出國報告(出國類別: 進修)

精準醫學與資料科學進修報告

服務機關: 臺中榮民總醫院醫學研究部

姓名職稱: 陳一銘 / 主治醫師

派赴國家 / 地區: 美國 / 國家衛生研究院腦中風中心生物資訊小組

出國期間: 108 年 7 月 1 日至 109 年 6 月 30 日

報告日期: 109 年 7 月 23 日

目　次

一、摘要

為發展本院精準醫學，本人於 2019 年 7 月至 2020 年 6 月前往美國國家衛生研究院(NIH NINDS Bioinformatics Section)進修 data science, genomic data analysis 以及 genetic counselling。

進修期間學習 variant analysis of NGS data，包括 pre-processing, alignment 以及 variant calling，在 NIH hands on 課程學習。

在指導老師協助下，申請 NIH database BTRIS，瞭解資料申請步驟、流程與管理介面，研究倫理規範以及如何結合基因資料，同時也有機會認識 All of Us 這個美國大力推行的精準醫學計畫。

此外分析大量資料需用到人工智慧之演算法，在美期間也學習利用 Jupyter Notebook 的 Python 介面，以 machine learning 的方法分析資料。

在 NINDS 期間也有機會接觸醫療影像人工智慧分析，學習使用 XNAT 平台，從 PACS 系統中截取影像與報告，加以標示，並以演算法進行分類運算分析。

最後是基因諮商在臨床實務上，也有機會到診間與基因諮商師以及基因遺傳專科醫師共同訪視病人，學習基因資訊可能對病人治療的影響。

本次進修雖然後期因為 COVID-19 肆虐，美國多數機構包括 NIH 都行 work from home，以致原本已註冊的 precision medicine 課程停開，臨床基因諮商門診暫停，但在美國國家衛生研究院擔任交換學者期間，仍能獲得許多值得本院在推展精準醫學時之寶貴經驗，可為未來臨床運用及研究發展之重要參考。

關鍵字: 精準醫學、資料科學、基因諮商、人工智慧

二、目的

精準醫療之進展日新月異，近兩年來已成為全球頂尖醫學中心聚焦發展之新興學門。為發展本院精準醫療，配合中央研究院台灣精準醫療計畫及人體生物資料庫100KPM 檢體收集計畫，透過人體基因資料庫進行比對及分析，從中找出風險基因，並利用基因資訊發現適合病患的治療方法與藥品，或避免副作用之產生，本院需培育人才，學習基因資料定序、品質控制、臨床資訊管理、並結合基因資料與電子病歷資料，產出基因報告、基因諮商、解釋檢測結果、告知病患並提供相關治療策略，因此前往 NIH NINDS 進修 data science, genomic data analysis, genetic counselling。

三、過程

A. 基因資料定序分析

進修期間參加 NIH FAES 開立的 Variant analysis of NGS data，在 NIH hands on 課程學習。分析的 workflow 如下:

Pre-processing: 需要 fastqc quality check，也學習使用 skewer 這支程式進行 adapter trimming (圖一)。

Alignment: 使用 BWA-MEM 這支程式，對 SAM/BAM file format 進行 alignment。

Variant calling: 使用 HaplotypeCaller 進行 variant calling，再接著進行 variant filtering and annotation。

FAES 課程事先會提供 cookbook，把實作課程會用到的 source code 先寄給學員，在上課時即使從沒有使用過 linux 系統，也可以用簡單的 cope paste 把程式語言貼上再執行，就可以使用範例 database 的進行運算。

