出國報告（出國類別：進修）

2020 美國國家衛生研究院國家神經疾病和中風研究所進修心得

服務機關：高雄榮民總醫院/高齡醫學中心
姓名職稱：梁志光主治醫師
派赴國家/地區：美國/馬里蘭州
出國期間：2019/09/01-2020/08/31
報告日期：2020/09/24

摘要

人工智慧領域的發展一瞬千里，但於醫療領域仍處理開發階段，雖已有許多演算法應用，但因醫療的特殊性，從發展到實際應用之間仍有很大鴻溝。然而，人工智慧醫療領域可發展的方向很多，可使用的數據來源也很廣，如臨床訊號源(如 EEG, EKG, vital signs 等)、影像資料、視覺圖像資料、語音資料與電子病歷資料等，均會產生不同數據資料可供後續如臨床決策支持系統、疾病診斷輔助、檢查篩檢輔助、自然語言處理和理解、預後預測建模、機器人和圖像引導手術、個人化的影像和治療建議等領域發展。台灣已有多家醫院正在全力發展中，因此次透過於國家衛生研究院的一年進修期間，學習了解現今美國人工智慧領域的發展狀況與方法，並建立未來持續合作的關係與交流管道。

關鍵字

人工智慧，機器學習

目次

一、目的

人工智慧領域於這些年因科技進步與數據儲存成本下降，因而發展迅速。人工智慧運用的領域很廣泛，醫療應用範圍更是其中最受矚目的部分，臨床應用部分，包括可以協助病理、皮膚與眼科等疾病診斷、標記與判讀影像資料、個人化癌症治療選擇或電子病歷語意萃取等，而基礎研究亦有應用的方向，如基因體分析、蛋白質體分析及藥物設計與發展等。在國內，各政府部門均對於人工智慧研究投注豐富資源，而美國國家衛生研究院更是美國最頂尖的政府研究單位，美國國家衛生研究院現今支持的研究案中，多數都跟生醫資訊或大數據分析有關且密切結合。因大數據研究仍持續處於熱浪中，還有許多領域可以研究發展，因此職希望透過此次於美國國家衛生研究院的一年進修期間，學習了解現今美國人工智慧領域的發展狀況與方法，並建立未來持續合作的關係與交流管道。

二、過程

申請過程

感謝高榮院長、副院長、院部長官們，還有高齡醫學中心林育德主任等長官的推薦，讓職取得出國進修一年的難得機會。於獲准出國進修通知後，即開始聯繫出國進修的單位與行政事宜。因職對於資料分析與人工智慧與機器學習分析領域有興趣，也透過台中榮總陳一銘醫師(現任研究部轉譯醫學科代理主任)協助，聯繫上美國國家衛生研究院國家神經疾病和中風研究所的范揚政教授，范教授本身也是台灣出生，在美國主修 Computational Chemistry，主要研究領域涵蓋生物資訊、人工智慧、機器學習、程式語言、電腦運算、資料庫設置與管理等，目前為美國國家衛生研究院國家神經疾病和中風研究所 IT 部門主任。范教授也與中央研究院、國家衛生研究院、國立陽明大學、臺北醫學大學、中國醫藥大學、交通大學與林口長庚醫院等長期合作，每年固定時間會至上述單位進行研究協助與討論，亦經常協助安排台灣研究人員至美國衛生研究院進修與參訪事宜，因此與國內研究單位的合作相當密切，也很歡迎台灣學者至美國衛生研究院學習。透過范教授的熱情協助，讓職有機會進入到美國國家最大的研究部門，進行人工智慧與機器學習領域的學習與研究。

美國國家衛生研究院與國家神經疾病和中風研究所

美國國家衛生研究院位於馬里蘭州的貝塞斯達，由 27 個中心(centers)或院所(Institutes)所組成，佔地約 367,251 坪，園中有 75 座建築物，與約 1,200 位主要研究員(Principal Investigators)。其為全美最頂尖的政府研究單位，目前已經有 156 位獲得了 NIH 的支持的研究者獲得諾貝爾獎，也是全世界最重量級與研究經費最豐富的部門，每年有超過 300 多億

美元的研究經費支援，八成是與院外研究單位合作。國家神經疾病和中風研究所則是於 1950 年成立，單位內包含神經科學研究部門(Division of Neuroscience)、院外研究部門(Division of Extramural Activities)、臨床研究部門(Division of Clinical Research)、轉譯研究部門(Division of Translational Research)與院內研究部門(Division of Intramural Research)。目前國家神經疾病和中風研究所主要負責研究有關大腦和神經系統的相關知識，並利用該知識減輕神經系統疾病的負擔，主要研究內容涵蓋基礎研究、臨床研究與數據分析，主要項目包含修復與可塑性(Repair and Plasticity)、系統與認知神經科學(Systems and Cognitive Neuroscience)、通道突觸與迴路(Channels, Synapses, and Circuits)、神經基因學(Neurogenetics)、神經環境(Neural Environment)、大腦影像(brain image)、神經退化(Neurodegeneration)與大數據分析等。

進修單位介紹

職所進修的單位屬於國家神經疾病和中風研究所內的院內研究部門(Division of Intramural research)，而范教授除是 IT 部分主管外，亦是院內信息技術和生物信息計劃主任(Director, Intramural IT and Bioinformatics Program)，其下除了負責 IT 部門外，另也領導兩個研究團隊，一為生物信息研究團隊，此團隊主要成員包括 Amar Yavatkar(Bioinformatics Programmer)與 Kory Johnson(Ph.D., Bioinformatics Scientist)等，主要負責國家神經疾病和中風研究所內外生物信息研究的協助，包括實驗研究設計、蛋白質體分析、次世代基因分析、序列分析與統計分析等，另一為人工智慧研究團隊，此團隊研究方向多數為機器學習、神經網路與自然語言分析處理等領域，主要成員以許凱程醫師(現任中國醫藥大學附設醫院人工智慧醫學診斷中心主任)與林敬恆博士研究員為主，林敬恆博士研究員為國立陽明大學生物醫學資訊研究所博士畢業，已數次到美國國家衛生研究院學習。而職到職時，同單位同時有台北榮總放射腫瘤科胡育文醫師與台中榮總教研部陳一銘醫師一同受訓。

報到後，辦公室位於 10 號樓 3B05，單位也為每位研究人員均提供辦公室座位，並配備一台筆電與 29 吋螢幕使用。

報到與環境適應

職於 09/02 到達國家衛生研究院所在地馬里蘭州蒙哥馬利縣的貝塞斯達，貝塞斯達除了國家衛生研究院外，亦是美國國家海軍醫療中心、洛克希德.馬丁公司總部與萬豪國際酒店集團總部，美國國家海軍醫療中心是美國歷任總統體檢的醫療中心，也位在國家衛生研究院對面。貝塞斯達超過八成為白人，亞洲人僅佔 7%左右，因多數住民為政府雇員，教育水準高，治安好。

國家衛生研究院是屬於美國國立政府單位,因而對於員工與進修者的身份調查相當嚴格,雖於申請進修時已繳交許多資料，但報到後初期欲進入園區內，每日均須經由訪客中心進行身份確認，取得臨時通行證才能進入，且臨時通行證需每日重新申請，因此九到十月每日均需繞道到訪客大門申請臨時通行證件。職於此過程中，同時經過填寫身家調查、照相、確認保險額度等過程，經由一個半月後才取得正式員工證，才可從任何大門經由刷卡進入，且取得正式身份後，方可使用整個院區的資源服務。

各月學習過程

2019/09

參加每月研究部門會議 (monthly AI team meeting)

