整合多基因風險分數提升對類風濕性關節炎的預測性
Integrating Polygenic Risk Score to Enhance Predictive Ability for Rheumatoid Arthritis

實習單位:台中榮民總醫院醫學研究部

實習學生: 張震奕 指導老師: 林敬桓博士 蕭自宏博士 盧子彬導師

Abstract

類風濕性關節炎（RA）是一種自體免疫疾病，屬於慢性炎症性疾病。過去的研究主要集中在治療方面，嘗試使用藥物和非藥物方法來減輕症狀，緩解患者的痛苦。然而，隨著藥物治療的不斷成熟，我們已經具備一定的能力來降低疾病活動，並提升患者的生活品質。

在目前這個階段，我們面臨的挑戰是如何深入探究類風濕性關節炎的發病機制，以及如何預測該疾病的發生。隨著藥物療法的不斷發展，我們逐漸意識到，更深入了解RA的根本原因，以及在早期階段對其進行預測，將對疾病的管理和治療產生重大影響。

此次研究主要使用基因來對 RA 進行預測。其中會使用 PRS，做為不同個體的 RA 遺傳風險。並期望能透過台灣精準醫療資料庫建立屬於台灣對於 RA 的 GWAS 、 PRS，且更進一步的透過整合多個 PRS，以提升對 RA 的預測能力。最後也期望這新型的預測工具，能讓 RA 患者能夠提早診斷提早治療，進而提高臨床結果與生活質量。

Material & Method

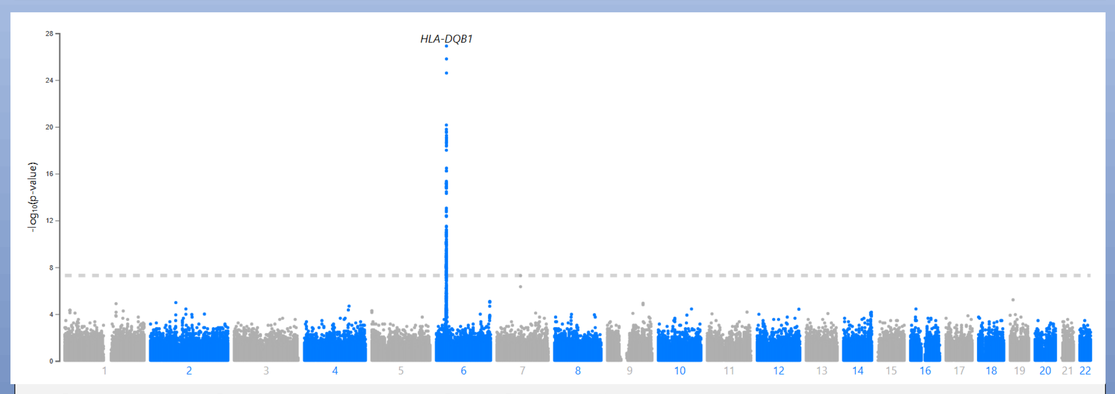
- TPMI data

台灣精準醫療計畫 (TPMI) 的總樣本數有 63542，而在 GWAS 的 QC 後剩下的個體為 63505筆。而其中的性別比為 0.82，年齡大多都落在中壯年 (年齡平均為 56 歲)。在 Data 中使用 ICD-Code 來進行 RA 的辨識。

- Genome-wide association study (GWAS)

此次使用 plink 來進行 GWAS ，再透過R套件顯示與 RA 相關的 significance SNP。此次找到的高相關基因 HLA-DQB1，屬於 HLA 的第二型(主要與細胞外病原體的胜肽段相關)。

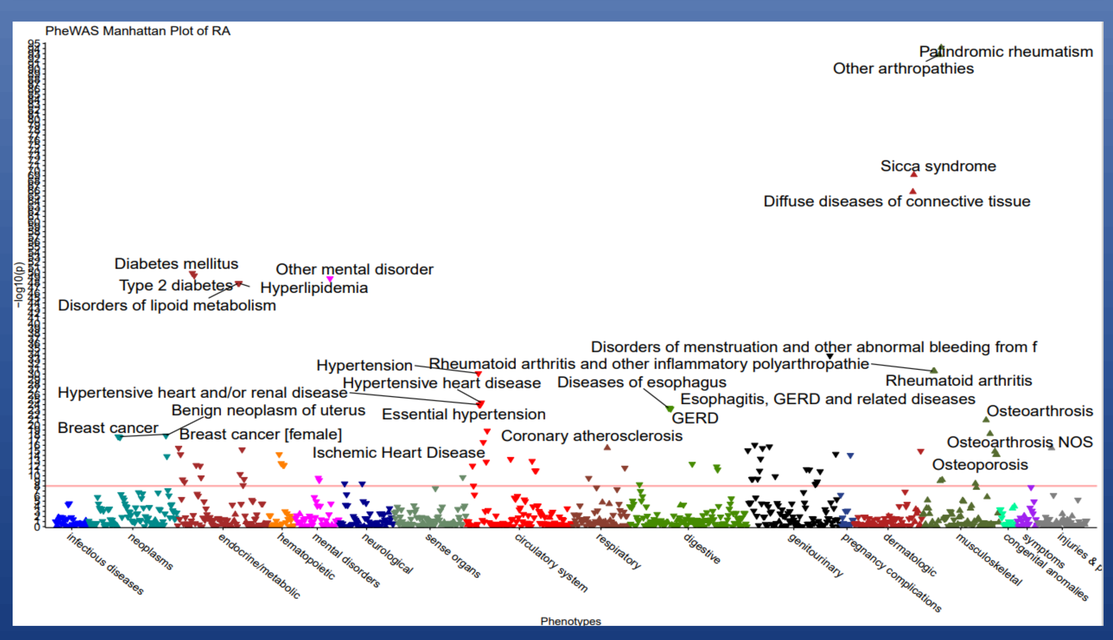
而過去以 RA 為 phenotype 的 GWAS 大多與 HLA 基因有顯著相關，其中 HLA-DQB1 也在其中。後續會以此 GWAS 結果來進行 polygenic risk score。



