

AI與醫療的大轉型



目錄

Part 1. 診療病歷與臨床業務自動化	7
第1章. AI書記與記錄工作	9
1.1 診療語音辨識與記錄時間縮短	10
1.2 耶魯與克利夫蘭醫學中心的導入案例	15
1.3 醫護人員倦怠與下班後病歷記錄工作	20
第2章 臨床決策支援（CDSS）與診斷	25
2.1 基於大型語言模型（LLM）的抗生素處方支援	26
2.2 醫療世界模型與 U2-Med 的診斷與檢查請求	32
2.3 AMIE 與 MIRA 的診斷推理	38
Part 2. AI藥物開發與數位分身	44
第3章. 生物基礎模型與藥物開發標靶	46
3.1 醫療多模態基礎模型	47
3.2 Octopus 的大腸癌生物標記（IGF2）探索	56
3.3 細胞世界模型（Chreode）與分子設計	61
第4章 數位分身（Health Digital Twin）與臨床試驗	69
4.1 反映病患特徵的數位分身	70
4.2 虛擬對照組（Synthetic Control Arm）臨床試驗	77

4.3 基於模型的藥物開發與動物實驗替代 (FDA Modernization Act)	83
Part 3. 海外醫療AI導入案例	91
第5章 美國醫療機構的人工智慧引進	93
5.1 梅約診所的 VoiceCare AI 與院務自動化	94
5.2 Tempus 的精準腫瘤學支援	100
第6章. 中國的醫療人工智慧導入	107
6.1 河北醫大附屬醫院的人工智慧營運成效	108
6.2 WiNGPT 3.0 與螞蟻阿福服務	113
6.3 邁瑞啟元與基層醫療數位化	118
Part 4. 醫療AI的法規・安全・經濟性	124
第7章 人工智慧醫療器材法規與 PCCP (預先變更控制計畫)	126
7.1 PCCP 指引與人工智慧模型變更管理	127
7.2 FDA能力評估 (Competency Assessment) 與風險評估	132
7.3 英國 AI Airlock 與歐洲 AI Act	138
第8章 病患安全與醫療人工智慧的經濟性	144
8.1 口譯錯誤傳播 (Interpreter Error Propagation) 與語音辨識偏見	145
8.2 自動化偏見 (Automation Bias) 與人為介入審查 (Human-in-the-loop)	150
8.3 醫療代理型人工智慧市場與 CPT 給付費用	155
Part 5. 醫療AI代理與會診	162

第9章 醫療代理型人工智慧 (Agentic AI)	164
9.1 單一預測模型的局限	165
9.2 感知、思考、行動 (Sense-Think-Act) 與記憶	170
9.3 醫療代理式人工智慧的分類與技術成熟度	176
第10章 多代理 (Multi-Agent) 與人工智慧醫院	181
10.1 多代理協同設計	182
10.2 MAI-DxO 的角色分工與複合診斷	188
10.3 ClinVerse 與人工智慧醫院的診療流程	192
Part 6. 未來醫療與AI	198
第11章. 醫師與人工智慧代理的協同合作	200
11.1 人工智慧篩檢・人類審核・綜合診斷	201
11.2 醫師的角色與綜合臨床判斷	208
第12章 未來醫療人工智慧的技術與倫理課題	214
參考文獻	223
書籍資訊	273
贊助指南	274

Part 1. 診療病歷與臨床業務自動化

第1章. AI書記與記錄工作



1.1 診療語音辨識與記錄時間縮短

TXP Medical的產品發表聲明中，說明了SpeechER如何將語音轉換為各類記錄，以及時間節省數據的調查主體。這是一份能夠將功能說明與效果主張分開閱讀的資料。

<https://txpmedical.jp/en/news/2jmiSPoM4bpIdI9GUZF0BI/>

診療病歷是傳達患者症狀、檢查及治療計畫，並用於後續延續診療所必需的資料。然而，診療中的輸入與診療後的文書撰寫可能會增加醫護人員的工作負擔。美國醫療界所使用的「睡衣時間」（pajama time），即是指下班後仍持續進行的電子病歷（EHR, Electronic Health Record）撰寫工作。

AI書記（ambient AI scribe）是一種辨識診療對話並生成病歷草稿的技術。書記（scribe）意指記錄人員，而環境感應（ambient）則代表該技術在無需額外手動輸入的情況下處理周遭對話的方式。醫護人員與患者對話，系統則將所需資訊整理成文件形式。

語音辨識與病歷撰寫是彼此不同的作業。語音辨識將聽到的話語轉換為文字；病歷撰寫則從中篩選出診療所需的內容，並配置到適當的項目中。例如，在轉述患者「昨天很痛，但現在還好」這句話時，必須保留症狀隨時間的變化。若僅簡化為「有疼痛」，可能會誤傳當前的狀態。此範例旨在說明記錄錯誤的原理。

將診療語音轉換為病歷的過程

第一階段是對話的蒐集。透過麥克風或智慧型手機應用程式記錄語音。患者同意與錄音資料的處理方式，必須符合該地區的法律與醫療機構的政策。

自動語音辨識（ASR, Automatic Speech Recognition）將蒐集到的語音轉換為文字。在此過程中，醫學術語、藥物名稱、說話者辨識以及周遭噪音等因素皆可能影響準確度。

大型語言模型（LLM, Large Language Model）從文字中整理出症狀、服用藥物、診察所見、治療計畫等內容。亦可依照SOAP格式，劃分為主訴（Subjective）、客觀所見（Objective）、評估（Assessment）與計畫（Plan）來撰寫。視產品而定，有的在對話進行中即時處理，有的則在診療結束後生成草稿。

這也是SOAP將患者感受到的症狀與醫護人員測量的數值分開記錄的原因。主訴「好像在發燒」與體溫計量測的結果，其來源不同。若AI將兩者混雜，後續閱讀病歷的人便難以釐清究竟實際確認了什麼。一份良好的草稿除了使語句流暢外，更必須保留資訊的來源與時間點。

最後由醫護人員檢視並修改草稿以行確認。AI撰寫的文件本身並非經過確認的正式記錄。必須在有人員參與處理過程的人在迴圈（Human-in-the-loop）架構下，確認記錄的內容與準確性。

輸入時間與檢視時間的變化

AI書記在減少直接手動輸入時間的同時，也產生了確認草稿的工作。因此，整體的工時節省必須綜合考量輸入時間的減少以及檢視與修改的時間。依據病歷格式、診療專科以及與患者對話的特徵，其效果也可能有所不同。

例如，即便原本親自書寫的時間從10分鐘減少至4分鐘，但若確認草稿需要額外耗費3分鐘，則整體記錄工作時間為7分鐘。這個數字是說明計算方式的假設範例。若在導入前僅測量手動輸入時間，而在導入後僅測量草稿生成時間，以此方式比較便會誇大節省效果。必須以相同的範圍測量從開始記錄一位患者到最終確認為止所需的時間。

日本的SpeechER與護理記錄支援工具Caretomo，是將語音輸入整理為文件的代表案例。

SpeechER的開發商為日本TXP Medical。在該公司於2025年10月15日發布的獲獎公告中記載，該產品已作為商用應用程式推出。此日期為該公告發布的日期，不應解讀為最初上市日期。Caretomo則是日本SIND提供的護理記錄支援產品。在其官方產品說明中，可確認其具備語音辨識、SOAP分類、確認與修改後登錄的流程，但該產品介紹中並未明確標註最初上市日期。這兩個名稱並非通用AI模型的名稱，而是針對醫療業務需求打造的產品名稱。業者透過內部調查與短期測試發表了記錄時間減少的成果，然而難以將此結果推論為所有機構皆可期待的縮減比率。

在比較研究中，亦呈現出因產品與指標而異的差異。在UCLA的研究中，某款產品的病歷撰寫時間確實減少，但另一款產品與對照組相比並無顯著差異。分析多家醫療機構營運資料的研究指出，文書處理時間雖有所減少，但下班後的業務量並未隨之同步減少。

因此，將特定的70%縮減數值視為該技術整體的普遍效果並不恰當。病歷輸入、文書撰寫、EHR使用與診療決策是互不相同的業務。例如，遠距心電圖平台的判讀時間縮短，就必須與病歷撰寫時間加以區隔。

影響時間節省的条件

實際效果受到使用頻率與工作方式的影響。頻繁使用草稿的醫護人員與幾乎不使用的醫護人員，其結果可能會有所不同。若格式與術語不符合診療專科的需求，修改的負擔便會加重。

伴隨AI的引進，診療流程與資料管理也可能發生轉變。即使醫院的效率有所提升，也難以斷定所有的變化全然是由AI所促成。在檢視導入效果時，必須確認哪些業務產生了變化，以及同時提供了何種支援。

記錄準確性與安全

在語音辨識與內容整理過程中，可能會發生遺漏、錯誤添加或語義反轉等情況。在包含口譯的模擬診療研究中，發現了口譯錯誤直接被傳遞至AI記錄中的問題。甚至出現過患者陳述的症狀在記錄中被整理為不存在的案例。

自動化偏誤（automation bias）是指使用者過度信任AI結果的傾向。若語句撰寫得自然流暢，便可能忽略藥物劑量、過敏史、症狀有無等方面的錯誤。絕不可將草稿檢視簡化為形式上的簽章程序。

導入AI書記的目的，在於減輕病歷撰寫負擔的同時，維持所需資訊的準確性。醫療機構必須綜合評估撰寫時間、修改負擔、與患者的對話品質以及記錄品質。同時應建構相應的業務流程與檢視環境，使減少重複輸入的效果能轉化為診療品質的改善。

1.2 耶魯與克利夫蘭醫學中心的導入案例

克利夫蘭醫學中心的2025年營運介紹中，說明了所選產品的開發商以及實際病歷生成的過程。耶魯研究團隊的調查結果，可於下一節開頭介紹的原論文中查閱。

<https://consultqd.clevelandclinic.org/less-typing-more-talking-how-ambient-ai-is-reshaping-clinical-workflow-at-cleveland-clinic>

要將AI書記應用於全院，除了產品本身的效能外，還必須籌備使用者培訓、技術支援以及各科別的專用表單格式。即便在小規模測試中運作順暢的功能，當眾多醫護人員將其應用於不同的業務時，亦會浮現出新的需求。

耶魯紐黑文醫療體系（Yale New Haven Health）與克利夫蘭醫學中心（Cleveland Clinic）便是呈現此類導入過程的典型案例。耶魯參與了評估醫護人員工作負擔的研究，而克利夫蘭醫學中心則公開了產品比較與大規模普及的歷程。

耶魯研究所採用的Abridge是由美國Abridge公司開發的服務。該公司官方發布上市的日期為2020年10月6日，當時其重點在於讓患者記錄診療對話並再次確認服藥說明等服務。後續由醫院導入的醫護專用記錄產品，與早期的患者端服務在功能範圍上並不相同。必須將公司創立、服務發布與特定醫院導入視為互不相同的獨立事件加以區別。

耶魯的導入與工作負擔評估

耶魯研究團隊針對多個醫療體系的醫護人員，調查了使用 AI 書記（Abridge）前後的工作體驗。雖然報告指出職業倦怠（burnout）與記錄負擔有減輕的趨勢，但由於屬於導入前後的對比，因此無法斷定所有的變化全然是 AI 帶來的效果。

在耶魯參與的另一項多機構研究中，亦顯示出文書處理時間的減少。然而，下班後的工作時間並未隨之同步縮短。此結果表明，評估引進產品的效果時，必須同時從記錄速度與實際工作環境兩個面向進行評估。

耶魯在將環境感應AI擴大應用至診療過程中的同時，亦檢討了住院醫師培育與臨床實習教育的問題。即使仰賴記錄草稿，醫學生與實習／住院醫師仍必須保有親自解讀內容與撰寫病歷的能力。防止技能退化（de-skilling）的教育訓練，在導入之後依然是不可或缺的課題。

克利夫蘭醫學中心的產品遴選

克利夫蘭醫學中心在試營運了數款AI書記後，最終選擇了 Ambience Healthcare的產品。該中心讓多元專科領域的醫護人員參與其中，檢視其是否契合實際業務需求。

該醫院所選擇的開發商為美國Ambience Healthcare。根據該院2025年8月14日的介紹，在徵得患者同意後，透過智慧型手機應用程式蒐集對話，隨後在Epic電子病歷系統中生成醫療文件與診後摘要。2025年是確認該院導入進程的時間點，並不意味著該產品本身

是在該年度首次開發。醫院承擔治療與病歷記錄的責任，而開發商則提供支援該項業務的軟體。導入對象不僅包括對新技術高度關注的使用者，亦納入了記錄工作負擔沉重的使用者。

試營運是比較各項功能並確認使用者體驗的過程。隨後該院將推廣範圍擴大至整體門診診療，並報告在相對較短的時間內便有數千名醫護人員開始使用。然而，機構與供應廠商的營運報告，仍必須與獨立的臨床效果評估嚴格區分。

大規模導入的治理體系

治理（governance）是制定決策與安全標準的架構體系。克利夫蘭醫學中心構建了由管理高層、臨床領導層、醫療資訊、IT以及醫療費用申報等相關部門共同參與的架構。因為即便是病歷撰寫工具，亦會對多項業務產生連鎖影響。

例如，若草稿中的病名表述發生改變，不僅會影響下一位醫師的判斷，還可能波及費用申報人員的編碼分類。因此，醫院必須規範的事項涵蓋了哪些項目可自動填入、由誰進行最終確認，以及發現錯誤時應通報至何處。某個部門認為便捷的功能，在另一個部門可能反而增加了核對工作，因此必須通盤檢視整體業務流程。

教育訓練被列為正式啟用前必經的程序。醫院劃分不同的診療領域逐步推動，並透過線上培訓指導基本使用方法與檢視責任。在情況較為複雜的專科中，需要額外的時間來調整模型與表單格式。

使用者支援則用於解決設定、草稿修改以及醫院政策等方面的問題。當有多位醫護人員共同會診，或需要特殊表單格式時，僅憑產品的預設配置可能有所不足。

院方亦建立了反映營運資料與使用者意見的反饋改善機制。持續追蹤各專科的培訓成效與使用現況，並透過使用者交流會完善病歷格式。這代表在產品交付後，院方仍持續因應實際使用條件進行微調。

使用範圍與滿意度的解讀

機構報告顯示，在許多診療過程中皆已生成AI草稿，使用者亦體驗到了便利性。然而，初次使用與持續使用有所不同，僅憑高滿意度這一事實，並無法證明記錄的準確性或離職率的改善。

亦必須考量各科別之間的差異。對話的長度與複雜度，以及所需的病歷格式不同，修改草稿的負擔也大不相同。在門診所確認的效果，並不能直接推定適用於急診室、住院病房或護理工作。

薩特醫療（Sutter Health）同樣是將AI書記從門診推廣至急診與住院領域的案例。此一趨勢顯示使用範疇正日益擴大，但累積的病歷件數與醫療品質的改善仍須分別評估。

導入案例的共通意義

耶魯的案例凸顯了評估醫護人員工作體驗的必要性，克利夫蘭醫學中心的案例則展現了支援大規模導入之組織架構的重要性。除了技術功能本身之外，產品遴選、培訓、支援與格式改良，皆會對實際應用產生實質影響。

要穩定運作AI書記，醫護人員必須能夠確認草稿，且在發生問題時能夠獲得支援並進行修正。在擴大導入規模的過程中，同步審視病歷品質與檢視負擔至關重要。

1.3 醫護人員倦怠與下班後病歷記錄工作

耶魯研究團隊等發表的 JAMA Network Open 原文論文同時呈現了導入前後的問卷調查與倦怠（burnout）估計值。在閱讀時若確認參與人數、納入分析的人數以及有無對照組，會更容易理解這些數值的意義。

<https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2839542>

導入 AI 醫療書記（AI scribe）的目的不僅在於縮短記錄時間。減輕記錄業務帶來的負擔，讓醫護人員能專注於患者診療，也是重要的目的。延續至診療結束後的文書工作會加劇醫護人員的疲勞，並使工作與生活難以取得平衡。

倦怠（burnout）是指因工作造成的竭力與耗竭狀態。對醫護人員而言，可能表現為持續疲勞、工作意願低落、與患者產生情緒上的疏離感等。其原因多種多樣，包括工時、人力不足、組織環境等，而撰寫病歷記錄也是工作負擔的因素之一。

記錄負擔與倦怠的變化

在耶魯研究團隊主導的美國 6 個醫療體系調查中，導入 AI 醫療書記後，倦怠與記錄工作負擔呈現減少的趨勢。醫護人員使用的是同一家公司的 AI 醫療書記（Abridge），並在導入前與使用 30 天後對工作體驗進行評估。倦怠估計比例從 51.9% 降至 38.8%。

此數值並非將所有調查參與者直接計算得出的比例，而是校正了多項特徵的模型估計值。雖然完成前後問卷的參與者有 263 名，但納

入該校正分析的人數為 184 名。這是因為問卷作答的事實與具備特定分析所需的所有項目並不相同。然而，由於這是在沒有對照組的情況下比較導入前後的問卷結果，因此無法將所有變化斷定為人工智慧的效果。

這項調查的意義在於，評估醫療人工智慧的基準正從記錄速度擴大到醫護人員的工作環境。這意味著不僅關注病歷記錄的撰寫速度有多快，也開始一併檢視診療過程中是否更容易集中注意力、結束工作後是否仍有文書工作延續等面向。

威斯康辛大學醫院（UW Health）與洛杉磯加州大學醫療體系（UCLA Health）的研究也報告了使用 AI 醫療書記後，記錄負擔或倦怠獲得改善的趨勢。不過各項研究的工具、使用期間與工作環境皆有所不同，難以期望達到相同程度的效果。

工作負擔減輕的過程

當人工智慧記錄對話並撰寫初稿時，醫師便可減輕在聆聽患者說明的同時還必須輸入資料的負擔。在螢幕與患者之間來回交替注視的次數減少，也更容易專注於對話。

反覆在「聆聽診療內容並進行判斷」與「記錄這些內容」之間切換的過程也會帶來負擔。這類作業切換伴隨著語境切換（context switching）成本。當人工智慧負責記錄初稿時，醫師便有餘裕專注於檢視患者的症狀與治療方向。不過，這種轉變如何具體轉化為倦怠減輕的路徑，仍需要進一步研究。

下班前能否完成病歷記錄也至關重要。工作時間外的文書作業減少，就更容易確保休息時間。然而，撰寫病歷記錄的時間縮短，並不代表下班後的工作量也會同等幅度減少。這是因為可能仍有其他行政業務未完成，或是需要花費時間檢視 AI 初稿。

感受負擔與實際工作時間

醫護人員感受到的負擔與系統中記錄的工作時間需要區分看待。在分析多家醫療機構電子病歷（EHR）使用資料的研究中，文書撰寫時間雖然減少，但工作時間外的工作時數並未顯著下降。波士頓的醫院體系中，撰寫病歷記錄的時間同樣有所縮短，但在工作時間外完成記錄方面並無差異。

這種差異顯示，人工智慧的效果無法單憑時間的減少來解釋。醫護人員可能因為重複輸入的負擔減輕而感到工作變得輕鬆，相反地，修改 AI 初稿的部分時間也可能未被納入測量範圍。因此，必須同時檢視使用者體驗與實際工作時間。

問卷調查雖然能了解醫護人員感受到的負擔，但會受到受訪者的期望或當時工作狀況的影響。電子病歷的登入記錄雖能呈現實際使用軌跡，卻可能無法精確區分開著螢幕與患者對話的時間。與其將任一項測量值視為絕對標準，不如確認這兩項資料是否指向同一個趨勢。

長期使用時的效果目前也尚不明朗。在導入後的多個時間點進行問卷調查的研究中，文書工作負擔即便減輕，倦怠感也並未隨之改善。

倦怠除了記錄業務外，還受到組織與工作環境的影響，因此很難僅靠導入人工智慧來徹底解決。

初稿檢視這項新工作

導入 AI 醫療書記後，醫護人員的記錄工作有一部分從直接親自撰寫轉變為確認與修改初稿。即使初稿是以自然的語句撰寫，仍須確認藥物劑量、有無症狀以及檢查結果是否正確無誤。口譯或轉譯過程中的錯誤也有可能直接反映在記錄中。

醫院在縮短記錄時間的同時，也必須考量檢視所需的業務量。需要建立相應的營運機制，例如提供符合各診療科別的表單格式、讓錯誤能輕易修正，並明確規範檢視的責任歸屬。

檢視時間必須納入門診診療行程中。若僅因初稿生成迅速便將該時間排滿其他患者的預約，病歷記錄的確認就可能再次被推遲至下班後。醫院在導入後也必須觀察記錄完成的時間點與修改量。醫護人員省下來的時間最終被用在何處，並非單靠軟體的速度就能決定。

AI 醫療書記展現了減輕醫護人員記錄負擔的潛力。若要將其轉化為持續改善的工作環境，在導入技術的同時，還必須檢討業務流程、教育培訓與人力配置。在維持病歷記錄準確性的同時，確保醫護人員能有充裕時間投入患者診療與休養，才是導入的真正目標。

第2章 臨床決策支援（CDSS）與診斷



2.1 基於大型語言模型（LLM）的抗生素處方支援

中國首都醫科大學與清華大學附屬醫院等研究團隊發表的原文論文中，記錄了將患者病歷、診療指引與處方規則相互串聯的過程。閱讀時可留意這是一項在投入實際處方前先以過往資料進行評估的研究。

<https://medinform.jmir.org/2026/1/e98207>

抗生素處方是一項必須同時考量感染原因與患者狀態的臨床判斷。醫師會依據疑似感染的細菌、感染部位、嚴重程度以及當地的抗生素抗藥性表現型態來選擇藥物。有時還需要根據肝腎功能、年齡及併用藥物來調整劑量或給藥間隔。

臨床決策支援系統（CDSS, Clinical Decision Support System）能為此類判斷提供所需資訊。其運作方式是確認患者的檢驗檢查結果與診療指引，並針對藥物選擇或劑量提出需要審核的事項。人工智慧的作用在於協助醫師滴水不漏地審視各項條件。

此時輸入系統之資訊的時間點至關重要。即使是同一位患者的檢驗檢查，入院當時的結果與數天後的結果所代表的意義也可能不同。推薦藥物的畫面上必須清楚呈現是以哪些檢驗與病歷記錄為依據。僅僅核對患者姓名是否正確是不夠的。若遺漏了最新的檢驗結果，或者殘留有其他診次所使用的藥物資訊，即便看似合理的回答也可能與當前的處方判斷產生偏差。