9/6 日至 35 號樓 IT 辦公室參與九月份人工智慧團隊月會，范揚政教授於會議中介紹與其研究團隊認識,並說明底下有兩組研究團隊,一為生物信息團隊,一為人工智慧研究團隊,我參與的為後者，亦說明目前美國國家衛生研究院支持的研究案中，多數以上都跟生醫資訊有關且結合，因大數據研究仍持續處於熱浪中，還有許多領域可以研究發展，因此建議先從人工智慧所需的技能開始熟悉，可先了解需要使用的程式語言，教授並提供幾場院區內主辦與人工智慧研究領域相關研討會，希望透過參與會議了解國家衛生研究院目前於人工智慧領域的相關研究方向與進展。會議中也知悉各研究人員的研究方向，與可與之學習的領域。許凱程醫師對於 Python 與 R 語言熟悉,且研究方向涵蓋機器學習、神經網絡、生物信息分析等，林敬恆博士研究員，亦精通上述領域且熟悉自然語言處理。

2019/09/13

參與 Biomedical Translational Research Informatics(BTRIS)會議

范教授為我們新進研究人員安排至生物醫學轉譯研究信息學(BTRIS)辦公室，學習國家衛生研究院所建立的BTRIS資料庫，此會議由Andrea Beri主講，她是BTRIS資料分析單位主管，BTRIS是國家衛生研究院內的一個整合式研究資源系統,此資料庫數據從1976年收集至今且資料仍持續增加中。它將臨床中心病歷和其他國家衛生研究院研究所和中心的相關臨床研究的數據與資料均匯總在同一系統中，內容包含所有研究計畫與臨床試驗的說明、個案資訊與所收集的資料,同時亦涵蓋臨床病歷、診斷資料、臨床檢查檔與報告(如影像資料、心臟超音波、心電圖、微生物檢查與病理資料等)、實驗室檢查報告與藥物等，並持續整合新資料來源，未來亦計畫從社會安全資料庫中將死亡資料檔串連進BTRIS系統內。同時BTRIS也提供一個整合式環境讓臨床試驗研究人員將自身的臨床試驗個案資料與研究資料整合進系統內，後續計畫中之研究人員可自行根據研究方向與所需內容，來透過BTRIS網路訪視自身研究符合條件的受試者可識別數據，而不需像過去自行手動從原始資料中去搜尋研究所需資料。此外，因同

一個案可能進行不同臨床研究，亦可透過同一個案其個人資訊，串聯所有過去研究資料一起提供使用。

BTRIS同時針對可釋出之資料轉成無需個人識別碼的數據，提供所以園區內有興趣且經核准的研究人員申請使用，而IRB則由BTRIS辦公室協助。BTRIS辦公室所提供的協助還包括將不同資料來源做整合、協助研究人員釐清所需資料、取得資料、清理資料、轉換資料格式、資料分析與資料視覺化等。除了研究方面外，並可提供臨床試驗定時進度報告與不良反應資料表，以協助臨床試驗監控研究過程。然而，BTRIS亦面臨部分困境，如不同研究試驗的資料格式不同，資料註記不夠清楚，尤其是不同資料的時序不容易釐清，如症狀、診斷、檢查與預後間的先後時序等，還有文字資料的處理部分，仍持續在解決中。

2019/09/27

參加國家神經疾病和中風研究所主辦之 NINDS Stroke Branch Symposium，主題為 Celebrating Three Decades of Research in Stroke Organization(s)

講題內容涵蓋腦中風過去三十年研究的精華，主題從過去、現在，一路談到未來腦中風潛在的治療藥物發展，因此涵蓋了靜脈 r-TPA 治療，經動脈取栓治療，腦部中風後神經發炎，腦中風及失智之關係，SUMOylation 發展，及如何減少腦中風二次傷害等。SUMOylation 是一種有潛力的中風治療藥物，這是由 John Hallenbeck 博士所帶領的團隊，從一種會冬眠的地松鼠體內發現，稱為 SUMO 化(SUMOylation) 的細胞機制過程，這跟動物在冬眠時大腦血流量減少之下仍能存活下來的機制有關，亦被認為有神經保護作用，演講者哈佛大學 Joshua Bernstock 教授亦在 2020 於 Current Issues in Molecular Biology 期刊發表一篇回顧，提到已有大量實證證實 SUMOylation 會於腦缺血後分泌，而這也跟缺血後神經修復的過程有關，但目前仍需要尋找作用於此機制的適合藥物。而低溫治療也被發現可降低腦中風過程中因缺氧導致的腦細胞損傷。此外，免疫療法亦是一種具有潛力的未來中風治療方式，腦中風發生後，因缺氧導致腦細胞損傷，而後會帶入大量發炎性細胞，雖發炎細胞在某個階段會促進細胞修復，但同時若產生過量發炎反應，反而會導致梗塞範圍擴大，而延長發炎反應時間將導致神經損傷過久，無法自行修復，目前針對 tumor necrosis factor，interleukin (IL)-1，IL-6 與 IL-10 均有研究在進行，只是發炎反應有利有弊，何時使用免疫療法介入，並避免全身性不良反應是目前免疫療法待突破的地方。

2019/10

2019/10/1

參與 Artificial Intelligence Healthcare - From Prevention & Diagnostics to Treatments 研討會

會議是由 the NIH Artificial Intelligence Interest Group 與 the NIH AI Working Group for Autonomous Therapeutics 兩個小組共同主辦，此小組從 2003 開始運作，目前兩個小組除了國家衛生研究院成員外，也擴展到 FDA、全美各大學和業界對於人工智慧於醫療應用有興趣的人員，本場會議由主席 June Lee 女士主持，會議中有兩場主題演講，題目為國防高等研究計劃署(The Defense Advanced Research Projects Agency (DARPA))及以 AI 為基礎的外科手術方式。此外，另有 AI 預防醫學與診斷，自主療法(AUTONOMOUS THERAPEUTICS)，優化臨床預後的生成(OPTIMIZING CLINICAL OUTCOME GENERATION)與優化臨床實證的產生(OPTIMIZING CLINICAL EVIDENCE GENERATION)四場會議主題分別進行討論。

人工智慧於醫療應用範圍相當廣泛，臨床部分包括協助病理、皮膚與眼科疾病診斷、影像標記與判讀、癌症個人化治療選擇或電子病歷語意萃取等，而基礎研究亦有基因分析、蛋白質體分析及藥物設計與發展等應用範圍。而國家衛生研究院於 2018 年當年就投入 2 億 8 千多萬美元在人工智慧研究領域，而國家衛生研究院的自我角色定位包括提供跨領域合作機會，讓生物醫學與電腦科學領域合作與交流，並提供教育訓練予包括人工智慧領域專業人士與生物醫療研究人員等，此外亦主導發展人工智慧研究倫理標準，提供所有研究人員遵守。而外科在人工智慧的發展部分，目前已在使用與發展的部分包括術中即時影像人工判讀與機器人手術過程輔助人工智慧決策、虛擬術前手術模擬與教學、穿戴裝置即時監控與警示、臨床資訊病患預後與病患安全預先決策系統等，而未來還有 3D 組織與器官列印與非侵入性分子生物能量手術已在發展中。

國家衛生研究院國家兒童健康與人類發展研究所也正朝向發展不昂貴、且可快速篩檢或診斷與隨身使用的人工智慧模組，目前已有的應用包括結合電子病歷與次世代基因分析預測先天性疾病新生兒，人工智慧模組的預測能力較單純只使用全基因體定序與全外顯子定序的五成診斷率高出甚多，且可提早 22 小時診斷出。另外手機應用程式 BiliCam 作為新生兒黃疸值偵測工具，偵測敏感性超過 80%。其他正在發展的部分還有使用腦部結構與功能性 MRI 影像資料預測非常早期早產兒是否有認知障礙，分析腦部神經網絡來區別出閱讀障礙與整體發展障礙孩童，與偵測出家暴幼兒的肋骨骨折等。然而雖然人工智慧研究如火如荼發展，但目前仍著重在明確且單一任務的預測，而實際醫院中需面臨臨床環境裡複雜的多重訊號與資訊來源，這時所需的多重智慧即時監控，因同時涵蓋太多複雜變數，雖研究已有部分發展，但仍會有較高的誤判機會，仍待未來技術突破來改進。

