醫療機構應用生成式人工智慧指引

115 年 5 月 29 日衛部醫字第 1151663164 號函頒

一、本指引目的、適用範圍

(一)目的

人工智慧(artificial intelligence, AI)在醫療場域的應用日漸普及，新科技的使用能協助臨床醫療，減輕醫事人員的負擔，並提升醫療品質。生成式人工智慧近年來是許多醫院積極建置，以優化醫療流程的系統。然而生成式人工智慧的運作方式與其他人工智慧系統不同，除可能帶來資訊正確性、資料安全與倫理風險等新挑戰外，亦涉及病歷、影像、語音等敏感資料之蒐集、處理與傳輸。若缺乏妥善的風險治理與隱私保護配套，可能產生誤用、過度依賴、個人資料不當利用或資訊安全等問題。

本指引之目的，是協助醫療機構在導入生成式人工智慧之全生命周期，有明確的檢視重點與評估指標，掌握系統導入所需之準備、應注意之風險及後續管理機制，並作為醫療機構內治理、內稽內控與教育訓練制度設計之參考。

除前述發展背景外，本指引係配合人工智慧基本法賦予目的事業主管機關就各該領域人工智慧應用風險管理目標，訂定以風險為基礎之管理規範及協助產業訂定指引與行為規範之權責。本部就醫療機構生成式人工智慧之應用提出行政指導建議，落實「以人為本」與隱私保護、風險治理等基本原則。藉由提供一套實務可行的參考架構，期能推動醫療場域安全、負責任地運用新科技，讓人工智慧真正成為臨床決策與照護工作的輔助工具，同時兼顧病人安全、醫療品質與病人隱私。整體精神以「負責任創新」為核心，並落實風險治理與適應性

治理的原則，讓科技導入能在可控且能即時調整回應風險的環境中進行。本指引屬行政指導性質，旨在提供原則性與實務性之參考，並非強制性規定，醫療機構仍得依其實際情況與相關法令規定，採取適當之導入與管理措施。為因應科技變遷並完善醫療風險治理，本部將視施行情形，適時更新本指引或另行頒布細部注意事項。

(二)名詞定義

1、**生成式人工智慧**：指能夠根據輸入提示或條件，學習數據的統計模式和特徵，從而生成新的內容的人工智慧技術或模型。

2、**生成式人工智慧系統**：指整合了生成式人工智慧模型，並包含數據處理、使用者介面、安全控制、輸出管理等組件，用於實際導入於醫療機構環境中，並提供生成式人工智慧功能的完整技術系統。

3、**人工智慧代理系統(AI Agent System)**：指能夠感知環境、自主決策、並採取行動以達成特定目標的人工智慧系統，具備一定程度的自主性、反應性、主動性和社交能力。在醫療照護情境中，指能夠自主或半自主地在授權範疇內執行醫療相關任務，並可能對患者健康結果產生直接或間接影響的人工智慧系統。

4、**資料保護評估**：指在生成式人工智慧系統導入或使用前，檢視其在蒐集、處理、儲存及利用個人資料過程中，對隱私與資料安全可能造成的風險與影響，並評估其資料利用合規性且確認已採取適當防護措施的程序。

(三)適用範圍

本指引適用範圍為預備導入或已經導入生成式人工智慧系統之醫療機構，包括公私立醫院及診所。其適用情境包含生成式人工智慧系統在醫療或管理作業中的應用，例如病歷撰寫輔助、臨床決策支援、

行政文書撰擬及病人溝通工具。

人工智慧代理系統，目前尚未普遍應用於醫療照護情境，且其自主決策與採取行動之特性可能引發更高之病人安全、醫療責任與合規性風險。相關規範與適法性仍有待主管機關及各界持續探討，尚不列入本指引之適用對象。

本指引不適用於非醫療場域或非受醫療監管之生成式人工智慧應用。涉及醫療器材者，應回歸依醫療器材相關法規及其主管機關規定辦理。若為個人研究、教育示範或僅限功能驗證之內部測試，可視情況參照本指引原則辦理。惟若測試內容涉及真實病人資料或臨床作業流程，例如於門診或病房實際使用、系統與機構內資料介接，即屬生成式人工智慧系統導入之階段，仍建議依本指引相關規範執行。