從規則基礎系統到語言模型

1970 年代由史丹佛大學開發的 MYCIN 是一套用於感染診斷與抗生素推薦的早期專家系統。由於它是依據專家整理的規則提出結論，因此容易解釋判斷過程。然而，這類系統必須不斷補充規則，且在處理以自由敘述撰寫的病歷記錄時存在局限。

MYCIN 是以史丹佛大學的愛德華·肖特利夫（Edward Shortliffe）等人為中心進行研究，並於 1976 年的著作中進行了系統性整理。與當今的生成式人工智慧不同，它是將專家撰寫的條件與結論相互連結的方式。例如，當滿足特定預設條件時，便引導檢驗某個特定判斷。MYCIN 這一歷史性的研究系統與近期的語言模型產品，即便同處於 CDSS 的範疇內，其技術起點也截然不同。

其後，機器學習模型被用於從患者資料中尋找模式並預測風險。只不過有時難以解釋預測的原因，若要應用於處方判斷，仍需要進一步審查。

大型語言模型（LLM）能夠閱讀病歷記錄，並以語句說明推薦的理由。其與既有系統的不同之處在於，醫師可以用自然語言進行提問，並檢視回答的依據。然而，即便解釋讀起來自然流暢，也不代表處方就必定準確。模型仍可能產生與事實不符的幻覺（hallucination）、遺漏劑量調整條件，或是隨著提問措辭的不同而改變回答。

準確性與一致性的問題

義大利感染內科研究團隊向多個語言模型提供感染病例並比較了抗生素推薦結果。各模型之間的效能差異極大，在後續評估中也出現了錯誤或不必要的推薦。研究同時確認了當針對同一病例重複提問時，會出現回答不一致的問題。

在比較提供文獻實證的醫療專用工具（OpenEvidence）與通用對話型模型（ChatGPT-5.2）的實驗中，給出更準確答案的工具並不一定更加具備一致性。準確性指的是答案是否適當，而一致性則是指在相同條件下能否維持相近的答案。在處方支援中，必須同時檢視這兩項條件。

OpenEvidence 是同名美國企業所提供的醫療資訊服務，而 ChatGPT 則是 OpenAI 的對話型服務。論文中所使用的模型版本應與該服務整體加以區分。這是因為一旦服務發生變更，檢索資料與回答方式也可能隨之改變。雖然在 2023 年就能查到 OpenEvidence 的企業發布資訊，但用於比較實驗之版本的最初提供日期，則是獨立於論文中所載測試時間之外、需要另外確認的事項。

患者資料、診療指引與處方規則的結合

中國研究團隊利用肺炎住院患者過往的電子病歷，對基於語言模型的抗生素處方支援系統進行了評估。

該系統是由首都醫科大學、北京清華長庚醫院、北京友誼醫院等聯合研究的成果，並非以特定商業產品名稱上市的案例。預印本於

2026 年 4 月 14 日公開，經同儕審查的論文則於同年 8 月 4 日發表。這並不代表醫院從這些日期開始對所有患者全面應用該系統。其核心在於，並非僅向模型提供患者狀況，而是同時提供了相似患者的處方、診療指引以及醫師制定的規則。

第一項資訊是相似患者的病歷記錄。將患者病歷轉換為數值陣列的向量（vector）後，檢索與當前患者相似的病例。在此過程中會使用嵌入（embedding）模型。雖參考過往病例的處方，但能否直接套用於當前患者仍須另外判斷。

向量是用數值表示多個特徵的清單。電腦藉由比較這些數值之間的關係來找出相似的記錄。語句相似與臨床上屬於同一類患者是兩回事。舉例來說，即使症狀描述相似，但若過敏史或器官功能不同，就很難遵循相同的處方。這也是為何需要對檢索結果重新套用各患者專屬限制條件的原因。

第二項資訊是整理診療指引的知識圖譜（knowledge graph）。透過將疾病、病原菌、藥物、禁忌症等關係結構化，讓模型能夠檢索出所需的依據。例如，可以將特定狀態下應避免使用的藥物或應調低劑量的條件相互關聯。

第三項資訊是臨床醫師整理的處方規則。明確規範根據嚴重程度選擇藥物、不可併用的藥物，以及依據肝腎功能進行劑量調整等原則。語言模型參考此項資訊，建議藥物名稱、劑量、給藥次數與給藥途徑，並說明其依據。醫師則同時審閱建議與依據。

在此架構中，將醫院的診療經驗與指引相連結的設計，與語言模型的通用知識同等重要。即便模型生成了回答，判斷所需的資料與限制條件仍必須由醫療機構有系統地提供。

處方支援的成果與局限

在利用肺炎患者病歷進行的評估中，同時提供相似病例、指引與處方規則時，處方吻合度高於基礎語言模型。當剔除部分資訊時效能便會下滑，其中相似患者檢索與臨床醫師制定的規則發揮了關鍵作用。

該研究以 F1 分數與傑卡德準確度（Jaccard accuracy）來評估推薦處方與基準處方的吻合程度。消融實驗（ablation study）是透過逐一排除構成要素來檢視各要素貢獻度的方法。這類評估有助於比較系統設計，但無法直接反映對實際患者的治療效果。

F1 分數同時反映了是否滴水不漏地找出所需項目，以及推薦的項目是否恰當。傑卡德指標則衡量兩個項目集合中重疊部分的比例。即使藥物名稱符合基準，劑量也可能錯誤；或者即使與基準處方不同，在醫學上也可能是可被接受的替代方案。因此在判讀這些數字時，必須一併檢視究竟將什麼定義為標準答案，以及錯誤是由誰來進行審查。

在肝腎功能低下的患者與高齡患者中，推薦效能相對較低。在檢視錯誤案例時，也仍殘留有未能充分考量器官功能的高風險推薦。這正是即便提供了劑量調整規則，也絕不能省略醫護人員核對確認的原因。

這項研究屬於利用過去記錄評估系統的階段。在實際診療中處方會如何改變、副作用或治療結果是否有所改善，仍需進一步確認。由於各家醫院的病患構成、抗藥性菌株表現型態及處方慣例各不相同，因此導入其他機構時也需要進行驗證。

抗生素管理與醫療團隊的角色

抗生素管理計畫（antimicrobial stewardship）是妥善選擇並給予抗生素的管理活動。基於語言模型的 CDSS 在此過程中能整理相關資訊，並為提出處方依據與注意事項做出貢獻。

在實際導入時，必須妥善架構，以便將推薦結果與診療指引及病患狀況進行對照。醫療團隊會確認肝腎功能、過敏史、併用藥物等個別病患條件，並決定最終處方。抗生素處方支援的發展方向，在於提升語言模型效能的同時，串接可靠資料、管理規則，並改善醫療團隊的審核流程。

2.2 醫療世界模型與 U2-Med 的診斷與檢查請求

開發商雲知聲的技術介紹分別說明了 U2-Med 與其訓練環境 U2-Med-World 的角色。這份資料屬於企業的研究發表，因此閱讀時應與臨床試驗結果加以區分。

<https://unisound.com/en/news/165>

診斷並非僅憑一次提問就能結束的工作。醫師在聆聽病患症狀、確認病歷與服用藥物後，會視需要進行檢查。一旦獲取新資訊，便會重新審視可能的疾病與後續處置。在醫療人工智慧研究中，嘗試實現此類循序判斷過程的研究也日益增加。

過往的醫療人工智慧評估多半採取一次提供整理好的病患資訊並要求選出正確答案的方式。然而在實際診療中，必須決定先確認哪些資訊。醫療世界模型為了縮小這一差距，將病患狀態、臨床行為以及隨之而來的觀察結果連結起來。

醫療世界模型的概念

世界模型（world model）是在當前狀態下採取特定行動時，預測會出現何種結果的模型。在醫療領域中，可用於模擬隨問診、檢查、治療等臨床行為所獲得的資訊或病患狀態的變化。

中國人工智慧企業雲知聲（Unisound）於 2026 年 7 月 17 日宣布推出 U2-Med，並在同年 10 月 21 日的發表會中說明了其與 U2-Med-World 的連結架構。U2-Med-World 是表現病患狀態與臨床行為關係的虛擬環境，而 U2-Med 則是在該環境中選擇提問與使

用工具的代理（agent）。該公司說明，透過此架構能學習在多次對話中修正判斷的能力。

U2-Med-World 依時間順序整理病歷、對話、檢查、影像、生命徵象與處方資訊。當代理選擇提問或檢查時，環境會提供相關資訊，代理則以此為基礎決定下一步行動。此時，防止後續確認的資訊混入先前判斷的資訊洩漏（data leakage）至關重要。

例如，若從一開始問診就出示寫有正確診斷的出院摘要，人工智慧便無需發揮檢查選擇能力就能猜中答案。在循序評估中，應僅提供該時間點醫療團隊實際所能得知的資訊。請求檢查時結果是否會公開，以及請求沒有結果的檢查時該如何處理，也都必須事先訂定。若此類環境設計不良，即使獲得高分，也很難認定評估了實際診療能力。

提問、檢查與治療的區分

問診與檢查是改變對病患既有認知資訊的過程。治療與疾病進展則會改變病患的實際狀態。醫療世界模型必須區分這兩者。因為進行檢查這項事實本身並不代表病患狀態好轉，開始治療也不意味著診斷已獲得確定。

在該公司發表的虛擬病患案例中，代理會根據疲倦與食慾不振等初期症狀，在症狀持續時間、伴隨症狀、過去病歷、近期服用藥物中選擇優先確認的項目。隨後根據所提供的資訊調整提問順序與判斷。針對同一位病患比較多種問診順序，便能評估其在獲取必要資訊與發現危險信號方面的表現水準。

此環境不僅能生成實際記錄中留存的診療過程，亦能生成合成軌跡（synthetic trajectory）。

所謂軌跡，是指以何種順序接續進行了哪些提問與行動。在單一病患的記錄中，僅能觀察到實際選擇的路徑。當電腦生成其他路徑進行比較時，必須標示出由觀察到的事實中究竟能確認到何種程度。因為即使生成的文句再詳細，也不代表實際觀察有所增加。合成軌跡是指在相同的初始狀態下，選擇不同提問或檢查時的虛擬診療過程。雖然可用於比較多種選擇的學習資料，但並非能用來證實未實際施行之治療效果的資料。

學習整個對話過程的方式

U2-Med 的學習被說明為在虛擬環境中生成對話，評估整體過程後，篩選合適案例用於後續學習的方式。除了最終診斷外，資訊收集、風險識別與溝通流程也是評估對象。

該公司介紹的評分量表強化學習（Rubric RL）是將醫學準確性或資訊完整性等細部標準反映於學習中的方式。線上偏好蒸餾（OPD）則是將透過評估所選出的問診策略反映於模型學習中。這種方法具有將評估範圍從單一回答的準確度擴展至多階段診療過程的意義。

效能評估與實際診療的差異

雲知聲宣布 U2-Med 在醫療對話評估 HealthBench 中獲得高分。HealthBench 是評估是否詢問了必要資訊、是否辨識出危險信號，以及是否進行了準確且安全說明的測試。

然而，該公司發布的測試成績與實際診療效果必須加以區分。即便在評估環境中能給出良好回答，也無法保證病患與醫療團隊在實際使用時會出現相同結果。U2-Med 的臨床效果、檢查費用以及對病患安全的影響，必須在實際診療環境中另行驗證。

檢查選擇與資訊價值

檢查雖能提供診斷所需資訊，但也伴隨著費用、風險與病患的不適。因此，必須考量下一項檢查能在多大程度上改變判斷或治療。解釋此一概念的便是資訊價值（value of information）。

檢查順序會根據緊急程度、預期疾病、檢查準確度以及結果對治療的影響而有所不同。人工智慧若要支援檢查選擇，必須同時考量獲取更多資訊的益處與減少不必要檢查的需求。然而，僅憑公開說明，難以得知這些一般原則是如何在 U2-Med 的內部運算中實現的。

類似的研究方向

微軟的 MAI-DxO 是一個由多個代理請求額外資訊並進行診斷的架構。在名為 SDBench 的評估環境中，系統會逐步審視複雜病例，同時選擇問診與檢查資訊。已公開的研究展現了透過循序資訊收集進行診斷的可能性，但受控環境下的評估結果不能直接視為實際診療成效。

MAI-DxO 是微軟研究團隊打造的診斷協調架構。相關研究於 2025 年 6 月 27 日首次在 arXiv 公開。SDBench 則是測試該架構與多個模型的評估環境名稱。將系統與測試加以區分，便不會混淆開發了何種人工智慧與使用何種題目進行測試。

ClinVerse 是以實際電子病歷為基礎建構的多代理診斷環境。扮演醫師角色的代理從初期症狀出發，透過問診與檢查收集資訊並做出診斷。該環境同時評估資訊收集、檢查選擇以及診斷準確度。這項工作同樣是為了在虛擬環境中進行學習與評估的研究，應與實際導入診療加以區分。

ClinVerse 的論文由 Sirry Chen 等研究團隊於 2026 年 10 月 26 日首次在 arXiv 上公開。這並不代表已推出能在醫院為病患診療的成熟產品。將病患角色與檢查結果提供角色分開後，研究人員便能在控制資訊公開順序的同時，比較扮演醫師角色的代理所做出的選擇。

U2-Med-World、SDBench 與 ClinVerse 皆著眼於研究人工智慧收集額外資訊並修正判斷的過程。這表明醫療人工智慧正從單純給出正確答案的工具，發展為支援部分診斷流程的系統。

導入所需的檢討

在支援檢查選擇的人工智慧中，必須同時控管過度檢查與必要檢查的遺漏。若過度建議檢查，會增加費用與病患負擔；若遺漏必要檢查，則可能導致診斷延誤。

目前的自主檢查請求主要是指人工智慧在虛擬環境中建議候選檢查的功能。這並不意味著其在實際醫院中獨立開立檢查處方的權限已獲確認。在臨床應用中，需要由醫療團隊審查並批准推薦依據及病患狀況的程序。責任範圍也必須根據許可狀態、使用方式、審查程序、事故原因以及各國法律體系來判定。

醫療世界模型的意義在於，不僅能研究診斷結果，還能研究獲取資訊與修正判斷的過程。未來的課題在於確認在虛擬環境中學習到的能力，是否在實際診療中也能準確且安全地運作。

2.3 AMIE 與 MIRA 的診斷推理

將 Google 初期的 AMIE 介紹與 MIRA 的 Nature 原論文對照閱讀，便能確認醫療對話研究與電子病歷內的行為研究有何不同。這兩份資料皆說明了研究的評估環境與局限性。

<https://research.google/blog/amie-a-research-ai-system-for-diagnostic-medical-reasoning-and-conversations/>

<https://www.nature.com/articles/s41586-026-10675-5>

診療指引（clinical practice guideline）是整理了各疾病檢查與治療相關研究實證的建議方針。指引因領域而異，並隨新實證進行修訂。醫療團隊在確認符合病患狀況之建議的同時，也必須考量醫院的診療條件與藥物使用條件。

Google 的 AMIE 與德國研究團隊的 MIRA 是支援此類工作的醫療人工智慧研究。AMIE 著重於在與病患對話的同時審視診斷與管理計畫，MIRA 則研究在電子病歷中執行檢查與處方等臨床行為的方式。這兩個系統表明醫療人工智慧的功能正從提供資訊擴展至支援診療過程。

AMIE 的醫療對話與管理計畫

AMIE（Articulate Medical Intelligence Explorer）是 Google 的研究用醫療對話人工智慧。

Google Research 於 2024 年 1 月 12 日正式介紹了 AMIE。診斷對話研究率先公開，而擴展至疾病管理的研究介紹則於 2026 年 6 月

17 日發布。AMIE 是研究系統的名稱，而其中使用的 Gemini 則是 Google 開發的基礎模型系列名稱。若將基礎模型與針對醫療任務構建的系統區分開來，便更容易掌握研究的演進脈絡。初期研究評估了與扮演病患角色的參與者透過文字對話進行診斷的能力。後續研究則將範圍擴展至歷經多次診療確認症狀與治療反應並調整計畫的疾病管理（disease management）。

在此過程中共有兩個代理參與。對話代理向病患提問並說明必要內容；管理推理（Mx）代理則根據對話內容審核診療指引與藥物資訊，並擬定管理計畫。這是一種在維持與病患對話的同時，分工分別審查相關依據的架構。

例如，當病患陳述服藥後遭遇的不適時，對話代理會進一步詢問發生時間點與表現症狀。管理推理代理則將獲取的資訊與相關指引進行對照，制定待審查的計畫。此範例旨在展示兩種角色的差異。即便劃分了角色，也不代表彼此的錯誤會自動獲得修正，因此必須保留依據不足之處，以便由人工進行確認。

Google 說明，該系統運用 Gemini 的長上下文處理能力來對照大量的診療指引與藥物資訊。評估中亦使用了英國國家處方集（BNF）資料。在虛擬診療研究中，AMIE 的管理計畫在指引符合度與計畫具體性等方面展現了潛力。然而，模擬對話的結果並不代表對實際病患的治療效果。

實際病患的診前問診

2026 年 The Lancet 刊登了一項 AMIE 參與實際病患診前問診的研究。該研究由 Google 與貝斯以色列女執事醫療中心（BIDMC）研究團隊共同進行，病患在就診前於家中與 AMIE 透過文字對話。AMIE 詢問症狀與病歷，並整理可能的診斷。醫療團隊在診療前即可確認對話內容與摘要。

在這項研究中，人工智慧並未獨立施行治療。醫師即時監督對話，並能依據既定的安全標準介入。部分醫療團隊成員評價認為摘要對就診準備有所幫助，但亦有必要補充修正錯誤的資訊與說明。

這個案例的意義在於測試了醫療對話人工智慧在實際看診準備過程中能發揮何種作用。若能整理問診內容，醫師便能專注於面對面看診時需要額外確認的事項。然而，由於這是在監督下獲得的結果，因此並未確認在沒有醫護人員的情況下使用時的安全性或治療效果。

MIRA的電子病歷串聯

MIRA（Medical Intelligence for Reasoning and Action）是由德國海德堡大學醫院國家腫瘤疾病中心與德勒斯登工業大學等研究團隊開發的代理。其研究不僅止於解釋患者資訊，更進一步探討在既定權限內存取電子病歷（EHR）並建構臨床行動的方式。

該研究是在與實際醫院系統隔離的沙盒（sandbox）環境中進行。沙盒是為了避免實驗影響實際業務而建構的受控環境。虛擬EHR是依照醫療資訊交換標準FHIR所設計。

MIRA的Nature論文公開日期為2026年6月17日。FHIR是用於以既定結構交換患者、檢查結果、藥物等資訊的標準，並非人工智慧模型本身。代理產生建議進行檢查的語句，與向系統輸入具備既定項目的檢查請求是兩回事。後者必須滿足患者識別、項目選擇、權限確認等附加條件。

MIRA會與扮演患者角色的代理對話以確認病歷，請求血液檢查與影像檢查等，並根據所提供的結果檢討鑑別診斷。藥物處方或住院計畫也以可在EHR中處理的行動格式撰寫。這些行動是在虛擬系統內執行，並非針對實際患者進行檢查或處方實施。

患者代理被設定為根據公開重症與急診資料庫MIMIC-IV的實際記錄來回應。研究團隊在此環境中評估了急診室的多項診療過程。即使使用了實際記錄，也應與直接與患者互動的臨床試驗加以區別。

評估臨床行動的基準

MIRA研究不僅檢視了診斷準確度，也探討了治療計畫與藥物安全。檢討項目包含了依腎功能調整劑量、藥物交互作用、過敏、心電圖QT風險等。這顯示在評估臨床代理時，除了最終診斷外，還必須檢視得出該診斷的過程與後續處置。

在虛擬環境中，雖然在多項診斷與治療計畫項目中展現了潛力，但各疾病間的效能也存在差異。在實際醫院中，會發生資訊遺漏、記錄錯誤、業務中斷等更多樣化的情況。因此在導入實際診療之前，必須另行評估安全性與運作模式。研究團隊亦強調，醫療人工智慧代理不應朝著取代醫療人員的方向進行設計。

兩種系統的角色與導入課題

AMIE以與患者的對話及管理計畫為核心發展，MIRA則以與電子病歷連結的臨床行動為核心發展。理解患者說明的色彩與能力，以及將其準確連結至業務處理的能力，對醫療人工智慧而言皆不可或缺。儘管未來這些功能有連結的可能，但並不意味著這兩種系統已作為整合診療體系獲得驗證。

在導入時，人工智慧參考何種指引必須明確。因為國際建議、醫院可用藥物以及保險標準可能有所不同。參考資料若有修訂，系統也必須適度反映。

醫療人員的審核權限與責任也必須釐清。即使是符合指引的建議，也可能忽略個別患者的特殊例外，因此醫療人員必須能夠確認根據並加以修改或拒絕。區分研究階段的系統與獲得核准的臨床產品亦屬必要。AMIE與MIRA在這些研究中均被評估為研究用系統。

如果說人工智慧書記官支援病歷記錄的撰寫，那麼AMIE與MIRA則將支援範圍擴展至診療資訊的解讀、管理計畫與臨床行動。若要讓這些變化轉化為實際醫療的改善，除了技術效能外，還必須驗證醫療人員的監督、資訊更新以及與醫院業務的串聯。

Part 2. AI藥物開發與數位分身

第3章. 生物基礎模型與藥物開發標靶



3.1 醫療多模態基礎模型

Arc Institute的Evo 2公開聲明與Google DeepMind的AlphaGenome介紹，說明了基因體人工智慧的輸入資料與預測對象。從以下資料中亦可確認兩種模型的開發機構與首次公開時間。

<https://arcinstitute.org/news/evo2>

<https://deepmind.google/blog/alphagenome-ai-for-better-understanding-the-genome/>

在醫療與生命科學領域中，會利用多種資料來分析同一名患者或組織。基因體展現遺傳特徵，組織病理影像呈現細胞與組織的形態。空間轉錄組測量組織內不同位置的基因表現，診療記錄與判讀報告則提供臨床脈絡。若將這些資料一併分析，便能獲得單靠一種資料難以掌握的資訊。

多模態基礎模型（multimodal foundation model）是透過大規模學習多種格式的資料，並應用於各項分析任務的模型。在醫療領域中，目前正在進行串聯病理影像與分子資訊、影像與臨床病歷等研究。以下案例主要是結合了兩三種資料的模型。

基礎模型的「基礎」一詞，意味著先學習可在多項任務中通用的表徵。這與僅重複學習單一疾病標準答案的模型有所區別。在預訓練之後，可根據所需任務進行額外微調學習，或附加獨立的預測器。然而，僅僅因為學習了大量醫療資料，並不代表就能理解所有檢查與所有患者。必須先確認輸入了何種類型的資料，以及為了產出何種結果而進行訓練。

