具隱私保護暨安全資料探勘之醫療資料倉儲系統[1]

謝承翰*、范俊逸**

壹、前言

伴隨著醫療逐漸朝向全面數位化，電子病歷詳細記載病患的各類病徵細節及醫師診斷，然而各項數據都是個人極為隱私的資訊，因此醫療數據的安全性常被放大檢視。近年來 AI 技術盛行，在各領域嶄露頭角，甚至在癌症檢測的準確率已經超越人類的判讀（Martin Stumpe & Lily Peng, 2017），透過分析大量數據，找出過往未被發現的潛在因子，以加速各種病症檢測速度與準確度。目前世界大型 AI 模型所使用的訓練數據集大小高達數百兆位元組，如同擁有 1,750 億個參數的 GPT-3（Generative Pre-trained Transformer 3）自然語言模型（李建興，2022）。建立醫療模型目的是為找出真正的疾病因子，且該醫療模型要求必須是高準確性並提供可靠的結果，大量訓練數據就成為建構強健醫療模型的重要基礎。因此，每一筆訓練數據對於醫療模型都是非常珍貴且必要的，然而醫療數據的使用牽涉的隱私法規層面廣泛，如未妥善進行

[1]「具隱私保護暨安全資料探勘之醫療資料倉儲系統」是由本文作者群隸屬團隊—國立中山大學資訊工程學系執行行政院國家科學及技術委員會「110 年前瞻資安科技專案計畫」之成果，並將同時於 2022 年臺灣資安大會（CYBERSEC 2022）展示。

* 國立中山大學資訊工程學系博士生。
**國立中山大學資訊工程學系教授兼資訊安全研究中心主任。

數據處理，數據使用者將面臨嚴重的裁罰或是病患無止盡的控訴。雖然個人的隱私保護觀念日漸被重視，但仍導致多數珍貴的醫療數據難以開發或被應用。而近年來許多團體或國家開始探討個人隱私的議題，甚至頒布與之相關的法令及規範，如歐盟於 2016 年頒布目前世界上最為嚴格的隱私保護規範，名為一般資料保護規範（General Data Protection Regulation，GDPR），不再是固定裁罰金額，而是以該企業全球營業額的百分比進行計算，然而如此嚴苛的懲罰仍發生如資料庫洩漏等事件，導致用戶資料外洩的新聞層出不窮，如果用戶涉及國家特定人員，更可能導致國安或是外交危機。因此，本團隊選定醫療數據的倉儲系統為基礎，結合多種加密技術，保證醫療數據產生後的儲存、取用、操作及運算皆以密文形式進行，設計一個兼顧病患隱私與醫療數據應用的具隱私保護暨安全資料探勘之醫療資料倉儲系統。

本文將從基礎的加密技術開始介紹，且對比使用單純加密機制來保護數據之作法，本團隊運用更為進階與不同特性的加密技術，使得數據經加密後仍可進行存取控制、搜尋及運算，改變以往加密後就難以操作數據的僵化思維。雲環境不僅提供大量的儲存空間，更提供可隨需求調整的運算能力，在上雲需求日漸增加的趨勢下，仍須兼顧個人隱私的保護，同時這也是本系統設計初衷與最終目標。

貳、 加解密技術

一、密碼機制概述

加密技術已廣泛應用於各種領域，包含銀行提款卡及信用卡，或是現代人每天都會使用的線上服務，如網頁瀏覽、影音串流服務、社群或雲硬碟等，只要是與個人數位資產或是隱私相關的數據，皆會使用到加密技術來保護數據在靜態儲存（數據存放於硬碟中）及傳輸的安全性。而加密技術主要可以分為兩大種類，分別為對稱式加密（Symmetric Encryption）與非對稱式加密（Asymmetric Encryption）。

進階加密標準（Advanced Encryption Standard，AES）是目前主流使用的加密方法之一，被歸類為「對稱式」加密意味其加密機制僅包含一把金鑰，並以此進行訊息加密與解密。「加密」亦即把訊息轉換為對於第三者無法分析出原始訊息含義的訊息，該訊息即被稱為「密文」，而解密則使用同一把金鑰即可將密文還原為原始訊息。圖 1 顯示對稱式加密的流程，Alice 想要傳送一個檔案給 Cindy，使用對稱式加密前，需事先共享同一把金鑰，Alice 利用共享金鑰加密檔案後，傳送給 Cindy，而 Cindy 可使用相同共享金鑰以解開加密檔案，最後得到原始檔案。加解密運算快速是對稱式加密的優點，然而事先需要共享同一把金鑰是其主要問

