出國報告（出國類別：研習）

赴全球應用流行病學訓練網路全球研討會參加 R 語言疫情分析應用工作坊研習

服務機關：衛生福利部疾病管制署
姓名職稱：鄭皓元 防疫醫師
派赴國家/地區：美國/亞特蘭大
出國期間：108/10/26-108/11/04
報告日期：108/12/31

摘要

RECON 組織為一跨國國際非營利組織，由倫敦帝國理工學院，衛生與熱帶醫學院之公衛專家組成，組成緣由主要為 2014 年伊波拉疫情爆發時，公衛專家有感於缺乏一個整合具彈性可讓疫情現場利用的資訊工作流程，藉以提供後續疫情基本分析與模型建置使用，故聯合不同領域專家成立 RECON 組織開發 R 語言套件並推廣使用。本次工作坊以實際疫情之資料為例，教授如何利用 R 語言收集疫情及疫調資料，進行資料分析，傳染病模型推估與結果呈現，以協助疫情應變。疫情中心近年來積極推動使用 R 語言進行自動化疫情分析與報表呈現，故希望利用此次研習機會，進一步學習 R 語言在傳染病疫情疫情爆發可應用之分析，同時與國外專家交流實務經驗。

目次

工作報告

出國行程表

日期	工作日誌	地　　點	行　程　內　容
108/10/26	**啟程-抵達**	**台北→亞特蘭大**	**路程-抵達**
108/10/27-29	**研習**	**亞特蘭大**	**RECON R 語言疫情資料分析工作坊**
108/10/29-30	路程	亞特蘭大-薩爾茲堡	路程
108/10/30-11/3	研討會	薩爾茲堡	參加薩爾茲堡全球研討會(另由主辦單位補助)
11/3-4	返程	薩爾茲堡→台北	路程

緣起與目的

R Epidemics Consortium (RECON)組織為一跨國非營利組織，主要是由來自倫敦帝國理工學院，倫敦衛生與熱帶醫學院及其他歐洲各地公衛及傳染病模型專家組成，成員亦包含熟稔程式語言撰寫之軟體工程師。該組織之成立主要是有感於近幾年來持續有新興或再浮現傳染病於非洲或其他開發中國家爆發嚴重疫情（如 2014 年西非伊波拉疫情爆發），然而在第一線的疫情應變時，從資料收集，整理，清理，到後續的分析，往往受限於第一線人員的訓練，以及資訊基礎建設的程度而破碎難以整合。缺乏良好資料可以提供分析，也讓傳染病模型專家難以利用這些疫情資料建置模型評估，提供第一線的資訊。

因此，該組織希望能夠利用 R 語言的開源，跨平台且易擴充的特性，開發符合第一線公衛人員資料分析需求的套件，同時間將傳染病疫情分析（Outbreak analysis）的流程模組化，讓公衛人員能夠在最短的時間上手進行資料分析，甚至是進一步帶入基本之傳染病模型，藉此評估疫情的傳播速度，後續走勢及疫情規模，以供疫情應變參考。也因為這樣，該組織一方面除了開發傳染病資料分析的 R 語言套件之外，亦積極在各傳染病相關研討會舉辦 R 語言工作坊，提供手把手的小組教學，讓即使之前沒有 R 語言學習經驗的人也可夠快速入門。另一方面，該組織亦撰寫大量實用教案，以真實疫情資料為例，用 case study 的方式引導學員一步步完成資料的載入與分析，這些也都公開在該組織經營之

教育網站上可供自學或複習。

此次趁著全球應用流行病學訓練網絡（TEPHINET）全球大會在亞特蘭大召開的機會，RECON 亦在研討會前舉辦兩天之 R 語言工作坊，以實際的食品中毒調查案疫情資料為例子，教授如何利用 R 語言收集疫情及疫調資料，進行資料分析，傳染病模型推估與結果呈現，以協助疫情應變。由於疫情中心近年來積極推動使用 R 語言進行自動化疫情分析與報表呈現，利用此次研習機會，希望能夠進一步學習 R 語言在傳染病疫情爆發時可應用之分析，同時與國外專家交流實務經驗。