圖一、使用 skewer 進行 adapter trimming 之操作情況

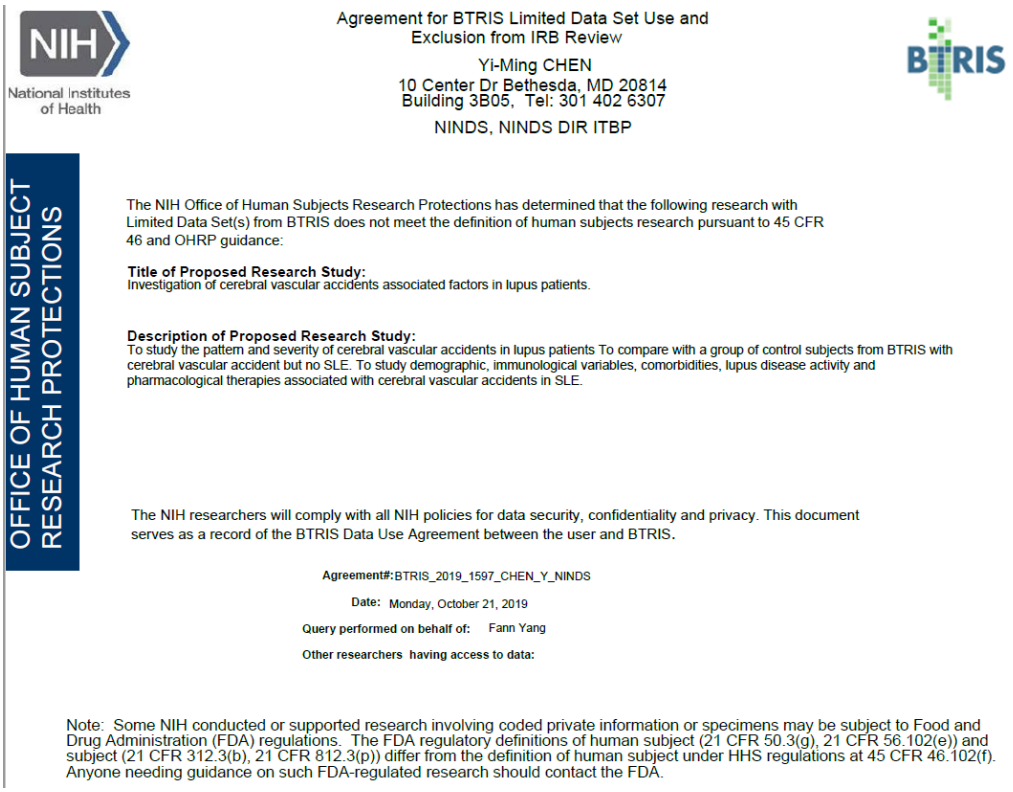
```
skewer v0.2.2 [April 4, 2016]
Parameters used:
-- 3' end adapter sequences in file (-x):    adapters1.fa
A:   AGATCGGAAGAG
-- paired 3' end adapter sequences in file (-y):
adapters2.fa
01:  AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
-- maximum error ratio allowed (-r):    0.100
-- maximum indel error ratio allowed (-d):   0.030
-- minimum read length allowed after trimming (-l):    25
-- file format (-f):          Unknown format (auto detected)
Sat Aug 17 15:24:55 2019 >> started
|================================================>| (100.00%)
Sat Aug 17 15:24:56 2019 >> done (0.267s)
10000 read pairs processed; of these:
     0 ( 0.00%) degenerative read pairs filtered out
     0 ( 0.00%) short read pairs filtered out after trimming by
size control
```

B. NIH 院內資料 BTRIS

在指導老師 Dr. Fann 的協助下，我有機會申請 NIH 的院內資料庫 BTRIS，瞭解資料申請步驟、流程與管理介面，研究倫理規範以及如何結合基因資料。

這個資料其實跟本院臨資申請資料類似，但流程上方便許多，最主要是一開始申請時不需要 IRB 審核，就可以先瞭解資料庫中的資料型態是否足以後續分析，簡單的申請 exclusion from IRB review 就可以先 access data (圖二)，等到資料真的適合發表再來提出正式的 IRB 申請

圖二、BTRIS Limited Data Set 免 IRB 審查申請表



NIH
National Institutes of Health

Agreement for BTRIS Limited Data Set Use and Exclusion from IRB Review

Yi-Ming CHEN
10 Center Dr Bethesda, MD 20814
Building 3B05, Tel: 301 402 6307
NINDS, NINDS DIR ITBP

BTRIS

OFFICE OF HUMAN SUBJECT RESEARCH PROTECTIONS

The NIH Office of Human Subjects Research Protections has determined that the following research with Limited Data Set(s) from BTRIS does not meet the definition of human subjects research pursuant to 45 CFR 46 and OHRP guidance:

Title of Proposed Research Study:
Investigation of cerebral vascular accidents associated factors in lupus patients.

Description of Proposed Research Study:
To study the pattern and severity of cerebral vascular accidents in lupus patients To compare with a group of control subjects from BTRIS with cerebral vascular accident but no SLE. To study demographic, immunological variables, comorbidities, lupus disease activity and pharmacological therapies associated with cerebral vascular accidents in SLE.