題，因為此把共享金鑰是加解密檔案之關鍵，不可能直接公開傳送。非對稱式加密則可解決此問題，然而效率將大幅降低。因此透過對稱與非對稱兩項技術搭配即可同時解決上述兩大問題，這也是日常使用的服務（如網頁瀏覽網站、串流服務等）最常採取的方法之一。

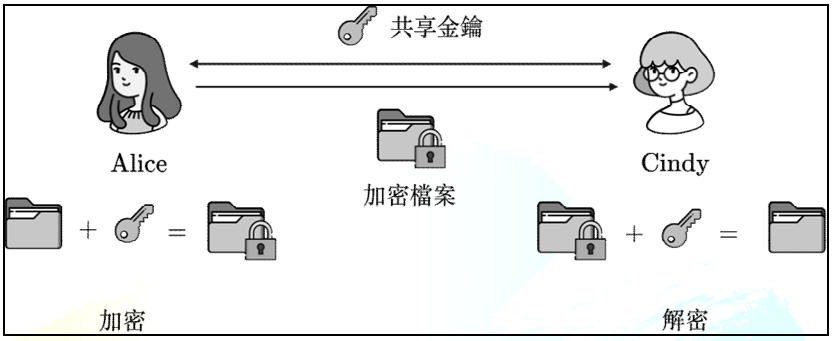

圖 1　對稱式加密

另一種加密種類則為「非對稱式加密」（又可稱為 Public-Key Encryption 公鑰加密系統），常見的技術如由 Ron Rivest、Adi Shamir、Leonard Adleman 三位學者於 1977 年共同提出的 RSA，為目前最常使用的非對稱加密技術，是基於對極大整數做質因數分解的難度來確保其安全性。非對稱即意味此加密法擁有一對金鑰，包含公開金鑰（簡稱公鑰）與私密金鑰（簡稱私鑰）組成之金鑰對（Key Pair），此金鑰對是互相搭配且隨之生成，通常金鑰對的擁有者會公開分享公鑰。如圖 2 所示，假設今天 Alice 需要加密訊息給 Cindy 時，Alice 就可以利用 Cindy 分享的公鑰進行加密，Cindy 收到加密的訊息後則利用自身的私鑰進行解密。因

此，可以從對稱式加密與非對稱式加密的特性發現，在使用對稱式加密時，需要雙方先溝通好共享金鑰才可進行訊息加解密，而非對稱式加密則可直接公布公鑰，只有私鑰的擁有者可以解密訊息。

然而非對稱式加密之效率較差，因此常會運用兩種互相搭配使用，即利用非對稱加密保護對稱式加密的金鑰，以此方式交換對稱式的共享金鑰，後續就可以利用此把共享金鑰進行訊息的加解密。非對稱式加密解決對稱式加密需事先秘密協議好一把共同金鑰的問題，接著使用對稱式加密避免需持續使用非對稱式加密保護訊息而導致犧牲效率。

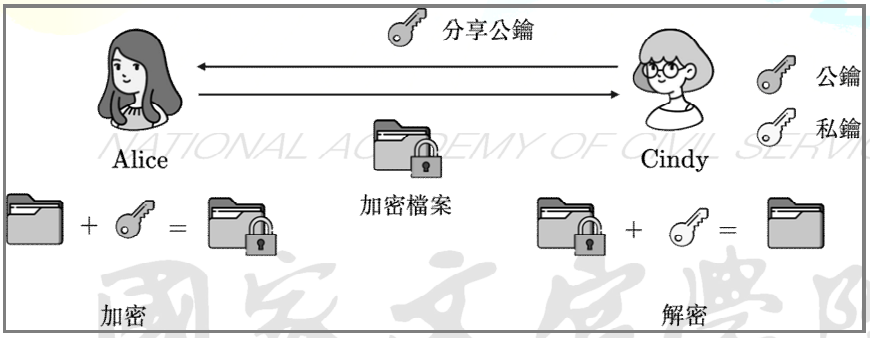

圖 2　非對稱式加密

二、擁有特殊屬性之加密機制

加密技術不只保護訊息傳輸的安全性，包含電子公文使用的數位簽章或是可以為檔案建立數位指紋的雜湊函式，都是加密技術的延伸應用。然而本研究主軸之一在加密訊息上，因此，本文僅談及加密技術如