2019/10/03

參與國家衛生研究院演講，題目為 BTRIS: Translating Data into Results，此演講除了之前針對 BTRIS 資料庫結構的介紹外，透過今日的演講主題，了解 BTRIS 資料庫如何應用於臨床研究，從主題形成至產出文章的過程。

2019/10/16

參加 Grand Round，主題：Radiology in the Era of Artificial Intelligence

主講者為史丹佛大學醫學系 Radiology and Biomedical Informatics 學科 Curtis P. Langlotz 教授。他也是史丹佛大學 Center for Artificial Intelligence in Medicine and Imaging 單位主任。

演講中提到，目前放射影像學於人工智慧領域發展的方向包括影像標記(Image labeling)、跨學科合作(Interdisciplinary Collaboration)、精準健康(Precision Health)、照護轉型(Care Transformation)、強化放射科醫師量能(Radiologist Augmentation)與建立公共大數據集(Public Data Sets)等。而實際臨床應用上，已有包括針對肺結節做偵測判讀，使用 HeadXNet 模組偵測腦血管瘤，Xray4All 手機應用程式判讀急診肺炎，透過人工智慧運算提升影像解析度，降低使用顯影劑與 PET 檢查的放射線曝露量等許多臨床應用。而一個影像人工智慧應用的發展過程會包含著多重步驟，從原始資料匯入、影像資料標記、非影像資料萃取、資料前置處理、合併影像與非影像來源資料進行機器學習運算、產生決策模組、臨床應用、輸入新資料並持續即時優化決策模組等。而影像標記、資料萃取與前處理部分需花費至少六成以上時間，而資料的收集也需要花費兩成左右的時間，這兩個前段步驟確實耗人耗時，須有足夠人力協助，而資料完整建立與確認正確性後，後續決策模組發展才能順利進行，最後於臨床運用與即時回饋部分則需要投注足夠的硬體資源協助，因次演講中也再次提及有完整的人工智慧基礎建設(Infrastructure for AI)很重要，當有標準化的影像與非影像資料儲存平台，後續才能利用固定模式去做資料轉換與資料萃取，減少前段耗費的時間，而前置基礎建設完善，後續資料分析就能快速發展。最後，仍不免提到，雖現今已有多項診斷與預測模組於臨床應用，但醫師的角色仍相當重要，因目前仍無法完全信任且將判讀的重責落於電腦系統。

2019/10/26

受華府國建聯誼會(CAPA)翁毓廷會長邀請，至國建會會員聚會中分享失智症現況與台灣照護經驗，此次會員聚會場地在前會長吳東麟博士家中，前會長(1998 年度)莊德茂院士也和大家做近期的分享，最後再由我做醫學新知的分享。

2019/11～12

10 月份主要學習目標為參與不同研討會，從中熟悉國家衛生研究院環境與資源，並學習目前人工智慧發展現況。11~12 月份則主要針對研究所需的基本能力開始進行學習，主要學習方向分為兩部分，第一部份為程式語言於機器學習領域的入門，主要是熟悉所需程式語言的語法，如 R 語言，第二部份則學習人工智慧與機器學習相關知識。針對上述兩項目標，職除了跟隨研究團隊熟悉如何思考機器學習的分析方法，這跟原本過去所學的研究與統計方式與思

路並不同，同時並參與國家衛生研究院所辦的教育課程資源。此外，職亦搜尋線上可用的學習資源，如並購買多人使用的線上課程，學習 R 語言的開發環境，語法格式與機器學習相關套件使用。

於學習人工智慧與機器學習相關知識部分，職空閒時也額外透過台大電機系及台大人工智慧與機器人研究中心李鴻毅老師的線上課程：機器學習(Machine Learning)來學習。機器學習是一種人工智慧的演算技術，透過機器自身學習，可將大數據資料依據不同演算法去學習到其中的固定規則，當模型訓練完成後，當加入新數據於模型中，則模型會依據過去學習到的模型來判斷新數據。而機器學習領域的學習方式主要區分為監督式學習(supervised learning)，非監督式學習(unsupervised learning)與半監督式學習(semi-supervised learning)三種模式。監督式學習(supervised learning)主要是有一明確定義的學習對象(即為標籤，label)，如是否有疾病，是否有病兆，也可分為類別與數值兩種，各自有決策樹，隨機森林，K-近鄰演算法及支援向量機與線性迴歸及多項式迴歸等不同演算模型。非監督式學習(unsupervised learning)則是不設定學習對象，即不給予人為判斷過的結果，直接由機器依照設定的機器學習方式進行歸類或分組，將整體數據依相似特徵的歸類，可使用集群分析來學習。半監督式學習(semi-supervised learning)則是結合前兩項，因很多數據量大，不易花費人力進行標記，因此利用部分標記好的數據進行學習，而後在使用無標記數據進行無監督式學習。

11 與 12 月份的研究會議中主要討論未來研究方向，因職本身同時也是神經內科醫師，范教授提供了台灣腦中風登錄資料庫大數據資料供我使用，且許凱程醫師與林敬恆研究員也剛使用此資料庫完成研究題目與投稿，因此可透過他們之前使用此數據集研究的方法來熟悉此資料庫，並從而了解機器學習分析步驟。因此從 12 月份開始，除每月團隊會議外，與許凱程醫師與林敬恆研究員固定每週一次聚會定期討論研究方向與內容。

2020/01

本月份因許凱程醫師邀請，參與 Lawrence Latour 團隊討論，並學習腦部 CT 與 MRI 的標記方式。Lawrence Latour 教授為國家神經疾病和中風研究所裡 Center for Neuroscience and Regenerative Medicine 及 Acute Cerebrovasular Diagnostics Unit 的主要負責人，他本身專長於腦中風與腦部外傷的 MRI 研究。范教授研究團隊取得 Lawrence Latour 教授部分腦部 MRI 影像，欲進行近一步研究區分不同腦中風的類型。因參與影像標記，需加入研究團隊才能使用後續資料，因此會議後，需接受國家衛生研究院 IRB 相關課程，取得資格後，才能取得協助資料處理的資格。

針對研究方向討論，目前團隊手上正進行的研究主題為使用自然語言處理從放射線報告中萃取出頸動脈與顱內血管疾病標記，並分析此模型的準確度，同時團隊手上亦有某國內大

型院內大數據資料，亦欲透過自然語言處理模式來預測再中風機會，而主要使用的工具為 Google 所發展的 Bidirectional Encoder Representations from Transformers(BERT)技術。

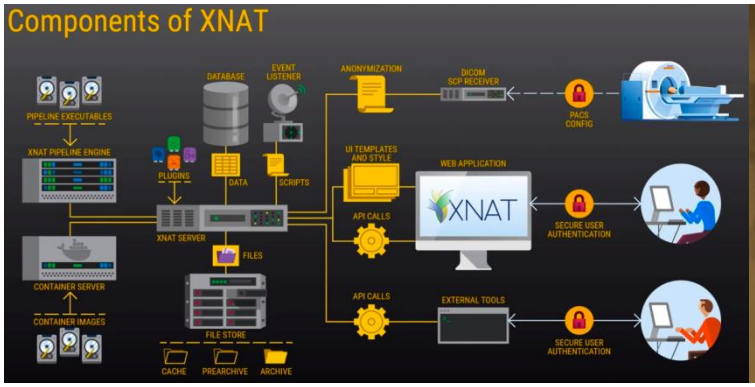
本月確認未來亦會使用到 Python 工具，因此在接下來開始也同時透過單位課程與網路課程學習 Python 語言語法，如 Udemy 的 Machine Learning A-Z 與 Python 程式設計入門課程，及 DataCamp 網站的 Machine Learning 課程與 Importing and Cleaning Data 課程。