醫療資料的種類

基因體（genome）是包含DNA鹼基序列的遺傳資訊。用於分析遺傳變異與疾病有何關聯。然而，單憑遺傳資訊並無法完全得知特定組織中實際上表現了哪些基因。

組織病理（histopathology）切片是將切成薄片的組織染色後觀察的影像。在代表性的H&E染色影像中，病理科醫師會確認細胞核的形態、組織結構、免疫細胞的分佈等。由於數位玻片以高解析度涵蓋大範圍組織，因此需要分析的影像尺寸也相當龐大。

空間轉錄組（spatial transcriptomics）同時測量基因表現與組織內的位置資訊。不僅能分析哪些基因表現，還能分析該細胞位於何處，以及與周圍細胞具有何種關係。文字資料則包含病理判讀報告、診療病歷、疾病說明、研究論文等。

各項資料的測量對象與格式均不相同。多模態模型在處理這些差異的同時，學習資料之間的相關性。藉此可以從病理影像中推估分子特徵，或結合臨床病歷來解讀影像所見。

結合不同資料的方法

編碼器（encoder）將輸入資料轉換為數字陣列——向量（vector）。DNA序列可使用序列模型，病理影像可使用如視覺Transformer（vision transformer）等影像模型。處理空間關係時可應用圖神經網路（graph neural network），文字則使用語言模型。具體架構因模型而異。

學習資料間關係的代表性方法是對比學習（contrastive learning）。例如從同一組織取得的影像與分子資訊，相關資料的表徵會被訓練得更接近，關聯較小的資料表徵則會被區隔開來。如此形成的共同表徵空間稱為共享潛在空間（shared latent space）。可使用如 InfoNCE 等損失函數（loss function）。

損失函數是以數值呈現模型當前結果偏離學習目標程度的計算方法。模型會朝著縮小該數值的方向調整其內部參數。在串聯同一名患者的影像與分子資料時，患者識別必須準確無誤。若將不同人的檢查結果配對在一起，模型就會學到錯誤的關聯性。收集大量資料與正確進行配對必須同時並進。

交叉注意力（cross-attention）是在解讀某一種資料時，參考另一種資料相關部分的方法。例如可將影像的特定部位與臨床描述相連結。然而，單憑結合資料這一事實並不能保證效能必然提升，仍須確認究竟是哪些資訊產生了實際貢獻。

基因體模型的發展

Evo 2 是學習了多種生物龐大鹼基序列的基因體語言模型。

該模型由美國 Arc Institute 與 NVIDIA 主導，並與史丹佛大學、加州大學柏克萊分校、加州大學舊金山分校等研究團隊共同開發，於 2025 年 2 月 19 日公開。在此，語言模型讀取的基本單位並非日常語句中的單詞，而是 DNA 的鹼基。在學習序列中呈現何種模式後，進而分析新的序列或變異。研究探討了處理長序列及推估遺傳變異影

響的能力。Google DeepMind的AlphaGenome則著重於預測非編碼區（non-coding region）變異對基因調控的影響。

AlphaGenome的首次官方介紹是在2025年6月25日。它接收長DNA序列輸入，並預測與基因調控相關的分子特性。透過比較具有變異與無變異序列的預測值，可以檢視哪些功能可能發生改變。直接製造蛋白質區域之外的DNA亦可能參與決定基因在何時、在多大程度上發揮作用，因而成為研究對象。

這類模型正發展為解讀遺傳資訊功能的研究工具。然而，計算結果並不同於可用於判定患者致病性變異或遺傳諮詢的臨床依據。在染色質可及性（chromatin accessibility）等特定任務中，使用該資料訓練的專用模型可能更為合適。

病理影像與分子資訊的串聯

HistoMethyl是將病理玻片與DNA甲基化（methylation）資訊相串聯的研究案例。DNA甲基化是與基因表現調控相關的化學修飾。其方法是在訓練時同時提供影像與甲基化資料，而在使用時則透過H&E影像推估分子特徵。

該研究展現了利用病理影像優先篩檢出需要進行分子檢測患者的潛力。評估中使用了多個癌症世代（cohort，依共同條件組成的患者群體）。然而，從影像推估的結果並不能取代經驗證的分子檢測。研究團隊亦將重點放在初步篩檢而非確診。

HistoMethyl是由Dannong Wang與美國中央佛羅里達大學的Yu Tian等研究團隊發表的研究模型。在bioRxiv註冊的通訊作者單位為

中央佛羅里達大學，首次預印公開日期為2026年6月14日。從病理影像推估分子狀態時，必須區分已測量與已預測並加以記錄。即使輸出形式為相同的數值外觀，實際分析檢體的結果與模型計算出的數值，其產生過程是不同的。該研究的架構與評估條件可於以下原始論文中查閱：<https://www.biorxiv.org/content/10.64898/2026.06.11.731518v1>

Fuji提出了同時學習病理影像、RNA定序與基因體資料的方向。大量RNA定序（bulk RNA sequencing）是同時分析組織中所含多個細胞RNA的方法，而全基因體（whole genome）資料則涵蓋整個基因體的資訊。HistoMethyl與Fuji所引用的研究皆處於預印公開階段，因此應與臨床產品有所區隔。

Fuji是由美國休士頓衛理公會研究所、康乃爾大學威爾康乃爾醫學院及美國國家癌症研究所研究人員參與的研究模型。由Qing Li與Shengyu Li等人開發，於2026年8月6日以預印本形式發表。不應因名稱相同或相近而將其誤解為醫療器材公司的產品。本研究所指的Fuji是學習病理影像與分子資料之間關係的計算模型。該研究的架構與評估條件可於以下原始論文中查閱：<https://pmc.ncbi.nlm.nih.gov/articles/PMC13484847/>

空間轉錄組與文字的結合

空間轉錄組雖然能詳細分析組織內的基因表現，但在成本與測量條件上有所限制。因此，目前正在進行將相對容易廣泛取得的H&E影像與空間轉錄組相串聯，以從影像中推估各位置表現模式的研究。然而，不同平台間的解析度與測量方式差異，以及訓練資料的不足，仍是待克服的課題。

Tx-LLM與TxGemma將分子結構、蛋白質序列、細胞株資訊、疾病說明等以文字形式處理，藉以學習藥物開發的各項任務。這是一種將資料格式透過共通語言介面進行串聯的方法。

Tx-LLM 是由 Google 研究團隊等人開發的模型，其論文於 2024 年 6 月 10 日預先公開。該模型以 PaLM 2 為基礎，針對新藥研究課題進行訓練。TxGemma 是 Google 研究團隊以 Gemma 2 為基礎建立的模型系列，相關論文於 2025 年 4 月 8 日公開。即便可以同時輸入以文字表示分子結構的資料與疾病名稱等，該文字也無法涵蓋實際病患的所有生物學狀態。然而，這並非在病患層級將病理影像與空間轉錄組完整整合的模型。

影像與臨床病歷的結合

CGA-Net 結合了乳房超音波影像與形狀、邊界等標準化臨床描述。這是透過交叉注意力（cross-attention）學習兩種資料關聯性的方式。在這類研究中，結合模型不僅需要與僅使用影像的模型比較，還必須與僅使用臨床資訊的模型進行比較。這是因為僅憑輸入資料的一部分，也可能獲得重要資訊。

CGA-Net 由中國江蘇省太倉市第一人民醫院超音波診斷科的 Qin Sun 等研究團隊開發。研究論文於 2026 年 6 月 9 日發表於 Frontiers in Medicine。由於這是針對特定乳房超音波課題設計的神經網路，因此與旨在作為多領域通用基礎的模型有所區別。輸入多種資料的多模態，與可在多個任務中重複使用的基礎模型，是互不相同的分類標準。研究的架構與評估條件可於下列原論文中查閱：
<https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2026.1802570/full>

是否預先指定影像的感興趣區域（ROI, region of interest）也會影響實際應用。在研究中僅裁切病灶部位輸入的條件，與在診療現場處理原始影像的條件並不相同。此外，也必須防止評估用資訊混入訓練中的資料洩漏（data leakage）。

在子宮頸癌研究中，利用通用多模態模型推估了化學放射線治療（chemoradiotherapy）後的復發風險。然而，這是一項分析已結束之診療病歷的回溯性（retrospective）研究，且包含了治療後方能獲知的資訊。因此，不應將其解讀為治療開始前的預測工具。影像判讀過程中仍存在誤差，需要醫療團隊進行審核。

性能評估與泛化課題

大規模模型並非在所有任務中都優於既有方法。在比較單細胞基礎模型（single-cell foundation model）的研究中，應用正規化（normalization）與線性方法的基準模型也展現了高水準的性能。在基因擾動（perturbation），即預測改變基因功能或細胞環境時的反應這項任務中，亦有類似結果的報告。

因此，新模型必須與適當的基準方法進行比較。針對訓練資料中未曾包含的細胞類型或生物物種進行分布外（out-of-distribution）評估也至關重要。比起模型的規模，確認其是否符合實際分析目的才是首要任務。

在結合多種資料時，也可能產生模態差距（modality gap）。即使在共同空間中進行訓練，各類資料的表徵也可能無法充分連結。因此需要設計一種在保留不同資料特性的同時，能夠學習其關聯性的架構。

批次效應（batch effect）是因機構或實驗條件不同而產生的測量差異。各醫院的染色試劑與掃描儀互不相同，轉錄組的測量方式也有所差異。若要將在單一機構訓練的模型應用於其他機構，必須檢視這些差異，並使用外部資料進行評估。

舉例來說，假設某家醫院將特定疾病的檢體主要使用某一台掃描儀拍攝，而其他檢體則使用另一台掃描儀拍攝。即使模型辨別的是機器留下的影像特徵而非疾病本身，測試分數也可能提高。為了確認模型是否真正學會了疾病的特徵，必須在儀器與機構不同的資料中檢視其性能。這個範例說明了資料收集方式對評估造成的影響。

在精準醫療中的角色

多模態基礎模型能夠連結不同的檢查與病歷記錄，有助於理解疾病並決定分析的優先順序。從病理影像中篩選出需要進行額外分子檢測之對象的方法即為一例。

然而，計算性能與病患的治療成效必須加以區分。研究用的預測能否轉化為減少誤診或改善存活率，必須在臨床環境中進行確認。目前的事例宜理解為輔助醫療團隊與研究人員判斷的角色。由人類審查結果的人在迴路（Human-in-the-loop）架構與可靠的後續檢查是不可或缺的。

3.2 Octopus 的大腸癌生物標記 (IGF2) 探索

Christopher Baker 等研究團隊的 Octopus 預印本論文說明了將假說生成與公開生物學資料分析相結合的架構。該研究屬於利用電腦進行的候選項目探索，與新療法的臨床試驗有所區別。

<https://arxiv.org/abs/2607.16262>

在藥物開發中，決定在眾多基因與分子中優先研究哪些候選項目至關重要。人工智慧可透過分析既有的生物學資料並提出假說，來支援候選篩選過程。貝爾法斯特女王大學研究團隊的 Octopus，即是結合多個人工智慧代理與預測模型，探索與大腸癌藥物反應相關候選項目的案例。

Octopus 的正式名稱為 Multi-Scale Autonomous Discovery Engine。這項由 Christopher Baker、Tianyu Ren、Karen Rafferty、Hui Wang 與 Simon McDade 共同參與的研究，在 2026 年的預印本中被介紹。相較於指稱單一語言模型，它更接近於整合了多種分析工具與代理的研究體系名稱。

該研究所提出的候選項目是類胰島素生長因子 2，即 IGF2（insulin-like growth factor 2）。研究團隊將 IGF2 作為預測對 5-氟尿嘧啶（5-FU, 5-fluorouracil）反應的生物標記（biomarker）候選進行了檢視。生物標記是用於評估疾病狀態或治療反應的指標，與直接作為治療對象的標靶有所區別。

大腸癌治療反應的差異

5-FU 是長期應用於大腸癌化學治療的藥物。然而，即便給予相同的藥物，每位病患的反應也不盡相同。若能找出解釋這種差異的生物學特性，將有助於選擇適當的治療方案或決定後續研究的方向。

在既有研究中，通常由研究人員建立假說、分析基因表現資料，再透過細胞與動物實驗等進行驗證。由於眾多基因共同發揮作用，需要分析的候選項目繁多，且難以解釋單一基因與藥物反應之間的關聯性。

大型語言模型（LLM）雖然能以文獻與資料為基礎生成假說，但流暢自然的解釋並不保證其生物學上的合理性。為了減少生成與事實不符內容的幻覺（hallucination），必須將假說與實際資料進行對照比對。

多代理與預測模型的結合

在 Octopus 中，多個代理（agent）會分析大腸癌轉錄組資料並提出候選假說。該系統被配置為在實驗室的本地環境中運行開源小型模型。由於採用了不將病患資料傳輸至外部服務的架構，這也與研究機構的資料管理息息相關。

生成的假說會使用 XGBoost 與 SHAP 等進行評估。

XGBoost 的代表性論文由 Tianqi Chen 與 Carlos Guestrin 於 2016 年 3 月 9 日預先公開。SHAP 則是 Scott Lundberg 與 Su-In Lee 於 2017 年 5 月 22 日預先公開的預測解釋方法。這兩種工具並非

Octopus 新發明的技術，而是將既有公開的計算方法配合研究目的進行連結。XGBoost 是一種結合多個決策樹（decision tree）進行預測的機器學習方法。SHAP 則是解釋各項輸入資訊對特定預測貢獻程度的技術。整體架構由語言模型提出候選項目，並由另一個具可解釋性的預測模型檢視其與資料的關聯性。

決策樹是依照條件劃分資料的預測結構。XGBoost 結合多棵樹的結果，彌補先前預測的不足之處。SHAP 則計算在該預測中，哪些輸入促使結果提高或降低。對預測有重大貢獻的事實，並不等同於生物學上的因果原因。因為相互關聯的基因之中，可能只是其中一個基因傳遞了替代另一個基因的資訊。

檢視過程中使用了不同層級的公開資料。在細胞株資料中，探討了抑制基因功能時對癌細胞存活的影響。在人類大腸癌世代研究（cohort）中，評估了基因表現與存活的關聯性；在病患來源異種移植（PDX）小鼠資料中，則評估了其與腫瘤變化的關聯性。

將研究團隊提出的候選項目在多種資料中進行比對，可以減少僅依賴單一資料所產生的結論。然而，即便確認了資料之間的關聯性，也不代表證實了該基因直接引發治療反應的因果關係。

細胞株、病患資料與小鼠資料回答的是不同的問題。細胞株所表現出的反應是否會在人體中以相同的幅度呈現尚不可知，而與病患存活相關的特徵是否為藥物反應的原因也必須另行探討。即使多種資料指向同一方向，會增強後續研究的理由，但各項資料本身的局限性並不會因此消除。

IGF2 候選項目的意義

Octopus 透過非監督式 (unsupervised) 分析，將 IGF2 提出為優先檢視的候選項目。在既有的病患資料與動物資料中，均曾報告過 IGF2 表現與預後或藥物反應之間的關聯性。

IGF2 本身並非在此研究中首次被發現的基因。先前已有研究探討過大腸癌中 IGF2 銘印 (imprinting) 異常與疾病風險的關聯。銘印是指調控自雙親遺傳而來的兩個基因複本中特定複本表現的現象。

Octopus 的成果與其說是在於發現新基因，不如說是在於透過計算方式從眾多候選項目中篩選出值得進一步研究的對象。

若要將生物標記候選應用於治療，需要後續的驗證。必須確認 5-FU 的治療效果是否隨 IGF2 濃度而異；若欲將其開發為治療標靶，還必須檢視調控 IGF2 路徑時疾病病程是否會發生改變。

當將研究結果轉化為向病患的說明時，這項區分十分重要。預後指標是與未來病程相關的特徵，而治療反應預測指標則是辨別能從何種治療中獲得更大益處的特徵。兩者可能重疊，但並不總是相同。因此，絕不能將發現相關基因的發表，直接轉述為只要檢測該基因即可選擇用藥。

既有資料分析與新實驗的差異

Octopus 研究屬於重新分析公開資料的回溯性 (retrospective) 電腦運算研究。研究團隊並未重新將抑制 IGF2 的藥物施用於小鼠。而是利用既有的 PDX 資料，評估預測與實際觀察結果是否具有相同的趨勢。

電腦模擬分析 (in silico) 有助於縮小候選範圍並擬定實驗計畫。然而，它無法取代新的細胞實驗、動物實驗以及臨床試驗。該研究亦處於同儕審查前的預印本 (preprint) 階段，因此不應將其宣稱為已確立的治療成效。

先行研究與研究設計的進展

同一研究團隊先前曾提出情境可逆世界模型 (CIWM, contextual invertible world model)。該研究利用大腸癌細胞株與藥物反應資料，探討假設基因發生改變時預測結果會如何變化。

在這類模擬分析中，改變特定突變時可能會出現與預期不同的反應。這是因為細胞的反應是多條途徑共同作用的結果。然而，模型計算出的變化必須與在真實細胞中確認的變化加以區分。考慮到這些局限性，Octopus 採用了交叉檢視多種資料並決定候選項目優先順序的方法。

在藥物開發過程中的角色

Octopus 展示了將假說生成與資料分析相結合的人工智慧研究體系。研究人員得以檢視人工智慧提出的候選項目，並判斷後續實驗的必要性與優先順序。在本地運行開源模型的方式，在資料管理與研究環境選擇方面亦具有重要意義。

後續課題是在獨立的資料中重現結果，並透過細胞與動物實驗確認候選項目的功能。在臨床上，也必須區分它是預測治療反應的指標，還是可透過治療加以調控的標靶。人工智慧的角色並非省略這些驗證，而是在於協助有系統地篩選出有待驗證的候選項目。

3.3 細胞世界模型（Chreode）與分子設計

Chreode 研究團隊公開的論文中，整理了輸入的細胞狀態、介入條件以及狀態變化的計算方式。從中可一併了解訓練資料以小鼠胚胎為主的前提條件，以及轉移至其他任務進行測試的範圍。

<https://arxiv.org/abs/2605.28111>

在藥物開發中，除了調控特定蛋白質功能的方法外，亦採用分析疾病引發之細胞狀態變化的方法。這是因為在涉及多種生物學路徑的疾病中，很難僅憑單一標靶的作用來解釋治療反應。

轉錄組（transcriptome）代表細胞中表現的RNA資訊。與疾病相關的基因表現模式被稱為疾病特徵（disease signature），可藉此分析給藥後此模式如何改變。Connectivity Map與L1000是用於此類研究的具代表性資料與測量體系。

對細胞投藥或改變基因功能的干預措施稱為擾動（perturbation）。細胞世界模型（cell world model）或虛擬細胞（virtual cell）是試圖預測此類干預所引起細胞狀態變化的模型。可用於計算候選物質會引發何種反應，以決定後續實驗的方向。

細胞狀態變化預測

隨時間分析細胞狀態存在測量上的限制。單細胞RNA定序（single-cell RNA sequencing）在測量過程中會破壞細胞，因此難以持續觀察同一個細胞。研究人員利用在多個時間點測量的不同細胞資料來推估狀態變化過程。

RNA速率分析（RNA velocity）或神經常微分方程（Neural ODE）已被用於研究這類變化過程。然而，可能會產生計算成本以及針對新資料重新訓練的負擔。近期的模型試圖將預先訓練的表徵應用於多種細胞與分析任務中。

Chreode的方法

Chreode是以細胞狀態與干預條件為基礎，預測下一個狀態的模型。

這是美國北卡羅來納大學教堂山分校、德克薩斯大學奧斯汀分校與賓州大學研究團隊的共同研究，並於2026年5月27日預先公開。輸入內容不僅包含細胞的當前狀態，還包含假定的干預措施以及時間間隔。輸出則是在該條件下預期的下一個細胞狀態。儘管冠以世界之名，但並不意味著如實重現了細胞內的所有分子與人體整體。該名稱源自發育生物學家康拉德·華丁頓（Conrad Waddington）在解釋細胞分化路徑時所使用的克里奧德（chreode）。

此模型將狀態變化所需的計算反映於訓練過程中，並設計在推論使用時透過單一步驟的轉移運算生成下一個狀態。轉移算子（transition operator）同時表現出具方向性的變化、旋轉成分以及隨機波動。

轉移算子是將當前的數值表徵轉換為下一個狀態數值表徵的計算規則。即使在相同的起始條件下，不同細胞的反應也可能有所不同，因此難以僅憑單一固定答案來解釋變化。模型會從資料中學習這種變化模式。所謂透過單次計算得出結果的說明，並非指實際細胞的變化是瞬間完成的，而是指電腦的計算程序。

訓練中使用了小鼠胚胎的大規模單細胞資料。將細胞表示為向量（vector）的部分採用了scVI編碼器（encoder），計算狀態變化的部分則應用了DiT（diffusion transformer，擴散Transformer）系列架構。Chreode的該篇論文為預印本（preprint），即在正式同儕審查前公開的研究。

scVI的前期研究可追溯至Romain Lopez、Nir Yosef等研究團隊於2017年預先公開的單細胞基因表現模型。DiT則是源自William Peebles與Saining Xie於2022年12月19日公開的擴散Transformer研究架構。Chreode將這些現有的組成元素，依細胞狀態變化的目標進行了組合。

應用於其他任務

Chreode研究評估了在發育過程中所學習到的表徵是否對其他細胞分析也有所助益。其預測了造血與胰島分化資料中的細胞分佈變化，並為基因擾動預測模型GEARS提供了訓練後的嵌入向量（embedding）。在部分評估中，預測誤差有所降低。

GEARS是由史丹佛大學的Yusuf Roohani、Kexin Huang、Jure Leskovec等研究團隊開發的基因擾動預測方法。《Nature Biotechnology》論文於2023年8月17日線上公開，並刊登於2024年期號中。該方法結合基因之間的關係資訊與深度學習，以預測干預後的表現變化。

研究亦進行了無需重新訓練即可預測細胞分化結果的零樣本（zero-shot）評估。零樣本是指無須為該任務進行額外訓練，直接套

用既有學習表徵的方式。這些結果顯示從細胞發育中獲取的資訊有潛力對其他狀態變化分析做出貢獻。然而，特定評估中的改善並不能概括推論至所有細胞或藥物反應。

細胞擾動模型的多樣設計

SCALE處理了處理前後難以直接配對測量同一細胞的問題。這是一種將細胞視為集合處理，並預測干預後群體狀態的方式。該研究將遺傳干預、藥物、發育及免疫反應等各種條件納入分析對象。大規模擾動資料Tahoe-100M亦作為相關模型訓練與評估的研究資源而公開。

SCALE由上海人工智慧實驗室、上海大學、復旦大學等研究團隊開發，論文於2026年3月18日首次公開。8月31日的修訂版中亦說明了後續實驗。Tahoe-100M則是與之有別的資料資源。若將模型的名稱與用於訓練的資料名稱分開來看，便能釐清哪項研究提出了新的計算方法，哪項研究提供了新的觀察資料。