何保護你我的數據。另一個主軸則為使用擁有特殊屬性的加密機制，其加密技術已經不僅單純加密訊息，相關之數學性質也讓研究人員設計出許多擁有不同特性的加密技術，如基於身分之加密機制（Identity-Based Encryption，IBE）、屬性加密機制（Attribute-Based Encryption，ABE）、可搜尋式加密機制（Searchable Encryption，SE）與同態加密（Homomorphic Encryption，HE）等，也因為擁有許多不一樣的特性，讓許多新型態應用或是構想在確保數據安全的同時得以實現。

非對稱加密中金鑰對如表 1 所示，非對稱式演算法 RSA 的安全性是植基於數學難題與金鑰長度，且每個人皆可自行生成獨一無二的金鑰對。當加密時，僅需提供自身公鑰給欲傳送訊息的一方，如表 1 右邊表格所示，對方利用公鑰即可進行訊息的加密。

以目前最常見的實際應用來說明，公鑰即可代表一個人的身分，因為擁有對應私鑰才能進行解密，只要有第三方公正單位記錄公鑰與身分的對應關係，就可以此來建立相對應的系統，如 MOICA 內政部憑證管理中心所發行的自然人憑證，使用公鑰系統與簽章的機制，提供如電子公文簽章的服務，並保證利用私鑰簽署完成後之公文具不可否認性，因為公鑰已由內政部記錄，並對應國民之身分，只要利用公鑰即可針對簽署完的電子公文進行驗證。另外，自然人憑證卡中記錄如表 1 的非對稱式加解密金鑰對，免除每個人需要背誦對應金鑰對的困擾，僅需插入卡

片，電腦即可自動帶入該私鑰進行簽章作業，其中在進行簽章需利用到私鑰，而所有解密或簽章操作運行皆在憑證卡晶片上，其中晶片提供基於硬體的安全性，包含只要經過外部量測或監聽，晶片會啟動消除內部數據的機制，也因此外部難以取得內部私鑰，以此確保每個人私鑰的安全性。許多服務使用非對稱加密，但憑證卡與相關設施建置成本高，加密前使用者也需透過一些方式之公鑰與對應擁有者身分進行核對。目前線上服務多採用帳號密碼驗證用戶，以確認該公鑰對應之身分。而基於身分之加密與屬性加密可取代難以記憶的字串，取而代之的是身分代號及屬性。以「屬性」加密為例進行說明，尤其屬性加密也是本系統的核心機制之一。

表 1　RSA 金鑰對

-----BEGIN RSA PRIVATE KEY----- MIICWgIBAAKBgG1eTl2AJu62PTfLb2Qown0ZR13AgtENG+jjYZy5jAUUbJY8xYYM PwEO3VxFcK9fHdLRQU7p4TVhSGjPWkl5n26QhsxUALmm…（省略） -----END RSA PRIVATE KEY-----	-----BEGIN PUBLIC KEY----- MIGeMA0GCSqGSIb3DQEBAQUAA4GMADCBiAKBgG1eTl2AJu62PTfLb2Qown0ZR13A gtENG+jjYZy5jAUUbJY8xYYMPwEO3VxFcK9fH…（省略） -----END PUBLIC KEY-----