The NIH researchers will comply with all NIH policies for data security, confidentiality and privacy. This document serves as a record of the BTRIS Data Use Agreement between the user and BTRIS.

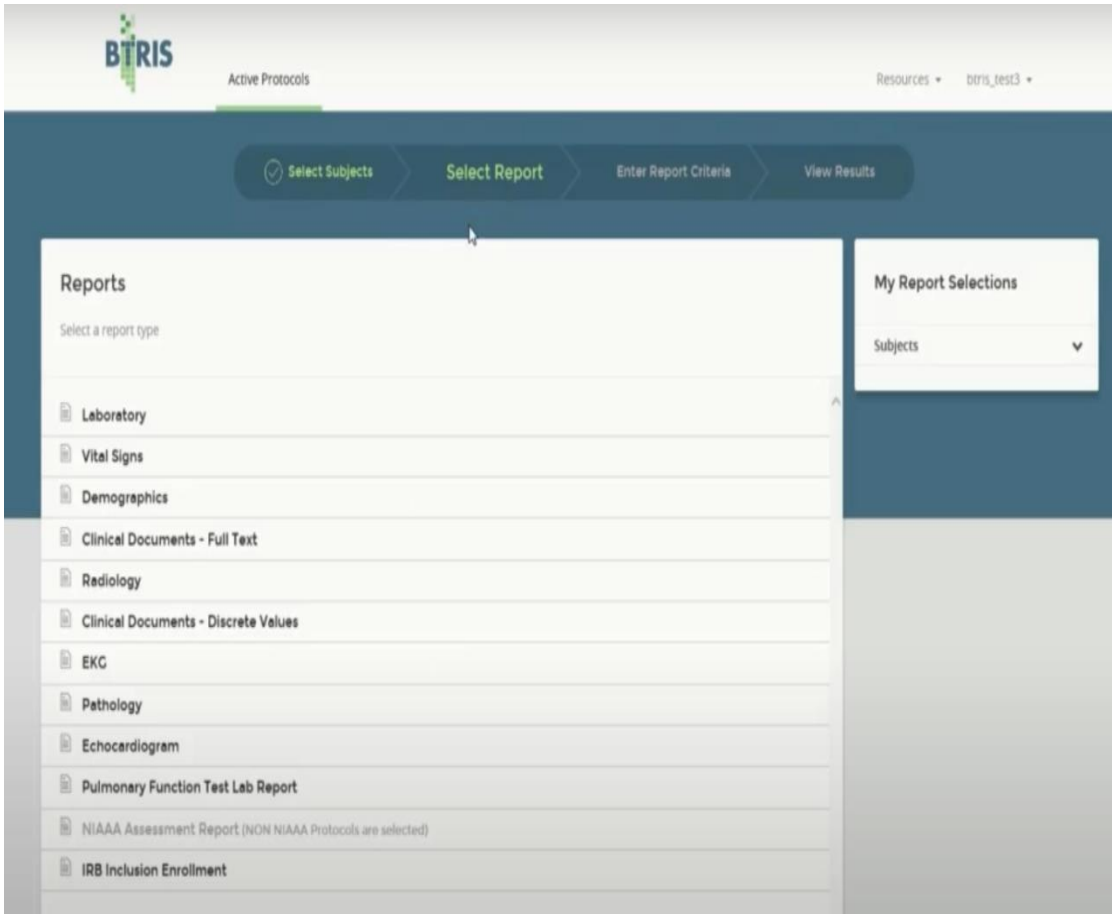
Agreement#:BTRIS_2019_1597_CHEN_Y_NINDS
Date: Monday, October 21, 2019
Query performed on behalf of: Fann Yang
Other researchers having access to data:

Note: Some NIH conducted or supported research involving coded private information or specimens may be subject to Food and Drug Administration (FDA) regulations. The FDA regulatory definitions of human subject (21 CFR 50.3(g), 21 CFR 56.102(e)) and subject (21 CFR 312.3(b), 21 CFR 812.3(p)) differ from the definition of human subject under HHS regulations at 45 CFR 46.102(f). Anyone needing guidance on such FDA-regulated research should contact the FDA.