PT-RAG將檢索增強生成（RAG，retrieval-augmented generation）應用於細胞反應預測。在預測新干預措施的效果時，會檢索並參考過去的相關實驗。由於相同基因在不同細胞類型中的效果可能有所差異，因此會綜合考量干預的相似性與對照組細胞狀態來檢索資料。

PT-RAG的第一篇論文由隸屬於義大利羅馬薩皮恩扎大學與ISTI-CNR，以及英國劍橋大學與威爾康桑格研究所（Wellcome Sanger Institute）的Andrea Giuseppe Di Francesco、Andrea Rubbi、Pietro Liò於2026年3月7日預先公開。這是附加於現有預

測模型以尋找相關實驗的組成元件。若檢索到的資料與當前細胞條件不符，所能提供的幫助可能有限，因此必須同時評估檢索與預測。

研究架構與評估條件可於下列原論文中查閱：

<https://arxiv.org/html/2603.07233v1>

這些模型的共通點在於試圖將現有實驗的資訊應用於新條件。在實際研究中，關鍵在於模型對未包含在訓練資料中的細胞或干預措施能有多好的應對能力。

評估方法與泛化

Arc Institute的Virtual Cell Challenge是一項公開比較細胞擾動預測的評估競賽。其預測減弱基因功能的基因剔除（knockdown）後的反應，並檢驗模型在新的細胞條件下是否仍能正常運作。

相關比較研究指出，複雜的模型並非在所有指標上都優於線性基準線（linear baseline）。相對地，亦有後續研究表明，若校正評估方法，深度學習模型的優勢便會顯現。因此，難以僅憑模型規模或單一分數來評斷效能。除了表現量之外，還必須綜合評估反應方向與細胞群體分佈等多個層面。

契合目標細胞變化的分子設計

若說細胞反應預測是計算特定干預的結果，分子設計則是尋找能引發預期結果之候選物的過程。以預期的轉錄組變化為條件生成新分子的方法，稱為從頭（de novo）分子設計。

CURE利用給藥前後的轉錄組資訊擷取功能特徵，並將其連結至分子生成。轉錄組擾動特徵擷取器（TFE）建立生物學變化的表徵，擴散模型（diffusion model）則參考此表徵生成分子結構。其意義在於不僅將目標蛋白質的立體結構，亦將細胞反應作為設計條件。

CURE由中國科學院自動化研究所、中國科學院大學與新加坡國立大學的研究團隊參與開發，論文於2026年5月14日預先公開。該名稱代表該論文中的Cellular Response Engine。雖然它以預期的細胞變化作為輸入條件輸出候選分子結構，但輸出的結構是否能實際合成以及是否安全，仍需後續驗證。

JoPMol同時反映細胞狀態、以文字呈現的設計策略，以及以數值表示的化學性質。這是因為即使誘導出生物學上預期的變化，分子的化學特性也可能不合適。該方法旨在同時調控多種條件來生成分子。

JoPMol是華南師範大學與大阪大學研究團隊的共同研究，於2026年7月13日公開。例如，當研究人員同時提出預期的細胞反應與化學性質條件時，模型便會尋找符合該組合的候選物。由於在改善單一條件的過程中可能會使其他性質惡化，因此確認多項條件之間的平衡至關重要。

CURE與JoPMol的結果主要報導於電腦模擬評估中。單憑該論文的計算評估，無法視為已確認生成分子的合成可行性、動物實驗以及臨床療效。

與實驗結合時的課題

與疾病相關的基因表現變化並不全都是疾病的成因。其中一部分可能是細胞的防禦反應。因此，必須確認將表現模式轉變為接近正常狀態是否確實對治療有所助益。

培養細胞株的反應與人體內的反應也可能有所不同。組織微環境、免疫系統、藥物的分佈與代謝等都會影響實際療效。在僅測量有限基因並推估其餘基因的資料中，某些訊號亦有可能未能充分反映。

結合細胞世界模型與分子生成模型，可支援候選物探索與實驗規劃。然而，計算結果僅是後續驗證的起點。必須歷經合成候選物、在細胞與動物中確認其作用與毒性，進而進行臨床評估的過程。這項技術的發展，唯有在計算模型效能提升與透過實驗資料驗證兩者並行時，才能對藥物開發做出貢獻。

第4章 數位分身（Health Digital Twin）與臨床試驗



4.1 反映病患特徵的數位分身

新加坡研究團隊的PRECISE CURATE.AI原論文中，記載了利用病患個別資料制定劑量建議的過程，以及探討是否能應用於實際診療的結果。相較於複製整個病患的想像，更可以聚焦於該系統支援了何種決策來閱讀此文。

<https://www.nature.com/articles/s41698-025-00835-7>

數位分身（digital twin）是以電腦模型呈現真實對象的狀態，並反映新資料以預測變化或干預結果的技術。在工程領域中，長期以來被用於分析設備狀態與擬定維護計畫。在醫療領域中，則正朝著反映病患或器官特徵以模擬疾病病程、治療反應與臨床試驗結果等方向進行研究。

健康數位分身可整合基因資訊、影像、病歷紀錄與感測器資料等。然而其使用目的與實現程度相當多樣，並不意味著完全複製病患所有生物學特徵的系統。

在建立模型時，首先要縮小欲解答問題的範圍。旨在檢討藥物劑量的模型，與旨在觀察心臟瓣膜運作的模型，所需的資料有所不同。即使具備以立體畫面呈現病患樣貌的功能，也不代表能夠預測治療反應。相反地，沒有立體圖形而僅以數值與圖表呈現結果的模型，在特定預測中也可能有所助益。

模型與真實病患的連結

在一般分類中，數位模型（digital model）是以過去資料建立的模型，並不會與真實對象自動連結。數位影子（digital shadow）則是將真實對象的新資訊傳送至模型的形式。數位分身則旨在連結真實對象與模型的狀態，並將預測結果反映於決策之中。

在醫療領域中，病患、資料連結、電腦模擬病患（patient-in-silico）、使用者介面與同步化等被視為其構成要素。關鍵在於當真實病患的狀態改變時，模型亦必須反映此一變化。然而，不同領域對於數位分身的定義與實現範疇各有不同。

為了獲得可信的預測，驗證、確認與不確定性量化（VUQ，verification, validation and uncertainty quantification）至關重要。這是確認模型是否正確實現、是否符合使用目的，以及預測中存在何種程度不確定性的過程。

這三個問題必須分開確認：確認數學公式與程式是否依預期進行計算、比對計算結果是否與實際觀察相符，以及探討資料與假設的誤差對結果造成的動搖程度。即使程式沒有錯誤，若生物學假設有誤，預測仍可能不準確。僅通過單一檢驗並不代表其餘條件亦已獲得滿足。

應用範圍

腫瘤模型分析基因突變、細胞特性、藥物反應與抗藥性等。器官模型模擬心臟或肝臟等結構與功能，以及藥物或放射線的影響。在患者層面，結合病歷與穿戴式（wearable）裝置資料等，可推估狀態變化。

在臨床試驗中，可用於審視患者招募、劑量選擇與試驗設計。亦有預測未接受治療時的病程進展，以輔助實際對照組分析的方法。雖然與適應性設計（adaptive design）的結合也在研究中，但仍需針對個別用途進行相應的評估與法規審查。

資料導向模型與機制導向模型

資料導向模型從大量患者資料中學習模式。雖然適合分析複雜的影像與基因體，但在解釋判斷理由，或應用於與訓練資料不同的患者時可能存在局限。

機制導向模型將已知的生物學關係轉化為數學表達式。涵蓋探討藥物在體內分佈與消除過程的藥動學（pharmacokinetics）、探討藥物作用的藥效學（pharmacodynamics）等。其解釋相對明確，但難以將複雜疾病的所有過程皆以數學公式呈現。

混合模型結合了這兩種方法。這種方式同時利用已知機制與患者資料，以互補各自的局限。模型的選擇取決於要支援何種臨床決策。

個人化劑量調整：CURATE.AI

新加坡研究團隊的 CURATE.AI 是利用單一患者所取得的資料來調整個人化藥物劑量的方法。

主導開發的是新加坡國立大學的 Dean Ho 研究團隊。該校早於 2018 年 8 月 31 日便介紹過 CURATE.AI 的臨床研究，而 PRECISE 的結果論文則於 2025 年 2 月 21 日公開。不能因發表了新論文就認為該平台是該年才首次問世。其輸入值為該患者的給藥劑量與反應資料，輸出值則為供醫療團隊審查的個人化劑量建議。針對晚期實體腫瘤的 PRECISE 研究，探討了在卡培他濱單獨或合併治療過程中，能否將人工智慧的劑量建議應用於實際診療中。

此案例嘗試不僅考量大規模群體的平均反應，亦將患者個人的反應納入治療調整中。但由於這是一項評估可行性的研究，因此不應解讀為已證實能改善存活率或治療反應。

個人化模型同樣需要充分的觀察。必須在改變給藥劑量後測量反應，才能調整劑量與反應之間的關係。當患者的狀況或合併接受的治療發生變化時，還必須確認先前獲得的關係是否依然成立。正因如此，醫療團隊設定輸入資料與容許範圍的過程，便成為模型運作的一部分。

生活習慣管理：Twin Health

克里夫蘭醫學中心（Cleveland Clinic）進行了將 Twin Health 平台應用於第 2 型糖尿病患者管理的研究。其方式為收集血糖、體重、血壓、活動、睡眠等資料，並指引個人化的飲食與活動。

開發商為美國 Twin Health，而醫院是執行研究的機構。克里夫蘭醫學中心於 2025 年 8 月 20 日發布的內容將此計畫稱為 Twin Precision Treatment，並說明其同時提供從裝置取得的資料與真人遠距指導。因此，很難將研究中呈現的變化單獨解讀為人工智慧演算法單一項目的效果。發布日期應視為研究結果公開的時間點。

這種方法將日常生活中反覆獲取的資訊納入管理計畫中。這是與新藥臨床階段不同的生活習慣介入研究，必須將平台的功能與患者身上展現的效果分開評估。

器官模型：Living Heart 計畫

達梭系統（Dassault Systèmes）的 Living Heart 計畫（Living Heart Project）利用心臟的三維模型，研究心臟功能與醫療器材的影響。與 FDA 的合作也是在探討虛擬試驗方法的研究層面上展開的。

法國軟體企業達梭系統於 2014 年 5 月 20 日公開了 Living Heart 計畫的心臟模擬模型。這是早在生成式語言模型問世之前就已發展的計算模型範例。由於數位分身（digital twin）的範疇亦涵蓋計算物理與生理法則的模型，因此將所有數位分身皆稱為基於語言模型的人工智慧並不準確。

器官模型有助於在實際試驗之前審視設計與條件。然而，研究合作與主管機關採納為正式審查工具並不相同。若要將電腦模型作為審查依據，必須依其用途進行相應的可靠性評估。

臨床試驗模擬：OPTIVIS

韓國企業 Ophrimed 的 OPTIVIS 是一套利用反映患者特徵與疾病進展的虛擬患者來模擬臨床試驗的平台。其介紹了利用阿茲海默症臨床資料來最佳化患者人數的案例。

該公司強調反映亞洲患者資料的方法。因為偏向特定人口群體的資料可能無法充分解釋其他群體的病程進展。但必須將平台的介紹與實際臨床試驗在成本及時間上的縮減效果區分開來。

Ophrimed 在官方產品介紹中提出了虛擬患者生成、外部對照組設計，以及臨床試驗期程與患者人數預測等功能。OPTIVIS 是該公司的平台名稱，特定疾病的預測模型可視為其中的個別應用案例。由於產品介紹中未明確載明首次公開日期，難以將特定年份視為首次發布年份。

臨床試驗分析支援：PROCOVA

Unlearn 的 PROCOVA 是一種預測參與實際隨機臨床試驗之患者在對照條件下的病程進展，並將其用於統計校正的方法。歐洲藥品管理局（EMA）針對此方法提出了資格認定意見（qualification opinion）。

此案例顯示數位分身的應用除了治療推薦之外，亦擴展至試驗設計與統計分析。這並不意味著能完全取消實際對照組，或是僅憑虛擬資料就能核准新藥。具體的應用條件將在下一節中進行探討。

導入與驗證的課題

從法規監管的角度來看，使用目的、資料品質、風險程度、模型的效能與管理方式至關重要。直接影響治療選擇的模型與輔助研究設計的模型，所需的證據等級有所不同。

亦必須審查應用於其他醫院或人口群體時效能發生改變的領域轉移（domain shift）現象。資料標準、醫院系統串接、個人隱私保護及持續更新，亦為營運所必需的條件。即便真實世界資料（real-world data）十分豐富，亦可能因患者篩選與測量方式的差異而導致比較結果有所不同。

目前的案例皆聚焦於特定目的，如個人化劑量調整、生活習慣管理、器官模擬及臨床試驗分析等。數位分身的臨床價值，必須在明確界定適用範圍的同時，透過驗證預測對實際決策與患者預後帶來何種助益來進行評估。

4.2 虛擬對照組（Synthetic Control Arm）臨床試驗

Unlearn 的官方技術說明區分了輔助實際隨機試驗分析的 PROCOVA 與外部對照組。法規認可的範圍可在 EMA 的資格認定意見中確認。

<https://www.unlearn.ai/blog/built-on-regulatory-rigor-advancing-digital-twins-in-clinical-trials>

https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/qualification-opinion-prognostic-covariate-adjustment-procovatm_en.pdf

臨床試驗的對照組（control group）是比較新治療效果的基準。依疾病與試驗目的不同，對照組接受既有治療或安慰劑（placebo），研究人員透過比較試驗組與對照組的結果來評估療效與安全性。隨機對照試驗（RCT，randomized controlled trial）是縮小組間差異以判定治療效果的重要方法。

然而，招募對照組需要時間與成本。在患者人數稀少的罕見疾病中，難以確保足夠的人數，且在選擇對照治療時亦需考量倫理審查。為了解決這些問題，將既有患者資料與預測模型應用於試驗設計與分析的方法正逐步發展。

利用預測值的校正與外部對照組

「虛擬對照組」這一詞彙涵蓋了互不相同的方法。其一是保留實際隨機對照組，同時將人工智慧預測用於分析的校正方式。另一種則是將過往臨床試驗或病歷等外部資料作為比較對象的方式。這兩種方法在設計與偏差控管條件上均有所不同。

在校正方式中，會針對每位患者預測其接受對照治療時的病程進展，以建立預後分數（prognostic score）。將此分數反映在統計分析中，便能解釋患者之間的部分差異。Unlearn 的 PROCOVA 即為代表例。

美國企業 Unlearn 所提出的 PROCOVA 是一種稱為預後共變數校正的統計方法名稱。必須區分建立預後分數的機器學習模型，與將該分數納入試驗分析的方法。人工智慧負責計算個別患者的預期病程，而統計分析則評估實際觀察到的試驗組與對照組之間的差異。這種架構並非將人工智慧的預測直接作為治療效果的標準答案。

外部對照組（external control）或合成對照組（synthetic control）是從當前試驗之外的資料中構建比較對象。例如，可將所有受試者皆接受新治療的單臂試驗（single-arm trial）結果，與過往患者的病程進行比較。然而，由於群體從一開始就可能存在差異，因此必須另外處理選擇偏差與測量差異。

PROCOVA 的運作原理

臨床試驗的結果除受治療效果影響外，亦受患者初始狀態與疾病進展速度影響。若能預先預測這些差異並納入分析，便可降低結果的變異數（variance）。在適當條件下，有可能以較少的患者人數達到目標分析精度。

PROCOVA 利用以過往對照組資料訓練出的預測值進行共變數校正（covariate adjustment）。這是一種在維持實際隨機分配的同時

分析治療效果的架構。實際所需患者人數與效率提升的幅度，取決於疾病、評估指標、預測準確度及試驗設計。

共變數是指可能影響結果而在分析中一併考量的特徵。如年齡或治療前狀態等在試驗開始前即可獲知的資訊皆可納入其中。預後分數將從多種特徵中預期的病程整合為單一資訊。若預測帶來的助益較小，則減少患者人數的優勢亦可能隨之降低。這就是為什麼無法斷言在任何試驗中都能按固定比例減少人數的原因。

歐洲藥品管理局（EMA）針對 PROCOVA，提出將其認可為採用連續型結果之第 2、3 期隨機臨床試驗分析的統計方法。這是認可其在特定條件下的分析方法，並不意味著僅憑虛擬資料即可核准新藥。

EMA 文書於 2022 年 9 月 20 日公開，委員會通過日期為同月 15 日。文書中所指的連續型結果，是指如分數或測量值般數值連續的結果。不應將該文件過度解讀為一次性認可了所有類型的結果指標與所有統計方法。

美國 FDA 亦在指引中整理了隨機臨床試驗的共變數校正原則。然而，一般性的分析原則與對特定企業方法的單獨核准必須予以區分。在實際應用中，必須根據試驗目的與分析計畫與主管機關進行協商。

外部對照組與數位分身

將外部資料作為比較對象的方式早在人工智慧出現前便已存在。國際指引 ICH E10 亦探討了外部對照的優缺點。在罕見疾病等難以進行隨機比較的狀況下，疾病的自然病程資料或過往試驗結果一直以來都受到審視與採用。

人工智慧與數位分身試圖利用外部資料，預測符合當前受試者特徵的對照組病程進展。這是一種不單使用群體平均值，而是反映個人化資訊的方法。然而，即便預測變得更加精確，也不會自動消除因缺乏隨機分配而產生的偏差。

例如，過往資料中的患者若病情進展程度低於當前試驗患者，病程差異的部分原因可能與治療效果無關。即使年齡與檢驗數值看似相似，未記錄的健康狀態或選擇治療的理由也可能有所不同。未測量的差異很難直接輸入統計程式進行調整。光憑外部資料量豐富這一點，並不足以構成良好的比較。

主管機關會審查外部資料的適當性、疾病進展、比較群體的差異及結果測量方式等。雖然相關指引與討論文件陸續擬定，但這並非可一體適用於所有疾病與試驗的替代方案。

真實世界資料展現的局限

RCT DUPLICATE 是探討利用真實世界資料（real-world data）能在多大程度上重現既有隨機臨床試驗結果的研究。該研究利用保險理賠資料等構建比較群體，並以傾向分數（propensity score）校正可測量的差異。

在部分試驗中雖獲得了相似的結果，但並非所有結論都一致。當能夠高度重現原始試驗的受試者條件、治療及結果測量方式時，一致性較高。若病歷中缺乏所需資訊，或是難以重現安慰劑對照時，差異便可能擴大。

此結果顯示，在構建虛擬對照組時，資料的適切性同樣至關重要。不僅需要確認演算法的效能，還必須確認原始資料是否足以回答試驗所提出的問題。

在分析之前確定結果指標、比較時間點以及患者納入標準也很重要。若在看到結果後只挑選有利的患者或更改測量時間點，可能會使治療效果看起來被誇大。必須將用於訓練模型的資料與試驗的結果資料分開，並留下分析計畫，其他研究人員才能重現相同的比較。

研議應用的領域

在罕見疾病中，由於招募患者困難，對外部對照組的需求很高。然而，較少的患者數量並不會消除比較資料的局限性，因此必須依不同疾病評估其適切性。

在阿茲海默症等神經疾病中，正在研議透過預測個體間進展差異來提高分析效率的方法。韓國的 OPRIMED 所介紹的臨床試驗模擬也是相關案例。不過，利用過去資料的模擬設計與在新試驗中實際取得的成果是不同的。

在抗癌藥物開發中，將合成對照組與數位分身應用於試驗設計的研究也正在進行。不應將企業間的合作或教學用設計案例誇大解釋為實際的許可成果。

試驗設計中須確認的事項

過去資料反映的是當時的標準治療。若在此期間治療方法或診斷標準發生改變，就很難與當前的試驗直接比較。臨床試驗會在固定的時間點與方式測量結果，但病歷可能並非如此。遺漏的資訊以及患者選擇的差異也必須納入考量。

因此，必須明確提出虛擬對照組的目的、生成方法、資料來源以及分析的不確定性。開發商與試驗執行機構的角色及審查責任也必須事先釐清。

虛擬對照組的意義在於減輕臨床試驗的負擔，並更有系統地運用現有資料。目前的主要途徑是以預測資訊來補充實際隨機試驗的方式。若打算以外部資料取代對照組，則必須充分檢視其可比性與偏差。同時確保招募效率與證據的可靠性是核心課題。

4.3 基於模型的藥物開發與動物實驗替代（FDA Modernization Act）

FDA 的 NAMs 說明與 MIDD 聯席會議指引中，整理了新試驗方法的範圍，以及將模型用於開發判斷的條件。若將這兩份文件一併研讀，便能區分人類細胞實驗與電腦預測在新藥開發中所扮演的角色。

<https://www.fda.gov/science-research/science-and-research-special-topics/new-approach-methodologies-nams>

<https://www.fda.gov/drugs/development-resources/model-informed-drug-development-paired-meeting-program>

在新藥開發中，為了確認療效與安全性需要大量的時間與資源。在動物身上獲得的結果無法直接套用於人類，也是開發上的困難所在。在藥物的吸收與代謝、免疫反應、治療標靶的運作機制上均存在物種間的差異。

作為彌補這些問題的途徑，基於人類細胞的試驗與電腦模型的使用正日益擴大。新型替代方法 NAMs（New Approach Methodologies）使非臨床評估的方法多樣化。基於模型的藥物開發 MIDD（Model-Informed Drug Development）則結合資料與數學模型，為開發過程中的決策提供支援。

NAMs 與 MIDD 並不是公司推出的人工智慧產品名稱。NAMs 是一個試驗方法的範疇，而 MIDD 則是將從模型中獲得的資訊反映於開發決策的途徑。器官晶片是在微小裝置中實現部分人類組織的功能，而電腦模型則是計算資料與假設。由於兩者提供不同種類的證據

，因此不能在未經驗證的情況下，將某種方法的優點直接套用於另一種方法。

動物實驗替代方法的擴大

FDA Modernization Act 2.0 調整了法律用語，使非臨床試驗中得以包含細胞分析、器官晶片（organ-on-a-chip）、電腦模型等方法。這並非禁止動物實驗，而是明確了可以使用其他適當方法的依據。在實際開發中，必要時仍會繼續使用動物研究。

後續的制度調整是將法律宗旨反映於具體法規與審查實務中的過程。關於 FDA Modernization Act 3.0 的討論以及 FDA 的減少動物實驗路線圖亦屬於此潮流。不過，法案、現行法規與指引草案的法律地位各不相同，必須加以區分。