研習過程

本次工作坊主要分為兩天，課程安排如下表：

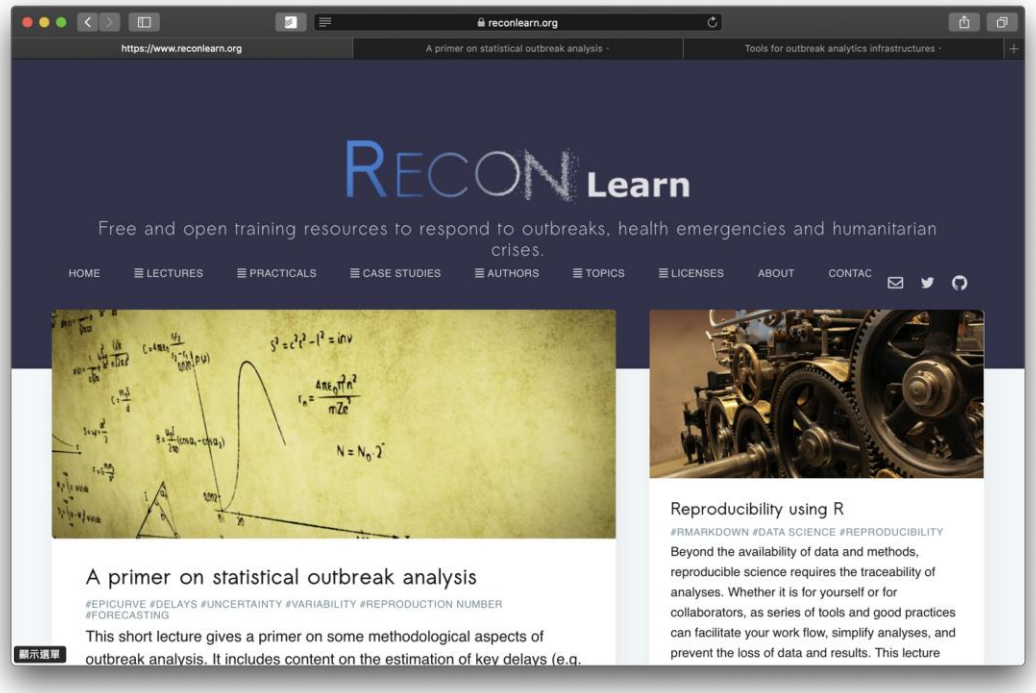
Day 1, Sunday, 27 October 2019			
Time start	**Time end**	**Duration**	**Subject**
08:30	09:30	01:00	Registration and coffee + installing R/deployer
09:30	09:45	00:15	Welcome to the course, announcements
09:45	10:00	00:15	Introduction into RECON and the course
10:00	10:30	00:30	Quick introduction to R
10:30	10:45	00:15	Setting expectations for day 1
10:45	11:15	00:30	Coffee break
11:15	12:45	01:30	Introduction to RStudio
12:45	13:45	01:00	Lunch break
13:45	14:00	00:15	Introducing the Stegen case study
14:00	14:30	00:30	Importing data into R
14:30	15:00	00:30	Stegen: overview and summaries
15:00	15:25	00:25	Coffee break
15:25	15:50	00:25	Stegen: summarising and data exploration
15:50	16:50	01:00	Stegen: graphical exploration
16:50	17:00	00:10	wrap-up, outlook to day 2
Day 2, Monday 28 October 2019			
Time start	**Time end**	**Duration**	**Subject**
09:00	09:10	00:10	Recap day 1, outlook for day 2
09:10	09:20	00:10	Setting expectations for day 2
09:20	10:20	01:00	Good practices for data science/reproducibility
10:20	10:45	00:25	Coffee break
10:45	12:10	01:25	Stegen: descriptive analysis + functions
12:10	13:10	01:00	Lunch break
13:10	15:00	01:50	Stegen: spatial analysis + maps
15:00	15:20	00:20	Case study wrap-up
15:20	15:50	00:30	Coffee break
15:50	16:20	00:30	Reproducibility in Rmarkdown (R4Epis)
16:20	16:35	00:15	RECON packages in the field
16:35	17:00	00:25	Closing lecture + evaluation