二、生成式人工智慧應用原則

(一)生成式人工智慧之風險類型

生成式人工智慧雖可提高醫療機構營運效率，但其資料驅動的學習方式、內容生成能力、廣泛應用及技術複雜性等特性，也帶來潛在風險，風險來源可分為以下六類：

1、**基礎模型風險**：係指於模型設計、訓練與系統建置過程中，因模型架構選擇、訓練策略、參數調整、訓練資料組成及微調流程等因素，使模型在完成建置時即已內嵌系統性偏差。於醫療場域，基礎模型偏差可能導致對特定疾病、特定族群或特定臨床情境的判斷產生系統性偏誤，影響診療建議的正確性與公平性。

2、**資料來源風險**：係指生成式 AI 系統於建置完成後之實際運作過程中，因連結取用外部資料來源（如檢索增強生成所介接之知識庫、即時介接之電子病歷、外部醫學資料庫或網路資料等）時，因資料

品質不一、來源可信度不足、資料時效性過期或隱私保護不足等因素，導致產出結果之正確性與可信度受到影響。此類風險有別於基礎模型建置階段之偏差，係於系統運作時動態產生，且可能因資料來源變動而持續改變。於醫療場域，取用過時、不正確或不完整之外部資料可能直接影響臨床建議品質。

3、**輸出結果風險**：生成式 AI 可能產出看似合理但實際不正確之內容，包括編造不存在的醫學文獻、產生錯誤的藥物交互作用資訊，或提供與現行臨床指引不符的建議。此類幻覺在醫療場域具有直接危害病人安全之潛在風險。

4、**資安攻擊風險**：生成式 AI 系統可能面臨提示詞注入攻擊（prompt injection）、對抗性樣本攻擊、模型竊取、資料中毒或資料外洩等資安威脅。醫療場域之資安事件可能同時影響病患安全與個人資料保護。

5、**使用者依賴風險**：醫事人員可能因對 AI 系統產生過度信任，而降低自主臨床判斷之審慎程度、忽略反常訊號，或將 AI 產出當作最終結論。此類行為變化可能導致臨床判斷能力退化(deskilling)，並在 AI 運作異常時無法有效回應。

6、**服務中斷風險**：多數生成式人工智慧解決方案仰賴外部模型服務商或雲端 API。實務上可能遭遇服務停止、模型下架、API 介面或行為變更、版本更新造成輸出漂移、速率限制、定價政策調整或合約條款變動等問題，進而影響機構內流程的可用性與成本可控性。

在醫療情境下，前述各類風險往往並非單一發生，而是多重風險交互作用，其影響涉及生命健康安全，尤須特別重視。因此，導入生成式人工智慧系統時，醫療機構應針對前述各種風險類型，以整體系

統觀點（人、流程、技術與外部依賴）進行風險辨識與管理，以兼顧效能與安全之精神，建立檢核與持續監測機制，確保技術應用符合安全與品質要求。

(二)核心實施原則

醫療機構在導入生成式人工智慧應用時，應由機構層級秉持以下核心實施原則，作為管理機制與人員培訓之指導原則。

1、指派主責單位或人員與風險辨識

指派主責單位或人員，確保生成式人工智慧系統導入及使用過程中，風險可被辨識、控管與管理。主責單位之組成方式可依機構層級規模採多元模式，包括委員會或專責人員，並鼓勵跨機構之專業治理資源共享。

2、完成資訊安全與資料保護評估

導入前進行資訊安全與資料保護評估，確保系統運作安全並符合個資保護法規要求。

3、規劃系統整合與作業持續

確保生成式人工智慧系統能順利與既有醫療資訊系統及臨床工作流程整合，並可追蹤輸出結果與使用軌跡。本於醫療機構作業持續原則，預先規劃 AI 系統失效、服務/供應鏈中斷或緊急停用時之替代流程(如既有人工作業流程)、復原程序與責任分工，以維持服務不中斷並保障病人安全與作業穩定性。

4、落實負責任的組織文化

系統使用應符合法規與相關標準，建立明確的使用範圍與權限管理機制。機構應強化對人員之培訓，使其充分了解系統功能、限制與正確操作方式，防止誤用或濫用。

5、持續監控與改進

導入生成式人工智慧後，建立監測、通報、事件處置及改進流程，且納入版本控制與變更管理機制，包含變更評估、測試驗證、核准上線與版本回復程序，並持續檢視輸出品質、偏誤、資安事件與作業中斷風險，以確保系統運作安全性、持續性、穩定性且風險可控。

