結合大型資料庫與小型確證資料的二階段校正 (TSC) 統計分析法

中央研究院統計科學研究所 程毅豪

大型電子化健康數據資料庫如健保資料庫分析在近年的醫學研究中日漸普及。這樣的研究具有省時省力，且可避免回溯性研究之回憶偏差等優點。然而此類研究的重要侷限之一，是這些大型資料庫的資料收集並非針對學術研究目的，因此其往往缺乏較詳盡的關於個人之干擾因子(confounder)如吸菸、飲酒、飲食習慣與職業暴露等，以及生物標記(biomarker)測量如血壓血糖等資訊，因此這些資料庫無法產出較精準的研究成果。

一個解決此問題的方法是設法由一些專門的研究或調查資料庫中取得較詳盡完整的資料，其包含了干擾因子及生物標記資料。這樣的資料我們可稱之為確證樣本資料 (validation data)。此確證樣本可提供校正了干擾因子及生物標記資訊的研究結果，因而可得出較為準確的研究結論。但相較於前述的大型資料庫，此類專門性的研究或調查資料庫往往所收集的個案數規模小了許多，因而影響其統計上的效力 (power)。

在此介紹一個二階段校正統計分析方法 (Two Stage Calibration, TSC)，其基本想法是將上述兩類資料於分析時加以截長補短，做適當的結合，並校正在主樣本 (大型資料庫資料) 分析中因為缺乏干擾因子及生物標記資料所可能產生的偏誤。在第一階段中，TSC 將主樣本與確證樣本結合 (總樣本數為主樣本數 N1 + 確證樣本數 N2)，並取出同時出現於兩樣本中的變數 (排除確證樣本中的干擾因子及生物標記變數，因其未出現於主樣本中)，並根據此整併之資料進行統計分析。由於此階段並未利用干擾因子及生物標記資料 (因大型資料庫無此項資料)，因此雖然此時有較大樣本數 (N1+N2)，但只能以較簡略的分析模式 (如疾病與處置相關性的迴歸分析，未校正干擾因子與生物標記) 得到統計分析結果。

在第二階段中，TSC 僅對確證樣本進行分析，即樣本數僅為確證樣本的樣本數 N2。此時可進行兩種分析: 一為利用與第一階段相同的簡略分析模式 (未校正干擾因子與生物標記之疾病與處置相關性的迴歸分析)，只是資料由原先的整併資料 (主樣本+確證樣本) 改為僅用確證樣本資料，即樣本數由 N1+N2 變為 N2。另一為利用進一步校正干擾因子及生物標記的精確分析模式 (疾病與處置在校正干擾因子與生物標記後之相關性的迴歸分析)，並利用確證樣本資料 (主樣本因無干擾因子與生物標記資料故無法直接利用)，樣本數為 N2。根據數理統計學中的理論，此兩階段的分析結果為相依的多變數常態分布，因此利用下述公式：

$$\beta = \beta 2 - C \times (\alpha 1 - \alpha 2)$$

可得到最後的結合主樣本與確證樣本的 TSC 迴歸分析結果 β，其中 β2 代表利用確證樣本與精確分析模式 (校正干擾因子及生物標記資料)所得到的迴歸係數，α1 與 α2 為分別利用整併資料及確證資料分析粗略模式 (未校正干擾因子及生物標記資料) 所得到的迴歸係數，*C* 為 β2 與 α1 -α2 的共變異數。TSC 迴歸係數 β 與 β2 (確證樣本中精確分析模式之迴歸係數) 具有相同的意義，即他們都是校正干擾因子及生物標記資訊後得到的分析結果，因此較 α1 與 α2 的結果更為準確 (因 α1 與 α2 未校正干擾因子及生物標記資訊)。但 TSC 迴歸係數 β 與 β2 不同之處在於，後者僅利用確證樣本資料，而前者不僅利用確證樣本資料也同時(間接地)利用主樣本資料，因此前者較後者的統計效力 (power) 更高。事實上，由數理統計理論可知，TSC 迴歸係數 β 的估計標準誤 (standard error; SE) 為下列公式：

$$\mathrm{SE}(\beta) = \sqrt{\mathrm{SE}(\beta 2)^2 - \frac{\mathrm{N1}}{\mathrm{N1+N2}} \times \frac{C^2}{\mathrm{SE}(\alpha 1 - \alpha 2)^2}}$$