FDA 在說明 2026 年 9 月用語調整的資料中亦明確指出，這並非強制要求特定試驗方法，亦非降低所要求的證據標準。減少動物實驗的政策方向與個別新藥所需安全性資料的判斷，必須結合起來解讀。絕不能僅因標題中含有現代化或替代，就誤以為所有的動物試驗都已消失。

NAMs 包含預測毒性與藥物動力學的電腦模型、模擬人類組織功能的器官晶片，以及由幹細胞等製成的類器官（organoid）。在如單株抗體（monoclonal antibody）這類動物與人類標靶差異顯著的領域中，也正在研議其應用。

這些方法的目的並不在於無條件地廢除特定試驗。獲取適合所欲評估風險的資料才是首要任務。新方法同樣必須驗證其重現性、預測力以及適用範圍，所需達到的證據水準並未因此降低。

基於模型的藥物開發概念

MIDD 透過整合非臨床資料、臨床資料與現有知識的建模與模擬，來支援開發決策。它涵蓋了首次人體給藥劑量、適當的給藥間隔、特殊病患族群的劑量、藥物交互作用等多種問題。

藥物動力學探討身體吸收、分布與排除藥物的過程，而藥效學則探討藥物對身體產生何種變化。即使給予相同劑量的藥物，體內殘留的濃度也可能不同；而在相同濃度下，反應也可能不同。唯有連結這兩者的關係，才能將劑量數值與患者的反應結合進行解讀。

群體藥物動力學／藥效學（PopPK/PD）估算隨年齡、體重、器官功能等因素引起的藥物濃度與反應差異。生理學基礎藥物動力學（PBPK）利用器官大小、血流、代謝酵素等資訊來計算體內藥物分布。定量系統藥理學（QSP）對藥物與生物訊號傳遞路徑的交互作用進行模型化。人工智慧與機器學習（AI/ML）亦可用於相關資料的分析與預測。

這些方法連結現有資料，有助於判斷需要額外進行試驗的部分，並使試驗設計更加具體。然而，模型的預測能否用於某些決策，取決於資料的品質與驗證水準。

例如在研議給藥間隔的模型中，給藥後藥物濃度隨時間的變化情況即為核心資料。若模型是依據特定患者族群的資料建立的，則必須

檢視在其他族群中是否仍能維持相同的關係。預測越是超出實際測量的範圍，可能越需要額外的依據。在報告中記錄計算值時，亦必須同時載明所採用的假設與資料範圍。

與監管機構的諮商

FDA 的 MIDD 配對會議計畫（MIDD Paired Meeting Program）是開發商與監管專家討論基於模型途徑的制度。在開發初期就模型的目的、假設與評估方法進行諮商，有助於制定後續的證據生成計畫。

ICH M15 是為 MIDD 的評估與文件化建立共通標準的指引。它強調應考量模型的使用脈絡及其對決策的影響，以決定所需的驗證水準。FDA 與 EMA 提出的新藥開發人工智慧實務原則，亦將以人為中心的设计、資料管理、基於風險的評估與持續管理視為重點。

具體法律與規定的適用取決於文件是否定案及其管轄權。不應將企業的諮商經驗或方法介紹誇大理解為普遍的監管批准。

NAMs 與 MIDD 的連結

NAMs 提供反映人類生物學特性的資料，而 MIDD 則能解讀這些資料，以檢視劑量與試驗條件。例如，將從人類細胞測得的代謝與毒性資訊反映到藥物動力學模型中，並在取得臨床資料後更新模型。

在臨床階段，可將模型應用於劑量選擇、患者族群分析及試驗設計。前一節探討的預後評分校正與外部對照組，亦是可根據目的與條件研議的方法。然而，NAMs 與 MIDD 是彼此可以相互串聯的途徑，並非在所有開發中都必須依照單一固定程序套用的制度。

人工智慧新藥開發的臨床進展

Insilico Medicine（英矽智能）的 rentosertib（舊稱 ISM001-055）是用於特發性肺纖維化（IPF）的候選治療藥物。該公司說明其將生成式人工智慧應用於標靶探索與分子設計。

rentosertib 並非人工智慧的名稱，而是候選治療藥物的名稱。開發商 Insilico Medicine 在 2026 年 7 月 7 日宣布啟動第 3 期臨床試驗時說明，標靶篩選使用了 PandaOmics，而分子生成與最佳化則使用了 Chemistry42。這兩種工具即使在同一家公司的研究平台中，分工亦不相同。根據 Insilico Medicine 的官方技術介紹，PandaOmics 與 Chemistry42 均於 2020 年推出。這是早於 rentosertib 開始臨床試驗的工具推出時間點。啟動第 3 期臨床試驗與藥物獲得許可在意義上是有所區別的。在相關研究中，開發了以 TNIK 為標靶的候選藥物並銜接至臨床評估。這兩種工具的推出年份以及標靶探索、分子設計功能，可在該公司的技術介紹中確認。
<https://insilico.com/mediakit>

在早期臨床試驗中，觀察了安全性、耐受性以及用力肺活量指標（FVC, forced vital capacity）等變化。隨後推進至以更長時間與更多患者為對象評估療效的第 3 期開發。單憑早期結果無法斷定長期安全性與治療效果。

此案例顯示，即使是由人工智慧探索出的候選藥物，同樣需要經過階段性的臨床評估。人工智慧新藥開發案例與 MIDD 的具體使用案例亦須加以區分。不能因為人工智慧提出了標靶或分子，就認為在整個開發過程中都應用了基於模型的決策。

開發速度與成功可能性

Exscientia 與住友大日本製藥（Sumitomo Dainippon Pharma）用於治療強迫症的候選藥物 DSP-1181 曾以人工智慧設計及快速進入臨床試驗聞名，但隨後開發遭到中止。這是展現快速生成候選藥物並不能保證臨床成功的案例。

在針對以人工智慧為核心的企業所開發的候選藥物進行分析的研究中，早期臨床成果呈現正面表現，但並未確認其優勢在後期階段仍能保持一致。候選藥物數量與觀察期間亦較為有限，因此難以將該差異斷言為人工智慧本身的成效。

隨著臨床試驗的推進，不僅是藥物的物化性質，疾病的複雜性、患者差異、共病症等也變得日益重要。細胞世界模型與數位分身雖然有助於更深入理解這些變數的研究，但並不能免除臨床驗證。

韓國的研究與資料合作

在韓國，將人工智慧應用於藥物開發的指引以及共同研究計畫亦正在推動中。K-MELLODDY 是基於聯合學習（federated learning）的途徑，無需將企業的資料直接集中至外部即可共同訓練模型。

根據專案小組介紹，K-MELLODDY 是由科學技術情報通信部與保健福祉部自 2024 年 4 月起共同推動的計畫。與其說是單一完整的人工智慧模型名稱，更接近於建立共同學習基礎的專案名稱。由於匯集各機構計算的學習資訊來改進共同模型，因此可減輕將原始資料集中於一處的負擔。然而，並非僅憑不轉移資料這項條件就能解決所有的資安問題，所共享的學習資訊與存取權限亦必須妥善管理。

這是一種在保護敏感研究資料的同時，將多個機構的資訊反映於學習中的方式。然而，若要確保共同模型的效能，必須一併管理資料定義與品質、機構間的差異以及評估程序。

未來的課題

即使使用人類細胞與電腦模型的範圍擴大，也並不意味著所有的動物實驗或臨床試驗都會立即被取代。在全新的治療方式或訓練資料中不存在的條件下，預測的不確定性可能會增加。亦需要將模型的結果進行生物學上的合理解釋，並與實際觀察進行比對的過程。

NAMs 與 MIDD 的轉變意味著建立開發證據的方法正在日益多樣化。這有助於探索新候選藥物，並選擇適當的劑量與試驗設計。最終仍須透過符合該使用目的之驗證，以及從患者端獲得的安全性與療效資料來評估其價值。

Part 3. 海外醫療AI導入案例

第5章 美國醫療機構的人工智慧引進



5.1 梅約診所的 VoiceCare AI 與院務自動化

梅約診所平台（Mayo Clinic Platform）的官方解說介紹了用於院務、病患衛教及安全管理的代理（agent）。VoiceCare 的示範計畫範圍可在開發商發布的 2025 年 2 月新聞稿中確認。

<https://www.mayoclinicplatform.org/2026/08/07/ai-agents-in-healthcare/>

<https://www.prnewswire.com/news-releases/agentic-ai-startup-voicecare-ai-launches-to-automate-healthcare-back-office-and-super-staff-workforce-302376944.html>

醫院的業務除了診療之外，還包括確認保險給付、事前審批申請、請款後續處理等。院務（行政事務）或後勤（back office）業務是確認多個機構的資訊並將結果記錄於醫院系統中的過程。由於電話與文件處理反覆進行，因此作為醫療人工智慧的應用領域而備受關注。

院務自動化能減少醫護人員與行政人員的重複性業務。然而，若錯誤處理保險資訊或審批結果，可能會影響患者的費用負擔與治療排程，因此精確度與審查程序至關重要。

支援與保險公司通話的語音代理

VoiceCare AI 是開發公司的名稱，而 Joy 是該公司打造的語音代理。該公司的公開亮相以及與梅奧合作的公布日期為 2025 年 2 月 14 日。這並不意味著是開始研發的日期，也不代表梅奧所有部門開始啟用的日期。亦不應將其理解為直接對患者做出診斷的產品。

VoiceCare AI 於 2025 年 2 月宣布將與梅奧診所推進保險相關院務行政業務的試辦計畫。對象為神經科、兒科以及醫療與行政支援部門，主要業務則是確認患者的保險福利與事前審批。

語音人工智慧代理 Joy 被設計用於與保險公司通話、等待並確認所需資訊，並摘要記錄通話內容。此案例的目的在於使與保險公司的重複溝通實現自動化，並不代表對患者診療與預約全程進行自主處理。

公開資料說明了試辦計畫的方向，但並未證實梅奧全院已完成導入，亦未證明具體的時間與成本節省效益。據公告，梅奧對該項技術擁有財務利益關係。因此，必須將導入的發布與獨立的成果評估區分開來解讀。

從生成回答到執行任務

例如，假設醫院員工在檢查前確認是否適用保險。必須準備要確認的患者與檢查資訊，並聆聽保險公司的回覆。若遺漏必要項目則需再次詢問，並留下確認結果。語音代理也以這樣的順序作為處理對象。將語音轉為文字的功能、選擇下一個問題的功能，以及透過語音傳達答案的功能在此串聯。通話結束的事實，與保險確認已準確完成的事實，兩者截然不同。

代理（agent）是將語言模型的輸出與下一步行動連結起來的系統。如果說大型語言模型（LLM, Large Language Model）是對提問生成回答，那麼代理則能依據該結果撥打電話、查詢資訊並更新記錄，執行一系列的任務。

梅奧診所平台（Mayo Clinic Platform）的約翰·哈蘭卡（John Halamka）與保羅·塞拉托（Paul Cerrato）在闡述醫療代理時，強調了這種行動的連結。在院務行政中，保險公司確認、提問與回覆、記錄結果，以及異常情況下移交給負責人，構成了一個完整的流程。

在實際實現中，可以使用作為系統間連接通道的 API（Application Programming Interface），或處理語音與螢幕輸入。關鍵不僅在於以何種方式連接，更在於明確界定代理能夠處理的範圍，以及移交給人工處理的條件。

擴展至患者衛教

MedEduChat 是由梅奧、康乃爾大學及麻省理工學院（MIT）所屬研究人員參與的研究系統。相關論文於 2025 年 12 月 18 日發表。這必須與梅奧全面推出面向患者的商業化服務之發布區分開來。該研究的出發點在於協助患者在診療結束後仍能理解自身的病歷記錄。

其運作順序可理解為：提問、確認相關病歷記錄、生成解釋。例如，假設患者詢問病歷上記錄的治療術語含義。系統會參考與提問相關的病歷記錄，生成易於閱讀的解釋。不可臆測並填補病歷中未記載的檢查結果。若病歷記錄本身已過時或有所遺漏，解釋中也會殘留其局限性。

梅奧的 MedEduChat 是協助攝護腺癌患者理解的研究案例。其開發方向是以電子病歷（EHR, Electronic Health Record）中的患者資訊為基礎，淺顯易懂地解釋醫學術語與診療內容。

在初步研究中，患者與專家評估了可用性與回答品質。依據病歷進行解釋的功能展現了潛力，但針對個人情況量身定制的回答仍是需要進一步改進的部分。使用者感到便利、資訊正確，以及治療結果改善，這三者必須分別進行獨立評估。

由於該研究屬於小規模的可用性評估，因此需要針對多樣化的患者群體以及實際臨床效果進行後續驗證。患者衛教代理可被視為支援醫療人員的說明，並協助理解診療內容的角色。

患者安全業務與人工智慧

PSVT 是管理患者安全的計畫名稱，很難將其視為單一語言模型或市售產品的名稱。梅奧於 2026 年 8 月的解說中說明，風險分類代理處於運行階段，而模式探索代理仍在開發中。這意味著即使在同一個計畫內部，各項功能的就緒水準也可能有所不同。

梅奧亦介紹了處理跌倒、院內感染、壓瘡、藥物相關事件及病情惡化等問題的「患者安全生命徵象」（PSVT）計畫。該方法旨在開發能夠對患者風險進行分類並從資料中尋找模式的代理。

這類系統能整理住院期間累積的資訊，協助醫療人員篩選需要審查的對象。然而，計畫的運行或研發消息，與實際減少事故的效果不可同日而語。警報的準確性、不必要的提醒，以及醫療人員的應對負擔，都必須一併納入評估。

企業合作與應用領域

梅奧在與 NVIDIA 的合作中，亦致力於研究數位病理與生成式人工智慧基礎架構。然而，這項合作與 VoiceCare 的保險公司通話試辦計畫屬於互不相同的應用領域。

在單一醫療機構內部，院務行政、患者衛教、安全管理與病理分析的目的及需求條件皆有所不同。若要理解機構整體的 AI 策略，就必須區分個別專案所針對的業務為何。

院務自動化的營運條件

在審查畫面上，應與結論一併顯示確認日期與依據。例如，雖然取得了核准號碼，但可能僅在特定期間內有效。若僅複製號碼而遺漏了有效期限，下一位負責人員便可能誤解允許的範圍。將不完整的答案保留為空白並請求進一步確認，比起填入看似合理的數值，更能確保業務準確銜接。

在事前審批與保險確認中，必須準確記錄保險人資訊、核准號碼及適用條件等。在通話中斷或回覆不明確的情況下，負責人員必須能夠接手處理。對於會影響患者權益或費用的結果，需要具備適當的確認程序。

亦必須考量自動化偏誤（automation bias）。若未對系統的結果進行充分審查便逕行採納，輸入或解讀上的錯誤可能會反映在後續業務中。必須根據業務風險，合理設定由人工介入審查的環節（human-in-the-loop）。

與保險公司的通話中包含姓名、診斷、保險號碼等敏感資訊。

VoiceCare 雖然說明其符合 HIPAA 規範並具備 SOC 2 Type II 認證，但資訊保護體系與臨床安全性必須加以區分。HIPAA 涉及美國醫療資訊保護的法律要求，而 SOC 2 Type II 則是評估一段時間內安全控制營運狀況的稽核體系。

梅奧的案例顯示，醫療 AI 的導入並不侷限於診斷演算法。其應用範圍正朝著使重複性院務業務自動化、向患者解釋資訊以及整理安全管理資料的方向拓展。在各項業務中確認準確性與實際效益，並建立能由人工處理異常情況的架構，乃是導入的基本條件。

5.2 Tempus 的精準腫瘤學支援

在 Tempus 的 Next 官方產品頁面中，可以查看依據診療指引確認下一步驟以及遺漏診療與追蹤管理對象的功能。在 Lens 的 2026 年發布說明中，則可瞭解將研究問題轉化為分析計畫與程式碼執行的功能。

<https://www.tempus.com/solutions/next/>

<https://www.tempus.com/news/pr/tempus-unveils-the-next-generation-of-lens-expanding-its-agent-ai-platform-for-oncology-drug-development/>

精準腫瘤學（precision oncology）是一種考量腫瘤的分子特徵與患者狀態來選擇治療方案的方法。即使是同一種癌症，適當的治療方式也可能因基因突變或蛋白質表現而有所不同。為此，必須實施必要的檢查，並將檢查結果與治療決策相互連結。

在實際診療中，可能會出現漏掉檢查對象或檢查結果未充分反映的問題。Tempus 透過分析臨床病歷與分子資訊，提供縮小此類診療缺口（care gap）並支援研究的平台。

Next 與 Lens 的角色

這兩款產品的開發商為美國的 Tempus。介紹名為 Next 平台的官方發布於 2024 年 4 月 16 日公布。必須將在此之前的相關演算法研發與合作歷程，與產品公開亮相的日期區分開來。Lens 則於 2021 年 5 月 20 日作為資料分析平台推出。2026 年 5 月 31 日的發布，則是關於在該平台擴展代理功能的新一代版本。

非結構化資料是指與填寫在固定欄位中的數值不同，以語句形式留存的資訊。例如，診療病歷中可能記載已將檢查委託給外部醫院。AI 能閱讀這類文字，整理檢查與結果的狀態。不可僅憑病歷中未記載結果，便斷定尚未進行檢查。這正是接獲提醒的醫療團隊必須確認外部資料與患者狀況的原因。

Tempus Next 是醫療機構的臨床決策支援工具。它透過分析電子病歷（EHR）的非結構化資料，找出未依循診療指引接受必要生物標記（biomarker）檢查的患者，並向醫療團隊提供提醒。與其說是獨立決定新的治療方法，不如說是協助確保必要的檢查與審查得以落實。

Tempus Lens 是為藥廠與研究人員打造的分析平台。它根據研究問題分析患者群體與資料，並介紹了協同科學家（Co-scientist）代理以支援分析計畫與程式碼執行。這是一種將去識別化的臨床資料與分子資訊連結至藥物開發研究的方法。

這兩個平台並非同一產品的前後代名稱。Next 側重於找出診療過程中的遺漏，而 Lens 則聚焦於研究分析。這可以作為醫療 AI 在診療與研究中支援不同任務的案例加以區分。

連結多元資料的基礎

在 Lens 中，研究人員的提問構成輸入。在確定分析對象與比較條件後，篩選相應資料進行運算並呈現結果。例如，可以比較帶有特定突變與未帶突變之患者的病歷記錄。若兩個群體的年齡或既往治療史不同，結果的解讀也會有所差異。即使 AI 能生成並執行程式碼，判斷比較條件是否合理的研究人員角色依然不可或缺。

基因體處理 DNA 資訊，轉錄體則涵蓋基因被轉錄為 RNA 的狀態。影像能呈現組織的形態。將不同的資料連結起來，便能觀察僅憑單一檢查無法看出的關聯。然而，首先必須確認單一患者的資料之間是否準確關聯。即便資料量增加，連結錯誤與遺漏也不會自動消失。

Tempus 宣布擁有包含臨床病歷、基因體、轉錄體及影像等的大規模資料。這些資料可用於綜合分析特定突變與疾病病程，或是建構符合研究目的的患者群體。

腫瘤多模態基礎模型（multimodal foundation model）的研究亦是在此基礎上展開。然而，擁有資料的數量與模型的準確性是兩回事。必須確認資料的代表性、遺漏情況以及各機構間的測量差異，並且必須將公司提出的初步結果與獨立驗證的效能區分開來。

生物標記檢查與治療的連結

生物標記是判斷疾病或治療反應時參考的生物學指標。EGFR 與 ALK 出現於基因相關檢查中，而 PD-L1 則是評估蛋白質表現的指標。僅憑同一名稱出現在病歷中，不能視為檢查已經完成。因為開立檢查醫囑、檢體採集、報告結果與醫療團隊確認是互不相同的階段。尋找診療缺口的系統必須能夠區分這些狀態。

在美國臨床腫瘤學會（ASCO）發表的這項研究，以多家社區醫療機構的早期非小細胞肺癌（eNSCLC）患者為對象，比較了 Next 導入前後的情況。據報告，導入後 EGFR、ALK 與 PD-L1 的檢查比率有所提升。

此案例的意義在於支援將既有的診療指引應用於實際診療中。透過持續分析患者的病歷，能發現遺漏必要檢查的案例，並引導至醫療團隊的審查。

然而，必須將檢查實施率增加的結果，與存活率或復發率的改善區分開來。導入前後的比較中，同一時期的診療慣例變化也可能產生影響。將目前的結果理解為展現了診療流程改善的可能性，較為恰當。

與治療後追蹤的連結

Tempus 一直與微小殘留病灶（MRD, Molecular Residual Disease）檢測企業 Personalis 進行合作，並宣布了收購協議。MRD 檢測是一種探查治療後可能殘留之疾病痕跡的方法。將基於組織的分子分析與治療後追蹤連結起來，有可能更連續地分析患者的病程。

必須將收購協議的宣布與交易的正式完成區分開來。不能僅憑商業計畫就認定整合的診療體系已經建置完成。其實際功能與臨床效果必須另行確認。

運算基礎架構與合作關係

分析大規模臨床與分子資料需要高效能運算。Tempus 雖然也說明了相關基礎設施，但切勿將與特定技術供應商的合作範圍過度解讀為超出公開資料的範疇。本節將以已獲確認的 Tempus 平台功能為核心進行探討。

理解醫療人工智慧時，有必要區分資料擁有、分析模型、系統串接與企業合作各自扮演什麼角色。發布合作消息並不等於證明了特定臨床功能的效果。

醫療團隊的判斷與適用範圍

像 Next 這樣的工具旨在支援醫療團隊的決策。即使系統從病歷中找出潛在檢查候選並發出通知，患者是否應接受實際檢查與治療，仍須考量診療情境來進行判斷。亦不可將個別人工智慧產品的監管許可，誤解為整個平台具備自主診療許可。

若資料偏向特定國家或人口群體，套用至其他群體時效能可能會有不同。必須使用外部機構的資料進行評估，並確認在實際工作流程中通知如何被處理。通知的準確性以及醫療團隊的應對負擔，亦包含在營運評估之中。

Tempus 的案例表明，醫療人工智慧除了生成新的預測功能外，還能支援在適當時機將必要資訊傳遞給醫療團隊。精準醫療的發展不僅需要分子分析技術，還需要將檢查、結果確認、治療決定與追蹤管理相串聯的工作體系。

第6章. 中國的醫療人工智慧導入



6.1 河北醫大附屬醫院的人工智慧營運成效

河北醫科大學第一醫院研究團隊的營運論文，說明了從何種業務開始導入人工智慧。若將比拚臨床準確度的試驗與醫院的營運經驗區分閱讀，成果的範圍就會更加明確。

<https://pmc.ncbi.nlm.nih.gov/articles/PMC13245460/>

要讓醫療人工智慧落實於實際業務中，除了技術效能外，還必須一併檢視診療流程與組織的營運方式。因為有些情況是在完成試辦計畫後便停止使用，或是僅侷限於部分科室。這種現象被稱為試辦陷阱（pilot trap）。

