感度與特異度的隨機效果可透過羅吉斯迴歸估計，如式(23)與式(24)所示，其中$antilogit(\mu_{1k})$與$antilogit(\mu_{2k})$為敏感度與特異度的平均效果。η_{i1}與η_{i2}為隨機效果，其分別代表第 i 個研究中，每個研究的敏感度及特異度與平均敏感度及平均特異度之差值所構成的變異，此部分代表的即為不同研究間的組間變異，我們假設它們二者相關，服從二元常態分配，如式(25)，特別注意此處為使模型簡單，我們假設所有診斷工具在敏感度與特異度的組間變異皆一致。δ_{i1k}與δ_{i2k}則表示敏感度與特異度的在不同診斷工具中的組內變異，其代表

doi:10.6342/NTU202302826

的是若有某一個研究其結果之數值與第 k 個診斷工具平均敏感度及平均特異度具有較大差異，即其在該研究中該診斷工具表現特別好或特別不好時，其差值的變異即會反應於此，而此二者獨立，各自服從常態分配，如式(26)與式(27)所呈現，其中τ_{1k}^2表示其在第 k 個診斷工具中敏感度的變異程度，τ_{2k}^2則表其在第 k 個診斷工具中特異度的變異程度。

$$logit(\hat{p}_{i1k}) = \mu_{1k} + \eta_{i1} + \delta_{i1k},\ i = 1, \dots, I,\ k = 1, \dots, K \tag{23}$$

$$logit(\hat{p}_{i2k}) = \mu_{2k} + \eta_{i2} + \delta_{i2k},\ i = 1, \dots, I,\ k = 1, \dots, K \tag{24}$$

$$\begin{pmatrix}\eta_{i1}\\ \eta_{i2}\end{pmatrix} \sim MVN\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \rho_{12}\sigma_1\sigma_2\\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}\right) \tag{25}$$

$$\delta_{i1k} \sim N(0, \tau_{1k}^2) \tag{26}$$

$$\delta_{i2k} \sim N(0, \tau_{2k}^2) \tag{27}$$

由於在上述的式(26)與式(27)中，τ_{1k}^2與τ_{2k}^2會隨著診斷工具增加，而產生出過多的參數需要估計，這會使得模型變得非常複雜。因此在 Nyaga 的論文中亦建議，可以透過假設敏感度在不同診斷工具中皆有相同的異質性，這樣τ_{1k}^2就可簡化為僅需估計一個參數τ_1^2，特異度也可以此類推，假定其在不同診斷工具中也具有相同的異質性，τ_{2k}^2也就只需要估計一個τ_2^2即可。更可以再繼續簡化，假設敏感度與特異度兩者在所有的異質性上也皆一致，這樣更可以將兩者再次簡化成僅須估計τ^2即可。這樣的模型又稱之為 Variance Component Model 或 Reduced Model。

doi:10.6342/NTU202302826

第三章　方法

3.1 加入疾病盛行率後的診斷工具網絡統合分析貝氏模型

由於在 ANOVA 模型中，其並未考量黃金標準可能不完美這件事情，且考量模型估計參數設定上的便利性，此處我們以 2.5.3 小節所提到的 ANOVA 模型，並搭配 2.5.2 小節的潛在類別分析，以此方法納入疾病盛行率作為考量的參數，重新建立一個貝氏統計架構下的診斷工具網絡統合分析模型。

此處對於平均的敏感度與特異度，我們以μ_{1k}與μ_{2k}的估計，如式(28)、式(29)與式(30)，其中式(28)、式(29)的μ_{i1k}與μ_{2k}表第 k 個敏感度與特異度的隨機效果。估計後再透過$antilogit(\mu_{1k})$與$antilogit(\mu_{2k})$以轉換回每一個診斷工具其敏感度與特異度的平均效果。

在該 ANOVA 模型中所估計之敏感度與特異度的組間變異η_{i1}與η_{i2}，此處可以發現在此二者並沒有下標 k，表示在不同診斷工具之下，平均敏感度與平均特異度仍然服從相同之變異數與共變異數矩陣，亦表示其在不同診斷工具下仍只會估計出一個敏感度與一個特異度的組間變異。而針對η_{i1}及η_{i2}，我們給予其一個二元常態分配作為先驗分配，如式(30)所示，以估計其效果，而其餘相關性之各參數之估計，呈現於式(31)至式(34)。

我們亦假設敏感度與特異度的組內變異δ_{i1k}與δ_{i2k}，分別服從一常態分配，並假設其中之變異數服從齊一分配，如式(35)與式(36)所示。此處為簡化參數，我們假設在不同的診斷工具之中，仍然有相同的變異程度。不同於組間變異，δ_{i1k}與δ_{i2k}帶有下標 k，兩者獨立，表示在每個診斷工具之下，其皆針對敏感度與特異度分別估計組內變異。

$$logit(\hat{p}_{i1k}) = \mu_{1k} + \eta_{i1} + \delta_{i1k},\ i = 1, \dots, I,\ k = 1, \dots, K \tag{28}$$

doi:10.6342/NTU202302826

$$logit(\hat{p}_{i2k}) = \mu_{2k} + \eta_{i2} + \delta_{i2k},\ i = 1, \dots, I,\ k = 1, \dots, K \tag{29}$$

$$\begin{pmatrix} \eta_{i1} \\ \eta_{i2} \end{pmatrix} \sim MVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right) \tag{30}$$

$$\mu_{1k} \sim N(0, 4),\ \mu_{2k} \sim N(0, 4) \tag{31}$$

$$\sigma_1 = \sqrt{\sigma_1^2},\ \sigma_1 \sim Gamma(2, 0.5) \tag{32}$$

$$\sigma_2 = \sqrt{\sigma_2^2},\ \sigma_2 \sim Gamma(2, 0.5) \tag{33}$$

$$\rho_{12} \sim Unif(-1, 1) \tag{34}$$

$$\delta_{i1k} \sim N(0, \tau_1^2),\ \tau_1 = \sqrt{\tau_1^2},\ \tau_1 \sim Unif(0, 2) \tag{35}$$

$$\delta_{i2k} \sim N(0, \tau_2^2),\ \tau_2 = \sqrt{\tau_2^2},\ \tau_2 \sim Unif(0, 2) \tag{36}$$