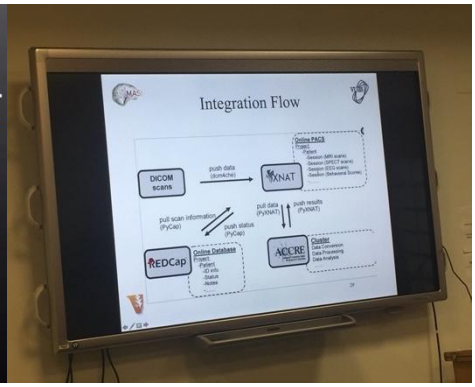
2020/02

本月開始協助撰寫 StrokeBERT: A Disease-Specific Language Representation Model for Cerebrovascular Disease Research 文章的部分內容。

2020/2/13

參與團隊與 IT 部門會議，討論使用 XNAT 開源工具建立影像資料庫並串聯電子病歷資料作為後續分析的平台。XNAT 是由華盛頓大學的神經信息研究小組(Neuroinformatics Research Group)所開發的開源成像信息學平台。XNAT 可簡化影像和相關數據的日常管理，提高資料處理效率，同時本身並具有影像品質確認與影像標記的工具可供線上使用。





2020/02/19

拜訪賓夕法尼亞大學佩雷爾曼醫學院王立三(Li-San Wang)教授

王立三教授主要研究方向為阿茲海默症與失智症之遺傳基因體研究與基因組學之大數據分析計算方法，拜訪原因主要是王教授所在單位為全美重要的失智症基因研究中心之一(National Alzheimer's Coordinating Center (NACC))，與全美各失智基因研究單位均有合作，且主要負責管理由美國國家老化中心支助的 Alzheimer's Disease Research Centers (ADRCs)所收集的資料，內容包括標準化臨床數據，包括參與者的人口統計數據、家族史、藥物、共病、身體檢查、幾種評估工具的結果(例如，臨床失智評分量表Clinical Dementia Rating (CDR)，老年性憂鬱症量表）、症狀和診斷的臨床評估以及神經心理學測試結果，也包括腦部解剖獲得的神經病理學資料。王教授為台大電機畢業，因緣際會對生物資訊產生興趣，於賓州大學做博後時開始接觸了基因體研究，當時做的人還不多，而剛好賓州大學正在尋找運用

生物資訊學分析老化基因體的人才，從此在這領域持續發展，王教授談到，目前歐美對於失智症的基因研究已經很豐富，但目前跨人種比較卻是較缺乏與有需求的，因此他也與台灣部分單位接觸過，未來若取得計畫，可與台灣進行合作，談話中王教授也提到，目前全美研究的趨勢多為合作關係，因此他們基因庫就是做跨單位研究團隊的資源共享，這樣的研究氛圍才能讓一直沒有新進展的失智症研究中有突破的機會。

2020/03

申請 Biowulf，並學習 Biowulf 操作方式。

Biowulf 本身是國家衛生研究院的超高效能電腦，從 1999 年建立後，Biowulf 已經增加了 21000+個 CPU，包括 56 個帶有 Nvidia V100 GPU 的 GPU 節點，每個節點 4 個 Nvlink Skylake 處理器，384 GB 內存和 3.2 / 1.6 TB SSD，供院內研究人員使用，而這樣的服務是需要每月付月費取得使用權。Biowulf 本身也內建許多雲端應用工具，如 R 語言，Python 語言，與機器學習或神經網絡分析工具，可以直接運用 Biowulf 本身的 CPU, GPU 與記憶體進行一般個人電腦無法進行的大數據資料分析。台灣部分，國家高速網路的臺灣 AI 雲(Taiwan Computing Cloud)AI 雲端平台 TWCC，共運用 2,016 個 NVIDIA Tesla V100 32GB GPU 服務，也是超級電腦等級，此平台亦內建「分析大師」（Data Analysis Service, DAS）可提供高速雲端運算引擎、海量資料的網路線上儲存及處理與一站式的資料整合與分析服務，范教授也與國家高速網路有合作關係，也提到會去申請使用此資源的單位目前不多，而目前高榮也已有使用此項服務。除此之外，亦可選擇自費或使用研究經費購買民間的 Google Cloud、Microsoft Azure 與 Amazon AWS 系統的使用權。

原本經范教授同意由部門提供報名費用 177 美元報名參加國家衛生研究院內科學高等教育基金會(Foundation for Advanced Education in the Sciences, FAES)的 7 週(2020/03/25~05/06)的精準醫學課程，FAES 是位於國家衛生研究院內的一個非營利組織，專門辦理研討會和大型會議與研究生課程，每期均有大量的院內與院外研究人員與學生報名課程，也有些台灣學者特地為上一堂課，不遠千里坐飛機來報名學習。然三月份為馬里蘭州，華盛頓 DC 與維吉尼亞洲 COVID-19 病患開始顯著增加的月份，因此整個馬里蘭州開始進行嚴格的隔離措施，我們單位也將會議調整為線上會議，並維持每週小團隊與每月整個 AI 團隊會議的模式，持續針對研究內容進行討論。而國家衛生研究院院長 Francis S. Collins, M.D., Ph.D.教授也持續每週五公開現場直播說明國家衛生研究院受疫情影響程度與後續措施。而報名之精準醫學課程也因此確認停課。

2020/04～05

協助團隊完成 StrokeBERT: A Disease-Specific Language Representation Model for Cerebrovascular Disease Research 文章。並確認自身研究方向與主題，預計透過腦中風登錄資料庫 250 多個特徵與上萬人次的數據集來分析：1. 運用機器學習分析預測出院後續追蹤之功能變化情形，是否改善、持平或持續惡化; 2. 運用機器學習分析預測住院治療後，出院時的 NIHSS 是否改善、持平或持續惡化。確認主題後，則從進行資料清理開始，並學習一步一步的進行資料前處理，包括缺失值的處理、類別資料的處理（有序、無序）與資料特徵縮放等步驟。

IT 團隊完成 XNAT 平台建置且匯入影像資料於 XNAT 內後，指導我們學習如何於 XNAT 系統中，查詢、匯入匯出影像與及影像附屬資料與於 XNAT 平台內進行影像標記。

2020/06

研究所需的腦中風登錄資料已經清理完畢，可分析資料為 47,000 多筆資料，並進行缺失值補值與特徵選取，但後續執行機器學習分析於預測多類別變項(三組)時卻有較低正確率，亦遇到不平衡數據導致較低正確率的問題，因而持續研究如何對資料進行過採樣與欠採樣，以提高結果預測正確率。

2020/06/06

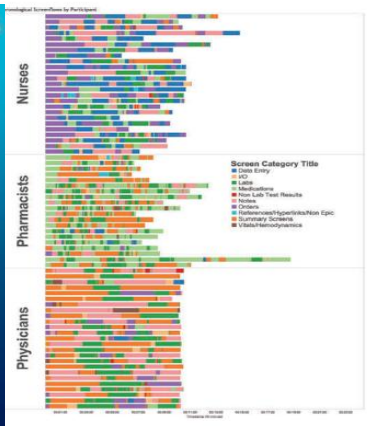
因對於數位醫療的興趣，剛好因疫情台灣許多會議也同時有線上轉播，也想了解目前台灣其他醫院對於數位醫療的發展，因而也報名參加數位醫療 2020 在臺北榮總的線上研討會 (Digital Healthcare 2020 at Taipei Veterans General Hospital)。從此次會議也看到人工智慧可以涵蓋的領域真的很廣，如病患安全方面，可透過大數據分析降低住院病人跌倒發生率，於智慧型醫院管理部分，如智慧化藥事作業、醫療設備與特材管理、麻醉電子病歷暨管理系統等，而疾病預測部分，則有如住院病患之病危預測、心血管風險預測、代謝體學預測缺血性中風功能、結直腸癌肺轉移預測等研究方向，而智慧型醫療教育部分，則有虛擬及擴增實境系統培訓醫學生職安及 OSCE 技能與臨床病歷紀錄模式改變等。雖然大部分的應用仍在研究中，實際能於臨床應用的還不多，但在病安與智慧醫院管理部分，應該是能先上線使用的項目，而與病人疾病、診斷與治療相關的部分，因牽涉甚廣則需要反覆確認與進一步認證後才能商品化使用。