此外，BTRIS 提供一個 user friendly 的介面，讓每一個使用者可以自己去調整需要的資料特性，包括實驗室數據、病人特徵、診斷碼等，再去篩選出來我們需要的病患族群(圖三)，以進行後續的資料分析。這樣的申請流程與系統可以減少臨

床研究者來回跟資料庫申請資料的時間，更有效率地產出可用的研究成果。

圖三、BTRIS 使用者操作介面



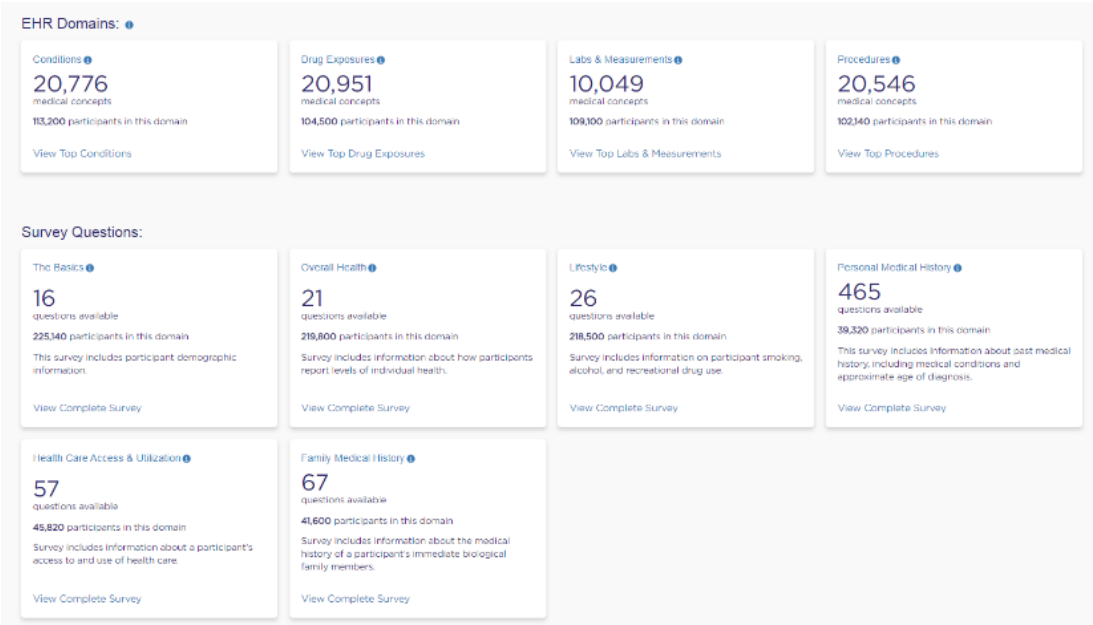
C. NIH All of Us 計畫

NIH 在美國政府支持下，於 2015 年 9 月開始推動 All of Us 精準醫學計畫，預計要收 100 萬人的大型研究，也會考慮 minority 的人種收案，這個計畫是跟 Taiwan biobank 計畫及本院要執行的 TPMI 計畫最為相近，但在深度廣度更甚。主要差異如下:

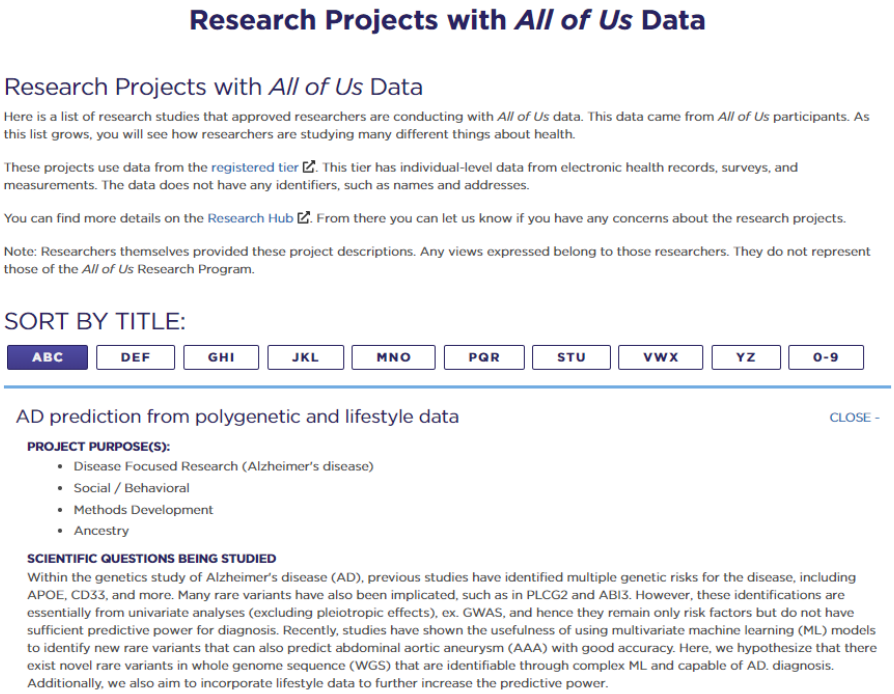
追蹤時間更久: 超過十年; 收案人數更多; 同意書取得增加電子平台，手機載具收案； 受試者均測量 physical measurements (BMI, vital signs, anthropometric measurements)； 生物檢體收集包括 blood, urine and/or saliva；除了電子病歷外，