而後我們仍保留潛在類別分析中，其假設一隨機變數π_i，作為估計發病率或疾病盛行率，以增加求算真陽性、偽陽性、偽陰性、真陰性四者時，對於不完美黃金標準的考量，呈現於式(37)至式(40)。針對概似函數，我們亦保留多項分佈的設計，其參數為先前估計出真陽性、偽陽性、偽陰性、真陰性以及其總人數，其呈現如式(41)。此外，我們在此處以Se_R與Sp_R分別表示不完美診斷工具的敏感度與特異度，並各給予其一範圍為 0.8 至 1 的均勻分配作為先驗分配，如式(42)，此處以固定效果進行估計。

$$Prob_{ik}^{TP} = \pi_{ik} \times Se_{ik} \times Se_R + (1 - \pi_{ik}) \times (1 - Sp_{ik}) \times (1 - Sp_R) \tag{37}$$

$$Prob_{ik}^{FP} = \pi_{ik} \times Se_{ik} \times (1 - Se_R) + (1 - \pi_{ik}) \times (1 - Sp_{ik}) \times Sp_R \tag{38}$$

doi:10.6342/NTU202302826

$$Prob_{ik}^{FN} = \pi_{ik} \times (1 - Se_{ik}) \times Se_R + (1 - \pi_{ik}) \times Sp_{ik} \times (1 - Sp_R) \quad (39)$$

$$Prob_{ik}^{TN} = \pi_{ik} \times (1 - Se_{ik}) \times (1 - Se_R) + (1 - \pi_{ik}) \times Sp_{ik} \times Sp_R \quad (40)$$

$$\boldsymbol{r_{ik}} \sim Multi\left(N_{ik}, (Prob_{ik}^{TP}, Prob_{ik}^{FP}, Prob_{ik}^{FN}, Prob_{ik}^{TN})\right) \quad (41)$$

$$Se_R \sim Unif(0.8, 1),\ Sp_R \sim Unif(0.8, 1) \quad (42)$$

此處特別注意，相較於先前於式(12)至式(15)及式(17)，在式(37)至式(41)中，其估計式皆多了下標 k。這是為了在考量多種不同診斷工具時，針對每個診斷工具我們都應分別考量其不同狀況，因此才會多出這個下標。

另外，此處我們為設定疾病盛行率之先驗分配，因此我們將其透過 logit 函數轉換，定義符號為φ_{ik}，並令其服從常態分配，呈現於式(44)中，而其餘常態分配中的參數則如式(45)所呈現。

$$logit(\pi_{ik}) = \varphi_{ik} \quad (43)$$

$$\varphi_{ik} \sim N(\nu, \zeta^2) \quad (44)$$

$$\nu \sim N(0, 4),\ \zeta = \sqrt{\zeta^2},\ \zeta \sim Gamma(2, 0.5) \quad (45)$$

將以上分析經過馬可夫鍊蒙地卡羅法迭代後，計算出之μ_{1k}與μ_{2k}透過 *antilogit* 函數經過計算，如式(46)與式(47)，即可求出各個不同診斷工具下的敏感度μ_{1k}與特異度μ_{2k}。

$$p_{1k} = antilogit(\mu_{1k}) = \frac{e^{\mu_{1k}}}{1 + e^{\mu_{1k}}} \quad (46)$$

doi:10.6342/NTU202302826

$$p_{2k} = antilogit(\mu_{2k}) = \frac{e^{\mu_{2k}}}{1 + e^{\mu_{2k}}} \quad (47)$$

3.2 不同診斷工具效果優劣排名

3.2.1 以累積排行曲線下的面積（SUCRA）進行優劣排名

由於 Nyaga et al.所提出的 ANOVA 模型，其原型是基於基準模型（Baseline Model）[9]，因此此處我們亦以基準模型所計算累積排行曲線下的面積（Surface Under the Cumulative Ranking Curve, SUCRA）的方式來對不同診斷工具之效果優劣進行評估。

計算 SUCRA 值的方式如下[19]，我們首先計算第 i 個診斷工具其排名落在第 j 位的機率，並將其以$P(i,j)$表示，而後我們計算該診斷工具其排名的的累積分配函數，其可透過式(48)可求得其累積分配函數。

$$F(i,k) = \sum_{j=1}^{k} P(i,j) \quad (48)$$

求得累積分配函數後，即可透過其求得 SUCRA 值，如式(49)所示。SUCRA 值可做為衡量診斷工具優劣之指標。若其數值愈高，則代表其為較佳的診斷工具的可能性愈大。

$$SUCRA(i) = \frac{\sum_{k=1}^{n-1} F(i,k)}{n-1} \quad (49)$$

3.2.2 以優勢指標（Superiority Index）進行優劣排名

在 Deutsch et al.於 2009 年發表的論文中則提供了另一指標 Superiority Index 進行優劣排名的計算[20]。其計算公式如式(50)所示，其中S_i表第 i 個診斷工具的 Superiority Index，a_i其表現優於第 i 個診斷工具的診斷工具的數量，b_i為表現劣於第 i 個診斷工具的診斷工具的數量，c_i為表現等於第 i 個診斷工具的診斷

doi:10.6342/NTU202302826

工具的數量。

$$S_i = \frac{a_i + c_i/2}{b_i + c_i/2} = \frac{2a_i + c_i}{2b_i + c_i} \tag{50}$$

若 Superiority Index 愈大，表示該診斷工具的表現愈好，因此最大的則為透過該指標所選出之最佳診斷工具，而第二高者則次之，依此類推。

3.2.3 以約登指標（Youden Index）進行優劣排名

另一經常用於評估敏感度與特異度之結果之指標為類別資料分析中常使用的 Youden Index，此為 W. J. Youden 於 1950 年所提出指標[21]，其計算公式呈現於式(51)。其中J_i表第 i 個診斷工具的 Youden Index，Sen_i與Spe_i則分別為第 i 個診斷工具的敏感度與特異度。由於在類別資料分析中，接收者操作特徵曲線（Receiver Operating Characteristic Curve, ROC Curve）通常為一上凸之曲線，此公式可計算出該曲線離對角線之最大值，進而求出最理想之分界點，因此亦可將其應用於診斷工具的網絡統合分析之中，評估敏感度與特異度之分析結果，以作為優劣排名之指標。

$$J_i = Sen_i + Spe_i - 1 \tag{51}$$

Youden Index 為一介於 0 至 1 之間之數值，若其愈大，表示該診斷工具的表現愈好，因此較大者為透過該指標所選出之最佳診斷工具，而第二高者則次之，依此類推。

doi:10.6342/NTU202302826

第四章　實例分析

在該章節中，我們將分別考慮 Hoyer and Kuss 在 2018 年所發表的研究中的資料[8]以及 Veroniki et al.在 2021 年所發表的研究中的資料[22]進行實例分析，以檢驗上述章節中所提出之模型之分析結果的合理性。

4.1 Hoyer and Kuss 的第二型糖尿病檢測資料

此處我們以 Hoyer and Kuss 在 2018 年發表的研究中所使用之資料來進行分析[8]，其資料主要透過 C. M. Bennett et al.以及 Satoru Kodama et al.兩篇論文的系統性回顧而來[23, 24]。該資料中主要透過三種方式對第二型糖尿病（Type 2 Diabetes Mellitus）進行檢測，分別為口服葡萄糖耐量試驗（Oral Glucose Tolerance Test, OGTT)、糖化血色素檢測(measurement of HbA_{1c}, HbA1c)，以及空腹血糖試驗（Fasting Plasma Glucose, FPG）。在該研究中口服葡萄糖耐量試驗被視為是參考準則，以比較檢驗糖化血色素測驗以及空腹血糖測試兩種診斷工具的準確度。