第一天課程

第一天上午主要為 R 語言入門，早上由講者先簡單分享 RECON 組織的由來和理念後，便開始進入安裝程式，實際操作和疑難排解。由於 R 語言雖然是個開源語言，且是為了統計需求而特別設計，是許多統計學家分析資料時愛用的語言，但實際入門的門檻對完全沒接觸程式語言的使用者來說仍是略高。因此課程設計上花了不少時間，還是從頭仔細解釋了整個 R 語言的語言邏輯，基本與法，包括基本的資料結構，流程控制，迴圈和函數的撰寫與使用。

雖然對於 R 語言的資料科學分析應用，已經有很多現成的套件如 dplyr 可以很輕鬆的學習使用，大幅降低入門者的學習門檻，學習曲線不至於太過陡峭，不過基本的資料結構和語法邏輯，還是需要一定程度的熟悉，才不會在程式出錯，需要除錯時容易卡關。

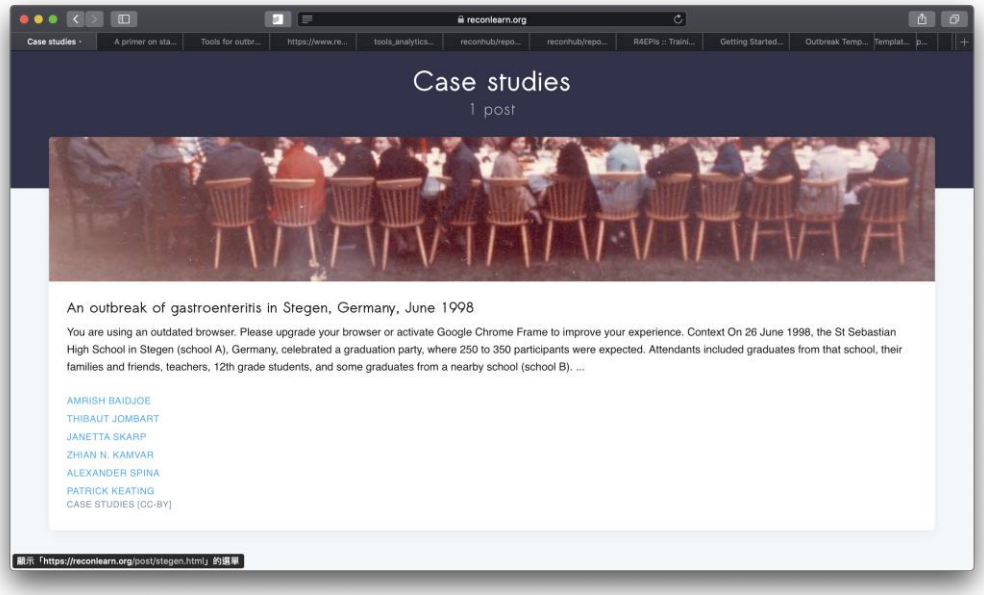
此次工作坊的特色就是所有的課程資料都是使用網路上公開的開源資料，是該組織事先撰寫的，所有課程資料也都公開在自家的網站上可供學習：RECON Learn (https://www.reconlearn.org)。從資料的下載，匯入開始，網頁上會有一步步的詳細解說，就算只是在家自學，也可以直接閱讀上頭的解說，一步步完成教案中提供的課題。即使一開始對程式碼不熟，沒辦法直接自己撰寫解題，也可以直接對照後頭提供的範例程式碼自己輸入測試，其實是有點像是翻轉教室的設計概念。而當天工作坊就是師父領進門，帶你設好環境設定之後，便一個段落一個段落的自己對照學習，卡關時助教會隨時在旁邊幫忙解惑。