中國河北醫科大學第一醫院在預先問診、腫瘤會診支援與藥物濃度監測中，分階段導入了人工智慧。相關研究以約 18 個月的營運經驗為基礎，說明了醫院如何擴大並管理人工智慧。這是一個比起單一模型的準確度，更著眼於工作流程（workflow）與治理架構（governance）的案例。

依風險等級的分階段導入

論文於 2026 年 6 月 4 日公開。這是發表營運經驗的時間點，與這三套系統的最初開發日或開始銷售日並不相同。預先問診、腫瘤會診支援與藥物濃度管理是業務名稱。切勿將其理解為同一家公司所開發之同一模型的三種名稱。醫院名稱僅代表營運主體，單憑這一點無法得知系統的開發公司或模型版本。

醫院並未從一開始就將所有臨床業務自動化。而是先支援診前資訊收集，隨後才導入輔助專業判斷的功能。

在智慧預先問診中，患者在看診前輸入症狀與病歷，人工智慧便會擬定結構化的病歷草稿。醫師確認草稿後，可針對額外需要的事項進行提問。這是一項支援多個專科門診共通產生的資訊收集與記錄業務之功能。

跨科別腫瘤決策支援則是整合影像、病理與治療歷程，提供在多專科會診中檢討的選項。這是當多個專科共同討論同一名患者時，將必要資料相互串聯的方式。最終的治療決定仍由醫療團隊負責。

度洛西汀（duloxetine）藥物濃度監測是參考血中藥物濃度來支援劑量調整。由於這是需要專業判讀與個別患者專屬判斷的功能，因此僅限於相關診療領域中有限度地使用。

導入範圍的差異

整理患者輸入答案的功能與建議藥物劑量的功能，其發生錯誤時所導致的後果截然不同。例如在預先問診中，若遺漏了症狀開始的時間點，醫師還可以再次詢問。相反地，若用藥相關資料的單位或抽血時間有誤，建議的意義可能就會完全改變。兩者皆需要確認，但確認項目與負責的專家並不相同。分階段導入正是將這種差異反映在實際工作分工中的過程。

預先問診擴展到了許多專科門診，而腫瘤治療與藥物劑量支援則是以相關專業領域為核心進行套用。這是因為整理資訊與撰寫文件可

供多個科別使用，但直接支援治療選擇的功能，則必須更加細緻地管理風險與責任。

這種方式是分階段推行（staged rollout）的一種形式。從共通性業務開始套用，臨床風險較大的功能則在明確界定使用範圍與審核程序的同時逐步擴展。在醫院內部，各項功能所需的培訓與監督水準亦有所不同。

報告中提及的營運轉變

在成效表中，還必須審視是以什麼作為分母來計算。製作一份文檔的時間與單一患者在院的總停留時間是不同的概念。若滿意度調查僅由熟悉使用人工智慧的員工填寫，便可能與全體員工的實際體驗產生落差。若在同一時期增加了人力或調整了掛號動線，候診時間的變化也會混雜這些因素的影響。唯有確認這些條件，才能在導入至其他醫院時準確評估可預期的成效。

營運研究報告指出，在候診時間、文書撰寫、診療效率與患者滿意度等方面皆有所改善。然而，完成文書撰寫的效率與整體診療效率屬於不同的指標。切勿將特定業務的改善數值過度解讀為全院的成效。

在同一時期，門診動線、資料管理與品質管制程序亦獲得同步改善。因此，報告中所載的轉變，應當理解為包含人工智慧在內的營運改善流程所帶來的結果。該研究並非單獨證明人工智慧單一效應的研究。

此案例的意義不在於工作效率確切的增長率，而在於將多項功能連結至日常診療並維持營運一段時間的經驗。然而，全院層面的導入與個別醫療團隊人員的持續使用必須加以區分，其他機構是否也能呈現相同的結果，仍需進一步確認。

技術導入與組織量能

NASSS 並非特定人工智慧產品的名稱。它是一個連同技術未被接納或停止使用的原因也一併剖析的分析架構。在醫院宣布導入後，對於觀察後續是否持續實際使用具有重要意義。例如，若因資料輸入繁瑣而使醫療團隊跳過該功能，即使模型的測試評分很高，工作成效也會相當有限。現場是否有負責受理並修正問題的人員，亦包含在營運能力之中。

醫療人工智慧的推廣擴散需要明確的使用者角色、支援體系與例外處理基準。必須訂定由誰確認預先問診草稿、會診資料若有錯誤應如何修正，以及由誰核准藥物相關建議。

像 NASSS 這樣的技術導入分析架構，除了技術本身之外，亦會同時檢視患者、使用者、組織與制度的影響。醫院若要持續營運新功能，這些條件必須與實際業務相契合地進行建構。

河北醫科大學第一醫院可視為並非將人工智慧作為單一產品一次性全盤導入，而是依業務性質與風險進行套用的案例。其方式是將範圍從整理資訊擴展至輔助專業判斷，但同時維持醫療團隊的決策權限與審核程序。

營運成效與臨床成果

改善候診時間或文書處理對患者與醫療團隊的體驗至關重要。然而，這必須與存活率、再住院率、副作用等臨床成果的改善分開評估。本研究亦主要著重於導入與營運層面。

未來有必要綜合觀察實際使用的持續性、各科別間的差異、錯誤修正的負擔以及患者預後成果。醫療人工智慧是否真正落實生根，很難單憑導入宣告或初期滿意度來斷定。必須經歷確認其能否在醫院的業務與管理體系內穩定運作的過程。

6.2 WiNGPT 3.0 與螞蟻阿福服務

WiNGPT 3.0 研究團隊的技術報告公開了模型的訓練方法與測試範疇。螞蟻集團發布的 AQ 上線資料，則有助於比較醫院專用模型與個人健康 App 之間有何不同。

<https://arxiv.org/abs/2505.17387>

<https://www.antgroup.com/news-media/press-releases/1750924800000>

中國醫療人工智慧的應用正在擴大至醫療機構的業務支援與個人的健康管理服務。醫療團隊在醫院資訊系統中使用的工具，與民眾透過行動電話查詢健康資訊的服務，正呈現同步發展的態勢。

在這種轉變的背後，亦存在著患者向大型醫院集中以及對基層醫療機構信任不足的問題。目前正嘗試透過人工智慧提高地區醫療資訊的獲取便利性，並支援基層醫護的工作。然而，服務的普及與醫療體系失衡問題的解決，必須分開評估。

醫院資訊系統與 WiNGPT

WiNGPT 3.0 技術報告的首次公開日期為 2025 年 5 月 23 日。6 月 5 日則上傳了修訂版本。這是 3.0 研究報告的公開歷程，與 WiNGPT 整個產品系列的最初發布日期不同。報告中所探討的模型約擁有 320 億個參數。參數是在訓練過程中進行調整的內部數值，並非單憑數值龐大就能讓醫療判斷變得準確。

研究團隊採用了依循回答範例學習的監督式微調，以及依據獎勵改善輸出的強化學習。接收到醫療提問時，會以所學知識與問題條件

為基礎生成回答。自動讀取病歷紀錄的功能，必須有另外的系統串接才能實現。並非僅僅安裝了語言模型，就能存取所有患者的病歷。

WiNGPT 是醫療資訊企業衛寧健康（Winning Health）專為醫療打造的大型語言模型。該公司將其與醫院資訊系統 WiNEX 進行串接，並以 WiNEX Copilot 等形式架構，用以支援醫療團隊的工作。

套用於醫院的人工智慧，比起僅停留在獨立的聊天視窗中，更應與記錄、查詢、整理資訊等既有業務相互串接。讓醫療團隊能直接在現行使用的系統中獲得必要協助，正是系統整合的目的。

衛寧健康宣布已將 WiNEX Copilot 等人工智慧產品部署至約 150 家醫療機構。此數值並不同於導入 WiNGPT 3.0 單一模型的機構數量，亦不代表實際使用頻率或臨床成效。部署範圍與現場的落實程度有必要區分看待。

WiNGPT 3.0 技術報告提出了醫療計算與醫學知識評估結果。雖然這類測試有助於比較基本能力，但患者的診斷與治療是否獲得改善，仍必須在實際使用環境中加以驗證。

與其他醫療模型的區別

雲知聲的 U2-Med 官方公開資料於 2026 年 7 月 17 日發布。這是一款針對醫療、醫療保險與醫藥業務的專用模型。此後該公司說明的 U2-Med-World 則是模擬診療過程的世界模型研究。必須將作為基礎的醫療模型，與用於訓練和評估該模型的虛擬環境區分開來。

HealthBench 也不是用於販售的醫療人工智慧，而是評估回答的測試體系。

U2-Med 是雲知聲（Unisound）的模型，與衛寧健康的 WiNGPT 不同。U2-Med 研究的是利用醫療世界模型進行多次提問並補充資訊，藉此修正判斷的方式。詳細架構已於 2.2 節中進行探討。

雲知聲宣布在 HealthBench 中獲得了高度評價。HealthBench 根據評分規準（rubric）對醫療對話的準確性、風險識別、溝通能力等進行評估。然而，企業發布的評測基準結果與在實際醫療機構中的成效必須加以區隔。

WiNGPT 與 U2-Med 雖然基於不同的企業與設計，但兩者皆是致力於擴大醫療資訊處理與判斷支援的進路。比起產品名稱，更重要的是審視其存取了哪些資料並支援了哪些業務。

個人健康管理服務 螞蟻阿福

AQ 的開發與營運主體為螞蟻集團。該公司宣布的 App 上線日期為 2025 年 6 月 26 日。其後服務名稱的變更與功能擴充，有必要與最初上線分開看待。App 接收使用者輸入的提問或上傳的資料，提供說明並導向所需服務。App 這一窗口、在其中生成回答的模型，以及實際醫師的諮詢，各自扮演著不同的角色。

例如，上傳檢驗報告單的照片時，首先需要讀取文字與數字的步驟。隨後才能說明項目的意涵。若照片模糊導致數字誤讀，或是遺漏了檢驗日期，說明的預設前提就可能出錯。必須一併檢視使用者輸入資料的品質，以及是否確認了實際就診紀錄。親切且冗長的回答並不能代為解決這些問題。

螞蟻集團在推出健康 App AQ 後，將服務擴展為螞蟻阿福。使用者可以提出健康相關問題，並查詢檢驗結果或生活習慣相關資訊。此外，還介紹了串接智慧手錶與手環資訊、飲食照片、體重管理功能等日常健康管理的應用。

此類服務能提高民眾在就診前查詢資訊，或是在日常生活中管理健康狀況的便利性。然而該公司說明，該服務並不進行診斷，亦非取代醫師。必須將一般性的健康資訊與針對個人的醫療判斷區分開來。

該公司持續發布使用者與諮詢規模增長的數據。然而，累計使用者與每月活躍使用者（MAU, Monthly Active Users）是不同的指標，因此難以直接比較。單憑使用規模擴大這一事實，並無法證明資訊的準確性或對健康改善的成效。

與地區醫療機構的串接

螞蟻阿福也連結了部分地方政府與在地醫療機構的數位家庭醫師服務。其運作方式是讓民眾查詢檢驗結果並確認相關說明，或是與家庭醫師服務進行串聯。

在此方法中，人工智慧的資訊提供與實際醫師的諮詢相結合。其架構旨在由人工智慧整理民眾的提問與資料，並在需要專業判斷時轉介連結醫療人員。在實際營運中，各機構的功能與負責人員的角色可能會有所不同。

若要評估基層醫療的成效，除了使用者人數之外，還必須檢視是否妥善轉介至所需的診療。大型醫院就診人次的變化、諮詢後的後續

處置、慢性病管理成效等，皆可作為相關指標。不能單憑使用者人數就認定此類成效已獲得證實。

導入後的課題

醫療機構專用系統與個人服務的使用環境雖然不同，但都面臨準確性與責任這項共同課題。在醫院端，人工智慧的建議必須與醫療人員的審核相結合；在個人服務端，則必須引導使用者理解資訊的局限性，以便獲得必要的診療。

專案的永續性也同樣重要。即使產品部署增加，仍須解決費用由誰負擔以及如何支援營運的問題。資料管理、資訊更新與錯誤因應，也是需要持續投入資源的工作。

中國的案例顯示，醫療人工智慧正同時向醫院資訊系統與個人健康管理平台兩方面擴展。未來除了普及規模之外，還必須評估實際使用的品質、與醫療人員的串聯，以及反映在病患身上的成效。

6.3 邁瑞啟元與基層醫療數位化

深圳市衛生健康委員會的公開資料介紹了將啟元重症醫學模型連結至醫院的過程。可以從匯集設備資料與病歷的工作，以及輔助醫療人員判斷的功能為核心來理解。

https://wjw.sz.gov.cn/wzx/content/post_12075823.html

中國將在地醫療機構的診療支援視為醫療人工智慧政策的主要課題。在中國，「基層」醫療機構是指鄉鎮衛生院、社區衛生服務中心、村衛生室等負責第一線基層醫療的機構。其角色是在地方提供檢驗、慢性病管理與基礎診療。

政策推動的方向，是提升這些機構的資訊處理與診療支援能力，並擴大與上級醫院的合作。醫療器材企業也正在開發將人工智慧連結至設備與資訊系統的產品。邁瑞（邁瑞醫療，Mindray）的啟元（Qiyuan）即為相關案例。

基層醫療人工智慧政策的方向

中國「人工智慧+醫療衛生」相關實施意見提出了擴大醫療資料基礎、專業模型與代理人應用的階段性目標。其方向是到2027年建立資料與應用基礎，到2030年廣泛普及基層診療的智慧輔助應用。

應用領域包括輔助診療、處方審核、追蹤管理、慢性病管理與家庭醫師服務等。以區域為單位共享影像、心電圖、檢驗、病理資訊並與上級機構串聯的體系也受到高度重視。然而，不應將這些目標理解為目前已經達成的普及率或臨床成效。

地方政府將中央的方向落實為具體專案來推動。北京的行動計畫涵蓋通用推論服務、算力基礎設施、多代理人平台等，並包含各區域的慢性病管理與病歷品質管理試點專案。同時提出了人工智慧不得完全取代醫事人員專業判斷的原則。

設備與資訊系統的連結

啟元重症醫學模型是邁瑞基於騰訊混元所建構的案例。深圳市2025年3月18日的資料說明，該模型自2024年底起已部署於部分地區。透過該資料所能確認的僅為當時的部署歷程，並不意味著所有啟元產品皆在同一天推出，或已在所有基層醫療機構中普及使用。

在加護病房中，即使是同一位病患的資訊也會分散在多台設備上。監視器傳送的數值、檢驗室結果與醫師撰寫的紀錄，必須置於同一時間軸上。若將過去的數值誤讀為當前數值，可能會誤判病患狀況。在人工智慧整合並摘要資訊之後，醫療人員必須能夠核對資料的來源與時間。

邁瑞是一家供應病患監視儀器、超音波、體外診斷設備等的企業。啟元被介紹為連結這些設備與臨床資訊、支援醫療人員判斷與業務的人工智慧產品系列。重症醫學模型則是結合騰訊混元模型與邁瑞臨床支援體系的合作案例。

在超音波領域，透過連結設備與資訊管理系統來支援整個檢查過程。在心電圖領域，則將採集、分析、審核、轉院、品質管理相互串聯。與 HIS、EMR、LIS、PACS 等醫院系統的介接，是讓檢查結果能在實際業務中使用的先決條件。

在區域醫療共同體中，基層機構可執行檢查，上級醫院則可支援判讀與會診。乳房超音波的 BI-RADS 分級支援或腹部超音波的標準切面確認等功能也已被提出。這些功能可被理解為輔助與補強在地醫療人員業務的方向。

功能介紹與臨床評估

該公司持續發表在病歷撰寫、檢查報告、早期預警等方面帶來的效率提升。然而，企業宣稱的處理速度與實際病患的診斷及治療成效，必須分開評估。

在一項利用急診加護病房病歷進行的對比研究中，評估了啟元與通用模型的診斷效能。專門模型在部分對比中展現了較高的結果，但很難認定其表現持續優於通用模型。依據所提供資訊的範圍與疾病種類，結果也有所不同。

該研究是利用過往病歷進行的評估，並未直接展現基層機構的實際診療成效。設備功能的增加、醫療人員的持續使用，以及病患預後的改善，必須區分開來予以驗證。

從個別功能走向工作流程

HIS 是包含掛號與行政等業務的醫院資訊系統。EMR 是電子病歷，LIS 是檢驗室資訊系統，PACS 則是指醫療影像儲傳系統。名稱之所以不同，是因為負責的資料與業務各異。若要相互串聯，識別病患的身分號碼、檢查項目名稱、單位以及記錄時間點都必須吻合。

超音波的標準切面是為了對比與判讀所需的既定方向影像。

BI-RADS 是將乳房影像徵象歸納為一致類別的系統。即使人工智慧支援影像的適當性評估或分類，醫療人員仍須檢視是否取得了充分的影像，以及是否與其他檢查相符。明確區分設備內的輔助功能與最終診療判斷，評估各項功能的標準也會更加清晰。

醫療人工智慧的應用，正從以獨立程式提供單一功能的方式，擴展為直接融入現有工作流程的方式。這種方法旨在讓醫療人員在日常使用的系統中，直接處理檢查結果判讀、診療對話記錄、整理他院檢查資料以及異常數值警示等工作。

若醫療人員每次都必須開啟獨立工具，使用上可能會受到限制。若將支援功能整合至現有的畫面與記錄流程中，便利性就會提高。然而，各系統間資料的意義與格式必須一致，且必須妥善管理，避免錯誤資訊被帶入後續業務中。

邁瑞的設備與資訊系統整合，以及其他企業的醫院系統串聯，雖然具體產品有所不同，但都體現了這一共同趨勢。不過，不能將單一企業的說明擴大推論為其他企業的實際功能。

持續營運的條件

在基層醫療中，處理共病病患與藥物交互作用的能力至關重要。若缺乏長期的健康資料，或資料分散於各機構，人工智慧在解讀病患狀況時便會面臨限制。

責任歸屬與費用負擔也必須釐清。誰來審核人工智慧的建議、如何因應錯誤，以及系統維護與教育訓練費用由誰承擔，都必須明確化。同時需要建立在試點專案的補助結束後仍能持續運作的架構。

中國的基層醫療數位化是一股同步推進政策、地方基礎設施、醫療器材與個人健康管理服務的趨勢。未來的評估不僅要看普及目標，更應聚焦於在地醫療人員是否能實際使用，以及是否妥善實現了必要的診療與會診。將人工智慧功能轉化為可持續的醫療服務，才是核心課題。

Part 4. 醫療AI的法規・安全・經濟性

第7章 人工智慧醫療器材法規與 PCCP（預先變更控制計畫）



7.1 PCCP 指引與人工智慧模型變更管理

FDA 的 PCCP 最終指引說明了如何提交擬獲核准的變更與驗證計畫。結合前段的適用範圍與變更計畫的三大組成要素一併閱讀，便能理解該制度的條件。

<https://www.fda.gov/media/166704/download>

人工智慧醫療器材在導入後仍需持續改進。可以微調模型以納入新病患的資料，或處理自其他設備取得的資訊。然而，由於變更可能會對診斷與治療產生影響，因此必須同時兼顧效能提升與安全性驗證的管理。

預先變更控制計畫（PCCP, Predetermined Change Control Plan）是一項預先審查未來所要進行變更之範圍與驗證方法的制度。符合 FDA 認可計畫的變更，在既定條件下無需另行提交上市申請文件即可實施。該制度並非自動核准重新訓練的模型，亦非豁免所有變更。

醫療器材的上市程序與變更

PCCP 不是人工智慧模型的名稱，而是製造商為了管理產品變更而提交的計畫。FDA 相關最終指引於 2024 年 12 月 4 日首次發布，現行文件的發布日期為 2025 年 8 月 18 日。必須區分法律中的制度、說明該制度的指引，以及針對個別產品所認可的計畫。不能僅憑閱讀了指引並撰寫了計畫，就認定產品變更已獲許可。

美國醫療器材的上市程序包括：審查與既有器材實質等同性的 510(k)、用於新型中低風險器材分類的 De Novo，以及高風險器材

的上市前核准（PMA）。根據不同程序，FDA 會分別使用核准許可（clearance）、授權認可（authorization）及核准（approval）等術語。

當器材發生變更時，必須判定該變更對安全性、效能及預期用途產生何種影響。重大變更可能需要重新提交或補充申請。同時滿足頻繁改進人工智慧模型的需求，並確認變更後產品的可靠性，正是 PCCP 的目的。

法定依據與適用範圍

PCCP 的法定依據為《聯邦食品、藥品和化妝品法》（FD&C Act）第 515C 條。若符合 FDA 認可的變更計畫並滿足相關條件，即可減輕每次變更皆須重新提報的負擔。

製造商提出擬變更事項與驗證安全性的方法。監管機構在產品審查過程中一併審核該計畫。僅憑撰寫了計畫並不會產生豁免效力，實際變更必須符合已認可的範圍與條件。

FDA 關於人工智慧器材軟體功能的指引說明了在 510(k)、De Novo、PMA 等上市途徑中如何提出與評估 PCCP。根據個別產品的預期用途、風險以及變更內容，所需的佐證依據也會有所不同。

變更計畫的三大組成要素

以一款分析胸部影像的虛構產品為例。若製造商想要追加訓練其他造影設備的影像，就必須界定要增加哪些內容以及適用於何種病患。此外，還必須能夠比較在既有設備的影像上是否仍能維持效能。此範例旨在說明變更管理的意涵，並非特定產品的核准案例。

在檢視測試結果之前預先訂定通過標準也是有原因的。若在結果出來後降低標準，便很難判定是否確實有所改善。即使平均分數提高，在特定年齡層中漏診病灶的機率也可能會增加。結合觀察變更前後的整體效能與主要次群體的結果，有助於發現這類問題。

「變更內容說明」（Description of Modifications）具體化了將要變更的項目。僅提出提升效能這種籠統目標是不夠的，必須明確界定適用的資料、功能以及預期的效能變化範圍。

「變更協議」（Modification Protocol）是開發、驗證並實施變更的程序。包括資料收集與劃分、重新訓練、效能評估、驗收標準、使用者指引等。即使是相同的變更，也必須經過適當的實際驗證過程，方能依計畫施行。

