巨量資料與統計分析Fall 2017

第三次作業（2017/11/24繳交）

* 本次作業仿造大數據比賽的模式，僅提供資料、不指定研究議題，由各組同學根據分析結果、結合專業領域的知識，發揮想像力、創造力，提出一個可實行、具有高附加價值的大數據研究企劃案。

🡪資料將於11/4（六）上線，11/7（二）上課時確定分組名單。

🡪每組同學3～4人，建議合併兩組現有的分組。

🡪本次作業將選出前三名，頒發獎金、獎狀；比賽結束後，前三名報告公佈至網站提供各組參考。

* 資料說明：台灣實施全民健康保險制度已逾20年（1995年至今），不僅提高大家就醫的便利性，也提高國民的健康與平均壽命，國際對我國全民健保評價也很高。全民健保資料庫記錄我國國民的就醫資料，因其資料量動輒超過GB層級，因此本次作業以全民健保資料庫為研究目標，讓同學練習讀取資料、清理資料、分析資料等之技巧。

🡪請至本課程專屬工作站，以雲端方式連結健保資料庫的承保資料檔(ID)資料，你們可讀取（而非下載）2010年的承保資料，2005年百萬人抽樣檔中就醫門診處方及治療明細檔(CD)。



* 可能的資料分析項目或方向：

🡪熟悉編碼簿是分析資料庫的首要步驟，請查詢2010年ID及CD兩個資料庫的編碼簿，整理對應的資料欄位（變數）及其定義，進行偵錯及清理。

🡪以承保資料檔(ID)資料，考量投保者的出生年月日（年齡）、性別、投保地區（縣市）等欄位，各組分析不同縣市的投保人口，並與內政部統計處(http://www.moi.gov.tw/stat)的資料比對，評估這兩份資料是否類似。（註：也可透過卡方檢定驗證樣本代表性。）

🡪分析民眾門診時間、門診地區、健保補助金額、就醫疾病（根據主要疾病代碼，或是ICD-9; International Classification of Diseases, Ninth Revision）等項目，疾病或可分為「慢性病處方箋」、「循環系統疾病」、「內分泌及代謝疾病」、「消化系統疾病」、「骨骼肌肉系統疾病」、「呼吸系統疾病」共六大類。

（1）慢性處方箋：

以「案件分類」（資料庫欄位為「CASE\_TYPE」）

為04、06、08、24、28的代碼，篩選慢性病處方箋

（2）各大類疾病：

其他五類疾病ICD9代碼分類方式如下：

--內分泌及代謝疾病：240 ~ 279

--循環系統疾病：390 ~ 459

--呼吸系統疾病：460 ~ 519

--消化系統疾病：520 ~ 579

--骨骼肌肉系統疾病：710 ~ 739