出國報告（出國類別：進修）

人工智慧於核子醫學影像造影之應用: 大型語言模型輔助骨骼造影掃描之語意報告產生

服務機關：成大醫院
姓名職稱：盧晞卉 醫師
派赴國家：美國
出國期間：2023.08.08 ~2024.08.06
報告日期：2024.09.02

摘要

本次進修主要探討人工智慧在核子醫學影像造影中的應用，特別是大型語言模型如何輔助骨骼造影掃描的語意報告生成。本報告涵蓋在美國哈佛醫學院的進修經歷，並介紹哈佛醫學院的生物資訊研究所如何利用人工智慧 (Artificial intelligence, AI) 技術提升影像判讀的精確度。目前使用放射性核種鍶-99m 甲基雙磷酸鹽進行骨骼掃描有其優勢及其局限性，特別是在判斷惡性病變時可能受到良性因素影響。為了克服這些挑戰，研究者嘗試通過訓練多模態語言模型如 LLaVA，以提高人工智慧對於醫學影像的識別與分析能力，並結合無監督式學習及半自動標註等技術來優化模型。然而，由於大型語言模型的微調過程依然複雜且所需資源密集，未來仍須持續試驗改進。

關鍵字：人工智慧，語言模型，醫學影像報告

目次

目的

對於惡性腫瘤的病患而言，各樣醫學影像在早期診斷、疾病分期、治療效果的評估以及長期追蹤腫瘤狀況的應用上，已是不可或缺的工具。其中全身骨骼掃描為現今評估病患是否有骨骼轉移與骨骼轉移程度變化相當經濟且敏感度高的影像學檢查。

目前普遍適用於全身骨骼掃描的放射性核種為同位素鍀-99m 甲基雙磷酸鹽(Tc-99m MDP)，其原理為透過靜脈注射，藥物由血液帶至全身經由化學吸附的方式聚積於骨骼結構。然而除了骨骼惡性腫瘤，惡性腫瘤骨轉移之外，其他各種良性因素包括：外力撞擊、骨折、關節發炎、退化性疾病、良性腫瘤等也都會造成藥物聚積，影響判別惡性疾病存在的正確性。專科醫師透過經驗以及訓練，考慮病患病史，相關血液檢驗，相對應的其他影像，也觀察病灶本身的分布、形狀、聚積強度等分析，進而判斷骨骼掃描病灶的良惡性。

因此，全身骨骼掃描的判讀，需要核醫專科訓練以及醫師的臨床經驗。隨著近年人工智慧的演進，人工智慧模型的能力也從分類等簡單任務進展到可產製出具完成語意的一整段文字。在此年度的研究中，我們實際訓練具圖像視覺功能的大型語言模型於全身骨骼掃描的判讀上。大型語言模型在近年的發展十分迅速，然而其識別圖像的能力尚在萌芽階段；針對這類大型模型的研究，皆需要動用可觀的運算資源與儲存資源，因此特別規劃這次出國進修研究前往美國哈佛醫學院之生物資訊研究所，獲得第一手的實戰經驗。

過程

指導教授與實驗室

本次進修很榮幸得以前往著明之美國波士頓哈佛醫學院。哈佛醫學院歷史悠久，是美國第三所醫學院，其研究與教學聞名世界。我們加入的部門「生物資訊研究所」，網羅生醫與資訊人才，並運用先進資訊技術研究生物醫學之難題。近年來的基礎醫學重要發展如「次世代基因定序」(Next-generation sequencing)，「多體學分析」(Multi-omics)等，皆受益與資訊人才的密切合作。有鑑於此，哈佛醫學院早在 2015 便已設立生物資訊研究所，以促進哈佛醫學院及其周邊醫院的眾多研究人員和機構緊密合作，以促進生物醫學領域的創新和發展。