4.1.1 資料介紹

該資料中總共納入了 38 篇包含以口服葡萄糖耐量的糖化血色素檢測以及空腹血糖試驗結果的測驗資料，且在這 38 篇研究中皆包含了上述兩種資料，且在每篇研究中，分別接受兩種測驗的檢測之人數相等，其網絡圖如呈現如圖 1 所示，而參考準則作為與兩種測驗的比較標準，則不會呈現於圖上。

doi:10.6342/NTU202302826

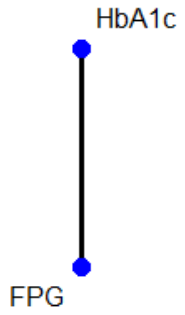

圖 1 檢測第二型糖尿病的方法網絡圖

由圖 1 中可以發現，由於所有文獻皆包含兩種檢測方法且人數相同，因此其兩個圓圈為相同大小。表 4 則呈現了該文獻中之原始資料。

表 4 Hoyer and Kuss (2018) 中之原始資料

編號	作者	TP1	FP1	FN1	TN1	TP2	FP2	FN2	TN2
1	Badings et al.	574	682	262	1389	633	465	203	1606
2	Choi et al.	489	1774	146	6966	445	524	190	8216
3	Li et al.	36	13	95	998	33	16	120	973
4	Schöttker et al.	338	29	2376	4060	266	101	1389	5047
5	Tahrani et al.	16	25	10	147	21	20	25	132
6	Wang et al.	424	192	121	2112	612	4	1281	952
7	Hu et al.	644	151	286	1217	648	147	293	1210
8	Zhang et al.	50	14	4	40	57	7	6	38
9	Zhou et al.	176	102	768	1286	206	72	823	1231
10	Kim et al.	72	16	46	258	75	13	35	269
11	Nakagami et al.	89	26	302	1382	74	41	79	1605
12	Salmasi et al.	23	7	5	109	16	14	21	93
13	Glümer et al.	181	71	1988	3877	198	54	721	5144
14	Anand et al., South Asia	25	2	45	243	24	3	60	228
15	Anand et al., China	12	2	25	268	12	2	59	234
16	Anand et al., Europe	13	6	35	260	9	10	40	255
17	Jesudason et al.	43	11	62	389	40	14	24	427
18	Tavintharan et al.	17	4	11	79	10	11	2	88
19	Ko et al.	575	52	1270	980	554	73	469	1781
20	Papoz et al.	100	12	108	381	77	35	103	386
21	Choi et al.	610	285	1692	3358	555	340	1667	3383

doi:10.6342/NTU202302826

表 4 Hoyer and Kuss (2018) 中之原始資料 (續)

編號	作者	TP1	FP1	FN1	TN1	TP2	FP2	FN2	TN2
22	Heianza et al.	184	154	638	5265	262	76	1418	4485
23	Law et al.	58	23	129	204	22	59	25	308
24	Mukai et al.	195	100	718	969	199	96	580	1107
25	Soulimane et al., Denmark	74	40	1156	3660	80	34	771	4045
26	Soulimane et al., Australia	145	41	1107	4719	121	65	641	5185
27	Soulimane et al., France	61	31	742	2950	69	23	876	2816
28	Cederberg et al.	21	43	36	284	14	50	24	296
29	Nakagami et al.	42	15	318	814	35	22	198	934
30	Sato et al.	392	267	1130	5015	541	118	2116	4029
31	Inoue et al.	187	181	1112	8562	328	40	2411	7263
32	Inoue et al.	9	8	37	395	15	2	71	361
33	Norberg et al.	88	76	39	265	82	82	33	271
34	Takahashi et al.	52	13	37	79	39	26	29	87
35	Ko et al.	22	22	35	129	19	25	20	144
36	Mannucci et al.	79	1	689	223	75	5	686	226
37	Wiener et al.	114	64	20	203	139	39	27	196
38	Tanaka et al.	135	43	96	592	93	85	0	688

4.1.2 分析結果

針對 3.1 節中所提出的完整模型，我們透過 OpenBUGS 以 3 個鍊（chain）迭代 100,000 次，並將其前 50,000 次結果預燒（burn-in），一共迭代 150,000 次，並搭配 R 語言中之 R2OpenBUGS 套件協助執行。

分析結果呈現於表 5 之中。透過該表可以發現，兩種治療的敏感度與特異度都有達到 0.7 以上，顯示出這兩種診斷工具皆能對於第二型糖尿病有著不錯的診斷效果，且作為準則之口服葡萄糖耐量之敏感度與特異度皆有達到 0.98，因此其作為參考準則是一合適之標準。

doi:10.6342/NTU202302826

表 5 以 OpenBUGS 對於兩種糖尿病診斷工具與參考準則的分析結果

項目		估計值	標準誤	95%可信區間	
糖化血色素	敏感度	0.7484	0.0338	0.6810	0.8140
	特異度	0.8082	0.0275	0.7471	0.8555
空腹血糖試驗	敏感度	0.7447	0.0335	0.6762	0.8083
	特異度	0.8411	0.0231	0.7935	0.8849
口服葡萄糖耐量	敏感度	0.9865	0.0127	0.9527	0.9996
(參考準則)	特異度	0.9973	0.0014	0.9946	0.9998

若改以 Stata 分析的 melogit 函數執行分析（於 Stata 中並未考量黃金標準不完美之議題），可以得到如表 6 之結果。可以看到下表的結果與使用 OpenBUGS 以貝氏統計分析的方法相近。

表 6 以 Stata 對於兩種糖尿病診斷工具的分析結果

項目		估計值	95%可信區間	
糖化血色素	敏感度	0.7227	0.6686	0.7710
	特異度	0.8095	0.7640	0.8480
空腹血糖試驗	敏感度	0.7331	0.6621	0.7938
	特異度	0.8409	0.7913	0.8805

4.1.3 比較不同診斷工具效果之優劣排名

考量 3.2 節中之比較不同診斷工具優劣的方式，此處分別針對敏感度與特異度考量其 SUCRA 之結果，其分別如表 7 與表 8 所示。此處可以看到其與分析結果上與前述相同，顯示出不論在敏感度與特異度上，口服葡萄糖耐量有最高的 SUCRA 值，因此其作為參考準則是十分適合的指標。而空腹血糖試驗與糖化血色素都有較高的 SUCRA 值，表示其在診斷效果上應較為準確。

doi:10.6342/NTU202302826

表 7 兩種糖尿病診斷工具與參考準則敏感度優劣之排名機率與 SUCRA 值

項目	Rank 1	Rank 2	Rank 3	SUCRA
糖化血色素	0.0000	0.5337	0.4663	0.2669
空腹血糖試驗	0.0000	0.4663	0.5337	0.2331
口服葡萄糖耐量(參考準則)	1.0000	0.0000	0.0000	1.0000

表 8 兩種糖尿病診斷工具與參考準則特異度優劣之排名機率與 SUCRA 值

項目	Rank 1	Rank 2	Rank 3	SUCRA
糖化血色素	0.0000	0.1848	0.8152	0.0924
空腹血糖試驗	0.0000	0.8152	0.1848	0.4076
口服葡萄糖耐量(參考準則)	1.0000	0.0000	0.0000	1.0000