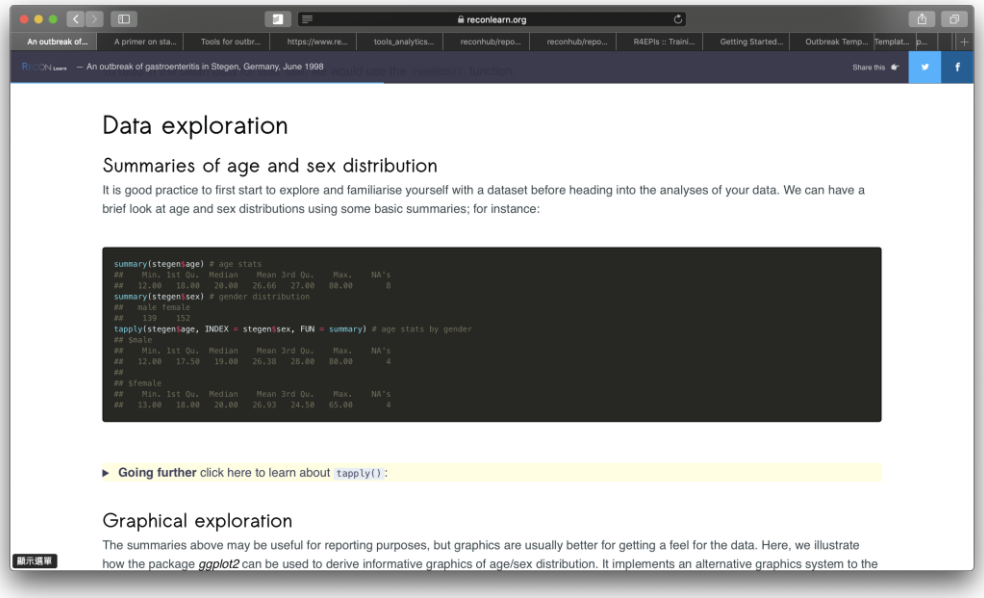


這種學習方式對於程式語言撰寫的學習其實是最有效率的一種學習方式，因為你必須自己思考怎麼利用剛學到的程式語法去拼湊出答案，同時 R studio 的環境可以讓你馬上看到撰寫好的程式碼的執行結果，所以可以馬上獲得回饋，跟疫情中心之前舉辦 R 語言和 Python 語言課程時老師課堂上的教學方式其實也十分類似。這樣的學習方式好處是，一來這樣的肌肉記憶學習方式可以讓大家印象更加深刻，同時完成的課題都是工作上可能會實際遇到的問題，所以可以馬上套用在工作的經驗中，二來是完成每個小段落都可以獲得回饋的成就感，也會讓人更樂於繼續學習。

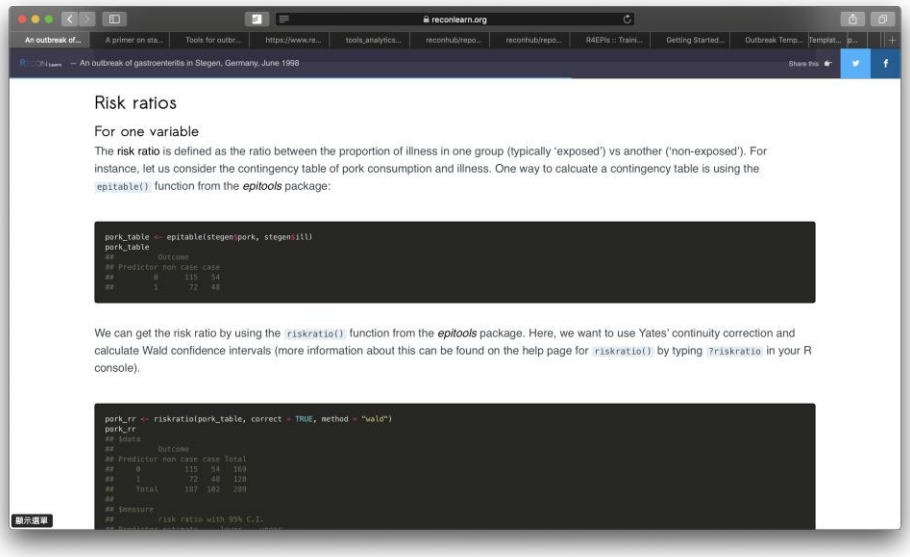
課程的另一個特色則是活用了 R 語言跨平台的特色，程式碼的撰寫皆以可以適應不同電腦環境、彈性可攜為核心理念，有注意到將程式碼攜帶到不同電腦時可能會有的問題，以不管在哪裡都可以順利執行為理念撰寫。這樣的特色主要是為了適應疫情應變現場工作可能會有的各種突發狀況，以免有類似換了電腦就完全無法執行，需要花上大量時間重寫或除蟲的狀況出現。同樣的理念也可見於之後該組成員的分享以及 R4EPIS 專案的開發。

