

人工智慧/機器學習技術之  
醫療器材獨立性能評估  
常見問答集

114 年 8 月

## 一、前言

為協助我國醫療器材業者瞭解人工智慧／機器學習（AI/ML）技術之醫療器材在申請查驗登記時之獨立性能評估（Standalone Performance Assessment）的審查重點及獨立性能評估之使用時機，衛生福利部食品藥物管理署特別撰擬「人工智慧／機器學習技術之醫療器材獨立性能評估常見問答集」，針對其常見問題進行說明，以協助業者準備申請查驗登記之資料。

## 二、常見問答集

### **(一)請問什麼是 AI/ML 醫療器材的獨立性能評估(Standalone Performance Assessment)?**

獨立性能評估是指在測試資料未再新增醫事人員介入判讀(如人工圈選、標註)的情況下，使用既有之資料進行該醫療器材之性能測試，這種評估主要聚焦於產品性能。

上述所稱「既有之資料」，係指於資料取得時即已具備可用於評估之關鍵資訊(如標註結果)之資料，不涉及再次由醫事人員進行臨床判讀。若資料原本即已包含標註資訊，後續進行影像清晰度確認、排除資料錯誤等處理，若未更動原始標註資訊或重新標註，仍屬「既有之資料」範疇。

「既有之資料」常見來源包括公開可取得且去識別化之資料庫、機構已建置完成並取得使用授權之回溯性臨床資料集。

### **(二)請問獨立性能評估與臨床評估(Clinical evaluation)有何差異？**

獨立性能評估是指在測試資料未再新增醫事人員介入判讀(如人工圈選、標註)的情況下，使用既有之資料進行該醫療器材之性能測試，這種評估主要聚焦於產品性能；臨床評估之目的為建立 AI/ML 醫療器材預期使用臨床症狀與臨床實證之關連，其資料特徵應由醫事人員進行確認或判定，以確保 AI/ML 醫療器材能提供預期之技術層面與臨床層面資訊。

### **(三)請問執行 AI/ML 醫療器材進行獨立性能評估時，是否需要倫理審查委員會(Institutional Review Board, IRB)的審查？**

執行獨立性能評估是否需要倫理審查委員會(Institutional Review Board, IRB)的審查，取決於是否進行「人體研究」且是否屬主管機關公

告的「得免倫理審查委員會審查之人體研究案件範圍」，並應符合《人體研究法》、《醫療器材管理法》、《醫療器材優良臨床試驗管理辦法》等相關規定。依據衛署醫字第 1010265075 號公告（101 年 7 月 5 日）「得免倫理審查委員會審查之人體研究案件範圍」第 2 款「使用已合法公開週知之資訊，且資訊之使用符合其公開週知之目的」得免送倫理審查委員會審查或由倫理審查委員會核發免審證明。如果獨立性能評估之測試資料來源倘涉人體研究法之規定，則需要 IRB 的審查。

**(四)請問執行 AI/ML 醫療器材進行獨立性能評估後，是否一定還要再提供完整之臨床證據資料？**

依《醫療器材許可證核發與登錄及年度申報準則》第 6 條及附表二說明九，已有類似品經中央主管機關核准上市者，除須以臨床證據佐證其安全及有效性者外，得免附臨床證據資料。

若 AI/ML 醫材為無類似品，則應提供臨床證據資料(包含學術理論依據與有關研究報告及資料、臨床評估報告或臨床試驗報告)，用以驗證產品在臨床使用時的安全性及有效性。於此情況，若僅進行獨立性能評估，無法完整驗證產品之安全及效能。

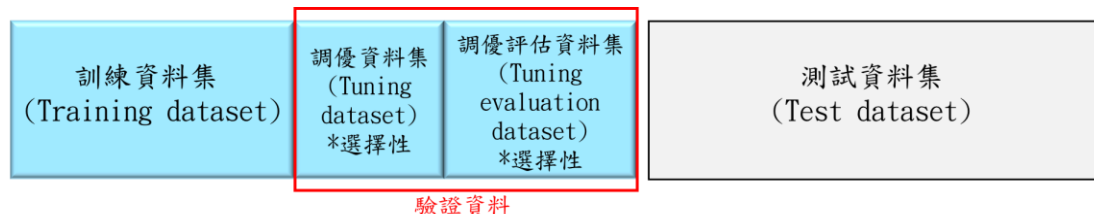
若提供之獨立性能評估不足以佐證產品安全及效能，則須檢附臨床證據資料加以佐證。

**(五)人工智慧醫療器材於產品開發與測試階段常見的資料集有哪些？**

在進行人工智慧醫療器材之開發與測試時，通常至少需準備訓練資料集(Training dataset)與測試資料集(Test dataset)。視開發需求，亦可增加驗證訓練結果之資料，如模型調優資料集(Tuning dataset)與調優評估資料集(Tuning evaluation dataset)等其他資料集。