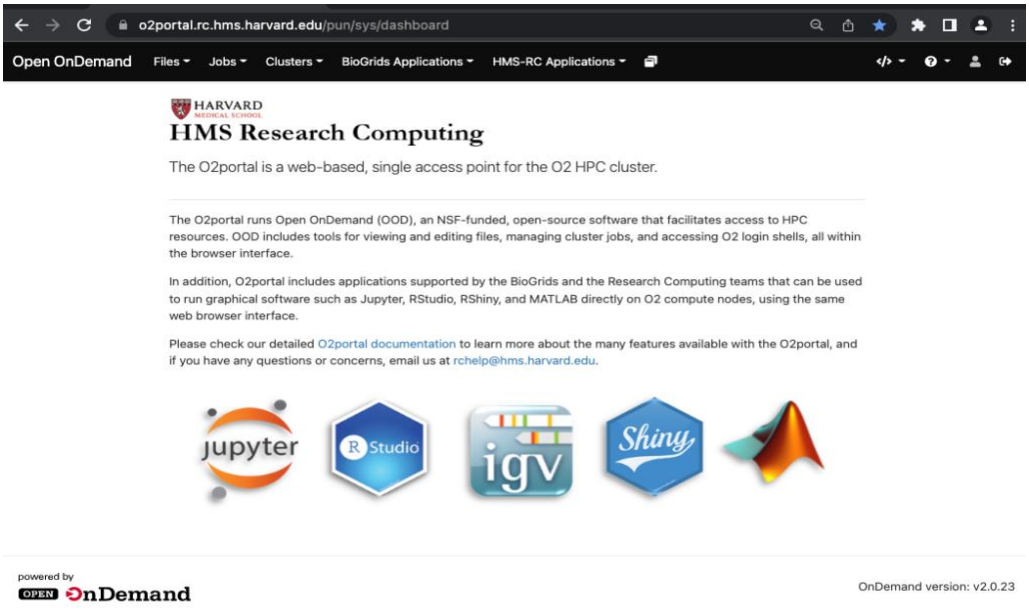
生物資訊研究所的研究領域廣泛，從罕見疾病，藥物開發，醫療系統，乃至於更廣泛的知識擷取與基礎演算法等，都是研究的範圍。本次進修研究的實驗室由余坤興 (Kun-Hsing Yu) 教授領導，則側重於數位病理分析，依此推斷疾病更多特性或預後等特性。實驗室過去的研究包括：由標準的 H&E 染色玻片推斷乳癌的賀爾蒙受體(ER)表現；從冷凍切片之玻片，直接預測腦瘤之分化程度；從標準 H&E 染色玻片推估癌細胞在多體學分析的表現等。

實驗室位置位在醫學院圖書館的上方。與傳統基礎醫學實驗室不同，生物資訊研究所的研究空間像是一個大辦公室，除了每個人有習慣的座位外，更規劃有會議室和討論空間。疫情之後視訊會議方興未艾，每層樓也都設計有適合視訊會議，宛如一人卡拉 OK 亭的一人隔音空間。

雲端上的實驗室

比起實體的空間，哈佛醫學院提供的「雲端」才是生物資訊研究所真正的研究進行的地方。哈佛醫學院架設有運算叢集，名為 O2；O2 由哈佛醫學院研究資訊部(Research Computing)所架設，提供所有醫學院研究人員使用。不論是基因定序、基因表現、分子動力學模擬、醫學影像分析等，都可以在 O2 運算叢集上進行。對於哈佛醫學院旗下的研究者而言，其儲存與運算不須額外支付任何費用。

O2 運算叢集同時也建置有具圖像運算單元(Graphics Processing Unit，GPU)之節點，供人工智慧相關研究使用。進行人工智慧的研究時，以目前訓練方式的特性，每批次(batch)若能看到越多資料，其訓練成效較佳；再者，近年興起的大型語言模型，已超過桌上型電腦 GPU 之記憶體大小。O2 運算叢集提供多張具有 48GB 視訊隨機存取記憶體(VRAM)之 RTX 8000 運算卡。此外，實驗室亦以本身資源購置數張具 80GB VRAM 之 A100 運算卡，僅供本實驗室使用，但仍透過 O2 運算叢集系統分配資源，以便達到最有效的利用。



圖一. 哈佛醫學院雲端頁面

訓練資料與困境

人工智慧模型的效能與訓練資料的質與量密不可分。早期的模型必須依靠人工標記，標記過程十分繁複且昂貴。對於人工智慧在醫療上應用而言，訓練資料的來源與品質更是一大瓶頸，因為正確的標記往往需要專科醫師的判讀。

全身骨骼掃描的研究亦碰上相同的挑戰。在本研究中，我們採用了去連結的兩萬多份全身骨骼掃描與其報告。全身骨骼掃描的報告為文字形式，但並無統一格式，不同醫師之間描述同一影像可能有很大的風格出入。相較於明確指出有病灶/無病灶之分類，可進行監督式學習的「強標記」資料，此等「弱標記」資料在訓練人工智慧模型時，往往需要百萬或千萬筆訓練樣本才有機會成功。緣此，人工智慧模型在沒有大量人工標注的問題上，如胸部 X 光或骨骼掃瞄，尚未取得成功。