第一天的課程是使用在德國發生的一次食品中毒事件的資料做教材，上午即是讓大家實際動手嘗試如何使用 R 語言進行從資料匯入，資料清理到基本敘述統計的每個過程。下午即接著帶到如何使用清理好的資料繪製疫情曲線圖，製作表格。

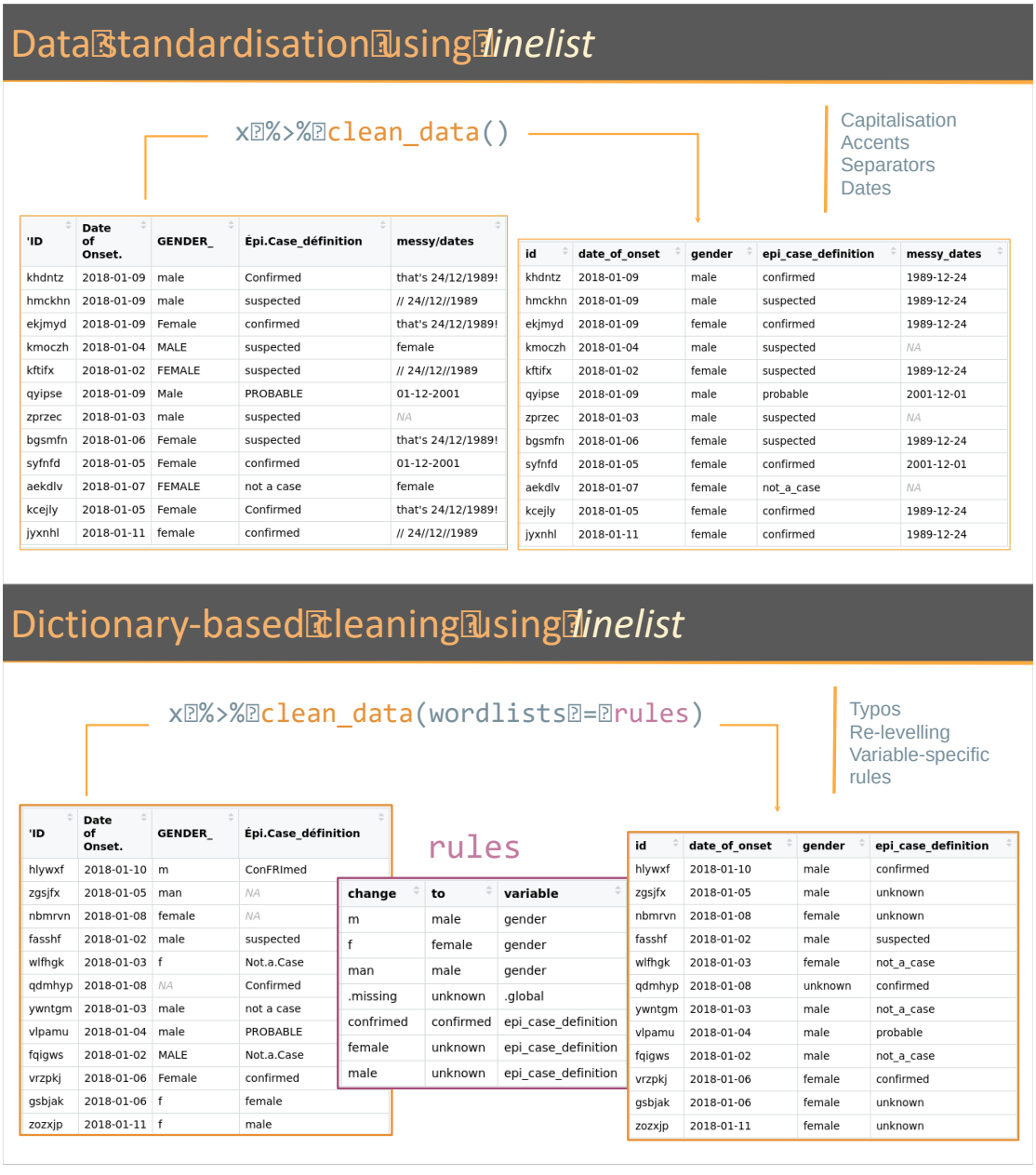




有了第一天的經驗，第二天上午便進入比較進階的疫情資料分析。計算每個風險食物的勝算比（Odds Ratio），找出最可能的風險食物。其實整個教案的設計就是跟著疫調的腳步在走，一步步教你怎麼用 R 完成在疫調的過程中會想要看到的各種資訊（例如病例數統計，疫情曲線圖，敘述統計表格，2 勝算比計算跟比較），對第一線的疫調人員來說應該是非常熟悉的工作流程，只是這次我們可以用 R 輕易地完成。相較於 FETP（應用流行病學訓練，或流行病學調查班）的課程常使用的 Epi-Info 此類的軟體，使用 R 語言不僅在客製化一些報表，數據分析或製圖上較為彈性，不限於 windows 系統以及相容環境等限制的特色，也讓這套流程的泛用性高上許多。在第二天上午課程的結束，同時也帶入了 GIS 資訊處理的部份，初步介紹了如何使用病例的地理位置資訊在 R 上繪製病例分佈地圖，也算是展示了 R 語言進階使用的可行性。