三、屬性加密（Attribute-Based Encryption，ABE）

屬性加密同樣是一種非對稱式加密，相較 RSA 這種一般的公鑰加密系統，除數學難題不同外，多了「屬性」賦予非對稱加密系統進行一對

多的加解密功能，圖 3 即呈現出一對多的特性。首先有個公正機構稱為金鑰授權機構，配發每名使用者包含其屬性的私鑰，數據擁有者在加密數據前，加入允許「父母親」與「醫生」存取該檔案的存取結構，意味著訊息加密後，「父母親」與「醫生」符合數據擁有者制定的存取結構，利用「父母親」與「醫生」自身的私鑰即可進行密文的解密，這也是與傳統加密與解密者通常是一對一的不同點。另外，屬性加密又可分為 Key-Policy ABE（KP-ABE）與 Ciphertext-Policy ABE（CP-ABE）兩種。KP-ABE 又名為金鑰策略之屬性加密機制，其存取結構則綁定於私鑰中，檔案綁定屬性集合，如圖 4 所示，左方檔案綁定「父母親」的屬性集合，這裡假設存取結構是以「OR」為主，意味著有「父母親」或「醫師」的屬性集合檔案都可以被其解密。而 CP-ABE 則相反，以目前主流的檔案管理系統為例，現階段主要是以單一檔案進行人員的存取管理。CP-ABE 把存取結構綁定在檔案上，如圖 4 所示，CP-ABE 的範例顯示檔案的存取結構為「父母親」或「醫生」。因此，擁有「父母親」屬性的使用者就擁有權限可以解開其文件。本團隊將以 CP-ABE 來建構系統。