2020/07

2020/07/15

七月份的Grand Round請到Keith A. Horvath醫師主講，主題為：It’s an Artificial Intelligence (AI) World and We Are All Just Living in It。此次演講主要在說明實際醫療環境中AI的應用與影響，演講的前半段，講師提到目前美國電子病歷(EHR)使用已經超過九成，也因為每分每秒均有大量的EHR資料儲存起來，EHR本身對於人工智慧醫院的形成佔有很重要的角色，但使用電子病歷也造成照護人員感覺佔用太多時間，且沒有常規訓練也造成資料輸入常有錯誤，也造成醫療錯誤或是事後分析產生偏差，因此美國有許多單位有開發模擬電子病歷(EHR simulation)，供訓練使用。此外也有許多輔助病歷撰寫的人工智慧APP協助，比如透過語音輸入轉換的方式，以提高效率。

此外，於不同專業人士身上也看到各自對於EHR裡不同項目的使用程度有很大的差異，如病歷輸入，I/O，藥物等，各自重視與觀察角度亦不同，因此如何結合並同時呈現各專業人士評估結果並將之視覺化亦是很有趣的嘗試，如右圖上可以看到一個病患有不同的安全評估項目，每個專業各自評估部分項目，再將之重要性視覺化，可看出同一項目各自判斷也不同。



Sakata. j ip care 2016

下半段演講則提到許多臨床應用，如HEART PATHWAY MOBILE APP，這個免費APP包含有年齡，症狀，生活習慣與共病及ECG輸入，並提供心臟病風險判斷與建議，但不可用在有心血管疾病與急性心肌梗塞患者，研究發現使用此APP可提升21%早期出院率與降低12小時住院時間; 管制用藥提醒系統，這在高榮很早就已有類似的應用;人工智慧結合大腸鏡影像檢查，提早息肉偵測機會;人工智慧系統協助非經專業訓練人士使用移動型心臟超音波取得高品質的影像; 於安寧緩和領域，透過過去病歷資料輸入判斷，早期篩選出存活時間3~12個月的個案，讓照護團隊主動出擊，不需等到轉介才介入。

2020/07/27-31

參加世界失智症協會2020國際研討會線上會議(2020 Alzheimer's Association International Conference)

本會議為每年度失智症國際性大會議，原在荷蘭阿姆斯特丹舉辦，但因COVID-19疫情改為全程線上會議，且完全免費，造福許多失智症領域的研究人員與臨床照護者。本會議一共

五天，每天均有不同主題做深入探討，包括基礎科學與病理機轉、生物標記、臨床表現與藥物發展、公共衛生與失智照護及專業發展等五大主題。簡要說明於會議中所學習到的新知，目前認為 β-類澱粉蛋白沈積是阿茲海默氏症的主要病理機轉，且在很早期就開始累積這樣的病理變化，隨之造成神經細胞損傷因而產生 tau 蛋白，接續造成認知功能障礙。而於主觀認知障礙至輕度認知障礙個案身上，隨著症狀越嚴重，會有越來越高比例會在腦中偵測出有 β-類澱粉蛋白沈積，且這樣的比例亦隨著年齡而上升。然而臨床診斷為阿茲海默氏症的患者，確會因年齡增加，反而有 β-類澱粉蛋白沈積個案比例下降的情形，且有約 15%臨床阿茲海默氏症診斷者沒有阿茲海默氏症的病理變化，其可能機轉為血管性病變或是 TDP-43 所造成的，而透過解剖的研究也發現多數臨床診斷阿茲海默氏症患者常並存著多重病理機轉。

研究也發現缺乏阿茲海默氏症病理變化的患者相較於具有病理變化者，年紀較長、沒帶有 ApoE4 基因、男性、較輕微的認知功能障礙且常伴隨多重病因。因此在阿茲海默氏症的治療中，區分出是否具有 β-類澱粉/Tau 蛋白的病理變化會影響治療選擇，而研究也發現如 Gantanerumab、BAN2401 與 Aducanumab 治療後均可減少腦中 β-類澱粉蛋白的量，也同時會降低 Tau 蛋白的量，然而目前對於藥物清除腦部的 β-類澱粉/tau 蛋白後是否能減輕或改善認知功能的退化，仍待進一步確認。研究也發現 Tau 蛋白常伴隨著 β-類澱粉蛋白後出現，而在無 β-類澱粉蛋白沈積的輕度認知障礙或阿茲海默氏症患者，卻很少發現有 tau 蛋白的存在，而在具有 β-類澱粉蛋白沈積但無認知功能障礙者身上，卻也發現約 40%的人也同時帶有 tau 蛋白的病理變化。β-類澱粉蛋白的量可能會於未有症狀的時候先有約 7~10 年的快速累積，而後累積量達到一個頂點後，則增加速度就開始減緩，再經 10 年的累積之後才開始出現臨床認知功能症狀，因此前一個 10 年的快速累積期可能是我們作為 primary prevention 的時機，而第二個 10 年在未出現症狀前則可作為 secondary prevention 的時機，而若帶有 ApoE4 基因型者，則相較於無帶基因者，腦部則更是會提早 15-19 年發現 β-類澱粉蛋白沈積。雖然如此，primary prevention 或 secondary prevention 是否能減少腦部病理變化與認知功能的衰退，目前仍未可知，而世界阿茲海默氏症協會與美國國家老化研究所共同贊助了 U.S.POINTER 研究，預計於美國收錄 2,000 位高風險個案，進行 2 年的多因子介入措施，想了解是否積極的預防措施可改善認知功能與腦部病理變化，這研究如果有很好的結果發表，將會對於目前失智預防措施提供很重要的實證。

2020/07～08

這兩個月也同時在協助團隊撰寫論文 Identifying carotid stenosis from angiography reports using natural language processing approaches 的部分內容。而於進修中研究進度部分，目前仍有預測準確率無法提升的困難尚待解決，持續學習使用不同機器學習分析方式期待能尋求好的演算模型。這幾個月的時間也因疫情關係，也抽空完成兩份使用出國前台灣資料的文章並進行投稿，且已有一篇使用健保資料庫資料探討住院中譫妄對於長期醫療耗

用，死亡情形與精神藥物使用影響的文章被接受，也同時協助其他兩位醫師文章的撰寫與修改，並開始針對第三篇於院內骨科病房進行非藥物多項目介入措施看長期預後成效的文章進行撰寫。

此外，也開始準備回國事宜，感覺雖然有一年的時間待在這個世界級的研究中心，但真的時光飛逝，感覺還在剛到美國開始尋找學習方向與目標，一晃眼已經要回國了，而且因疫情影響還有許多未學習與學習未完善的部分，包括雖然已初步運用 Python 與 R 語言於機器學習分析上，但對於此兩種程式語言的熟悉度仍然不足，而神經網絡分析方法也還未能有機會詳細學習，均仍有許多遺憾，目前台灣已有相關的學習資源，因此回國後也可持續在此領域繼續學習。而因為疫情也導致原本安排內部參訪其他單位的計畫也取消，且原本單位與日本與台灣許多研究單位的交流也暫緩，少了多方交流與認識的機會。范教授與台灣重要的研究單位均有長期合作，因此也於進修結束前邀請老師回台後能到高榮來看看，而原本研究室的同事，有大部分均回到台灣就職，遍佈在北中南各地，未來也將保持與范教授及其他同事的持續研究合作。

2020/02～05

報名科學高等教育基金會(Foundation for Advanced Education in the Sciences, FAES)的 Applied Machine Learning，講師為 Martin Skarzynski。