若採用 Superiority Index 作為指標，有別於 SUCRA 會將敏感度與特異度分開進行比較，Superiority Index 則會直接將兩者合併考量並計算出一個數值，來決定該診斷工具的優劣，其效果呈現於表 9 中。根據其數值呈現結果，我們可以知道空腹血糖試驗依舊較糖化血色素來得更為準確，但口服葡萄糖耐量仍有最高之數值，表示其作為一參考準則是合適的。

表 9 兩種糖尿病診斷工具與參考準則之 Superiority Index

項目	Superiority Index
糖化血色素	0.3236
空腹血糖試驗	0.5488
口服葡萄糖耐量(參考準則)	5.0000

以 Youden Index 作為指標時，如同 Superiority Index，其亦會將兩者合併考

doi:10.6342/NTU202302826

量並計算出一個數值，來決定該診斷工具的優劣，而其數值則應居於 0 至 1 之間。表 10 呈現了各項診斷工具透過 Youden Index 評估之排名結果，此處仍然反映出空腹血糖試驗為較糖化血色素更理想之診斷工具，但口服葡萄糖耐量仍有最高之數值，表示其作為一參考準則是合適的。

表 10 兩種糖尿病診斷工具與參考準則之 Youden Index

項目	Youden Index
糖化血色素	0.5566
空腹血糖試驗	0.5859
口服葡萄糖耐量 (參考準則)	0.9839

4.2 Veroniki et al.的 CIN2+子宮頸癌檢測資料

此處我們利用 Veroniki et al.在 2021 年發表的文章中的資料[22]，該資料主要包含三種檢測 CIN2+型的子宮頸癌的方法：細胞學（Cytology）、人類乳突病毒 DNA（Human Papillomavirus Deoxyribonucleic Acid, HPV DNA）與信使核糖核酸（Messenger Ribonucleic Acid, mRNA）與三種方法，其中細胞學則包含子宮頸上皮細胞異常（Atypical Squamous Cells of Undetermined Significance, ASCUS+）與低階病變（Low-Grade Squamous Intraepithelial Lesion, LSIL+）兩種。而在該研究中，其以陰道鏡檢查與組織學（Colposcopy and/or Histology）作為參考準則，比較前述三種診斷工具的準確度。

4.2.1 資料介紹

該資料中總共包含了 37 篇文獻，其中有 1 篇僅包含細胞學的結果、1 篇僅包含人類乳突病毒 DNA 的結果、31 篇為細胞學與人類乳突病毒 DNA 兩者結果的研究，以及 4 篇則為包含了三種結果的研究。其網絡圖如圖 2 所示。

doi:10.6342/NTU202302826

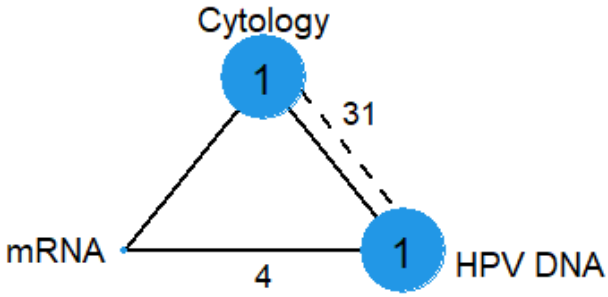

圖 2 檢測 CIN2+型的子宮頸癌的方法網絡圖

圖 2 中位於點上的數字，表示僅包含該類型診斷工具的研究篇數，虛線則表示同時包含相連的 2 種診斷工具的研究篇數，實線則為包含所有診斷工具的研究篇數。從圖 2 可以發現，受到研究篇數影響，因此在細胞學與人類乳突病毒 DNA 的診斷工具上，所收到樣本人數會比較多，而信使核糖核酸人數則較少。表 11 則為於 Veroniki et al.中之完整資料。

表 11 Veroniki et al. (2021) 中之原始資料

編號	作者/年份	方法	TP	FP	FN	TN
1	Agorastos 2005	Cytology	2	20	2	1272
1	Agorastos 2005	HPV DNA	3	34	1	1258
2	Agorastos 2015	Cytology	22	119	19	3683
2	Agorastos 2015	HPV DNA	41	368	0	3434
3	Belinson 2003	Cytology	331	1523	44	6599
3	Belinson 2003	HPV DNA	363	1652	12	6470
4	Belinson 2010	Cytology	19	63	11	886
4	Belinson 2010	HPV DNA	28	133	2	816
5	Bigras 2005	Cytology	48	445	34	13315
5	Bigras 2005	HPV DNA	80	1063	2	12697
6	Blumenthal 2001	Cytology	97	169	111	1822
6	Blumenthal 2001	HPV DNA	168	721	40	1270
7	Cardenas-Turanzas 2008	Cytology	7	52	9	767
7	Cardenas-Turanzas 2008	HPV DNA	11	55	5	764
8	Castle 2011a	Cytology	222	1964	209	5428
8	Castle 2011a	HPV DNA	380	3122	51	4270
9	Clavel 2001	Cytology	72	385	10	5184

doi:10.6342/NTU202302826

表 11 Veroniki et al. (2021) 中之原始資料 (續)

編號	作者/年份	方法	TP	FP	FN	TN
9	Clavel 2001	HPV DNA	129	1085	0	6718
10	Cuzick 1995	Cytology	45	83	36	1821
11	Cuzick 1999	Cytology	36	130	6	2816
11	Cuzick 1999	HPV DNA	31	146	11	2800
12	Cuzick 2003	Cytology	75	427	15	9841
12	Cuzick 2003	HPV DNA	87	697	3	9571
13	de Cremoux 2003	Cytology	32	189	9	1529
13	de Cremoux 2003	HPV DNA	150	331	31	1273
14	Depuydt 2011	Cytology	27	160	19	2699
14	Depuydt 2011	HPV DNA	44	429	2	2430
15	Ferreccio 2013	Cytology	33	99	63	8017
15	Ferreccio 2013	HPV DNA	91	742	5	7374
16	Gravitt 2010	Cytology	26	313	30	1862
16	Gravitt 2010	HPV DNA	34	200	22	1975
17	Hovland 2010	Cytology	11	9	4	277
17	Hovland 2010	HPV DNA	16	42	0	255
17	Hovland 2010	mRNA	15	18	1	279
18	Iftner 2015	Cytology	38	156	52	9205
18	Iftner 2015	HPV DNA	86	494	4	8867
18	Iftner 2015	mRNA	82	382	8	8979
19	Labani 2014	Cytology	14	118	18	4506
19	Labani 2014	HPV DNA	17	120	15	4504
20	Li 2009	Cytology	69	370	5	2118
20	Li 2009	HPV DNA	67	351	7	2137
21	Mahmud 2012	Cytology	22	94	6	1264
21	Mahmud 2012	HPV DNA	21	148	3	1180
22	McAdam 2010a	Cytology	16	14	5	446
22	McAdam 2010a	HPV DNA	14	34	7	436
23	McAdam 2010b	Cytology	23	51	4	434
23	McAdam 2010b	HPV DNA	22	31	5	454
24	Monsonego 2011	Cytology	70	355	31	3973
24	Monsonego 2011	HPV DNA	98	595	3	3733
24	Monsonego 2011	mRNA	93	363	8	3965
25	Moy 2010	Cytology	175	860	37	7816

doi:10.6342/NTU202302826

表 11 Veroniki et al. (2021) 中之原始資料 (續)