也同時收集穿載裝置資料、問卷調查 (圖四)。All of Us project 會做全基因定序，也會把基因報告發回給參加者，但目前尚未決定要 return 哪些 genetic information，研究計畫也可以在網路上查到 (圖五)。

圖四、All of Us 目前收案進度



圖五、All of Us Research projects



D. 人工智慧機器學習

分析大量資料需用到人工智慧之演算法，在美期間也學習利用 Jupyter Notebook 的 Python 介面，以 machine learning 的方法分析資料。包括使用 Biopython 做 bioinformatics 的分析，在分類問題上，也可以在 testing set 使用 sklearn 的 Linear regression / random forest／support vector machine (SVM)方法(圖六)，之後再不同的 cohort 裡面做 tranining/validation，比較 AUC 的不同，以判斷 algorithm 所使用的 features 分辨力的好壞。

圖六、Python 資料分析＆機器學習

Linear Classification

```
cv = sklearn.model_selection.StratifiedKFold(5, random_state=42)
lin_avg_r2 = Avg()
for i, (train, test) in enumerate(cv.split(X,y)):
          X_train, X_test, y_train, y_test = X[train], X[test], y[train], y[test]
          lr = sklearn.linear_model.LogisticRegression()
          lr.fit(X_train, y_train)
          test_r2 = lr.score(X_test, y_test)
          print("Cross fold ", i, ":", test_r2)
          lin_avg_r2(test_r2)
```

Random Forest Classifier

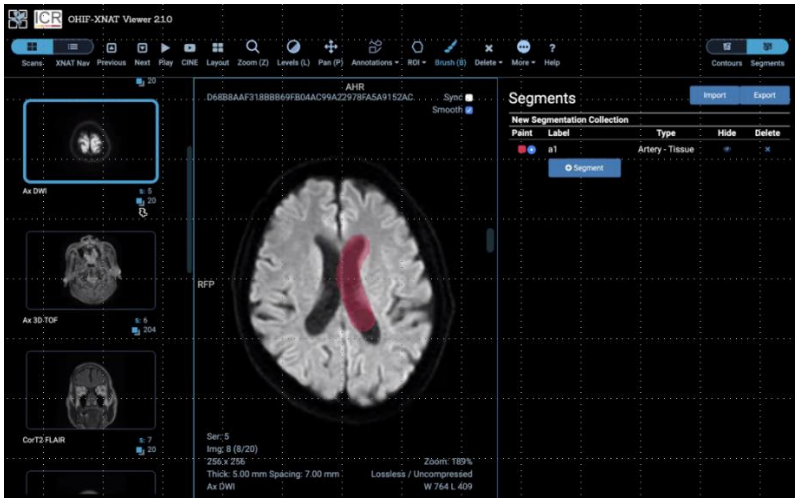
```
cv = sklearn.model_selection.StratifiedKFold(5, random_state=42)
lin_avg_r2 = Avg()
for i, (train, test) in enumerate(cv.split(X,y)):
          X_train, X_test, y_train, y_test = X[train], X[test], y[train], y[test]
          lr = sklearn.ensemble.RandomForestClassifier()
          lr.fit(X_train, y_train)
          test_r2 = lr.score(X_test, y_test)
          print("Cross fold ", i, ":", test_r2)
          lin_avg_r2(test_r2)
```

E．影像分析平台 XNat

人工智慧分析除了上述提到分類問題，自然語言處理文字型資料之外，另一個運用是在影像分析，我在 NINDS 期間也有機會接觸醫療影像人工智慧分析，學習使用 XNAT 平台，從 PACS 系統中截取影像與報告，加以標示 (圖七)，並以演算