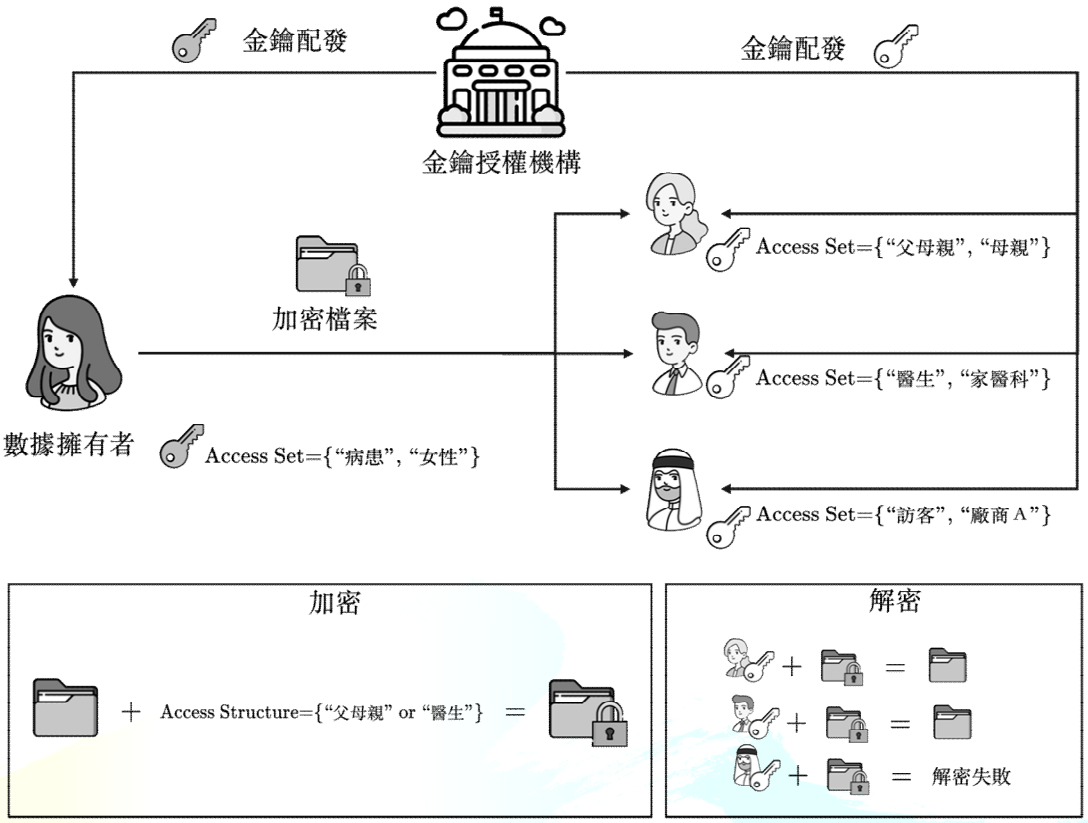

圖 3　屬性加密

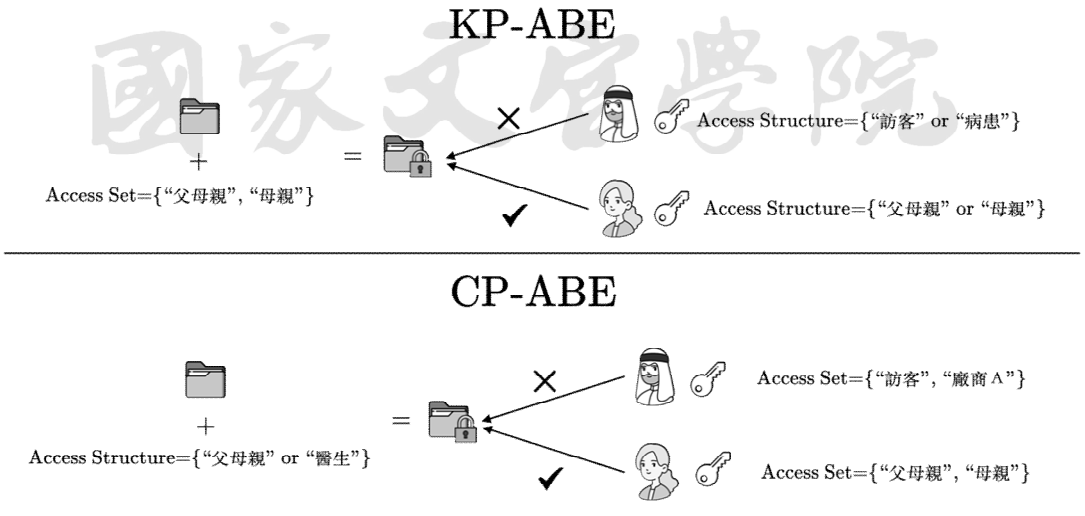

圖 4　KP-ABE 與 CP-ABE

四、可搜尋式加密（Searchable Encryption，SE）

明文與相對應的金鑰輸入至加密演算法後即可得到對應密文，只要沒有對應之解密金鑰則無法解回原始訊息，此為加密的核心概念。可搜尋式加密的出現允許授權使用者可以針對加密數據進行更多的操作，在資料庫中，除提供儲存功能外，搜尋功能也是必備的，而可搜尋式加密則提供在密文下的關鍵字搜尋。如圖 5 所示，Alice 利用 Cindy 的公鑰將數據與關鍵字「Covid」進行加密，隨之上傳至雲。Cindy 需取用數據時，可以透過自己的私鑰與有興趣的關鍵字生成搜尋權杖（Search Token），而雲伺服器會依序比對雲上的每個檔案，如符合關鍵字則會回傳相對密文，Cindy 即可利用私鑰進行解密，且公有雲服務提供商在搜尋過程中無法得知訊息及關鍵字內容，可搜尋式加密提供一個即使檔案加密後仍可以進行搜尋的一個解決方案。