編號	作者/年份	方法	TP	FP	FN	TN
25	Moy 2010	HPV DNA	204	1037	8	7268
26	Naucler 2009	Cytology	62	84	25	6023
26	Naucler 2009	HPV DNA	83	350	4	5652
27	Pan 2003	Cytology	81	432	5	1475
27	Pan 2003	HPV DNA	79	248	4	1505
28	Paraskevaidis 2001	Cytology	27	71	1	878
28	Paraskevaidis 2001	HPV DNA	25	24	3	925
29	Petry 2003	HPV DNA	45	369	1	7493
30	Qiao 2008	Cytology	58	69	12	2249
30	Qiao 2008	HPV DNA	68	333	2	1985
31	Ronco 2006	Cytology	54	850	19	15593
31	Ronco 2006	HPV DNA	73	1112	2	15223
32	Salmeron 2003	Cytology	60	127	41	7504
32	Salmeron 2003	HPV DNA	94	626	7	7205
33	Sankaranarayanan 2004a	Cytology	109	516	57	9909
33	Sankaranarayanan 2004a	HPV DNA	120	750	59	10589
34	Sarian 2005	Cytology	86	139	61	9852
34	Sarian 2005	HPV DNA	52	665	11	3467
35	Schneider 2000	Cytology	21	21	93	4626
35	Schneider 2000	HPV DNA	108	263	6	4384
36	Shipitsyna 2011	Cytology	5	13	1	760
36	Shipitsyna 2011	HPV DNA	6	101	0	716
37	Wu 2010	Cytology	18	89	9	1899
37	Wu 2010	HPV DNA	24	306	3	1682
37	Wu 2010	mRNA	27	174	0	1814

4.2.2 分析結果

此處我們仍針對 3.1 節中所提出的完整模型，透過 OpenBUGS 以 3 個鍊（chain）迭代 100,000 次，並將其前 50,000 次結果預燒（burn-in），一共迭代 150,000 次，並搭配 R 語言中之 R2OpenBUGS 套件協助執行。此處須特別留意，由於在 OpenBUGS 中，多項分佈有著不能以遺失值代入的限制，因此此處的程式部分在結構上需額外做調整（詳細可參考附錄之程式碼），以便於使用

doi:10.6342/NTU202302826

OpenBUGS 進行運算。

分析結果呈現於表 12 之中。透過該表可以發現，在不考慮參考準則的情況下，在敏感度的考量上，信使核糖核酸有較好的表現，而在特異度的部分，則以細胞學的表現情形較好。

表 12 以 OpenBUGS 對於三種子宮頸癌診斷工具與參考準則的分析結果

項目		估計值	標準誤	95%可信區間	
細胞學	敏感度	0.6937	0.0397	0.6120	0.7675
	特異度	0.9568	0.0076	0.9410	0.9706
人類乳突病毒 DNA	敏感度	0.9225	0.0148	0.8907	0.9487
	特異度	0.9106	0.0134	0.8804	0.9333
信使核糖核酸	敏感度	0.9309	0.0408	0.8277	0.9821
	特異度	0.9248	0.0327	0.8437	0.9694
陰道鏡檢查與組織學	敏感度	0.8468	0.0438	0.8012	0.9648
(參考準則)	特異度	0.9999	0.0001	0.9998	1.0000

若改以 Stata 分析的 meqrlogit 函數執行分析（於 Stata 中並未考量黃金標準不完美之議題），可以得到如表 13 之結果。可以看到下表的多數結果與使用 OpenBUGS 以貝氏統計分析的方法上沒有太大的差異，僅信使核醣核酸的敏感度在 OpenBUGS 與 Stata 的結果差異較大，此部分可能是受到信使核醣核酸僅有 4 篇研究包含該項目，且作為參考準則的陰道鏡檢查與組織學之敏感度也僅有約 0.85 所導致。若是要考量較高的敏感度，以人信使核糖核酸有較好的表現，但若要考量特異度，則以細胞學有較高的特異度。作為參考準則的陰道鏡檢查與組織學，敏感度約莫 0.84，但此處可能因收斂其況較差，僅反映先驗分配之資訊，特異度則來到 0.99，在子宮頸癌的檢測上作為參考標準仍可提供參考。

doi:10.6342/NTU202302826

doi:10.6342/NTU202302826

表 13 以 Stata 對三種子宮頸癌診斷工具的分析結果

項目		估計值	95%可信區間	
細胞學	敏感度	0.6688	0.6059	0.7261
	特異度	0.9470	0.9348	0.9570
人類乳突病毒 DNA	敏感度	0.9123	0.8871	0.9322
	特異度	0.9013	0.8800	0.9192
信使核糖核酸	敏感度	0.7869	0.7186	0.8422
	特異度	0.9273	0.9105	0.9411

4.2.3 比較不同診斷工具效果之優劣排名

此處我們仍考量 3.2 節中之比較不同診斷工具優劣的方式，分別針對敏感度與特異度考量其 SUCRA 之結果，其分別如表 14 與表 15 所示。此處可以看到其與前述之分析結果吻合，若考量敏感度，人類乳突病毒 DNA 有較高的 SUCRA 值，表示其為較準確之診斷工具，其次為信使核糖核酸，以及作為參考準則的陰道鏡檢查與組織學，但此處可能受其收斂情況較差所影響，最差則為細胞學；而若考量特異度，則以作為參考準則的陰道鏡檢查與組織學的 SUCRA 值最高，其次則以細胞學為較準確之診斷工具，再者為人類乳突病毒 DNA，最差者則為信使核糖核酸。

表 14 三種子宮頸癌診斷工具與參考準則敏感度優劣之排名機率與 SUCRA 值

項目	Rank 1	Rank 2	Rank 3	Rank 4	SUCRA
細胞學	0.0000	0.0000	0.0015	0.9985	0.0005
人類乳突病毒 DNA	0.3182	0.6216	0.0603	0.0000	0.7526
信使核糖核酸	0.6274	0.3009	0.0705	0.0012	0.8515
陰道鏡檢查與組織學 (參考準則)	0.0545	0.0775	0.8677	0.0003	0.3954

doi:10.6342/NTU202302826

表 15 三種子宮頸癌診斷工具與參考準則特異度優劣之排名機率與 SUCRA 值

項目	Rank 1	Rank 2	Rank 3	Rank 4	SUCRA
細胞學	0.0000	0.8556	0.1443	0.0001	0.6185
人類乳突病毒 DNA	0.0000	0.0004	0.2798	0.7198	0.0935
信使核糖核酸	0.0000	0.1440	0.5759	0.2801	0.2879
陰道鏡檢查與組織學 (參考準則)	1.0000	0.0000	0.0000	0.0000	1.0000