三、醫療機構使用生成式人工智慧應注意事項

本部分依據前述「核心實施原則」之內涵，提出九項具體應注意事項與補充說明。為協助機構按生成式人工智慧導入生命週期規劃與執行，這九項要點依導入流程分為三個階段：「導入前評估」、「導入與整合」及「導入後使用與監管」，各階段的重點說明如下。

(一)導入前評估階段

確認生成式人工智慧系統導入的可行性、潛在風險與法規遵循，確保導入與使用過程安全合規。

1、由主責單位或人員完成風險盤點

(1)指派對醫療品質、臨床安全具相當專業之單位或人員，負責全機構系統導入與使用過程的評估、審查及風險分級。

(2)生成式人工智慧系統之資料來源越複雜，其輸出可靠性越容易受影響。醫療機構應評估系統所整合資料來源之多元性、時效性與可靠性(如更新頻率、是否可追溯來源、是否經專業審核等)，作為風險分級之關鍵指標。

(3)導入前應先盤點擬導入之生成式人工智慧產品與系統，建立人工智慧產品與系統清單，並評估其內部風險(系統之可靠性：資料來源多元性、資訊安全與資料保護等)及外部風險(系統之臨床影響性：對醫療流程與臨床決策可能造成之影響)。據此建立

之風險分級機制，應作為後續緩解措施、監測頻率與稽核要求之基礎，以提升執行效率與一致性。

(4)醫療機構進行風險評估與分級時，得要求合作廠商提供合理可驗證之資訊，例如模型版本資訊、預定用途與限制、資料流向與保存政策、效能指標與已知風險、資安防護與事件通報機制、變更管理與支援承諾等，以支援醫療機構完成風險盤點與分級。

2、法規遵循檢視

(1)在導入或試行前，應確認生成式人工智慧系統之使用方式符合個人資料保護法有關個資利用目的限制等規範、醫療法有關病歷文書製作管理及病人隱私保護等相關規範。視系統預定用途與涉及之專業別，檢視涉及之個別醫事人員專門職業法規所定之執業範圍、專業責任與親自執行要求等相關規範。

(2)檢視生成式人工智慧系統的預定用途，若主要功能涉及診斷、治療、緩解或直接預防人類疾病，則應注意是否屬醫療器材，以及是否符合其核准用途，必要時應向食品藥物管理署申請進行屬性判斷；若非醫療器材，應透過適當設計或管理措施，避免系統輸出超出預定用途，例如限制輸出格式或內容、標註生成結果來源，並僅在人為監督或專業確認下使用，避免作為臨床診斷或治療依據。

(3)凡涉及臨床判斷、病人安全、病人溝通或醫療紀錄之使用情境，應由具相應專業資格且在其執業範圍內之醫事人員負最終確認、審核與責任。

3、資訊安全與資料保護評估措施

(1)在系統上線前，應進行整體安全評估，內容視情況可包含建立

「安全性說明文件」、「風險與事件記錄表」，並規劃「監測機制」，以利追蹤與改善。

(2)資料保護部分，應進行「資料保護評估」，依個人資料保護法及相關規範，檢視資料蒐集、處理、儲存與利用過程之合法性與安全性。評估過程中，應依據辨識出的風險制定對應的緩解措施，如技術控制（加密、存取權限管理）、流程管理（如人工監督、專業確認）、使用者教育與訓練，以及異常事件通報與應變流程等。

(二)導入與整合階段

聚焦於生成式人工智慧系統的部署、系統整合及合作廠商管理，確保導入後的運作穩定且原則上可追蹤。

1、系統整合、效能驗證及可擴充性考量

(1)規劃整合測試，確保生成式人工智慧系統能與醫療資訊系統(如EMR、HIS、PACS）及臨床流程順利串接。

(2)進行效能驗證、壓力測試及臨床安全測試，確認系統功能、負載能力與在代表性臨床情境下之輸出品質與安全性。臨床安全測試宜以具代表性之案例與工作流程，訂定明確驗收標準進行，經機構內專責單位確認後方可上線；並將測試流程、資料與結果完整紀錄，方便日後追蹤及稽核。