課程內容包括：

第一周：課程概述。機器學習主題簡介

第 2 週：使用 numpy，pandas，scikit-learn 進行數據檢索和可視化

第 3 週：數據整理，預處理和規範化

第 4 週：有監督的學習 1：回歸問題

第 5 週：有監督的學習 2：過度擬合，正則化，超參數優化和交叉驗證

第 6 週：有監督的學習 3：分類問題

第 7 週：無監督學習 1：聚類

第 8 週：無監督學習 2：潛在變量模型

第 9 週：無監督學習 3：降維和特徵選擇

第 10 週：深度學習 1：深度學習方法簡介

第 11 週：深度學習 2：實習

第 12 週：實施機器學習工作流程。常見的陷阱和最佳做法。

上課過程雖以實作為主，講師 Martin Skarzynski 教得相當仔細，但因疫情關係，課程後段內容無法有充足時間完成，但仍於課堂上學習到機器學習的分析方法及步驟與深度學習的基礎知識。

研究與發表：

職於進修期間跟隨范楊政教授與其團隊學習研究分析，參與兩個研究主題的討論與協助部分內容撰寫，1. 主題為 StrokeBERT: A Disease-Specific Language Representation Model for Cerebrovascular Disease Research，研究核心為使用腦中風電子病歷資料，將原生 BERT 模型重新訓練出適合腦中風的 StrokeBERT 模型，再透過是否能從放射線報告中正確分辨出顱內與顱外共 17 條動脈阻塞問題，與可否透過此模型從出院病摘中學習預測出院後再中風的能力，經此來判斷 StrokeBERT 的模型能力，此篇文章正在 IEEE: Journal of Biomedical and Health Informatics 審稿中。2.主題為 Identifying carotid stenosis from angiography reports using natural language processing approaches，此篇文章主要在於比較使用不同自然語言處理方式來辨別 11 條顱內動脈阻塞的能力差別，文章目前仍在修改中。

三、心得及建議

經過在美國國家衛生研究院國家神經疾病和中風研究所一年進修後，更是深深感謝院部長官與高齡醫學中心林育德主任給予的這次進修機會，還有中心其他同事協助分擔職原本的工作。在醫院日復一日的常規工作常會讓人落入習慣性的生活之中，而接觸的面向也相對單純，過去幾年因職有機會常出國參加許多國際會議，也發現唯有跨出去接觸外面世界才能看到自己缺乏的部分，也會受到其他國際學者的激勵，回國後才能持續產生動力而維持自己在研究方面的興趣。這次在國外研究單位進行持續長達一年的學習，也了解美國為何會這麼先進原因，因為匯集了全世界頂尖研究人員與豐富的研究經費，無論是硬體或是軟體均投入豐富的資源，且更新迅速。而國家衛生研究院也針對每一個領域每年均舉辦多次的專家會議，探討目前研究進展與未來擬定的研究方向，也根據擬定方向投與相對應適當的經費與資源。另外也投入資源將新的技術與知識對外分享，來提升整個國家的整體研究能力，並定期舉辦多領域的演講與研討會，邀請各領域頂尖專家授課，這些均有錄影且放置於網站與 NIH Youtube 頻道免費分享(https://www.youtube.com/c/NIHVideoCast/featured)供有興趣的人員觀賞聆聽。此外，在進修中也感受到跨領域合作與資源分享的氛圍，透過不同單位間的合作，更是能將既有的研究資料做更廣泛的探討與分析，對於科學領域的發展相當重要。這也是在台灣研究單位裡常感受不到的，當然這也需要自身有足夠的能力才能受到其他人重視，畢竟在國家衛生研究院裡的均是頂尖的研究人員。

經過這一年的進修有一些心得與建議如下：

1．高齡資料庫建立：

醫院內的資料數據龐大，也是目前人工智慧領域大家所珍惜且想探索的大數據資料，除了電子病歷(文字類型)外，醫院內還有數據、影像、訊號與圖像等不同資料來源，於國家衛生研究院進修時接觸到 BTRIS 資料庫，BTRIS 資料庫整合了所有臨床研究與病歷檢查等資料於一個資料庫內，這樣的方式對於後續無論是品質管理或是研究需求均有其便利性，但前端作業將原醫療資料轉成研究分析資料庫過程也有一些困難，如需有額外儲存空間，如何統一資料格式，如何匿名化轉換，如何建立管理與查詢平台等，職於進修時接觸到其他醫院所匯出的歷年病歷資料數據時也發現，目前多數做法仍是以個案申請需求，再透過單一窗口從醫療系統取得資料，匿名化後，再交由研究人員分析，好處是不需建置額外平台，但卻於每次取資料時均須重新整理成可分析資料，這步驟確實耗時，但也較不新增成本。而目前國際研究方向，已開始同時結合臨床資料、影像資料與生物資訊資料做合併分析，因此擴充資料庫納入所有可分析的資料源是未來的趨勢。職也發現於出國這一年中，高榮院內資料庫也已開始整合，作為提供院內研究使用。職未來將持續與范揚政教授團隊學習，試著學習整合高齡醫學中心過去研究資料內容，雖然資料數據非大數據，但也希望能便於後續研究的使用，提升中心研究動能，也將邀請范教授回台時到高榮與高齡醫學中心提供資料庫建立的建議。

2．跨界與跨專業領域合作：

人工智慧領域的研究持續在熱頭上，但發展也相對快速，一個想法馬上就會有許多團隊同時在進行，而且技術更是幾乎無時無刻都在更新，前一刻研究使用的技術，下一刻可能就落伍了。目前發展積極的中國附醫投注大量資源發展醫療人工智慧，下設四個ＡＩ團隊，長庚也於 2018 年成立醫療人工智能核心實驗室，均投注大量人力與資源，人員涵蓋有研發工程師(演算法開發)，系統開發人員(系統建置)，資料處理助理與數據分析師，但組成這樣團隊需要耗費太多資源，人工智慧領域並無法單打獨鬥，需跨專業合作，因此更需要尋找業界與學校資源。高榮也在品質管理中心與研創中心楊宗龍主任帶領下，起步也早，對外已與許多研發工程師與數據分析師合作，對內也已有許多科別參與人工智慧研究發展，因此已有很好的發展與創新，職過去也感謝楊宗龍主任的邀請，參與其手部震顫平板偵測系統開發的研究，學習收穫很多，也才知曉醫療人工智慧的豐富可能性。然而，因臨床人員平日工做繁忙，也不易有時間能對此新發展領域熟悉，但臨床人士的各領域專業知識是業界與學界所渴望合作的，因此小小建議有機會能創造讓各臨床單位人員與業界或學界有更直接的接觸，透過與不同屬性的專業小組合作討論，如醫學影像分析、訊號分析、生物資訊與文字分析等，應該會產生很多火花，並在各科室中埋下小小研發種子。