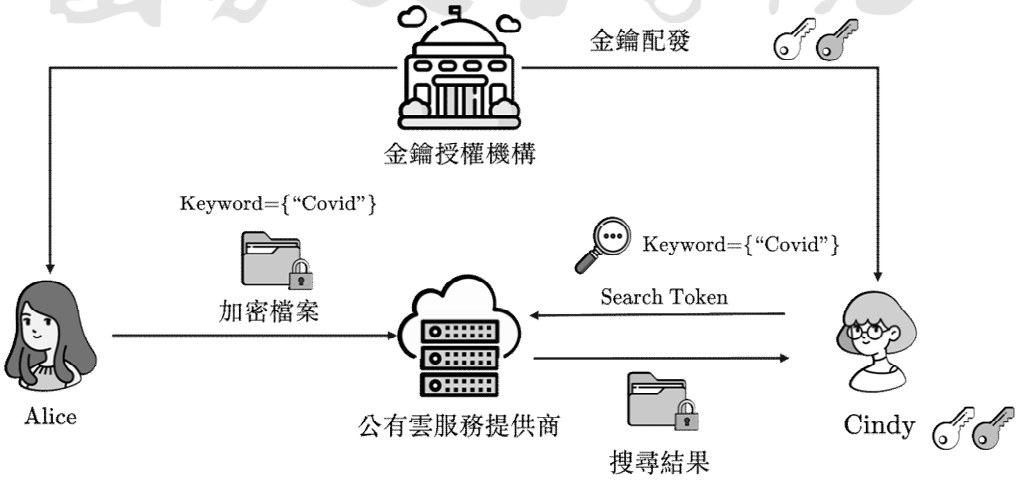

圖 5　可搜尋式加密

五、具關鍵字搜尋之屬性加密機制（Attribute-Based Encryption with Keyword Search）

可搜尋式加密最初的版本同樣以公鑰加密系統（Public-Key Encryption，PKE）為基礎。近年來，一些非對稱式加密系統也擁有關鍵字搜尋的特性，如 Q. Zheng, S. Xu, G. Ateniese 等人提出的 Verifiable Attribute-Based Keyword Search（VABKS）（Zheng, Qingji, Shouhuai Xu, and Giuseppe Ateniese, 2014），其基於屬性加密並附屬關鍵字搜尋的特性，透過屬性加密存取控制的特點，使符合權限的授權使用者以其私鑰與對應關鍵字即可生成搜尋權杖，如符合檔案的關鍵字與存取結構，即可取得對應的檔案並進行解密。相較於 PKE 的可搜尋式加密，結合屬性加密更符合雲檔案管理允許多人存取及搜尋的情境，因此本系統將以具關鍵字搜尋的屬性加密為基礎，提供醫療數據基於密文存取控制及關鍵字搜尋之特色。

六、同態加密（Homomorphic Encryption，HE）

近年來大數據分析與其應用盛行，數據生成快速導致數量龐大，使得數據漸漸交由第三方雲服務提供商進行數據清洗等前置處理作業，然而許多數據牽涉商業或個人機密，尤其病歷記載每個人最為隱私的數據，但同時許多醫療研究者也同時埋怨有限的數據限制醫療科技的發

展，因此開發可在隱私與數據應用上兼得的技術刻不容緩。數據大致有三種狀態，分別為儲存、傳輸與運算，而上面敘述數據如何使用對稱及非對稱加密來保證儲存及傳輸的安全，接著介紹更進階的加密機制，包含屬性加密與可搜尋式加密，其賦予加密數據存取控制及關鍵字搜尋能力，而運算則尚未提及如何保證其安全性，這也是目前許多研究者正在鑽研的領域之一，稱為機密運算（Confidential Computing）。滿足機密運算定義的方法有許多種類，接下來簡述本系統能保證運算機密的另一個元件，同態加密。

同態加密被譽為密碼學的「聖杯（Holy Grail）」，因為同態加密允許第三方在數據加密狀態下進行運算，而運算完成的結果經解密後可以得到與明文相互運算的結果相同。此外，同態被分為兩種，加法同態與乘法同態，即分別對應普通加法與乘法運算。同態在 1978 年由 Rivest 等學者提出，並稱為隱私同態，後續亦有學者提出相關研究，但多數仍為實現部分同態（即僅實現加法或乘法同態）。在 2009 年才由 Gentry 提出第一個全同態加密（Fully Homomorphic Encryption，FHE）（Gentry, Craig, and Dan Boneh, 2009），此機制可同時達成加法與乘法同態運算，爾後十年才真正開啟同態運算的研究及運用，並且擁有相對成熟的函式庫，包含 Google、Amazon 與 Microsoft 等都已應用於部分雲服務中，來保證用戶運算隱私，然而效率始終是全同態運算的通病，這也是研究者們努力的

方向之一。

參、具隱私保護暨安全資料探勘之醫療資料倉儲系統

醫療感測技術日益進步，藉由穿戴式裝置就可以獲取許多種類的體徵資訊，如心律、心電圖與血氧濃度等，同時廠商提供雲儲存服務存放用戶的醫療數據，方便使用者可以隨時隨地獲取該資訊，然而使用者數據可能遭未授權的使用或是販售給第三方，更可能因為成本導致產品擁有重大漏洞而使惡意攻擊者可以隨時存取用戶的數據，這也說明醫療數據沒有妥善保存仍可能經由不同途徑導致用戶隱私洩漏，此將導致用戶對於使用該類產品有一定的隱憂，且各國針對醫療數據的規範越來越嚴苛，顯示隱私越來越被重視，其重要性已不言而喻。然而除用戶擔憂隱私洩露的風險之外，企業或研究團隊也因數據應用受限而使得成果難以產出，本文所要介紹的正是在醫療數據隱私與數據應用兩者兼得的方案，命名為「具隱私保護暨安全資料探勘之醫療資料倉儲系統」。