訓練資料集(Training dataset)是用於訓練機器學習模型的一組資料集，透過使用訓練資料集建立或改進機器學習模型參數。模型調優資料

集(Tuning dataset)與調優評估資料集(Tuning evaluation dataset)用於調整其非可學習參數，並評估訓練模型。測試資料集(Test dataset)是用於評估訓練後機器學習模型的效能。上述資料集之概念如下圖：



(六)目前僅於《人工智慧/機器學習技術之電腦輔助偵測 (CADE) 及電腦輔助診斷 (CADx) 醫療器材查驗登記技術指引》中提及「獨立性能評估」，請問此評估方式是否僅適用於 CADe/CADx 類型的人工智慧醫療器材？

獨立性能評估為評估人工智慧醫療器材產品性能之方法，適用範圍不限於 CADe/CADx。雖目前僅於 CADe/CADx 技術指引中明確提及，但對於其他如醫學電腦輔助分流(Triage)與通知軟體等運用人工智慧/機器學習技術之醫療器材，同樣可透過獨立性能評估驗證產品性能。

(七)請問泛化性測試(Generalizability testing)、重複性與再現性測試有何不同？

泛化性測試是評估產品在面對不同條件時(例如：相同規格於不同品牌之設備)的性能穩定性。重複性與再現性測試則評估在相同或不同的資料擷取設備及操作條件下，對同一病患重複檢查，以衡量輸出結果差異。

(八)請問是否可以使用公開資料庫進行獨立性能評估？

使用公開資料庫進行獨立性能評估時，應確認符合產品的預期用途，建議注意下列議題：

- 資料代表性：雖然這些資料庫提供了大量資料，但它們可能不能完全代表所有群體或臨床情況，評估時應考慮資料的來源和

特點，例如是否涵蓋不同人種、性別、年齡等

- 資料品質：公開資料庫的品質可能參差不齊。在使用前，應確認符合業者產品的預期用途與輸入格式，例如確認資料是否為原始檔或經過處理。
- 時效性：某些資料庫可能包含較舊的資料，評估時應考慮資料的時效性。
- 法律和倫理考慮：使用這些資料庫時，必須遵守相關的資料使用協議和隱私保護規定。

常見之資料庫如下(僅供參考):

- The Cancer Imaging Archive (TCIA)
- NIH Chest X-ray Dataset
- MIMIC (Medical Information Mart for Intensive Care)
- MIT-BIH(The Massachusetts Institute of Technology – Beth Israel Hospital Arrhythmia Database)
- 全民健康保險學術研究資料庫
- 國家高速網路與計算中心生科雲醫療影像資料庫

**(九)請問使用 AI/ML 醫療器材進行獨立性能評估時，應注意哪些事項？**

- 進行獨立性能評估所使用的測試資料集，應與建立演算法模組的訓練資料集、驗證產品初步功能的驗證用資料不同，且應能代表目標族群及目標疾病、病症(Condition)或異常(Abnormality)，測試資料集應包含疾病案例、異常案例與正常案例。
- 應提供資料案例收集的計畫書(Protocol)。
- 可進行壓力測試 ( Stress testing )，刻意納入符合預期用途的病

例，但具備較罕見之資料特徵，這些資料是已知會對臨床醫師構成挑戰。

- 建議在執行獨立性能評估時，勿使用虛擬、合成資料。
- 為提升產品泛化能力(Generalizability)，建議測試資料來自不同的臨床機構。

**(十)請問獨立性能評估通常包括哪些主要評估指標？**

應依據產品特性訂定主要評估指標，常見的主要評估指標包括靈敏度(Sensitivity)、特異度(Specificity)、陽性預測值(Positive predictive value)、陰性預測值(Negative predictive value)、AUC-ROC 曲線(Area Under the Receiver Operating Characteristic curve)等。

### 三、參考資料

- (一) 《醫療器材管理法》
- (二) 《醫療器材優良臨床試驗管理辦法》
- (三) 《醫療器材許可證核發與登錄及年度申報準則》
- (四) 《人工智慧/機器學習技術之醫療器材軟體查驗登記技術指引》
- (五) 《人工智慧/機器學習技術之電腦輔助偵測(CADe)及電腦輔助診斷(CADx)醫療器材查驗登記技術指引》
- (六) 《人工智慧/機器學習技術之電腦輔助分流(Computer Aided Triage)醫療器材軟體查驗登記技術指引》
- (七) US FDA, Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data – Premarket Notification [510(k)] Submissions (2022)
- (八) US FDA, Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations (Draft Guidance for Industry and Food and Drug Administration Staff) (2025)
- (九) The EU Artificial Intelligence Act (2024)