法進行分類運算分析。

圖七、XNat 用以管理標示腦部 MRI 影像



F. 基因諮商

在 NINDS 受訓期間，有機會跟 genetic counselor Ms. Alice 一起學習在臨床實務上的基因諮商，也有機會到診間與基因諮商師以及基因遺傳專科醫師共同訪視病人，學習基因資訊可能對病人治療的影響。

進行實需要評估病患接受基因檢測的主要動機、個人病史與家族病史的收集尤其重要，也需要提供病患與家屬相關的基因、檢測資訊，回顧可提供的檢測選擇，在報告出來之後也能判讀結果，說明後續治療之選擇以及基因檢測的局限，對病患可能受到的工作、保險歧視也能預做準備。基因諮商的學習收獲頗豐。

四、心得

本次進修雖然後期因為 COVID-19 肆虐，美國多數機構，包括 NIH，從三月開始都行 work from home，以致原本已註冊的 precision medicine 課程停開，臨床基因諮商門診暫停。

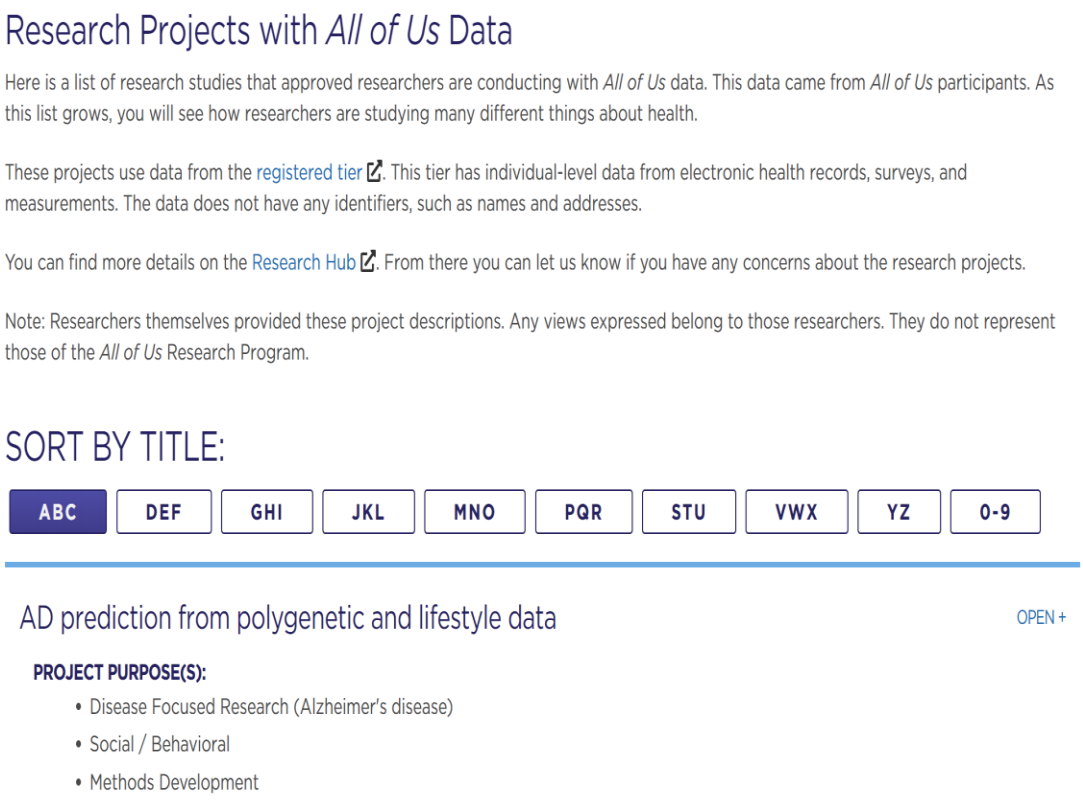
疫情期間雖然閉關在家，但和實驗室伙伴仍應每周定期視訊討論研究進度，機器學習 python 的 script 雖然自己不甚熟悉，但因 lock down 在家，時間很多可以互相學習，大家也在討論中給予很多研究分析的建議、實驗室伙伴也有統計專精的醫師，常常可以對團隊的整體統計概念有所提昇。

在美國國家衛生研究院擔任交換學者期間，我認為習得許多值得本院在推展精準醫學時之寶貴經驗，可為未來臨床運用及研究發展之重要參考。

五、建議 (包括改進做法)

A. 研究計畫提案機制

本院正在進行的精準醫學計畫，應可以參考 All of Us 計畫，對有興趣想使用這些基因/病歷資料的研究者開放，讓大家提出簡單的 research project，再經過委員會審查通過，我們可以看到 All of Us 在公開網域公布的 ongoing projects，不同疾病的 research ideas，也可以做為我們的參考。



Research Projects with *All of Us* Data

Here is a list of research studies that approved researchers are conducting with *All of Us* data. This data came from *All of Us* participants. As this list grows, you will see how researchers are studying many different things about health.