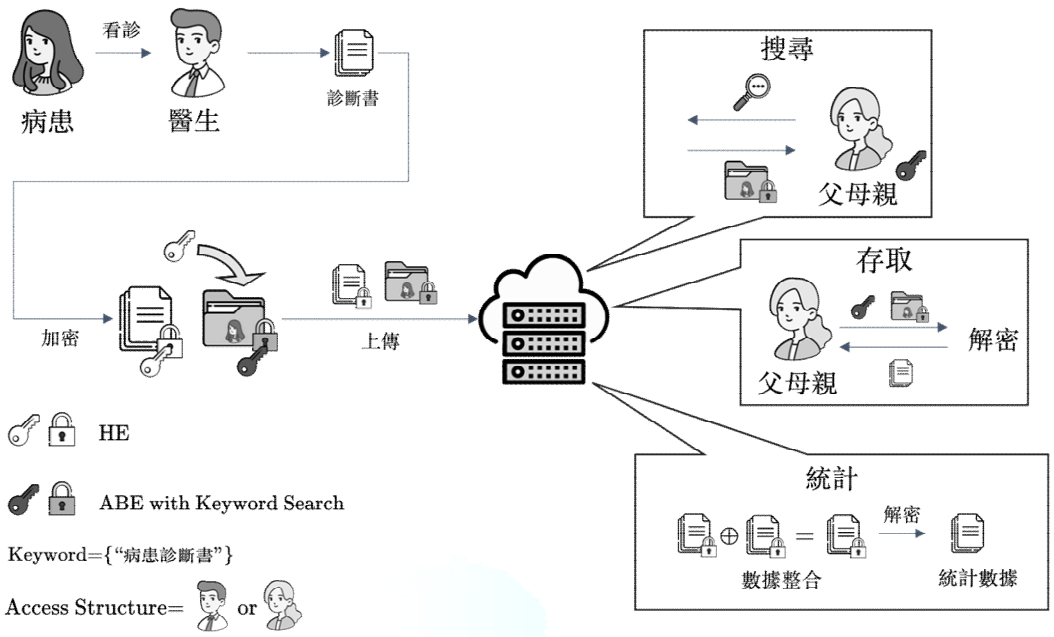

圖 6　具隱私保護暨安全資料探勘之醫療資料倉儲系統架構圖

本系統中的「資料倉儲（Data Warehouse）」[2]相較於與一般認知的資料庫不同的地方在於儲存的數據需進一步處理，以利於後續資料進行分析或應用。因此，除需提供數據儲存與搜尋外，數據運算和數據處理的操作也是必要功能。為滿足資料倉儲定義與保護用戶隱私，圖 6 展示整體系統各項功能，在病患到院看診的同時，醫生將診斷書加密後上傳至雲，爾後被授權者則允許進行存取、搜尋及統計等操作。接下來將說明如何運用上一章節所提及的技術並建構可保護用戶隱私之醫療數據倉儲系統。

[2] 維基百科：於 111 年 7 月 22 日取自 https://zh.wikipedia.org/zh-tw/資料倉儲。

一、數據上傳至雲

醫生診斷完成後，將診斷結果運用同態加密的公鑰加密診斷書，並將同態加密金鑰對透過病患制定的可存取對象結構「Access Structure」與關鍵字列表「Keyword」經具關鍵字搜尋的屬性加密進行診斷書的加密，完成後將加密的診斷書及加密的同態金鑰對上傳至雲。

二、使用者搜尋

只要使用者在病患所制定的可存取對象結構中，如圖 6 所示，存取結構中包含該位女性，則該位女性便可以利用「ABE with Keyword Search」的私鑰與關鍵字（如「病患診斷書」）製作搜尋權杖進行搜尋，雲伺服器配對成功即可取得該份加密診斷書及加密的「HE」金鑰對。

三、使用者存取

承接上個步驟，使用者獲取加密診斷書及加密的「HE」金鑰對之後，假如該使用者在病患制定的可存取對象結構中，則可利用自身的「ABE with Keyword Search」私鑰進行解密獲取當中之「HE」金鑰對，隨後進行診斷書的解密。

四、統計

因為診斷書內容是經由「HE」加密，因此可請求雲直接進行同態運

算，結果回傳並依據上步驟進行解密動作，而解密後的結果即為統計數據。

透過具關鍵字搜尋的屬性加密機制 ABE with Keyword Search 與 HE 同態加密，提供醫療數據，從存取、儲存、搜尋及統計，完全運作於密文與病患自行定義的存取權限之下，確保病患醫療數據在每個環節的隱私，最終建構出具隱私保護暨安全資料探勘之醫療資料倉儲系統。