因此 TSC 迴歸係數 β 較僅利用確證樣本資料的迴歸係數 β2 有較小的估計標準誤 (SE)，即較高的統計效力 (power)。

在實務運作中，若需校正的干擾因子及生物標記數量太多，我們可考慮先將這些干擾因子及生物標記以傾向分數 (propensity score) 的形式結合，再利用上述的 TSC 方法整合主樣本與確證樣本中的資料及分析結果。

我們以上述 TSC 方法分析台灣健保資料庫中取得之資料，探討慢性阻塞性肺病 (Chronic Obstructive Pulmonary Disease; COPD) 對帶狀泡疹(Herpse Zoster; HZ) 發病年齡的影響，並校正年齡、性別、共病 (comorbidity) 、 吸菸及飲酒等干擾因子。由於健保資料庫不包含吸菸及飲酒之資料，此兩項重要干擾因子之資料取自國家衛生研究院之國民健康訪問調查資料庫。在此應用中，主樣本資料為 8,486 名慢性阻塞性肺病患者 (2004—2006) 及 33,944 名非慢性阻塞性肺病患的對照 (1:4 年齡性別配對)，資料均來自健保資料庫。確證樣本資料為 244 名慢性阻塞性肺病患者 (2004—2006) 及 904 名非慢性阻塞性肺病患的對照 (1:4 年齡性別配對)，資料均來自國民健康訪問調查資料庫。在健保資料庫與國民健康訪問調查資料庫中，我們收集糖尿病、高血壓、冠心病等共病資料作為校正之干擾因子，而在國民健康訪問調查資料庫中，我們另收集吸菸及飲酒頻率資料作為進一步校正之干擾因子。分析之結果 (Cox regression hazards ratio for incidence of HZ: COPD vs. non-COPD, and the 95% confidence interval (95% CI)) 如下表：

分析	Hazards Ratio (95% CI)
α1 (健保資料庫+國民健康訪問調查; 未校正吸菸喝酒)	1.96 (1.72, 2.24)
β2 (國民健康訪問調查; 校正吸菸喝酒)	2.89 (0.96, 8.71)
β (TSC; 結合 α1 與 β2 且校正吸菸喝酒)	2.21 (1.79, 2.72)

由上表可看出： 僅由大型資料庫所得到的分析結果 (α1)，雖達統計顯著意義 (95% CI 不涵蓋 1)，但其所得到之 hazards ratio 的數值明顯低於校正吸菸喝酒後的數值 (如 β2 與 β)。而僅利用國民健康訪問調查資料庫的分析結果 (β2)，雖有校正吸菸喝酒且 hazards ratio 的數值較高，但卻未達統計顯著意義 (95% CI 涵蓋 1)，這是因其樣本數較少之故。而 TSC 方法所得到的整合分析結果 (β)，既校正了吸菸喝酒，因此其 hazards ratio 數值較高，且亦達到統計顯著意義。此一應用說明了 TSC 方法的特色，即其可將大型資料庫與小型確證資料兩者的優缺點互補，既利用大型資料庫樣本數「大」的優點，也結合了較精確仔細測量的「確證」資料，由此得出更具實際意義與統計效力的統計分析結果。

二階段校正 TSC 方法的 SAS Macro 分析程式及說明可取自作者網頁 (http://www.stat.sinica.edu.tw/yhchen/download.htm)

論文請參閱：

Lin, Hui-Wen. and Chen, Yi-Hau (2014). "Adjustment for missing confounders in studies based on observational databases: Two-stage calibration combining propensity scores from primary and validation data." American Journal of Epidemiology, vol. 180, pp. 308-317.