3. 人工智慧種子散播：

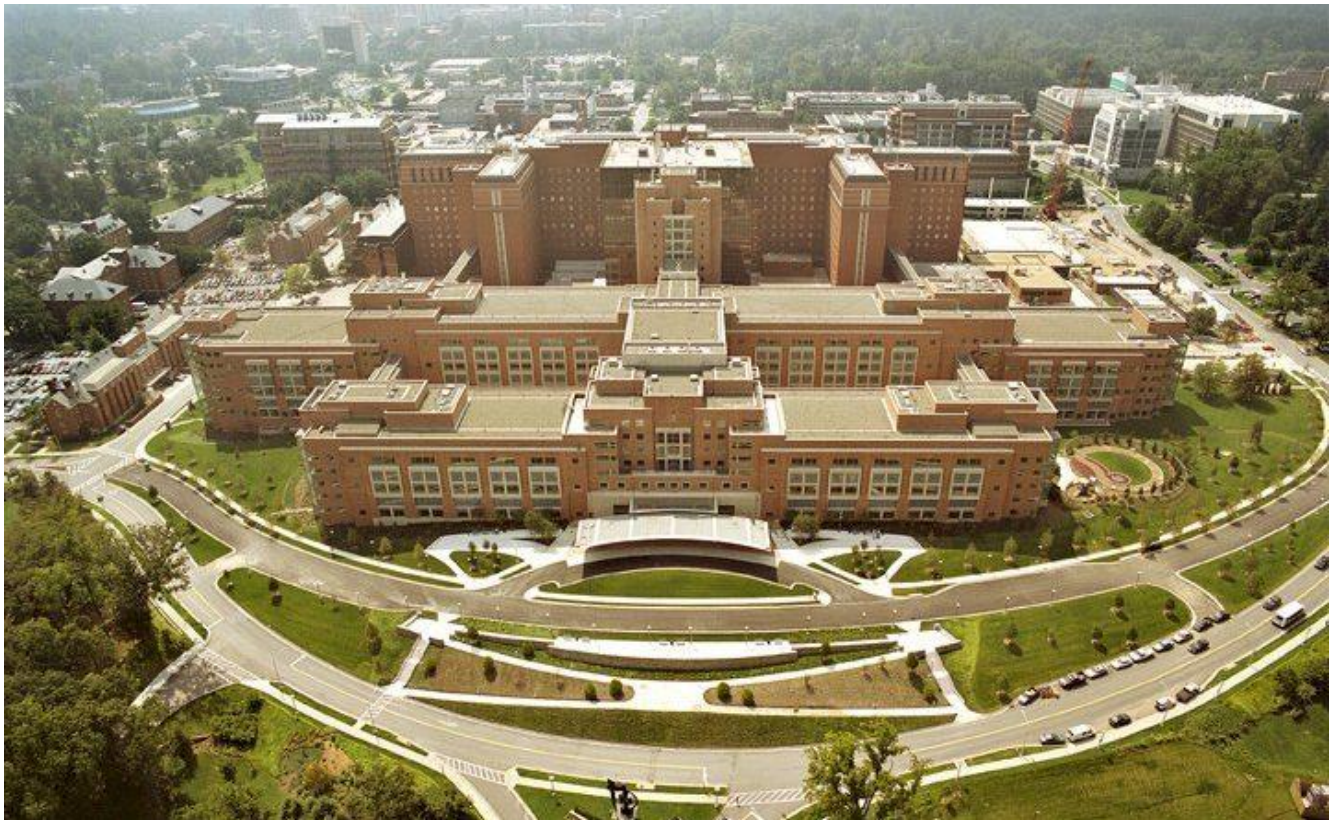
人工智慧醫療照護領域有相當廣泛的面相可以發展，而一般而言多數臨床人員並不清楚人工智慧對於各自領域有何幫助，實際上臨床服務過程中的訊號源(如 EEG, EKG, vital signs 等)、影像資料、視覺圖像資料、語音資料與電子病歷資料等，均會產生不同資料可供後續分析使用，而人工智慧的發展也有以下不同領域可依臨床需求發展，如臨床決策支持系統、疾病診斷輔助、檢查篩檢輔助、複雜的圖案和圖像分析、自然語言處理和理解、預後預測建模、機器人和圖像引導手術、個人化的影像和治療建議等。因各單位有不同的資料源與需求，因此建議可以依 1.整合單位所屬資料來源、2.議題確認與 3.團隊媒介三個步驟來協助，透過說明讓臨床人員理解單位所有的資料屬性與可發展的領域，再協助整理各單位特有數據資料來源與可發展領域，再由有興趣的臨床人員提出想法，並媒介符合需求的學校與業界資源。

4. 單位於高齡醫學與神經科學領域人工智慧的目前與未來研究方向：

職個人期待能於下列方向由中心與院內外專業人士持續合作發展，

- ✓ 臉部表情辨識和認知功能分析（合作單位:成大資訊工程系，收案中）
- ✓ 藥物大數據分析（合作單位:嘉南藥理大學，合作討論與分析中）
- ✓ 利用院內資料庫數據結合長照資料建立長照需求預測模型（資料整合階段）
- ✓ 建立人工智慧輔助標記極早期中風影像病兆與建立中風預後預測模型（未來發展）

附錄：

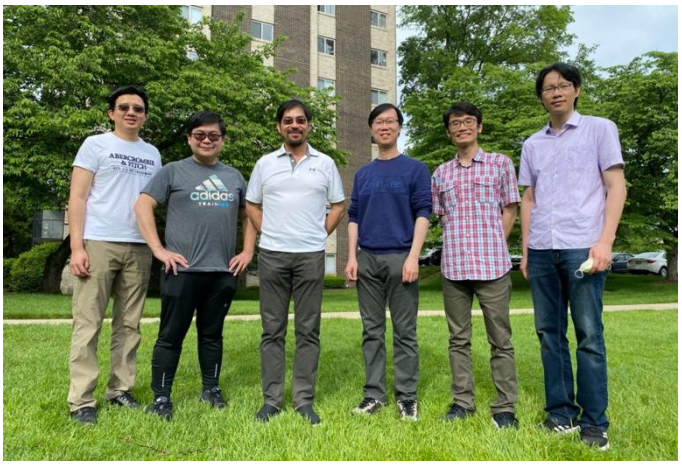


國家衛生研究院臨床中心，位於 10 號樓，為專門從事臨床研究的醫院，超過 1,600 個臨床試驗在這裡進行，COVID-19 vaccine 研究也在此

國家衛生研究院臨床中心的南側建築，為一研究型大樓

國家衛生研究院 1 號樓，於 1938 年建，又名 James H. Shannon，為行政中心，NIH 院長辦公室亦在此大樓

35 號樓，John Edward Porter Neuroscience Research Center
范楊政教授與國家神經疾病和中風研究所 IT 辦公室位於此



Bioinformatics Team

左:Amar Yavatkar(Bioinformatics Programmer)

右:Kory Johnson(Ph.D., Bioinformatics Scientist)