(3)評估生成式人工智慧系統對不同平台及臨床需求之支援能力，確保模型更新彈性、介面擴充性與資料交換標準化，並視需要保留替代模型或替代服務之切換可能性，以降低單一服務依賴風險。

(4)若經評估屬高風險應用，應採分階段導入。先於隔離測試環境完

成驗證，再於有限場域或特定科別試辦，並設定明確的擴大及暫緩門檻（go/no-go criteria）與監測期間。每一階段均應完成風險再評估與內部核准，逐步擴大使用範圍。

(5)建立版本更新、系統維護、模型校正與異常處理程序，並同步規劃回溯與降版（rollback）機制。此應包含版本鎖定、變更前後對照測試、異常偵測與通報、以及快速回復至既有穩定版本或啟動替代流程之步驟與責任分工，以降低對臨床作業與病人安全之影響。並應指定專責窗口負責模型表現追蹤與異常事件通報。

2、供應商管理與採購作業

(1)委託建置或維運之契約明確規範生成式人工智慧系統品質標準，例如模型準確度、偏誤管理、可追溯性等（其具體指標可依系統用途與風險等級訂定）。契約並應明訂資料使用權限、供應商持續支援要求、智慧財產權歸屬與管理等條款。

(2)契約應要求供應商揭露其使用之大語言模型（含模型服務商名稱、版本、部署型態、主要限制與更新通知機制）。另如系統包含硬體設備（如伺服器、邊緣運算設備、錄音錄影或其他關鍵裝置），契約應規範其設備製造地與供應鏈來源之揭露與法規遵循，並依醫院關鍵設備及資安相關法規要求辦理。

(3)供應商應提出系統維護計畫、版本更新與變更管理機制（含影響評估、測試驗證、上線通知與回復作法）、簽約時業界已有之評測報告或驗證報告、資安防護與事件通報及應變機制、資料流向與存取權限設計等文件。醫療機構應定期評估履約、技術支持與資安落實狀況，必要時要求改善或啟動替代方案。

(4)契約應載明供應商於蒐集、處理或利用醫療機構方、病人或其他

可識別個人之資料時，應遵守《個人資料保護法》及相關規範，並僅得於契約約定之目的與範圍內使用，且應配合機構內「資料保護評估」結果與資料治理要求。

(三)導入後使用與監管階段

著重於生成式人工智慧系統運作過程中的適度持續監控、偏誤管理與責任落實，確保系統穩定、安全且符合倫理原則。

1、持續監測與偏誤管理

(1)依導入前評估階段第 3 點「資訊安全與資料保護評估措施」所建立之監測機制，定期稽核生成式人工智慧系統輸出的正確性、偏誤與臨床適用性，並得納入監測指標，如錯誤或偏誤類型與發生率、異常事件通報、使用者覆核結果，以及臨床人員手動修改頻率等)，以掌握輸出品質與風險變化。監測結果可作為模型更新與系統改進依據，並保留紀錄以供稽核。

(2)依據風險分級管理原則，綜合評估系統對臨床決策之影響程度與病人安全關聯性，及系統所整合資料來源之複雜度，決定驗證強度與頻率。資料來源複雜度較高之系統，應設計使用者得以直接查閱系統生成內容所依據之原始資料或其出處之機制。

(3)當出現系統性偏誤或錯誤且無法於合理期間內完成修復，或可能造成病人安全影響時，醫療機構應依預先訂定之啟動條件與權責分工，採取暫停使用、限制適用範圍、回復既有流程或停止使用等措施，並完成內部通報、原因分析與必要之後續改善。

2、具風險意識之組織文化

(1)生成式人工智慧系統存在「惡意指令植入」風險，指攻擊者透過精心設計之輸入內容，操控系統產生非預定或有害之輸出，可能

導致錯誤醫療決策或資訊洩露。其發生原因包括外部使用者攻擊、跨系統資料污染、內部誤用或濫用。

(2)機構應建立輸入過濾與異常監測機制，並落實輸出內容審查程序。

(3)建立安全至上之組織文化，鼓勵員工在發現系統運作錯誤或人為操作不當時，勇於通報以避免風險持續發生。

3、責任分工與教育訓練

(1)醫療機構應與供應商訂定明確且全面的契約，界定各方於系統安全、維護、資料使用及偏誤修正等事項之責任分工。機構內則應進一步明確劃分使用者、管理單位及資訊部門之職責，確保責任與權限一致。