初步嘗試

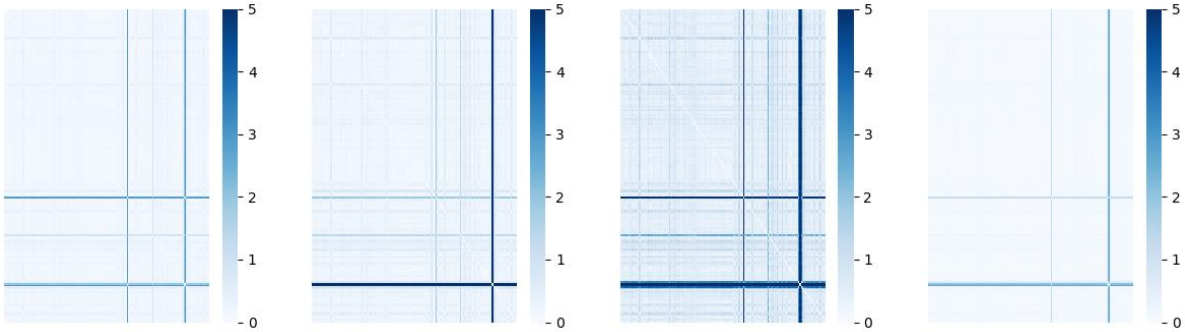
LLaVA 是 2023 年發表的一個多模態語言模型(multimodal language model)，具有同時處理語言與圖像的先進功能方法。LLaVA 與主要專注於文字或圖像之傳統模型不同，特點在於能夠無縫融合這兩個領域。這一特色使得模型能夠理解和解釋圖片與文字描述之間的複雜關係。

作為初步嘗試，我們試著使用微調(fine-tune)方式企圖讓 LLaVA 能從骨骼掃瞄產出文字報告內容。然而，對未曾看過骨骼掃瞄的模型而言，不同骨骼掃瞄的相異之處太小，以至於所有的骨骼掃瞄會產出相同的文字報告。是故，縱使大型語言模型的發展迅速，以目前情況而言仍無法用簡單微調方式應用於骨骼掃瞄之上。

無監督式學習

首先設法提高其圖片編碼器(Encoder)之性能。目前現有之多模態模型，其預訓練大多使用網路上的一般自然圖片。若要將模型應用於其他領域(如醫學影像)，則可以考慮近一步進行無監督式的預訓練。我們使用 Bootstrap Your Own Latent (BYOL)的方法，暫且略過圖片的報告，以骨骼掃瞄的圖片進行預訓練，以提高其對於骨骼掃瞄不同特徵的辨識度。

我們隨機選取了 200 筆影像，並計算其經編碼器萃取出特徵之幾何距離。下圖自左至右分別採用預訓練 4, 5, 6, 7 個世代的編碼器。可見預訓練在 4-6 世代中，對於區分出特徵的能力有所提升；然而，在第 7 世代後反而又下降。因此，我們採用預訓練 6 個世代的編碼器更進一步微調。



圖二. 經過 4, 5, 6, 7 個世代預訓練之編碼器相對表現

以語言模型輔助半自動標註

另一個提升模型表現的方向是用較小的模型初步進行病灶的位置識別，再交由大型語言模型總結。然而，此策略的難處在於缺乏訓練小模型的清楚標註。

所幸，大型語言模型的出現導致人工智慧的訓練方式有顯著的變化，這些大模

型的輸出可以用來作為小模型的輸入。如此一來，較小的模型能夠從大量的不完美標記資料中學習，減少了對人工標記的依賴；且這些訓練有效增加下游任務的表現和準確性。這一進展不僅提升了模型的效能，也為醫療領域的應用提供了新的機會。

我們採用大型語言模型從不定格式的報告中萃取其語意，進行半自動化的標註。成熟的大型語言模型可以辨識出報告背後的語意，並作適當的解析 (例如 Bilateral 相當於 Left + Right，或 L3-L5 相當於 L3, L4, L5 等)；這種理解語意的能力對於產出正確的半自動標註相當關鍵。在這一步中，我們也嘗試並評估了不同大小 (7B vs. 13B)，類別 (Mistral / LLaMA / Gemini / GPT-3.5 / GPT-4) 的模型在此任務上的表現。開源模型的表現雖然也不差，但仍略遜於大型的商業模型 (Gemini / GPT-4)。