- Phenome-wide association study (PheWAS)

PheWAS 主要透過 Genetic variants，來去尋找 associated phenotypes。其中是透過 R 套件來進行，並通過 ICD-code轉換成 Phecode 來計算各疾病 logistic regression 的 p-value。

此次使用 PheWAS 是以患有 RA 與否來對其他 phenotype 進行相關性的測試，並不是使用 Genic variant。後續會篩選這些 RA 相關的 PheCode 的 PRS。



- Polygenic Risk Score (PRS)

接續使用 GWAS 的結果，來計算 PRS。此次是使用 Clump + Threshold (C+T) 的方法來進行，先透過 clump 解決 Linkage Disequilibrium (LD 連鎖不平衡) 的問題，再透過界定不同的 p-value 為 threshold，來取得不同的 PRS，最後透過各個 model 相對於 Null model 的 R^2 的結果來抉擇使用何種 threshold 的 PRS。

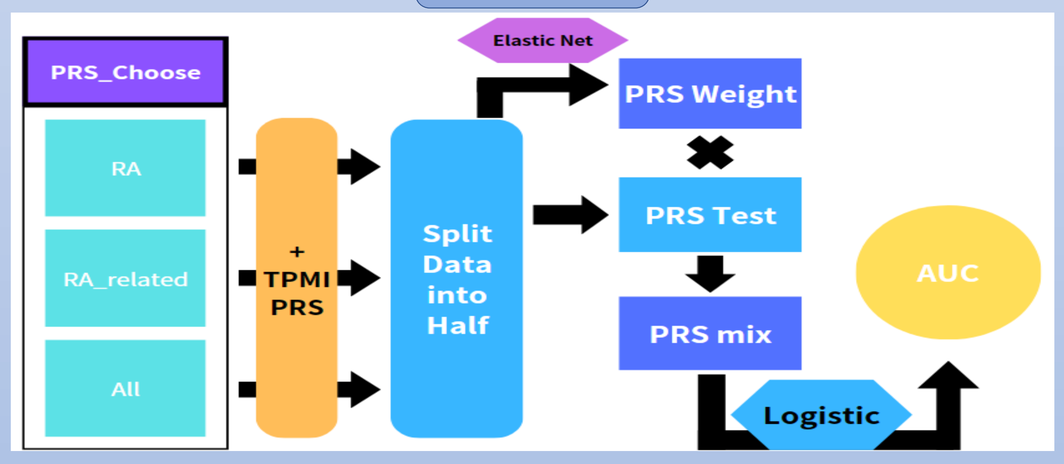
- Elastic Net

是一種正則化的方法，透過引入正則化項、正則化參數來調整，目的是避免模型的 overfitting。類似將 PRS 步驟中 GWAS shrinkage 的 Lasso regression 以及 ridge regression 合併，為此 Elastic Net 會有兩種正則化項、參數。

$$L(\lambda_1, \lambda_2, \beta) = |y - X\beta|^2 + \lambda_2|\beta|^2 + \lambda_1|\beta|_1$$

另外透過代數證明，也顯示了 Elastic Net 能夠克服 Lasso regression 的部分限制。(eq: Lasso 在 p > n 的情況下，最多只能選擇 n 個變量、當兩兩相關性非常高，那 Lasso 傾向選擇其中一個，並不在意選擇哪個。)

Result



此次研究目的，是測試能否透過 integrated PRS 的方法找到更好的對 target tratit 進行預測的方法。主要會透過計算各個 Single polygenic risk score 的模型 AUC，後續再以 integrated PRS 與其進行比較。

研究步驟:

1. 計算 Single PRS 的 logistic regression 的 AUC，並找出最佳的結果
2. 三種的 PRS 選擇以進行後續的 PRS integrated (RA 的 PRS、RA-related PRS、All PRS)
3. 進行 TPMI 的 GWAS 來得到屬於台灣 RA 的 PRS，並進行 logistic regression，依照 AUC 結果來決定是否也放入 integrated。
4. 將所有資料分成兩半，其中一半用於產生各個 PRS 在新的 Combine PRS 變相內的權重，另一半(test Sample)則是使用該權種來進行 Combine PRS 的計算
5. 使用 test Sample 來進行 AUC 計算，先將 Sample 分為 8:2 (train : test)，再透過 train 製作 model， test 進行 predict
6. 比較 Combine PRS 與 Best Single PRS ，來對結果進行審視
7. 後續放入 TPMI 的部分 Clinical 變相，試驗 Combine PRS 是否仍有解釋能力、顯著

	PRS	AUC
Best_single_PRS	PGS001875	0.6009

	Best_PRS	Group	Clinical
PRS_RA	0.6367	0.6188	0.6619
PRS_RA_related	0.6697	0.6543	0.6771
PRS_All	0.6851	0.6713	0.6914

Discussion

後續可以透過 Net Reclassification Improvement 或者是 Shuffle prediction 來對於新 model 對於原始的 RA prediction 的提升來去判定。

另外此次使用 TPMI 所製作的 PRS在進行 logistic regression 所產生的 AUC 並沒有更好的結果，為此在進行 integrated PRS 時並沒有將其匯入。可能在未來有更多的 基因樣本後能提升其預測效果。

而在此次所產生的 PRS weight 可能能作為未來使用相同方法預測的參考。

另外透過此次的研究，顯示了進行相似分析時，使用更多樣的基因分析，能夠對 prediction 的提升有幫助。