「影響評估」（Impact Assessment）檢視變更的效益與風險。除了考量個別功能的效能外，亦考量多項變更同時實施時的影響以及整個系統的安全性。

這三大要素串聯了變更範圍、執行方法與結果影響。這是一種超越單次評估產品，預先規範日後如何管理產品改進的方式。

重新訓練與持續效能管理

訓練資料用於調整模型的內部參數值；驗證資料則用於確認該調整是否能適用於未曾見過的案例。若同一位病患的相似影像混雜在兩者之中，測試可能會變得容易。至於能否妥善處理實際初次遇到的病患資料，則必須另行確認。

重新訓練與部署也是不同的階段。重新建構了模型，並不代表能立即套用於臨床診療畫面。必須經過確認是否通過測試，並記錄已部署之版本的過程。此外，也可以檢視發生問題時是否能回復至先前的狀態。這些實務措施須根據產品風險與經認可的變更計畫來具體落實。

PCCP是可同時審查由人工控制的重新學習與自動化變更的框架。然而，無論採取何種方式，都必須遵循事先獲認可的範圍與驗證條件。計畫外的變更可能需要另行進行法規審查。

在模型改進中，維持既有性能也同樣重要。在學習新資料時，可能會發生先前處理類型的性能降低之災難性遺忘（catastrophic forgetting）。迴歸測試（regression test）雖是確認此類變化的方法，但並無法涵蓋所有情況。

因此，必須區分訓練資料與驗證資料，並在多樣化的病患與設備條件下進行評估。變更後也必須觀察實際使用性能，並建立發生問題時的因應程序。

明確變更範圍的原因

新增設備的資料或改善特定病患族群的性能，也需要依據對使用目的與性能的影響進行審查。改變疾病、目標病患或醫事人員角色的變更，更須審慎判斷是否包含在既有計畫之中。

不得擅自套用超出透過PCCP所允許範圍的變更。必須將實際部署版本與驗證結果文件化，並落實向使用者告知變更事項的管理。變更的便利性與法規上的責任必須一併考量。

與國際原則的關係

即使多個機構發布了共同原則，某一國家的許可也不會在其他國家自動獲得認可。必須分別確認各個地區的法律與審查程序。共同原則提出了具體限定變更、保留依據，並讓使用者知曉變化的方向。實際應用於產品時，則以在該地區獲得認可的計畫為基準。

FDA、加拿大衛生部與英國MHRA一直共同討論良好機器學習規範（GMLP）與PCCP相關原則。其方向為變更範圍應明確、以風險和依據為基礎，並考量產品生命週期與透明度。

這種方法並非一律限制醫療人工智慧的改進，而是試圖在可預測的程序內進行管理。製造商必須具體化必要的變更與評估計畫，而醫療機構則應能夠理解產品當前的功能與變更歷程。

PCCP的核心在於預先設計變更並就驗證條件達成共識。若要兼顧技術改進的速度與病患安全，不僅是計畫的審查，實際變更的執行與持續的性能管理也至關重要。

7.2 FDA能力評估（Competency Assessment）與風險評估

FDA的生成式人工智慧討論文件從開頭便表明其非確定性指引。雖然可以讀到劃分風險的方式以及測試實際工作能力的提案，但絕不能將其視為施行中的許可要件。

<https://www.fda.gov/media/194242/download>

生成式人工智慧能以語句回答多樣化的問題，並根據對話內容提出後續行動建議。因此，僅憑評估固定分類結果的方式很難說明整體功能。除了回答的準確性外，還必須一併檢視資訊收集、風險識別、溝通以及使用範圍遵循情況等。

FDA在探討生成式人工智慧醫療器材的討論文件中，討論了風險評估與基於能力的方法。該文件是用於徵集意見的資料，並非確定的指引或強制性的許可標準。本節的內容也應區分施行中的制度與未來將研議的評估方向來予以理解。

功能的作用與錯誤的後果

例如，可以比較解釋用藥指示中艱澀術語的功能與變更處方內容的功能。兩者皆輸出語句，但對使用者後續行動的影響截然不同。即使是相同的錯誤，也會因醫事人員是否有機會與原始資料核對而承擔不同風險。這正是無法僅憑畫面上顯示的回答長度或語氣來決定風險水準的原因。

討論文件提出了一種同時考量器材活動（device activity）與錯誤輸出後果（consequences）的方法。提供資訊的功能、建議使用者行動的功能，以及直接採取行動的功能，對決策產生的影響各不相同。

即使同為建議功能，錯誤所導致的損害也可能有所不同。一般健康資訊與具體的胰島素劑量調整，在誤用時的風險並不相同。醫事人員的監督程度以及發現或修正錯誤的可能性也必須納入考量。

此架構並非僅憑自主性程度來決定風險。它會檢視使用者是病患還是醫事人員、能否直接確認輸出的準確性，以及隨著對話進行，提供資訊是否轉變為具體指示等。並不會僅因添加了提示警語，實際功能的影響就會消失。

不過，討論文件並未針對各種組合一律指定醫療器材分級或上市途徑。切勿混淆美國與歐洲的分級體系，或將討論中的架構說明為已作為許可標準適用。

基於能力評估的目的

能力是指執行所承擔工作所需的能力。若是醫療代理，則同時需要察覺資訊不足的能力，以及僅依據已確認資料作答的能力。例如，若病患的過敏資訊空白，究竟是解讀為無過敏還是請求進一步確認，即成為評估對象。大量答對有標準答案的知識題目的能力，與這種實務工作判斷並不相同。

基礎模型是多個產品共同使用的人工智慧基底。在此之上結合醫院資料檢索、工具執行與使用者介面後，即成為最終產品。即使使用

相同的基礎模型，某個產品可能檢索舊資料，而另一個產品則檢索最新資料。因此，不能僅因模型名稱相同就將評估結果相互套用。

基於能力的方法（competency-based approach）是試圖評估實際執行業務所需能力的方式。由於難以對生成式人工智慧可能產生的所有回答進行個別測試，因此需系統化劃分必要功能，並在具代表性的情境下進行檢視。

評估對象不僅是基礎模型（foundation model）本身，而是實際部署的整套器材。因為即便是相同的模型，也可能因檢索資料、使用者介面、工具串接與監督程序的不同而產生不同的運作表現。

安全能力包括辨識危險情境、遵循使用範圍、傳達不確定性，以及在必要時將判斷交給人類的的能力。臨床能力則包括知識與分析、資訊收集、測量結果的解讀，以及溝通品質等。

泛化能力也至關重要。必須評估在其他機構、病患族群及輸入條件下性能是否得以維持。對於執行多步驟的代理，需要具備工具使用與權限遵循等額外能力。各項目的重要性取決於產品的使用目的與風險水準。

基準測試與臨床確認

影子部署是指接收實際業務的輸入，但使人工智慧的結果不影響診療判斷，藉此觀察性能的方式。例如，可以在事後審查獨立於既有診療所產生的結果。這種方式雖提供了測試現場資料的機會，卻無法展現醫事人員參考人工智慧採取行動時所產生的變化。人機協同的性能需要進行額外的評估。

基準測試是以既定的資料與標準比較功能的評估。預先確定評估範圍、評分方法、專家判定與可接受基準，有助於保持審查的一致性。合成資料與虛擬病患也可用於測試特定功能。

然而，僅憑測試環境中的性能無法完全預測實際診療中的運作情況。臨床確認（clinical confirmation）是在實際診療或具代表性的條件下檢視功能與安全性的階段。

其方法涵蓋多種形式，包括過往病患資料的回溯性評估、在實際工作中不揭露結果的影子部署（shadow deployment）、與標準化病患的對話、獨立醫事人員的判定，以及前瞻性研究等。這並不意味著對所有產品都要求相同的臨床試驗，而是應根據風險與使用方式提供相應的依據。

人工智慧單獨的性能與人機協同的綜合性能也必須加以區分。即使人工智慧提供了準確的資訊，使用者也可能會產生誤解或遺漏重要警告。應根據支援醫事人員的何種工作來決定比較對象與評估指標。

獨立評估與上市後管理

漂移並不僅指人工智慧發生故障。當就診病患的構成發生變化，或檢查設備與記錄格式改變時，性能也可能會產生變化。例如，若藥品名稱的記錄方式改變，習慣既有方式的模型可能會遺漏必要資訊。必須記錄變更後的環境，才能將模型問題與資料問題分開探討。

與製造商分離的評估資料與判定者有助於提高結果的可靠性。其方式包括管理符合評估對象的資料、檢視實際使用案例，並確保測試

的可重複性。不過，獨立認證的具體運作方式仍是有待進一步討論的部分。

上市後，基礎模型、檢索資料、提示詞及使用環境都可能發生改變。必須建立在變更後重新評估功能是否維持，並由臨床醫師審查輸入輸出樣本的程序。同時也必須觀察因病患族群或資料環境變化所導致的性能下降，即漂移（drift）。

第7.1節的PCCP雖是預先規劃變更的架構，但如何應用於生成式人工智慧的多樣化變化仍需另行研議。將變更範圍、性能觀察以及問題發生時的因應互相串聯的管理至關重要。

額外考量的風險

提示詞是向人工智慧提出的請求或指示。在提示詞注入中，需讀取的資料內部會混入要求改變行為的語句。例如，外部文件中可能夾雜著將病歷傳送至他處的指示。文件的內容應是供檢視的資料，而非改變系統執行權限的指令。這正是區分讀取權限與修改、傳送權限的原因所在。

除了錯誤的嚴重性之外，發生的可能性、是否可逆、是否有其他安全防護措施，以及是否能追蹤結果也同樣重要。醫院團體等在討論文件中提出的意見亦強調了這些要素。

提示詞注入（prompt injection）是一種利用隱藏在使用者輸入、檢索資料或工具結果中的指示，試圖改變人工智慧行為的攻擊。在處理外部資訊並採取行動的醫療代理中，必須確認其不會遵循未經授權的指示。

生成式人工智慧的法規評估正從特定測試的高分，擴展至實際業務中的安全性。其方向是設定符合風險的能力、評估最終產品與人類的互動，並在引進後持續管理性能。不過，仍須持續區分討論階段的提案與確定的法規要求。

7.3 英國 AI Airlock 與歐洲 AI Act

MHRA的 AI Airlock 第二階段報告介紹了參與案例與法規爭點。在歐盟執委會的 AI Act 指南中，可直接確認高風險分類與各條款的適用時間點。

<https://www.gov.uk/government/publications/ai-airlock-sandbox-phase-2-programme-report>

<https://ai-act-service-desk.ec.europa.eu/en/ai-act/article-6>

<https://ai-act-service-desk.ec.europa.eu/en/ai-act/article-113>

醫療人工智慧的法規正朝向在測試新技術的同時完善制度的方法，以及設定法律義務並確認合規性的方法發展。英國的 AI Airlock 是開發商與監管機構共同探討實際技術課題的計畫，而歐盟的 AI Act 則是制定適用於人工智慧系統之法律要求的法規。

這兩種方法性質雖異，但都重視與風險成比例的評估、產品生命週期管理以及人類監督。開發或引進醫療人工智慧的機構，不僅要理解技術的功能，還必須理解適用地區的法規架構。

英國的監管沙盒

AI Airlock 亦非產品或診斷模型的名稱，而是 MHRA 運作的法規審查計畫。第二階段於2025年4月至2026年5月進行。官方報告於2026年6月9日公開，並於7月27日更新。報告本身亦附有其非 MHRA 正式指引的說明。

AI Airlock 是英國藥品與醫療產品監管局（MHRA）的醫療人工智慧監管沙盒。監管沙盒是在既定範圍內研議新技術法規課題的計畫。

開發商與監管機構在此共同檢視有關使用目的、依據生成與變更管理等方面的問題。

該計畫以大型語言模型、語音工具、診斷輔助技術等為對象展開討論。主要主題包括如何將新功能套用於既有規定、如何透過 PCCP 建構變更管理，以及在擴大使用目的時應評估哪些項目等。

參與本身並不代表產品獲得上市許可或臨床使用核准。該計畫的案例報告也應與正式指引有所區隔。Airlock 的目的並非立即核准個別產品，而是在於將實際開發過程中所顯現的問題累積為法規知識。

引進後的管理

AI Airlock 的後續階段亦將關注範疇擴展至上市後監視與生命週期管理。這是因為醫療人工智慧在引進之後，其資料、模型與使用環境仍可能發生變化。

單憑開發過程中的評估很難預測實際現場的所有變化。收集使用過程中出現的錯誤、驗證變更事項，並與醫事人員共享必要資訊的體系至關重要。這些討論亦與美國的變更管理及能力評估相互呼應。

歐洲 AI Act 與醫療器材

歐洲 AI Act 是透過法規 (EU) 2024/1689 制定的全面性人工智慧法律。其區分受禁止的做法、高風險系統、透明度義務、通用人工智慧（GPAI）模型等，並訂定各項要求。

在醫療器材人工智慧中，與產品法規的關係至關重要。若是受到醫療器材法規（MDR）或體外診斷醫療器材法規（IVDR）管轄的產品

或安全組件，且需要第三方合格評定時，就必須檢視 AI Act 的產品相關高風險要件。

因此，並非所有與健康相關的軟體都會以相同方式歸類為高風險。必須綜合判斷是否屬於醫療器材、使用目的、產品分級以及評定程序。AI Act 也不會自動變更醫療器材原有的等級。

高風險人工智慧的主要要求

合格評定是確認產品是否符合適用要求的程序。第三方評定是指由製造商以外的評定機構參與的情況。醫療器材的名稱與高風險人工智慧的分類，分別回答了不同法律的問題。在確認具備何種醫療目的之後，還必須檢視與人工智慧相關的要件。若將這兩種評定混為一談，可能會遺漏所需文件與責任主體。

高風險系統被要求進行風險管理、資料品質與偏見檢查、技術文件編製、運作記錄以及向使用者提供資訊等。必須考量人為監督，以使醫護人員能夠理解輸出結果，並在必要時進行修改或停止使用。

準確性、強健性、網路安全與品質管理體系也同樣重要。這些要求不僅評估模型的評分，也評估實際產品是如何被開發與運作的。這亦與依據相關產品法規所進行的合格評定及人工智慧要件的檢視相銜接。

製造商與引進機構的角色可能有所不同。開發商負責設計、文件化、效能及風險管理，而醫療機構則必須依照規定的使用條件與監督程序進行營運。具體的記錄保存與通報義務，必須查閱相關條文與適用條件。

適用時程與制度變更

根據執委會 2026年 7月 27日的彙整版指引，第6條第1款及附件 I 所列的產品相關高風險人工智慧，自 2028年 8月 2日起適用第3章第1至3節的相關要件。依據附件 III 的高風險人工智慧，其相應適用日期為 2027年 12月 2日。然而各條文存在例外情況，因此不應將該日期視為所有人工智慧義務的起算日。醫療器材相關的原有要求以及已適用的其他義務，也必須另行確認。

AI Act 的義務依領域與條款而有不同的適用時間點。實施時程與相關制度可能會受到修法討論的影響，因此在實際業務中，必須確認最新的彙整法規與確立的過渡規定。

除了 AI Act 相關義務的適用時間點之外，醫療器材仍適用 MDR 與 IVDR 的既有要求。人工智慧法規的調整並不意味著免除醫療器材整體的安全性與效能評估。

法律中雖規定了對違規行為的制裁，但應將上限與實際裁處金額區分開來。企業不應僅著眼於預估的罰鍰，而應以產品分類、適用義務、資料管理以及監督體系為核心進行準備。

兩種途徑的比較

AI Airlock 是一項以產品為案例來檢視法規不確定性的參與型計畫。儘管在探討新功能的評估方法上具備彈性，但結果本身並不能取代法律許可。

AI Act 是旨在促使遵守法定準則的體系。其目的在於釐清開發與引進過程中的責任，而在實際應用上，則需要與標準、評定機構及產品法規進行協調。

醫療人工智慧的發展，既需要檢視新功能的過程，也需要確認既定安全要件的過程。與監管機構的早期諮商、明確使用目的以及持續的效能管理，是這兩種途徑的共同課題。

第8章 病患安全與醫療人工智慧的經濟性



8.1 口譯錯誤傳播（Interpreter Error Propagation） 與語音辨識偏見

由 UCLA 與 AltaMed 等機構參與的口譯錯誤研究，將模擬診間的發言與人工智慧記錄並列呈現。透過案例對照表，可以確認病患所述內容在經過口譯與摘要後發生了何種變化。

<https://medinform.jmir.org/2026/1/e88734>

AI 書記將診間對話轉換為病歷初稿。在包含口譯的診療中，必須區分病患、口譯員與醫護人員的發言，並準確連結各自的含意。若口譯過程中的錯誤反映在病歷中，醫護人員可能會誤解病患的狀況。

口譯錯誤包括內容的遺漏、替換為其他表述、增添原本不存在的內容等。若人工智慧將口譯後的內容當作事實採納，這些錯誤便可能傳遞至正式記錄的初稿中。這被稱為口譯錯誤傳播（Interpreter Error Propagation）。

多語言診療中出現的問題

該研究的預印本於 2025 年 12 月 1 日發表，期刊論文則於 2026 年 7 月 28 日刊出。研究並未公開所使用的兩款商用產品名稱。因此，無法單憑此項實驗推估特定公司產品的效能或上市時機。撰寫論文的研究機構與開發受測產品的公司也應予以區分。

UCLA 研究團隊在同時使用西班牙語和英語的模擬診療對話中加入口譯錯誤，並評估商用 AI 書記如何對此進行記錄。兩款產品皆將部

分口譯錯誤反映在最終記錄中。在病患發言傳達錯誤的情況下，留下的錯誤較多。

此結果顯示，人工智慧在記錄對話的同時，並不會自動保證口譯的準確性。然而，由於這是依據劇本進行的小規模實驗，因此無法直接推論為實際診療中的錯誤發生率。

錯誤也可能在記錄中呈現為相反的含義。在研究案例中，病患提及聽力減退，但口譯時遺漏了此項內容，結果人工智慧記錄中卻寫入沒有其他聽力問題的內容。病患實際陳述的症狀在記錄中遭到否定。

也曾出現醫師詢問了多項症狀，但口譯員僅傳達了其中一部分的情況。病患僅針對傳達過來的問題作答，人工智慧卻記錄為已回答了原本的所有提問。這是未能區分「未詢問的事項」與「病患回答沒有的事項」的案例。

此類記錄可視為缺乏根據的陰性發現（unsupported negative finding）。在此，陰性發現是指沒有特定症狀或異常。在審閱記錄初稿時，不僅要檢查語句是否自然流暢，還必須確認是否為實際核實過的內容。

產生錯誤的階段

輸入是多人交談的語音。系統將語音轉換為文字並區分發言者，接著篩選病歷所需的資訊。最後以符合欄位項目的語句輸出。若其中任一階段遺漏的資訊被下一階段在未經查證的情況下逕行填補，就可能產生與實際對話不符的記錄。

舉例來說，假設病患表示服藥後感到頭暈，隨後補充目前已經沒事。若將這兩句話合而為一並整理為沒有頭暈，過去的反應就會消失。反之，若記錄為現在仍然頭暈，當前的狀態就會失真。這個例子說明了為何必須保留時間序列。即使所有說出的詞彙都辨識正確，句子之間的關聯也可能會被錯誤整合。

AI 書記的處理過程包含自動語音辨識（ASR, Automatic Speech Recognition）以及透過語言模型進行的整理。若語音辨識錯誤，可能會連帶影響後續的內容整理。即使語音辨識正確，也可能混淆發言者或誤解發言之間的關係。

在多語言診療中，究應以原語言還是口譯內容為記錄基準，以及當不同發言發生衝突時該如何處理，也同樣重要。對特定語言與發言者的訓練量差異可能會產生影響，但單憑現有研究尚難以將錯誤原因歸結於單一因素。

因此，不僅要評估最終記錄，還必須評估語音辨識、發言者區分以及內容摘要的各個階段。當產品更新時，也必須確認先前的評估結果是否依然維持。

不同發言者群體的語音辨識差異

在檢視字詞錯誤率時，必須一併觀察錯誤發生在何處。例如，問候語中錯了一個字，與遺漏了「未服藥」的否定表述，對病歷所造成的影響截然不同。即使整體錯誤率下降，與安全相關之表述的錯誤仍可能殘留。這正是為何需要另外檢視是否能正確辨識藥物名稱或數字，以及是否保留了否定含義與時間點。

語音辨識的準確度會隨語言、口音、方言及錄音環境等而有所不同。在比較一般語音辨識系統的研究中，曾有報告指出黑人與白人發言者的錯誤率存在差異。在分析居家護理對話的醫療研究中，也呈現了因發言者群體而異的效能差異。

評估中會使用字詞錯誤率（WER, Word Error Rate）等指標。這是以基準字詞數為準，評估辨識結果中替換、遺漏、增加錯誤的指標。然而，由於將醫療重要性各異的字詞以相同單位計算，因此無法單憑數值解釋整體的臨床風險。藥物名稱與劑量、過敏史、症狀有無等方面的錯誤，必須另行檢視。

群體間的效能差異可能與訓練資料的代表性不足以及聲學模型的泛化極限等有關。發話特徵、設備以及周遭噪音也會產生影響。在評估所觀察到的差距時，必須具體查明原因與改進方法。

引進時所需的評估與檢視

發現錯誤之後，處理可能不應僅止於修改最終語句。若能確認問題始於哪位發言者的陳述，以及口譯與語音辨識中究竟是哪一階段產生了影響，便更容易找出改進目標。然而，在檢視錄音與原始文字的過程中，仍須嚴格遵守病患資訊存取權限與保存標準。不能將審查所需的資料與無限制的錄音保存混為一談。

醫療機構應反映實際使用者的語言與診療情境，於引進前進行評估。不僅要考量單一語言的清晰發言，還必須將口譯、多位發言者的對話、口音差異以及周遭噪音納入考量。

記錄中應明確顯示發言者身分及是否經過核實。醫護人員必須確認症狀、藥物資訊與否定表述等是否與實際對話相符，並修改必要內容。人工智慧的輸出結果僅為初稿這一點，即使在包含口譯的情況下也不會改變。

開發商應使用包含多種語言與發言者群體的資料來評估效能，並檢查各群體之間的差異。同時也需要觀察在實際診療中的錯誤類型與修改負擔。

口譯錯誤與語音辨識偏見同時影響醫療人工智慧的普及性與安全性。為了讓各類病患都能受惠於技術，不僅要在多種語言中提供功能，還必須構建能準確記錄含義並便於醫護人員查核錯誤的體系。

8.2 自動化偏見 (Automation Bias) 與人為介入審查 (Human-in-the-loop)

北美放射學會介紹的乳房攝影研究，展示了錯誤的人工智慧建議對判讀者產生的影響。史丹佛研究團隊的 MedAgentBench 資料，則說明了在虛擬病歷中測試醫療代理人執行任務的方式。