以 ViT 模型進行病灶識別

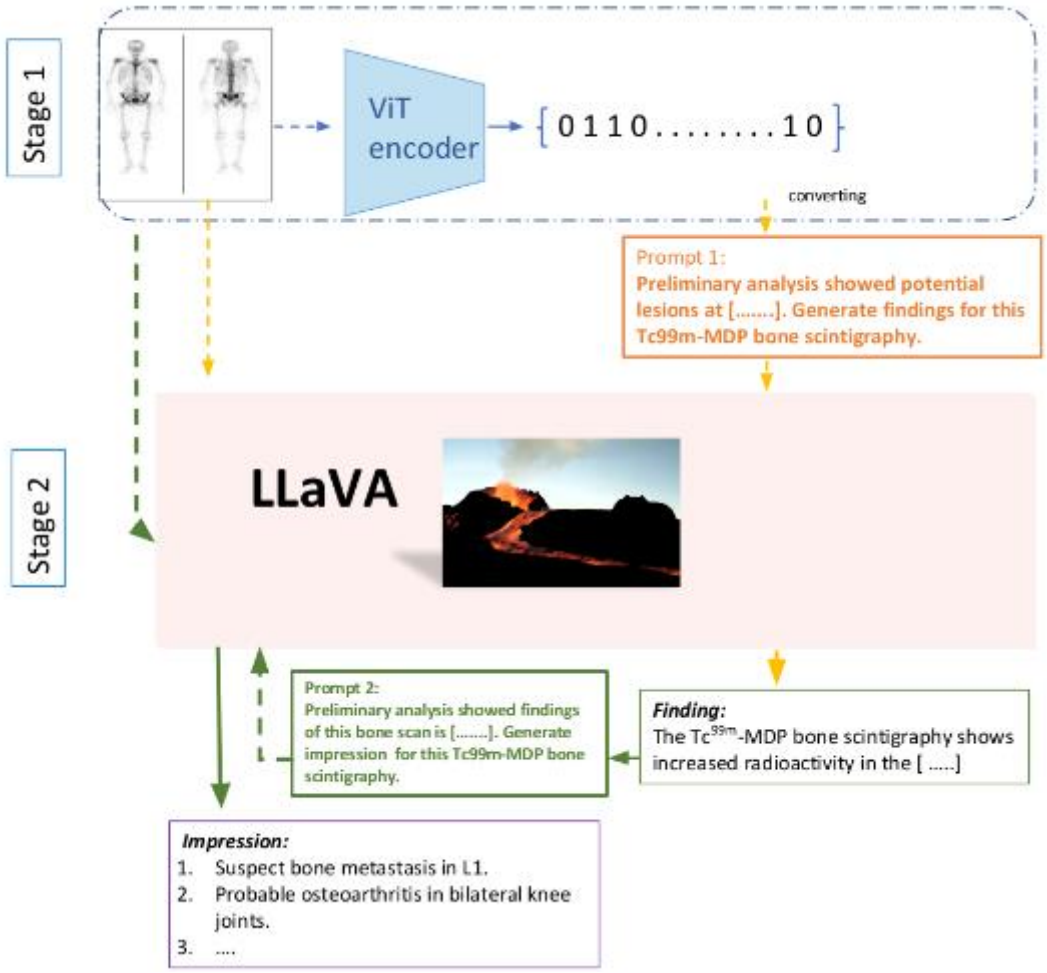
接著，我們將半自動標註轉換為適合訓練模型的標註：每份骨骼掃瞄檢查圖片，其正確答案標記為一串 91 個 0 或 1 數字，代表 91 個解剖位置是否有病灶。

類別中與類別之間有不等程度的類別不平均現象，對於模型的訓練有不良影響。首先，每個類別(即解剖位置)之中，無病灶的圖片其數量都大於有病灶的圖片。平均而言，僅有 15% 為有病灶，其餘 85% 皆為無病灶。因此，未經參數調整的模型訓練會傾向收斂至預測所有圖片皆無病灶。其次，不同類別(解剖位置)其具有病灶之比例亦不同，其範圍從最高者 25% 至最低者僅有 2%。