These projects use data from the registered tier. This tier has individual-level data from electronic health records, surveys, and measurements. The data does not have any identifiers, such as names and addresses.

You can find more details on the Research Hub. From there you can let us know if you have any concerns about the research projects.

Note: Researchers themselves provided these project descriptions. Any views expressed belong to those researchers. They do not represent those of the *All of Us* Research Program.

SORT BY TITLE:

ABC | DEF | GHI | JKL | MNO | PQR | STU | VWX | YZ | 0-9

AD prediction from polygenetic and lifestyle data OPEN +

PROJECT PURPOSE(S):

- Disease Focused Research (Alzheimer's disease)
- Social / Behavioral
- Methods Development

B. 擴充臨床資訊人才

我們醫院的精準醫學資料庫優點在於完整的實驗室追蹤與健保資料庫的連結，但並無像 All of Us 計畫有完整的身體評估/問卷調查，而且臨床資訊的整合仍然需要像 BTRIS 一樣使用者友善的系統以及足夠數量且有經驗的 clinical informatics 人才，協助研究者及時解決資料 retrieval 時遭遇到的大小問題。

我們在使用 BTRIS 資料庫時，NIH 有類似專案負責的資訊工程師跟我們聯絡討論，對我們的研究問題跟資料庫存取的技術問題十分熟悉，這方面非常值得我們學習效法。醫院目前臨床資訊的人員仍然分散在不同科別，除了人數不足以外，他們仍然需要分出時間來處理醫務管理或評鑑所需的資料，惶論可供使用者自行操作且好用的介面，這部份需要投入更多資源才能讓已經開始運轉的 TPMI/regeneron 計畫取得可以匹配的臨床資訊。

C. 定序技術提昇與生資人力擴充

雖然 TPMI 計畫廣泛收集臨床檢體進行 genotyping 分析，然而近年來研究主軸已經轉向 whole genome sequencing，然而在台灣因研究資源不如美國豐沛，本院與 regeneron 合作的 whole exon sequencing 已經是 compromised option，建議可以目前 TPMI 計畫收案之模式，全力產出研究成果，以做為爭取國家型計畫執行全基因定序的基礎；此外生資分析人力也不能全依賴外部研究機構，雖然本院不可能像 NIH 一樣，在各個臨床部科都有屬於自己單位的 bioinformatics 分析人員，但建立穩定且有分析臨床檢體經驗的 in-house 人力仍然是非常重要，可以更及時提供本院研究者在基因資料分析時的協助。

D. In-house 人工智慧分析量能

在 NIH 同一實驗室的交換學者同事已於七月份獲聘回台擔任中部某醫學中心人工智慧主任，在其科室內就有 30 名工程師，且該院目前已有四個類似規模的單位；據瞭解國內醫學中心近年來無不大張旗鼓擴編人工智慧人才，尤其是自身有 programming 能力且同時具有臨床研究經驗的研究者，甚或是由與業界有合作開發產品經驗的人才來領軍，在院內設立 in-house AI-ready 的工程師職缺，更進一步培育本院年青研究人才寫程式分析的能力，如此方可穩定提升機構的人工智慧研發能力。

E. 建立基因醫學臨床人力

有鑑於精準醫學在各科部之運用逐漸增加，建議院方可仿照其他醫學中心，培育本院 medical geneticist 以及 genetic counselor 人才，初期可以考慮和婦女醫學部/兒醫部現有的人力著手整合，從 perinatal screening 開始，擴展到 rare disease 以及癌症運用；此外也可考慮讓有興趣相關科別醫師，公費至國外接受至少兩年的 medical geneticist，鼓勵院內研究者參加 American College of Medical Genetics and Genomics (ACMG) / American Society of Human Genetics (ASHG)的線上課程與 annual conferences，如此可以深化本院基因醫學臨床服務，基因諮商與研究能量。