若採用 Superiority Index 作為指標，其效果呈現於表 16 中。根據其數值呈現結果，我們可以知道在這三種診斷工具之中，仍以陰道鏡檢查與組織學擁有最高的 Superiority Index，故以此作為參考準則是合理的。信使核糖核酸則為其餘三者中較準確之診斷工具，其次則是細胞學，最差者則為人類乳突病毒 DNA。

表 16 三種子宮頸癌診斷工具與參考準則之 Superiority Index

項目	Superiority Index
細胞學	0.3149
人類乳突病毒 DNA	0.8076
信使核糖核酸	2.0298
陰道鏡檢查與組織學 (參考準則)	3.3722

以 Youden Index 作為指標時，表 17 呈現了各項診斷工具透過 Youden Index 評估之排名結果。我們可以知道在這三種診斷工具之中，信使核糖核酸為較準確之診斷工具，但作為比較準則的陰道鏡檢查與組織學與人類乳突病毒 DNA，仍有與信使核醣核酸相近之 Youden Index，表示其仍有一定之準確度可參考，最差者則為細胞學。此處雖與 Superior Index 有所不同，但仍可加以確定信使核糖核酸為較佳之診斷工具。

doi:10.6342/NTU202302826

表 17 三種子宮頸癌診斷工具與參考準則 Youden Index

項目	Youden Index
細胞學	0.6505
人類乳突病毒 DNA	0.8332
信使核糖核酸	0.8557
陰道鏡檢查與組織學 (參考準則)	0.8468

doi:10.6342/NTU202302826

第五章　討論

5.1 疾病盛行率的意義

在本研究中所使用的模型中，與一般的二元模型與 ANOVA 模型不同，其特別加入了疾病盛行率作為考量的參數之一，因此實際上在統計軟體估計中，其資料是以疾病盛行率搭配資料提供之真陽性人數、偽陽性人數、偽陰性人數、真陰性人數，所推算出之真陽性人數、偽陽性人數、偽陰性人數、真陰性人數的比例，來對敏感度與特異度進行求算，而並非直接使用原來的數據。

若是黃金標準準確，那麼是否有考量疾病盛行率並不會對估計結果有太大差別。但若是黃金標準不準確，或是根本不存在黃金標準，此時就必須要透過納入疾病盛行率的方式，來考量診斷工具的網絡統合分析模型，方能估計出較準確的結果。

然而此處亦須留意，儘管在此處不完美黃金標準，其先驗分佈應可假設其範圍應可介於 0 至 1 之間，並假設其為均勻分配或 Beta 分配作為無資訊先驗分配（non-informative prior），或是對於該黃金標準的理解之合理先驗分布的假設。但有時在貝氏統計的計算上，可能受限於模型或是研究樣本提供之資訊不足，因此估計上出現不合理之結果，或是收斂情況不理想，而僅反映出其先驗分配所給予的資訊。儘管其為不完美黃金標準，但我們仍可相信其應該有較高的準確度，此時可調整其先驗分配給予其部分資訊，如前述方法中我們給予其範圍為 0.8 至 1，或是給予其範圍為 0.5 至 1 的先驗分配，但這樣的假設方式較為強硬，另一種作法則是以 Beta 分配作為不完美黃金標準的先驗分配，並將分配的高峰處設定在 0.7 至 0.9 左右的位置，增加先驗分配的資訊提供，以便於程式求得合適之估計結果。

doi:10.6342/NTU202302826

5.2 貝氏統計模型的影響

在診斷工具的網絡統合分析中，與頻率學派不同，貝氏統計模型最大的特色在於，其會透過系統性回顧所取得的資料，以更新先驗分配的資訊。一般而言系統性回顧所收集之資料並不會有過低的資訊量，因此其可以有效的更新先驗分配資訊，此時先驗分配的設定只要合理，並不會過於影響其結果的估計。但受到先驗分配的影響，其在敏感度與特異度的估計，變異程度可能會較使用傳統頻率學派的方式來的更大一些，但在點估計通常而言不會差太多。這樣的影響來自於先驗分配的變異程度也被考量進去，但儘管其變異可能使得可信區間變大，但大致不影響結果的判讀，因此並不會使得結果不穩定。此外，若在一個診斷工具中研究數較少，在頻率學派中可能其估計結果可信度可能較低，亦可能受到其自由度影響，若在診斷工具過多時，就可能使參數無法估計出結果。但在貝氏統計中由於其在參數的估中加入了先驗分配，此時儘管研究數少，我們仍然可透過調整先驗分配的方式，使得估計結果更得以信賴。

5.3 優劣排名結果的解讀及各項優劣排名指標之優缺點

在本研究中，考量了兩種不一樣的優劣排名指標，分別為 SUCRA 值與 Superior Index 兩個項目。兩者最大的差異在於由於 SUCRA 值只能考量單一數值，因此僅能將敏感度與特異度兩者分開考量，這樣的情況好處是若僅希望求取敏感度或特異度其中一個表現較好時，可以僅就其單一數值進行考量與下決策；但壞處則是，若敏感度與特異度所求出之 SUCRA 值不同，甚至是相互矛盾時，其決策便會不好決定。此時 Superiority Index 就是一個不錯的考量，由於其在計算公式上同時將敏感度與特異度皆考量於數值計算，因此其求算之數值便毋須擔心是針對敏感度或特異度，但由於 Superiority Index 不像 SUCRA 值必定介於 0 至 1 之間，其值解讀上僅為愈大愈好，因此我們無法知道其上限於何處，因此無法有效地從數值上即可推估可能其較其他診斷工具有多大程度的優

doi:10.6342/NTU202302826

勢或劣勢。

Youden Index 則同時具有兩者的優勢，其數值介於 0 至 1 之間，因此可以在一定範圍內比較診斷工具們之好壞。同時也將敏感度與特異度兩者納入考量，因此可以僅提供單一數值以評估診斷工具之好壞，不須像 SUCRA 值分開考慮二者。

但由於各項指標考量皆有所不同，SUCRA 值、Superiority Index 及 Youden Index 算出之結果在排名上亦可能有些微落差，但通常方向應一致。不過仍建議算出之排名應仍須與原始之敏感度與特異度進行對照比較，以確保解讀正確，不應僅仰賴這些指標所計算之排名結果進行解讀，而造成決策結果之錯誤。

5.4 OpenBUGS 程式的限制與調整

由於概似函數在其中考量了一個多項分配（Multinomial Distribution），以及 OpenBUGS 中程式碼的寫法，在附錄的程式碼之中，我們是以橫式（Wide Form）資料的方式進行資料分析。因此如 Hoyer and Kuss (2018)於表 4 中所提供之橫式資料，即可直接套入該程式碼中進行分析。但若是以 Veroniki et al. (2021)於表 11 所提供之直式（Long Form）資料，則必須先將資料經過轉換，將其調整為橫式資料，才能使資料得以分析。這樣的考量 OpenBUGS 中進行統合分析時也經常需要考量，因此應特別留意原始資料資料之形式，並進行調整，才能妥善應用於不同之程式碼。