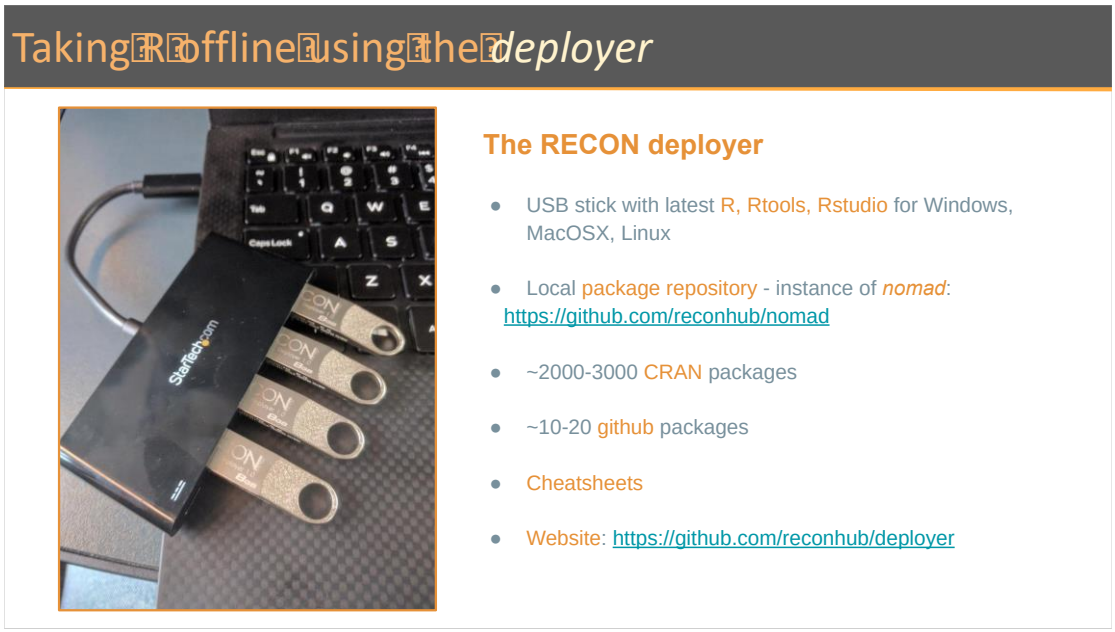


第二天則是最有趣的部分，安排了幾個 RECON 組織成員的分享。第一個分享是請實際到剛果支援 2018-19 年伊波拉疫情的成員，來分享些設計好的套件和工作流程，是怎麼樣實際的在現場應用。講者分享了許多在第一線推廣資料分析時可能會遇到的困境：資料的格式不一，沒有統一的資料分享平台和結構，缺乏一致的分析流程，並展示了他們為了這些難題開發的套件和解決方式。例為了可以快速清理病例清單資料的套件，他們開發了 linelist 套件，可以用比較直觀的語法，快速的將格式雜亂的病例清單資料清理成統一格式。



同時考慮到在疫情現場常常是需要一直更換電腦使用的不穩定工作環境，他們也設計了可攜式的資料處理 USB 隨身碟，裡頭有預載好的免安裝版的 R 語言程式以及所有相關套件（Windows 或 Mac 甚至 Linux 版本都有）和教學文件，並

做好可攜式的環境設定，所以要出任務時只要帶上一個這樣的隨身碟，在現場只要找到任意一台電腦就可以隨時開工進行資料處理。真正是做到彈性可攜又可無縫接軌的工作流程。



第二個則是 R4EPIS 計畫的展示，這個計畫主要是為了疫情應變時，常常會需要定時或不定時的產出疫調資料報告，提供決策者或相關單位參考。即便是以本署的資訊能力而言，為了維護這樣的疫情資料或疫調報告，往往也需要耗費莫大的人力和心力在上頭進行資料的更新確認與撰寫。所以利用 R 語言中的 shiny 套件，R4EPIS 專案希望能夠將資料分析，圖表產製到最後輸出報告文件檔的整個流程都統一後做到半自動化，大幅減少第一線人員花在維護和輸出這些疫情資料報告的時間和人力。這些套件的開發和使用的教學一樣都是直接開源公開在相關的網頁上，可以讓有興趣的人自行學習應用。