肆、 未來展望與結論

現今科技日新月異，每日產生的數據量已無法估算，從社群貼文、網路影片及物聯網設備，連廚房的電冰箱也可能每天上傳數據至雲來提供用戶更為方便的服務，隨需求調整的儲存空間及運算力，減少自行建置可能因為離峰服務的時段而導致運算力遭閒置的資源浪費問題，然而許多機敏數據因為各國法律規範，如在臺灣個人醫療數據已公告法源（王若樸，2022），有嚴苛的標準及複雜的審核機制限制醫療數據存放的環境，如醫療數據不能存放於境外資料中心等限制。另一方面，醫療數據也無法得到妥善運用，因為每一筆醫療數據須經由去識別化或使用者需經過嚴格審核才允許使用，這也是阻礙醫療技術發展的原因之一，但仍不可因此對病患隱私犧牲或妥協。

另外，現今加密技術保證數據在儲存與傳輸的安全，但在數據運算上仍

持續發展相關技術來保護隱私，因此近年來機密運算的概念孕育而生，數據儲存與傳輸有如 AES 與 TLS（Transport Layer Security）成熟技術保證其安全性，但仍未有通用或是主流的機制來保證異地運算的安全性。目前主流技術為可信任執行環境（Trusted Execution Environment，TEE），如中央處理器製造商推出的 Intel SGX 或 AMD SEV，此技術為基於硬體安全來保護數據運算時的安全。以往儲存或傳輸已經有成熟加密技術支援，而數據運算時必須於解密後進行，因此攻擊者可以透過旁通道攻擊（即量測電流來取得處理器或揮發記憶體的運作狀態）來獲取數據，TEE 則提供如安全記憶體加密或安全加密虛擬化等技術，來提供安全環境讓數據進行運算（李宗翰，2021），然而 TEE 硬體仍傳出含有漏洞可遭利用，進而造成隱私洩露，研究者仍試圖發展更為安全的替代方案。

目前技術可部分或完全符合機密運算定義的技術大致包含差分隱私（Differential Privacy）、多方運算（Multi-Party Computation）、混淆電路（Garbled Circuits）與同態運算等技術，上述部分方法可證明數據安全，但存在通訊成本或計算成本高等缺點，仍待研究者進一步優化。目前 TEE 依然是主流且成熟的技術來實現機密運算。本系統基於同態技術來保證數據運算的安全性，同態運算已被各大廠商（微軟 Microsoft SEAL、IBM 及 Google）實現或是開發相關函式庫，雖然仍有運算效率差的問題，然而相對其他方法，同態加密的方案提供可證明安全且完全在密文下進行運算操作的特色，也解

決 TEE 基於硬體保護可能含有漏洞的問題，目前研究者也正積極發展同態加密相關技術以達成真正的機密運算。

具隱私保護暨安全資料探勘之醫療資料倉儲系統結合多種加密技術，並因應醫療數據的特性與上雲之需求，包含儲存、搜尋及運算等操作皆在密文下進行，借助雲服務的優勢加速醫療技術的發展，且無需妥協病患的隱私保護，但效率、適配性與技術整合仍是目前本系統需持續調整與優化之處。

參考文獻

一、中文部分

王若樸（2022）。衛福部力推遠距醫療：電子病歷上雲辦法下周上路、通訊診療辦法提 4 大修正方向【iThome】。取自 https://www.ithome.com.tw/news/151811。

李宗翰（2021）。處理器與雲業者紛紛投入機密運算【iThome】。取自 https://www.ithome.com.tw/tech/145561。

李建興（2022）。OpenAI 改進 GPT-3 使其更能聽懂人類指示，並減少輸出有毒內容。取自 https://www.ithome.com.tw/news/149140

二、英文部分

Gentry, Craig, and Dan Boneh (2009), *A fully homomorphic encryption scheme*, Vol. 20, No. 9, Stanford: Stanford university.

Martin Stumpe & Lily Peng（2017）. Google 用 AI 檢測癌症，準確率 83%超越人類【INSIDE】。取自 https://www.inside.com.tw/article/8709-assisting-pathologists-in-detecting-cancer-with-deep-learning。

Zheng, Qingji, Shouhuai Xu, and Giuseppe Ateniese (2014), *VABKS: verifiable attribute-based keyword search over outsourced encrypted data*, IEEE INFOCOM 2014-IEEE Conference on Computer Communications, IEEE.