AI Team

左一許凱程醫師，左二陳一銘醫師，左三范揚政教授

右一胡育文醫師，右二林敬恒博士研究員

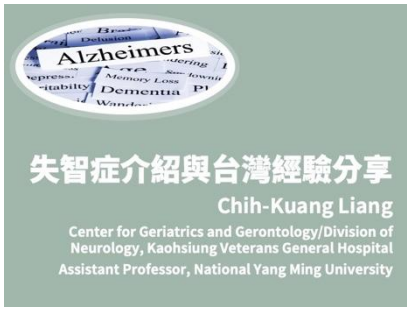
研究辦公室

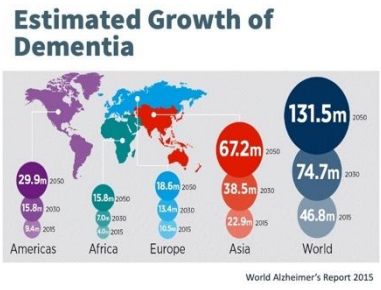
研究辦公室

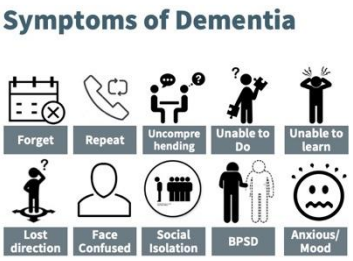
臨時通行證

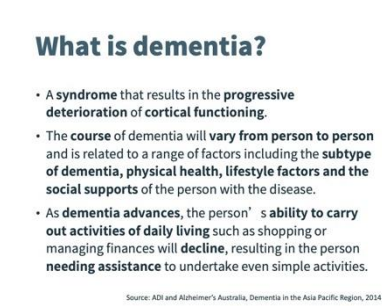
正式員工證

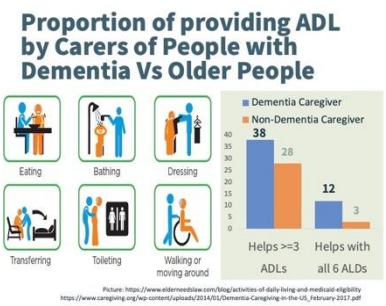
2019/10/26 華府國建聯誼會(CAPA)演講邀請

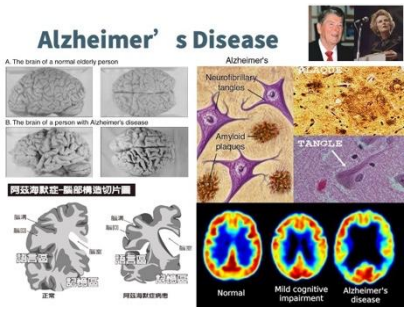


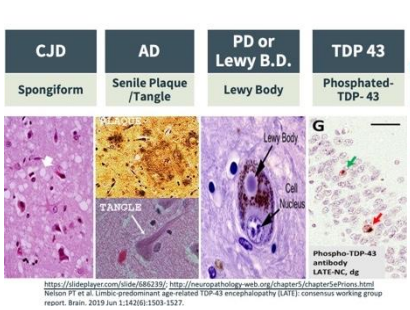


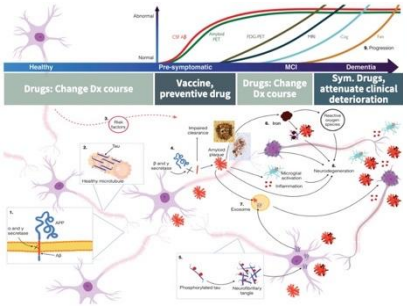


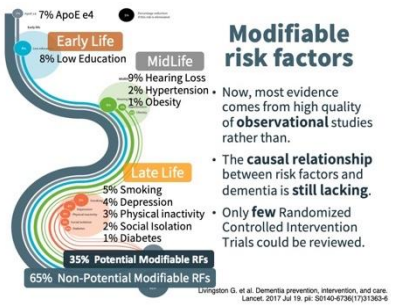


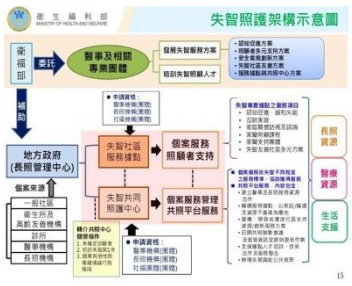












研究團隊投稿文章（審稿中）

StrokeBERT: A Disease-Specific Language Representation Model for Cerebrovascular Disease Research

Ching-Heng Lin, Kai-Cheng Hsu, Chih-Kuang Liang, Tsong-Hai Lee, Chia-Wei Liou, Jiann-Der Lee, Tsung-I Peng, Ching-Sen Shih, Yang C. Fann

***Abstract*—Effectively utilizing disease-relevant text information from clinical notes for medical research presents many challenges. BERT (Bidirectional Encoder Representation from Transformers) related models such as BioBERT and ClinicalBERT, pre-trained on biomedical corpora and general clinical information, have shown promising performance in various biomedical language processing tasks. In this study, we demonstrated that a BERT model pre-trained on disease-related clinical information can be more effective for cerebrovascular disease-relevant research. StrokeBERT was initialized from BioBERT and pre-trained on large-scale cerebrovascular disease related clinical text information. The pre-trained corpora contained 113,590 discharge notes, 105,743 radiology reports, and 38,199 neurological reports. Two empirical clinical tasks were conducted to validate StrokeBERT's performance. The first task identified extracranial and intracranial artery stenosis from two independent sets of radiology angiography reports. The second task predicted the risk of recurrent ischemic stroke based on patients' first discharge information. In stenosis detection, StrokeBERT showed similar or better performance on targeted carotid arteries, with an average AUC compared to that of ClinicalBERT of 0.968 ± 0.021 and 0.956 ± 0.018, respectively. In recurrent ischemic stroke prediction, after 10-fold cross-validation on 1,700 discharge information, StrokeBERT presented better prediction ability (AUC±SD = 0.838 ± 0.017) than ClinicalBERT (AUC±SD = 0.808 ± 0.045). The attention scores of StrokeBERT showed better ability to detect and associate cerebrovascular disease related terms than ClinicalBERT. This study shows that a disease-specific BERT model improved the performance and accuracy of various disease-specific language processing tasks and can readily be fine-tuned for enhancing cerebrovascular disease research and applications.

***Index terms*—BERT, specific language representation model, cerebrovascular disease**

This work is supported by funding from Intramural Research Program of National Institute of Neurological Disorders and Stroke, National Institutes of Health, USA and from the research grant of Chang Gung Memorial Hospital (CMRPG3F2211, CMRPG3J1161). This work utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov). The authors thank the NIH Library Writing Center for manuscript editing assistance. The authors thank Google for opening BERT source codes. CKL thank the Veterans Affair Council and Kaohsiung Veterans General Hospital, Taiwan, for financial support on study abroad. (Corresponding author: Yang C. Fann.)

Ching-Heng Lin is with the Center for Artificial Intelligence in Medicine, Chang Gung Memorial Hospital, 333 Taoyuan, Taiwan and also with the Bioinformatics Section, National Institute of Neurological Disorders and Stroke, National Institutes of Health, 20892 Maryland, United States (email: chingheng113@gmail.com)

Kai-Cheng Hsu is with the Department of Artificial Intelligence Center for Medical Diagnosis, China Medical University Hospital, 404 Taichung City, Taiwan, and also with the School of Medicine, College of Medicine, China Medical University, 404 Taichung City, Taiwan (e-mail: edwardfirst@gmail.com)

Chih-Kuang Liang is with Center for Geriatrics and Gerontology and Division of Neurology, Kaohsiung Veterans General Hospital, Kaohsiung City 81362, Taiwan, Bioinformatics Section, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, Maryland 20892, United States, and also with Aging and Health Research Center, National Yang Ming University, Taipei 112, Taiwan. (e-mail: ck.vghks@gmail.com).

Tsong-Hai Lee is with the Stroke Center and Department of Neurology, Chang Gung Memorial Hospital, Linkou Medical Center, 333 Taoyuan, Taiwan and also with the College of Medicine, Chang Gung University, 333 Taoyuan, Taiwan (email:thlee@adm.cgmh.org.tw)

Chia-Wei Liou is with the Department of Neurology, Kaohsiung Chang Gung Memorial Hospital, 833 Kaohsiung, Taiwan and also with the College of Medicine, Chang Gung University, 333 Taoyuan, Taiwan (email: cwliou@ms22.hinet.net)

Jiann-Der Lee is with the Department of Neurology, Chiayi Chang Gung Memorial Hospital, 613 Chiayi, Taiwan and also with the College of Medicine, Chang Gung University, 333 Taoyuan, Taiwan (email: jdlee540908@cgmh.org.tw)

Tsung-I Peng is with the Department of Neurology, Keelung Chang Gung Memorial Hospital, 204 Keelung, Taiwan and College of Medicine, Chang Gung University, 333 Taoyuan, Taiwan (email: tipeng@cgmh.org.tw)

Ching-Sen Shih is with Division of Neurology, Department of Internal Medicine, kaohsiung Veterans General Hospital, Kaohsiung City 81362, Taiwan. (e-mail: csshih@vghks.gov.tw)

Yang C. Fann is with the Bioinformatics Section, National Institute of Neurological Disorders and Stroke, National Institutes of Health, 20892 Maryland, United States (email: fann@ninds.nih.gov)

I. INTRODUCTION

Cerebrovascular disease is an important cause of mortality worldwide and a major source of morbidity, affecting 16.9 million cases in 2010 [1]. Cerebrovascular disease research relies on comprehensive clinical information, including images, laboratory data, and various clinical notes. Among these, clinical notes and radiology reports contain the most rich but under-utilized information; both are classified as unstructured electronic health records (EHR) data and usually contain jargon,