最後課程結束時，由於參加工作坊的成員多半都是各國的 FETP 指導員或學員，RECON 組織的副負責人 Amrish 除了歡迎大家繼續參與 RECON 的工作，例如教案撰寫或套件開發之外，也表示其實他們也很歡迎大家實際來試用或應用這些套件在第一線的防疫工作上，並給予他們更多的回饋意見，這樣他們也能夠根據這些寶貴的使用者經驗，來改善這些套件，讓整個套件和利用 R 語言進行疫情資料分析的使用變得更為簡單便利，真正做到減輕第一線工作人員的負擔。

心得與建議

由於 SAS 等統計軟體收費昂貴又更新緩慢，在新一代的資料科學領域，R 或 Python 此類開源免費，跨平台又有許多熱心使用者開發套件的程式語言已成為主流。然而先前由於大部分應用仍停留在其他領域如數學統計，商業資料分析等等，在疫情資料分析的領域，能有 RECON 這樣兼具應用流行病學和軟體開發背景的組織願意投入套件的開發，造福第一線的使用者，其實是防疫領域的一大福音。

疫情中心近年來亦陸陸續續投資不少資源，希望能夠把重心轉移到 R/Python 的應用，如本署現正使用之晨會疫情面板，即為將資料處理自動化後產出視覺化報表的代表作之一。如何結合像 RECON 組織已經投入的資料，利用他們開發的套件進一步提升並強化及簡化我們的疫情分析流程，甚至做到像是 R4EPIS 專案那樣的目標，半自動化的產出基本的疫情文件和報表，也是可以努力的目標之一。這些實際應用的經驗也都是該組織十分需要的寶貴經驗，屆時都是我們與外部專家交流時的珍貴資產。

建議

1. 後續可利用 RECON 組織提供之疫情教案，作為資料分析訓練參考。
2. 參考 RECON 組織開發套件經驗，持續改善疫情資料分析流程。