此外，由於該模型需估計的參數較多，若此時資料又複雜，其執行上可能就會花費更多時間。建議在以程式進行模型估計上，可以對其迭代給予適合的起始值，較能減少其在執行上迭代所耗費的時間，也能使迭代結果更為理想。

doi:10.6342/NTU202302826

第六章　結論

網絡統合分析在今日的影響力愈來愈大，除了其可以有效地透過系統性回顧蒐集各項研究之數據並加以整合外，亦提供了一個穩健的方法可以有效地比較多種不同的介入方式。診斷工具的網絡統合分析，則協助我們比較多種不同的診斷工具其準確度及優劣差異，但由於一般而言比較不同診斷工具我們須同時考量敏感度與特異度，因此不能以傳統介入方式的網絡統合分析進行分析。潛在類別分析與 ANOVA 模型的估計方式為診斷工具的統合分析提供了一個優良的統計架構，透過給予參數不同的變異與分配假設，進而評估多種的診斷工具的差異，搭配上疾病盛行率的考量，則可使得數據之結果更加可靠，並解決黃金標準可能不完美的問題。貝氏統計能夠針對分配假設給予參數各種不同的先驗分配，提供了一個更具有彈性的方式以比較不同的診斷工具。

近年來因為疫情與疾病的發展，快篩試劑、核酸檢測等各種用以判斷並患是否罹病的工具與指標也愈來愈受到重視，而如何衡量比較這些診斷工具之優劣性，也將成為一個重要的議題。我們透過原先只能用於評估一種診斷工具與黃金標準差異的二元模型，將其以 ANOVA 模型的形式進行擴展，以使其能過進而比較多種不同的診斷工具，以及判斷各種診斷工具的好壞，同時也搭配潛在類別分析，針對黃金標準可能不完美而進行估計上的考量，並以貝氏統計的方式，使得模型更有彈性，也更便於使用者能夠依照自己的需求對參數進行設定與調整。我們也期望這樣的研究，未來能夠使診斷工具的網絡統合分析能夠有更有效的模型，易於使用者在資料上的分析，進而使其在統計上的結果也更加精確與穩健，也使得診斷工具在公衛與醫藥領域的比較上，能夠有更理想的依循。

doi:10.6342/NTU202302826

參考文獻

[1] S. White, T. Schultz, and Y. A. K. Enuameh, *Synthesizing evidence of diagnostic accuracy*. Lippincott Wiliams & Wilkins Philadelphia, 2011.

[2] J. M. Campbell *et al.*, "Diagnostic test accuracy: methods for systematic review and meta-analysis," *JBI Evidence Implementation,* vol. 13, no. 3, pp. 154-162, 2015.

[3] A. S. Midgette, T. A. Stukel, and B. Littenberg, "A meta-analytic method for summarizing diagnostic test performances: receiver-operating-characteristic-summary point estimates," *Medical Decision Making,* vol. 13, no. 3, pp. 253-257, 1993.

[4] L. Irwig, P. Macaskill, P. Glasziou, and M. Fahey, "Meta-analytic methods for diagnostic test accuracy," *Journal of clinical epidemiology,* vol. 48, no. 1, pp. 119-130, 1995.

[5] R. Parikh, A. Mathai, S. Parikh, G. C. Sekhar, and R. Thomas, "Understanding and using sensitivity, specificity and predictive values," *Indian journal of ophthalmology,* vol. 56, no. 1, p. 45, 2008.

[6] R. Trevethan, "Sensitivity, specificity, and predictive values: foundations, pliabilities, and pitfalls in research and practice," *Frontiers in public health,* vol. 5, p. 307, 2017.

[7] D.-G. Chen and K. E. Peace, *Applied Meta-analysis with R and Stata*. Chapman & Hall/CRC, 2021.

[8] A. Hoyer and O. Kuss, "Meta-analysis for the comparison of two diagnostic tests to a common gold standard: a generalized linear mixed model approach," *Statistical methods in medical research,* vol. 27, no. 5, pp. 1410-1421, 2018.

[9] V. N. Nyaga, M. Aerts, and M. Arbyn, "ANOVA model for network meta-

doi:10.6342/NTU202302826

analysis of diagnostic test accuracy data," *Statistical methods in medical research,* vol. 27, no. 6, pp. 1766-1784, 2018.

[10] J. Menten, M. Boelaert, and E. Lesaffre, "Bayesian meta-analysis of diagnostic tests allowing for imperfect reference standards," *Statistics in medicine,* vol. 32, no. 30, pp. 5398-5413, 2013.

[11] J. P. Jansen *et al.*, "Indirect treatment comparison/network meta-analysis study questionnaire to assess relevance and credibility to inform health care decision making: an ISPOR-AMCP-NPC Good Practice Task Force report," *Value in Health,* vol. 17, no. 2, pp. 157-173, 2014.

[12] F. S. Tonin, I. Rotta, A. M. Mendes, and R. Pontarolo, "Network meta-analysis: a technique to gather evidence from direct and indirect comparisons," *Pharmacy Practice (Granada),* vol. 15, no. 1, 2017.

[13] P. Macaskill, C. Gatsonis, J. Deeks, R. Harbord, and Y. Takwoingi, "Cochrane handbook for systematic reviews of diagnostic test accuracy," ed: Version, 2010.

[14] K. Nishimura *et al.*, "Meta-analysis: diagnostic accuracy of anti–cyclic citrullinated peptide antibody and rheumatoid factor for rheumatoid arthritis," *Annals of internal medicine,* vol. 146, no. 11, pp. 797-808, 2007.

[15] M. S. Pepe and H. Janes, "Insights into latent class analysis of diagnostic test performance," *Biostatistics,* vol. 8, no. 2, pp. 474-484, 2007.

[16] X. Ma, Q. Lian, H. Chu, J. G. Ibrahim, and Y. Chen, "A Bayesian hierarchical model for network meta-analysis of multiple diagnostic tests," *Biostatistics,* vol. 19, no. 1, pp. 87-102, 2018.

[17] H. Chu, S. Chen, and T. A. Louis, "Random effects models in a meta-analysis of the accuracy of two diagnostic tests without a gold standard," *Journal of the American Statistical Association,* vol. 104, no. 486, pp. 512-523, 2009.

[18] J. Zhang *et al.*, "Network meta-analysis of randomized clinical trials: reporting

doi:10.6342/NTU202302826

the proper summaries," *Clinical Trials,* vol. 11, no. 2, pp. 246-262, 2014.

[19] C. H. Daly, B. Neupane, J. Beyene, L. Thabane, S. E. Straus, and J. S. Hamid, "Empirical evaluation of SUCRA-based treatment ranks in network meta-analysis: quantifying robustness using Cohen's kappa," *Bmj Open,* vol. 9, no. 9, p. e024625, 2019.

[20] R. Deutsch, M. R. Mindt, and R. Xu, "Quantifying relative superiority among many binary-valued diagnostic tests in the presence of a gold standard," *J Data Sci,* vol. 7, pp. 161-77, 2009.

[21] W. J. Youden, "Index for rating diagnostic tests," *Cancer,* vol. 3, no. 1, pp. 32-35, 1950.