(2)確保生成式人工智慧系統在醫療場域作為輔助工具，臨床醫師、行政主管或其他決策人員負有最終判斷責任。醫事人員製作病歷文書時，應遵守電子病歷製作規範確實審核與簽章。

(3)醫療機構應定期辦理教育訓練，使醫事人員充分了解生成式人工智慧系統的操作方式、功能限制與潛在風險，確保使用者能正確解讀並妥善運用其輸出結果。教育訓練內容除應涵蓋輸出品質風險（如幻覺、偏誤與不一致）及人為覆核要點外，亦應納入「輸入安全（Prompting Safety）」觀念與實務操作，並對應機構內資安與事件處置機制，提升整體操作知能與風險控管能力。

4、資訊透明以促進信任

(1)醫療機構應以適當方式告知病人及家屬生成式人工智慧系統之參與範圍、用途與限制，以維持透明與信任。在以生成式人工智慧系統直接和民眾對話互動之情境，應主動揭露此係以人工智

慧系統運作，並提醒其輸出可能出現幻覺、錯誤等風險，其內容僅供參考，必要時應由醫事人員進一步確認。

(2)因生成式人工智慧系統之使用而需在醫療照護過程中錄音、錄影時，應先以適當方式告知病人或家屬，說明錄音、錄影之必要性與用途限制。若病人或家屬明確拒絕時，則應停止錄音、錄影。前述錄音、錄影之資料僅得用於所告知之目的範圍用途。

参考文獻：

1. Office of the National Coordinator for the Health Information Technology(ONC), Health Data, Technology, and Interoperability (HTI-1) Certification Program(HTI-1 Rule), 2024 年 3 月
2. American Medical Association, Augmented intelligence development, deployment, and use in health care, 2024 年 11 月
3. NHS England, Guidance on the use of AI-enabled ambient scribing products in health and care settings, Version 2, 2026 年 3 月
4. NHS England, Using AI-enabled ambient scribing products in health and care settings, 2026 年 3 月
5. European Union, EU AI Act (Regulation (EU) 2024/1689), 2024 年 5 月
6. European Commission, Code of Practice on Marking and Labelling of AI-generated content, 2026 年 3 月
7. European Commission, Ethics guidelines for trustworthy AI, 2019 年 4 月
8. European Union Agency for Cybersecurity, eHealth security in the spotlight: A good practice guide for a robust and resilient EU health sector, 2025 年 9 月
9. HAIP（医療 AI プラットフォーム技術研究組合），《医療・ヘルスケア分野における生成 AI 利用ガイドライン（第 2 版）》（醫療・健康照護領域生成 AI 利用指引），2025 年 7 月
10. JaDHA（日本デジタルヘルスアライアンス），《ヘルスケア領域において生成 AI を活用したサービスを提供する事業者が参照するための自主ガイドライン（第 2 版）》（健康照護業者生成 AI 活用指南），2025 年 2 月
11. World Health Organization, Ethics and governance of artificial intelligence for health, 2021 年 6 月
12. World Health Organization, Regulatory considerations on artificial intelligence for health. 2023 年 10 月
13. NIST AI Risk Management Framework (AI RMF 1.0), 2023 年 1 月(USA)
14. NIST AI 600-1: Artificial Intelligence Risk Management Framework: Generative AI Profile, 2024 年 7 月(USA)

15. Testing and Experimentation Facility for Health AI and Robotics(TEF-Health), Technical and scientific support for Health AI providers and notified bodies, 網頁：https://tefhealth.eu/home (EU)
16. National Cyber Security Centre, Guidelines for secure AI system development, 2023 年 11 月(UK)
17. NHS England Digital, DCB0160: Clinical Risk Management: its Application in the Deployment and Use of Health IT Systems, 2018 年 6 月(UK)
18. 厚生労働省,《医療情報システムの安全管理に関するガイドライン 第 6.0 版(醫療資訊系統的安全管理指南 第 6.0 版)》，2023 年 5 月(Japan)
19. 厚生労働省，《医療機関等におけるサイバーセキュリティ対策チェックリスト（醫療機構等網路資安對策檢核表 2025 年版)》，2025 年 5 月(Japan)