為了解決這個問題，我們調整了損失函數參數，個別設定每個類別的正樣本權

重。如此一來，數量較少的正樣本可以有效調整模型訓練的梯度，避免過度收斂至預測無病灶。除了 ViT 模型外，我們亦嘗試了其他的現代模型，包括 ConvNext / SwinTransformer 等，但其表現與 ViT 無明顯差異。

最後，我們使用訓練過後的 ViT 模型處理骨骼掃描圖片(圖三)，得到 196 維的輸出，並將之轉換為文字。此模型可以初步輸出圖片病灶的位置，作為下一步大模型的輸入。



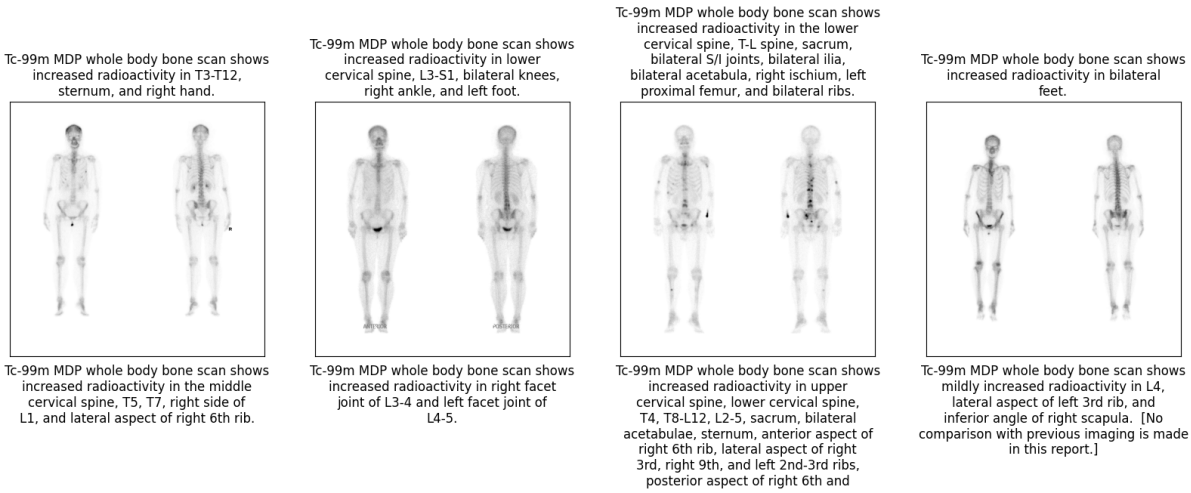
圖三. 本研究中結合小模型及大型多模態語言模型之骨架

微調 LLaVA 模型

有了初步病灶位置之後，我們最後一步再微調 LLaVA 模型 (圖三)，給予其從初步病灶位置產製出報告的能力。將為數眾多的病灶總結並化繁為簡，此一任務對於大型語言模型而言可駕輕就熟；另一方面，LLaVA 更可參考原骨骼掃瞄圖片，補上 ViT 小模型未成功定位出之病灶。

然而，大型多模態模型的微調仍在萌芽階段，相關技術仍陸續在開發中。以目前的情況，微調大型多模態模型並非簡單之事，光是載入模型與梯度就需要動用有 80 GB VRAM 的圖形運算單元，每個世代訓練更是需要 12 小時以上的時間。由於記憶體及時間的限制，因此不容易重複進行實驗來微調訓練超參數。

但另一方面，由於引進了大型語言模型，大幅度增加了任務的多樣性。舉例而言，我們可以輕易延伸模型處理針對影像的問答。憑藉模型本身所涵蓋的知識，未來甚至可以針對骨骼掃瞄結果提出可能的診斷，甚至建議的治療策略等。



圖三.多模態模型對於骨骼造影影像判讀輸出範例。
上方文字為預測結果，下方為原報告。

心得

近年來由於人工智慧迅速發展，在醫療方面的應用也如火如荼地進行。然而相對於其他領域，由於醫療的嚴謹性與對正確率的要求，人工智慧的應用更須廣受審慎檢視。我們一方面期待此項技術能夠減少醫護人員沉重的負擔，減低出錯率，也讓人員可以在降低繁瑣雜事的工作量後更有餘力專注於照護病患的身心靈；於此同時又需擔憂高科技的雙面刃，雖期望人工智慧能夠徹底翻覆醫療模式，但其能力是否真的可靠到足以取代過去數十載傳承的經驗，包含許多可能經由慘痛出錯代價學習而來的作業流程？在跳過漫長苦悶的工作訓練，預期將直接受惠於科技的年輕人員，是否有相對應能力駕馭這些科技，依舊秉持自身專業日常工作中的微枝末節，並在出錯時敏銳而從容的覺察並解決問題，是我們在全盤接受人工智慧帶來的便利前須思考的面向。