[22] A. A. Veroniki, S. Tsokani, E. Paraskevaidis, and D. Mavridis, "Evaluating multiple diagnostic tests: An application to cervical cancer," *HJOG,* vol. 20, pp. 11-24, 2021.

[23] C. Bennett, M. Guo, and S. Dharmage, "HbA1c as a screening tool for detection of type 2 diabetes: a systematic review," *Diabetic medicine,* vol. 24, no. 4, pp. 333-343, 2007.

[24] S. Kodama *et al.*, "Use of high-normal levels of haemoglobin A1C and fasting plasma glucose for diabetes screening and for prediction: a meta-analysis," *Diabetes/metabolism research and reviews,* vol. 29, no. 8, pp. 680-692, 2013.

doi:10.6342/NTU202302826

附錄（資料分析程式碼）

ANOVA 模型（3 種診斷工具）：

```
model{
 for(i in 1:ns){
  for(j in 1:nas[i]){
   prob[i, j, 1]<-pi[i]*(Se[i, test[i, j]]*s2)+(1-pi[i])*((1-Sp[i, test[i, j]])*(1-c2)) # TP
   prob[i, j, 2]<-pi[i]*(Se[i, test[i, j]]*(1-s2))+(1-pi[i])*((1-Sp[i, test[i, j]])*c2) # FP
   prob[i, j, 3]<-pi[i]*((1-Se[i, test[i, j]])*s2)+(1-pi[i])*(Sp[i, test[i, j]]*(1-c2)) # FN
   prob[i, j ,4]<-pi[i]*((1-Se[i, test[i, j]])*(1-s2))+(1-pi[i])*(Sp[i, test[i, j]]*c2) # TN

   n[i, j] <- tp[i, j]+fp[i, j]+fn[i, j]+tn[i, j]
   r[i, j, 1] <- tp[i, j]
   r[i, j, 2] <- fp[i, j]
   r[i, j, 3] <- fn[i, j]
   r[i, j, 4] <- tn[i, j]
   r[i, j, 1:4]~dmulti(prob[i, j, ], n[i, j])

   logit(Se[i, test[i, j]]) <- logitse[i, test[i, j]]+delse[i, test[i, j]]
   logit(Sp[i, test[i, j]]) <- logitsp[i, test[i, j]]+delsp[i, test[i, j]]

   delse[i, test[i, j]]~dnorm(0, taudel1)     # prior
   delsp[i, test[i, j]]~dnorm(0, taudel2)     # prior
  }

  logitse[i, 1] <- l[i, 1]
```


doi:10.6342/NTU202302826

```
  logitsp[i, 1] <- l[i, 2]
  logitse[i, 2] <- l[i, 3]
  logitsp[i, 2] <- l[i, 4]
  logitse[i, 3] <- l[i, 5]
  logitsp[i, 3] <- l[i, 6]

  l[i, 1:2]~dmnorm(mu1[], T[, ])
  l[i, 3:4]~dmnorm(mu2[], T[, ])
  l[i, 5:6]~dmnorm(mu3[], T[, ])

  logit(pi[i])<-logitpi[i]
  logitpi[i]~dnorm(mupi, taupi)
}

mupi~dnorm(0, 0.25)
precpi~dgamma(2, 0.5)
taupi <- 1/pow(precpi, 2)

taudel1 <- pow(SDtaudel1,-2)
SDtaudel1~dunif(0,2)
taudel2 <- pow(SDtaudel2,-2)
SDtaudel2~dunif(0,2)

mu1[1]~dnorm(0, 0.25)    # 第 1 個診斷工具
mu1[2]~dnorm(0, 0.25)    # 第 1 個診斷工具
mu2[1]~dnorm(0, 0.25)    # 第 2 個診斷工具
```


doi:10.6342/NTU202302826

```
mu2[2]~dnorm(0, 0.25)    # 第 2 個診斷工具
mu3[1]~dnorm(0, 0.25)    # 第 3 個診斷工具
mu3[2]~dnorm(0, 0.25)    # 第 3 個診斷工具

T[1:2, 1:2] <- inverse(S[, ])

S[1, 1] <- sigma1[1]*sigma1[1]
S[2, 2] <- sigma1[2]*sigma1[2]
S[1, 2] <- rho1*sigma1[1]*sigma1[2]
S[2, 1] <- rho1*sigma1[1]*sigma1[2]
sigma1[1] <- pow(prec1[1], -0.5)
sigma1[2] <- pow(prec1[2], -0.5)
prec1[1]~dgamma(2, 0.5)
prec1[2]~dgamma(2, 0.5)
rho1~dunif(-1, 1)

# Imperfect Gold Standard
s2~dunif(0.8, 1)    # prior
c2~dunif(0.8, 1)    # prior

# Sensitivity
Pooled_S[1] <- 1/(1+exp(-mu1[1]))
Pooled_S[2] <- 1/(1+exp(-mu2[1]))
Pooled_S[3] <- 1/(1+exp(-mu3[1]))
Pooled_S[4] <- s2
```


doi:10.6342/NTU202302826

```
# Specificity
Pooled_C[1] <- 1/(1+exp(-mu1[2]))
Pooled_C[2] <- 1/(1+exp(-mu2[2]))
Pooled_C[3] <- 1/(1+exp(-mu3[2]))
Pooled_C[4] <- c2

# Prevalence
preval<-exp(mupi)/(1+exp(mupi))

# SUCRA for sensitivity
for(z in 1:ntest+1){
  for(k in 1:ntest+1){
    rk.sen[z, k] <- ntest+2-rank(Pooled_S[], k)       # Good
    # rk.sen[z, k] <- rank(Pooled_S[], k)             # Bad
    best.sen[z, k] <- equals(rk.sen[z, k], z)
  }
}

# SUCRA for specificity
for(z in 1:ntest+1){
  for(k in 1:ntest+1){
    rk.spe[z, k] <- ntest+2-rank(Pooled_C[], k)       # Good
    # rk.spe[z, k] <- rank(Pooled_C[], k)             # Bad
    best.spe[z, k] <- equals(rk.spe[z, k], z)
  }
}
```


doi:10.6342/NTU202302826

```
  # Calculate Superiority index
  for(i in 1:ntest+1){
    for(j in 1:ntest+1){
      # Loop through total number of test/threshold combinatons
      M.1[i, j] <- step(step(Pooled_S[i]-Pooled_S[j])+step(Pooled_C[i]-Pooled_C[j])-2)
      M.2[i, j] <- step(step(Pooled_S[j]-Pooled_S[i])+step(Pooled_C[j]-Pooled_C[i])-2)
    }
    Sup.M.1[i] <- sum(M.1[i, ])-1
    Sup.M.2[i] <- sum(M.2[i, ])-1
    Sup.index[i] <- (2*Sup.M.1[i]+1)/(2*Sup.M.2[i]+1)
  }

  # Calculate Youden index
  for(i in 1:ntest+1){
    You.index[i] <- Pooled_S[i]+Pooled_C[i]-1
  }
}
```


doi:10.6342/NTU202302826