身處核子醫學科，我的臨床作業與電腦早已密不可分。從影像收取、影像重組分析到影像判讀，醫學影像電子化行之有年，在與人工智慧接軌上亦是比其他醫學應用更佔優勢。各種模型在醫學影像上的協助如火如荼，舉凡影像重組，病灶偵測，疾病判讀，影像優化等研究不計可數。有感於時代趨勢，我也進入成大醫學資訊研究所就讀，希望能結合醫學與資訊的能力做出貢獻。惟因資料量與儀器差異的限制，相較於放射影像，核醫可上市應用於臨床的模型尚待發展。這一次非常幸運有機會可以到舉世聞名的哈佛醫學院學習生醫資訊，從實際參與各種計畫中學習。在實驗室的研究人員不只是資訊工程的專業，更有來自數學、物理、生物等多專長領域。我深切感受到人工智慧在這個時代如何匯集多方領域的人才。我在實驗室中，因資訊工程能力尚難望他人項背，主要還是在提供醫學領域的知識協助，算是深刻體驗程式語言在這個世代的重要性。

在哈佛的這一年，另一項重要體認是人際間的連結在職涯中其實扮演很重要的角色。余老師的實驗室資源眾多，除了公開資料庫資料集，難能可貴之處更源於老師積極尋求合作的醫療院所，亦提供相當多寶貴的資料集進行合作。網路的發展讓世界無遠弗屆，一個人再怎麼優秀也不能再閉門造車，獨善其身。尤其在大數據的時代，多院所、多地區甚至多國家的資訊整合勢在必行。然而在積極加入智慧醫療聲浪下，醫療資訊的隱私與權益仍須確實被保障，才有其後的資源整合應用研究。

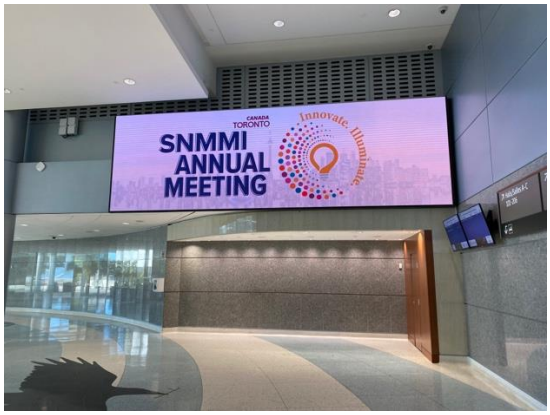
哈佛醫學院醫學圖書館

待一年的生物醫學資訊部門

位於圖書館四樓的實驗室

主持人余老師於年度會議分享研究成果

知名合作癌症醫院



將成果發表海報於多倫多舉辦之美國核醫年會

建議事項

1. 現今處於資訊發達，人工智慧廣泛運用於生活的時代，應用於醫療品質改善與輔助醫療端效率方面應與時俱進，唯受限於資料量以及資訊安全、病人隱私的問題，醫療資訊較無法有效整合。

2. 此次前往之哈佛醫學院生物醫學資訊實驗室，該實驗室著重數位病理影像的分析，取用資料包括美國癌症基因體資料庫(TCGA)及相關合作醫療院所，然其資料組成以高加索人種為大宗。國內專家學者所建置之台灣人體生物資料庫亦有相同目標，適合發展應亞洲人的診斷治療計畫；然而 TCGA 僅需線上申請即可下載公開資料且完全免費，相較之下台灣人體生物資料庫之數位資料仍需正式申請案並收費。

3. 因工作屬性需久坐於電腦前，對於眼睛以及脊椎皆有負擔。此次前往哈佛醫學院生物資訊研究室的實驗室，發現已廣設升降桌，可配合調整符合研究人員的需求，值得國人設計辦公室時納入考量。

4. 美國生活消費水平高，此行雖有補助款項，但須回國之後才能申請，緩不濟急，建議應考慮先撥予部分費用，以增加年輕學者出國的意願。