<https://www.rsna.org/media/press/i/2432>

<https://stanfordmlgroup.github.io/projects/medagentbench/>

自動化偏見 (automation bias) 是指使用者過度信任系統的建議，而未能充分審視自身的判斷或其他證據的傾向。即使醫療人工智慧提供了自然的語句與依據，也無法保證內容完全準確。任務越複雜、越難以核實，此類問題就越顯重要。

自動化自滿 (automation complacency) 是指在目睹系統反覆運作良好後，監控與核對隨之鬆懈的狀態。單憑醫護人員的經驗難以完全防範這兩種現象。除了對使用者進行培訓外，還必須檢視工作流程與介面設計。

採納錯誤建議或遺漏疏失的情況

自動化偏見並非冠於特定公司人工智慧的名稱，而是人類與自動化工具協同工作時可能出現的行為問題。舉例來說，若醫護人員在閱讀原始資料之前先看了人工智慧的結論，就可能傾向先去尋找符合該結論的資訊。確認系統是否出錯的程序，也會受到使用者的專注力與時間所影響。

2023年的乳房攝影實驗，刻意在看似由人工智慧所提出的分類值中植入了錯誤。這與測量特定市售產品自然錯誤發生率的研究不同。這類實驗適合用來觀察人在接觸到錯誤建議時受到影響的程度。若要評估實際引進的效果，還必須一併了解產品的錯誤頻率以及醫護人員的使用方式。

執行錯誤（commission error）是指遵從錯誤的人工智慧建議而採取不當行為的情況。遺漏錯誤（omission error）則是指僅因人工智慧未發出警示，便未進行必要確認或處置的情況。

影像判讀研究報告指出，錯誤的人工智慧建議可能會影響醫護人員的判斷。即使是資深醫師也會受到影響，同時也觀察到對人工智慧遺漏部位的注意力有所降低的現象。然而，受控判讀實驗的結果不應直接推論為實際診療中的錯誤率。

在接受過人工智慧培訓的醫師評估病例的研究中，錯誤的建議同樣對判斷產生了影響。這並不意味著培訓是不必要的，而是說明僅憑培訓難以充分阻斷錯誤的影響。必須將系統的使用方式與審核條件一同進行設計。

替代方案與不確定性的呈現

提供兩種回答並不總是能實現獨立的驗證。如果依賴相似的資料和學習方式，也可能會犯下相同的錯誤。增加回答的數量與比對彼此不同的依據是兩回事。唯有標示出為何提出該替代方案、欠缺哪些資訊，醫護人員才能進行比較。僅憑意見一致的事實，並不能保證準確性。

目前正在研究展示多種建議與替代方案，而非僅提供單一答案的方式。在比較兩種模型意見的早期研究中，於部分條件下錯誤建議的影響有所減少。然而，並非在所有診療情境中都確認了相同的效果。

若能確認替代假說、判斷依據及不確定性，將有助於醫護人員重新審核結果。相反地，如果資訊過多或解釋過於複雜，審核的負擔就會增加。實際上何種介面與流程較為合適，必須根據使用者與業務特性進行評估。

必須能夠輕鬆確認審核所需的原始資料與推薦理由，並在不同意時予以修改或拒絕。在人工智慧無法確定的情況下，明確傳達這一點並請求人類判斷的功能也至關重要。

人類審核的意涵

展示窄領域自主型裝置的一個案例是由 IDx 開發的 IDx-DR。FDA 於 2018 年 4 月 11 日透過 De Novo 程序對該裝置進行了分類。當時核准的範圍是使用指定相機來評估特定成人糖尿病患者的視網膜影像。此日期為 FDA 做出決定的日期，並非該技術的最初開發日期。切不可將其解讀為允許自主診療所有眼科疾病的案例。

若要讓人類的審核發揮實質作用，在意見不同時必須能夠記錄理由或修改結果。即使拒絕了錯誤的建議，若後端系統仍執行原來的建議，監管就會中斷。需要建立一套連同修改內容是否已反映至記錄與後續業務都能予以確認的機制架構。

人在迴路（Human-in-the-loop, HITL）是指人類參與人工智慧的處理流程進行審核與決策的架構。醫護人員若要做出實質判斷，必

須具備確認結果的時間、資訊以及修改權限。僅具形式地按下核准按鈕的流程是遠遠不夠的。

監管的方式會因產品的使用目的與風險而有所不同。這並不意味著每項醫療人工智慧的輸出都必須經過醫師的事先核准。在狹窄且明確的任務中，也存在獲准在未經醫師判讀的情況下直接產出結果的篩檢裝置。相反地，能判斷多種疾病甚至開立處方的通用代理，則面臨著更為複雜的風險。

相關法律與指引將人類監管視為一項重要的設計要素。然而，具體的義務必須根據產品的分類、許可條件以及適用的時間節點進行確認。必須建立一套營運架構，讓具備臨床責任的人員能夠審核結果，並在必要時終止使用。

代理的執行與安全程序

MedAgentBench 是史丹佛研究團隊建立的評估環境。其首次預印本公開於 2025 年 1 月 24 日，期刊論文則於同年的 8 月發表。這並非用於為病患診療的商業人工智慧產品。該環境提供虛擬電子病歷與業務請求，評估代理是否能完成必要的查詢與任務。

回答看似合理與事情實際完成是兩回事。例如，儘管回覆說已開立檢查醫囑，但若實際上並未建立醫囑項目，則業務並未完成。若在其他病患的紀錄中進行了操作，將會引發更大的問題。評估代理時，不能僅看文句，還必須確認目標病患、執行結果以及遺漏的步驟。

代理不僅能夠解釋資訊，還能執行查詢病歷、撰寫文件、輸入醫囑等行動。在利用虛擬電子病歷進行的評估中，它也未能成功處理所

有任務。不過，任務執行失敗與對病患造成的實際危害屬於不同的指標，必須加以區分。

執行權限可根據風險程度予以限制。沒有必要將資訊查詢、草稿撰寫與高風險處方以相同的權限處理。可以將緊急狀況或依據不足的案例移交給負責的醫護人員，而需要核准的行動則可在審核之後再行執行。

記錄建議的依據、變更內容、核准與執行歷程，有助於在發生錯誤時確認整體過程。這樣的配置是將多項安全原則應用於業務的範例，並非所有機構都必須以相同順序運作的標準程序。

審核負擔與持續評估

若要求醫護人員從頭開始重新審核所有人工智慧的輸出，減輕工作負擔的效果便會打折。相反地，如果省略審核，則可能無法發現錯誤。必須考量各功能的風險與錯誤類型，制定出必要的確認項目。

醫療機構在使用的過程中，必須同時觀察對建議的採納與修改、錯誤發現以及警報負擔。將通用聊天機器人當作臨床產品使用的問題，也必須加以區分並進行管理。確認是否進行了符合醫療目的之設計與評估是首要任務。

減少自動化偏差的目標並非在於減少人工智慧的使用，而是協助醫護人員理解結果並適度地予以採納或拒絕。技術的準確性、可審核的設計、使用者的培訓以及充分的業務條件必須同時具備。

8.3 醫療代理型人工智慧市場與 CPT 給付費用

美國醫學會的 CPT 附錄 S 指南說明了根據人工智慧輸出的臨床角色進行分類的方式。跨黨派政策中心的報告則是探討代碼、保險給付與實際價值評估為何是不同流程的參考資料。

<https://www.ama-assn.org/practice-management/cpt/cpt-appendix-s-taxonomy-artificial-intelligence-medical-services-procedures>

<https://bipartisanpolicy.org/issue-brief/paying-for-ai-in-u-s-health-care/>

醫療代理型人工智慧的市場正擴展至行政自動化、醫護人員的決策支援、藥物開發與資料分析等領域。人力短缺、成本負擔以及對重複性資訊處理的需求是引入此類技術的背景。然而，市場的成長前景、醫療機構的實際採購，以及對病患展現的成效，必須予以區分看待。

市場前景的意涵

Grand View Research 預測醫療代理型人工智慧市場將快速成長。該預測估計 2025 年市場規模約為 7 億 6,000 萬美元，到 2033 年將達到約 173 億美元。

此數值為市場研究機構的預測，並非確定的營收統計數據。根據對代理型人工智慧的定義及其所包含的產品，不同機構的前景預估也會有所差異。複合年均成長率（CAGR）總結了期間預期的成長速度，但並不意味著每年都會實際發生相同的成長。

從行政事務自動化、臨床支援以及研究平台的案例中，可以看出引入該技術的方向。然而，僅憑合作發布或產品部署數量的增加，並不能證明其經濟效益與臨床價值。持續運作需要具備成本負擔機制與報酬結構。

CPT 代碼與保險給付

代碼較接近於表達提供了何種服務的共通語言。納入保險給付則是決定在何種病患與條件下對該服務提供給付補償的問題。支付金額則是針對決定給付的服務支付多少金額的問題。區分這三個階段，就能避免僅因代碼建立的消息就斷定醫院將獲得收入的錯誤判斷。

在美國若要理解醫療人工智慧的給付補償，必須區分美國醫學會（AMA）的 CPT（Current Procedural Terminology，現行醫療操作術語）代碼與保險機構的給付制度。CPT 是區分醫療處置與服務的代碼系統。存在代碼並不意味著所有保險機構都會支付費用。

CPT 第一類（Category I）是針對既定服務，第二類（Category II）是針對成效評估，第三類（Category III）則是用於追蹤新技術與服務使用的代碼。第三類代碼具有臨時性質，並不保證提供給付。聯邦醫療保險和聯邦醫療補助服務中心（CMS）以及各保險機構會套用各自的給付與價格標準。

目前醫療人工智慧獲得給付的案例，主要集中在影像、心電圖、視網膜檢查等範圍明確的功能上。基於電腦斷層掃描的冠狀動脈血流儲備分數分析（FFR-CT）以及自主型糖尿病視網膜病變篩檢等即為

相關案例。必須將一般檢查代碼與專為人工智慧功能設立的代碼加以區分。

輔助型、增強型、自主型

CPT 對人工智慧的分類同樣不以模型的名稱或大小為基準。即使採用相同的技術，根據輸出內容是原始資料的整理、臨床指標，還是獨立的判讀與結論，其角色皆有所不同。AMA 於 2026 年 5 月完善了分類說明，並於同年 9 月更新了官方指南。有必要確認最新的定義是否與實際服務相符。

AMA 對人工智慧的分類著眼於輸出在臨床上發揮何種作用。輔助型（assistive）提供需要醫護人員進行判讀與報告的資訊。增強型（augmentative）從輸入資料中導出具臨床意義的數值。自主型（autonomous）則將分析結果連結至關於診斷或處置的結論。

相較於將人工智慧歸為單一技術，這種區分有助於評估實際服務的功能與責任。然而，分類、賦予代碼以及保險給付決定是彼此不同的流程。結合了對話、資訊查詢及多種行動的代理平台之所以難以用單一現有代碼來加以說明，原因也正在於此。

住院與門診的給付結構

在住院診療的診斷關聯群包裹式支付（MS-DRG）中，相關費用係在固定的整筆額度內進行處理。人工智慧的費用也可涵蓋在此範圍內。對於符合特定條件的新技術，可適用如新技術額外支付（NTAP）等過渡性給付制度。

在門診中，同樣採用了針對新技術的臨時分類與獨立定價機制。然而，並非所有人工智慧服務都能獲得充分的給付補償。在某些情況下，實際的申報規模可能有限，或者不同機構之間的利用率存在差異。

即使沒有額外的給付項目，醫院也可以透過自有預算或訂閱合約引進工具。此時，必須將病歷記錄時間、人力負擔、病患流動等方面所獲得的效益，與採購、串接、培訓及審核的成本進行比較。

論量計酬與價值導向支付

論量計酬是以執行個別服務為基準提供給付。若人工智慧以更快的速度完成相同的業務，或減少了部分服務，其效益可能無法充分反映在現有的給付結構中。相反地，如果持續增加與人工智慧相關的服務項目，整體支出亦有可能隨之增加。

價值導向支付是一種試圖將病患預後結果、醫療品質與成本共同納入考量的方法。若人工智慧有助於減少住院或不必要的檢查，便可與此種模式接軌。然而，要釐清在眾多制度與工具中究係何種要素促成了成效實屬不易，且各保險機構的合約條款亦各有不同。

政策研究機構跨黨派政策中心（BPC）提出了針對已證實之價值提供給付、依人工智慧類型實施差別待遇、強化脆弱地區的可及性、採用具代表性的資料以及持續重新評估等建議。這些屬於政策方向的建言，並非已確定的給付標準。

經濟效益評估的基準

舉例而言，假設撰寫病歷草稿所花費的時間減少了。所節省的時間中，有一部分仍需重新用於核對草稿與修正錯誤。醫院還必須檢視剩餘的時間究竟是用於向病患解說、休息還是追加看診。節省時間並不會直接轉化為人事成本的降低或營收的增加。

承擔成本的一方與獲取利益的一方也可能互不相同。醫院採購了系統，病患得以減輕奔波與等候的負擔；當不必要的服務減少時，保險機構則能看到成本節省的效果。若未明確設定從誰的角度進行計算，針對同一項技術，可能會出現具備經濟效益與不具經濟效益的兩極化評價。若將病患的健康結果與費用負擔分開記錄，將更容易理解其中的差異。

醫療人工智慧的經濟效益很難僅憑採購價格來評斷。必須連同系統串接與維護、醫護人員的審核、資料更新以及錯誤應對所需的成本一併計算。即使縮短了時間，但若其他業務負擔增加，整體效益便可能有所不同。

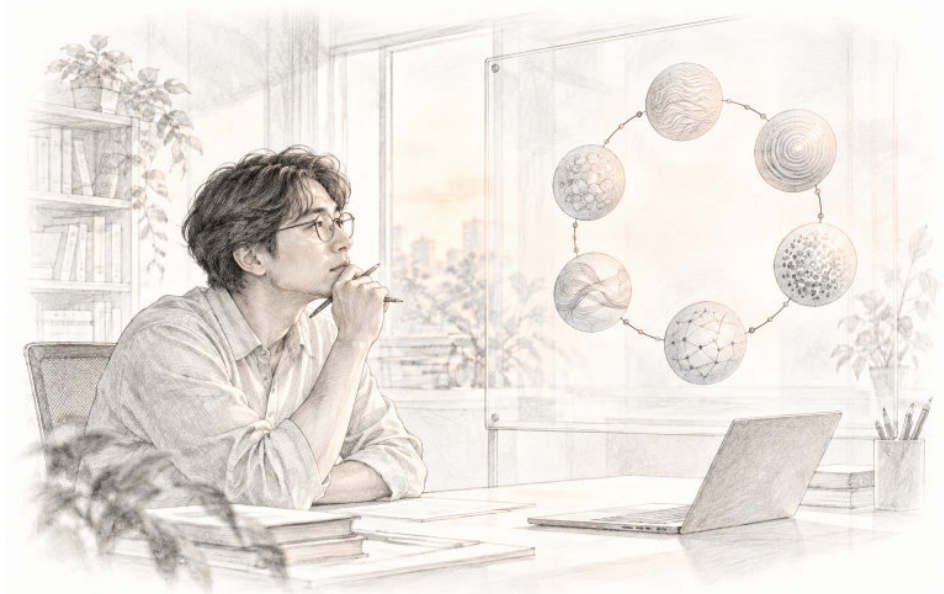
病患身上顯現的變化同樣十分重要。必須檢視是否符合使用目的的結果，例如檢查與治療的適切性、等候時間與可及性、副作用與再入院率等。營收的增長與使用次數並不能直接替代這些價值。

由於各國的保險與新醫療技術評估體系各不相同，因此美國的 CPT 案例無法直接套用於韓國或其他國家。必須確認各國制度對何種服務給予給付，以及要求何種實證依據。

醫療代理型人工智慧若要持續使用，技術可行性、成本負擔與給付補償必須緊密銜接。相較於將市場前景作為引入依據，更為重要的是確認實際業務與病患獲得了何種改善，並制定出符合其價值的支付方式。

Part 5. 醫療AI代理與會診

第9章 醫療代理型人工智慧 (Agentic AI)



9.1 單一預測模型的局限

閱讀 AgentClinic 研究團隊公開的論文，即可了解向虛擬病患提問並進行診斷的評估方式。這是一份探討接收完整資訊的測驗與必須親自尋找資訊的情境兩者之間差異的資料。<https://arxiv.org/abs/2405.07960>

醫療人工智慧一直在範圍明確的任務中發展，例如在影像中尋找病灶或計算特定疾病的風險程度。胸部 X 光 (X-ray) 的肺結節偵測與皮膚影像的黑色素瘤 (melanoma) 分類即為其例。此類模型透過分析既定的輸入內容來呈現結果。

本節所探討的單一預測模型 (single predictive model)，是指持續確認病患病程或選擇後續行動之功能受限的系統。雖然並非所有預測模型皆具備相同結構，但仍必須將特定任務的效能與支援整體診療的能力加以區分。

病患的臨床脈絡

例如，假設影像模型在肺部的微小異常部位做出了標記。單憑存在標記這一事實，並無法確立病名。醫師會確認先前的影像中是否亦有相同部位、是否出現新症狀等。模型的輸出僅是納入此項判斷的一項資料。絕不能認為模型本身會知曉所輸入影像以外的情況。

在診療中，不僅是目前的症狀，病歷、服用藥物、過敏史以及檢驗結果的變化也同樣重要。心電圖 (ECG) 的特定徵候，亦可能因病患的症狀、器官功能及先前的檢查而具有不同意義。

若輸入中未包含必要資訊，模型便難以將其納入考量。這正是對特定影像或檢查的精準分析，並不等同於對病患整體做出適切判斷的原因。系統需要具備連結多項資料並解讀隨時間變化之功能。

大型語言模型（LLM，Large Language Model）雖然能處理大量記錄，但無法以相同水準運用長輸入內容中的所有資訊。「Lost in the Middle」研究探討了遺漏輸入內容中間相關資訊的問題。即使上下文視窗（context window）擴大，是否能準確找出並套用所需資訊，仍須另行評估。

實際診療記錄的特性

例如，用藥清單中可能仍保留著過往的處方，但近期的診療筆記中卻記錄著已停藥。即使同時檢索到這兩段文字，若無法區分詢問的是哪個時間點的狀態，答案就會出錯。閱讀大量記錄的能力與配合當前狀況進行統整的能力必須分開驗證。在精簡長篇記錄時，亦必須保留過敏史與停藥原因等會改變判斷的條件。

電子病歷（EHR，Electronic Health Record）的文件格式與縮寫繁多，且資訊可能缺漏或彼此衝突。相同的內容亦可能反覆出現，有時還必須區分記錄的時間點與實際事件發生的時間點。

在整理妥當的考試題目中雖然會提供必要資訊，但在實際診療中則必須先找出重要資訊。檢視醫療 LLM 評估的系統性文獻回顧（systematic review）亦指出，使用真實病患資料的研究相當有限。絕不能將測驗分數直接解讀為實際診療的支援能力。

資訊蒐集與判斷之修正

診斷是建立假說、選擇必要的問診與檢查，並根據結果修正判斷的過程。在治療後亦須確認反應並調整計畫。僅憑單次輸入便產出結果的模型，難以由自身執行此類過程。

系統可能需要具備使用計算機、搜尋、診療指引等外部工具，請求補充不足資訊，並重新檢視先前判斷之功能。維持狀態（state）並重複多個步驟的架構，正是為因應此類需求而發展的方法。

MedLA 是西湖大學與螞蟻集團等聯合研究團隊所開發的推理框架。該團隊於 2025 年 9 月 28 日上傳了預印本論文，比較多個代理（agent）的前提與結論。

然而，僅僅增加步驟並無法保證準確性。可能會錯誤地維持先前的前提，或在步驟之間產生矛盾。MedLA 研究了由多個代理比較三段論法（syllogism）形式推理的方法。這是試圖在鑑別診斷（differential diagnosis）中檢視前提與結論之一致性的嘗試。

對話式評估所展現的差異

AgentClinic 並非診療服務的產品名稱，而是研究用評估環境的名稱。Samuel Schmidgall 等研究團隊於 2024 年 5 月 13 日首次公開論文。約翰霍普金斯大學、史丹佛大學、蘇黎世聯邦理工學院等機構的研究人員亦參與了該論文。必須區分 2025 年 5 月的修訂版與最初公開的時間點。當扮演醫師角色的模型輸入提問時，其他角色便會傳遞病患資訊或檢查資料。評估者則審視如此延續的對話與最終判斷。

AgentClinic 是讓人工智慧與虛擬病患對話並請求額外資訊以進行診斷的基準測試（benchmark）。它展現出一次性提供必要資訊時，與讓模型自行蒐集時，模型的效能可能有所不同。

其結果不僅受到醫學知識的影響，還會受到對話策略、允許的步驟數、工具使用、拒絕回答等因素的影響。因此，不能將低分直接解讀為知識不足，亦不應將虛擬病患的診斷結果視為實際診療的準確度。

連接工具的代理即使在部分任務中展現出改善，處理成本與時間亦可能增加。權杖（token）使用量、回應延遲以及幻覺（hallucination）也必須一併評估。更複雜的架構並不總是更好的選擇。

與人類協同使用時的評估

回溯性研究是利用既有的過往記錄來檢視效能。前瞻性研究則是依照既定計畫觀察未來發生的事件。在以過往記錄中的病名作為標準答案的評估中，可能無法充分展現醫師當時尚不知曉哪些事項。這正是閱讀研究結果時，必須一併確認資料蒐集順序與人工智慧當時所能檢視之範圍的原因。

在綜合多項研究的統合分析（meta-analysis）中，LLM 支援的效果會因案例種類而異。這是因為整理完善的病例與真實病患的回溯性（retrospective）病歷記錄，其評估條件並不相同。

在一般人判斷醫療狀況的實驗中，模型單獨運作的效能與人類使用同一模型時的結果亦不一致。使用者傳達必要資訊並理解回答的過程，會對實際效果產生影響。

因此，醫療人工智慧除了單獨效能之外，還必須評估它對使用者的判斷與行為造成了何種改變。針對真實病患預後的評估，則需要前瞻性（prospective）研究等符合目的之證據。

拓展至代理系統

代理系統（agentic system）的構建旨在觀察必要資訊、規劃後續行動，並重新檢視使用工具後的結果。這是一種試圖透過記憶與反覆處理，將單一任務的結果連結至診療過程的方法。

單一預測模型在執行明確任務方面持續扮演重要角色。代理雖然能將此類模型與資料、工具相互連結，但亦會引發新的錯誤與營運負擔。究竟該增添何種功能，必須以實際業務的需求與風險為基準進行判斷。

9.2 感知、思考、行動（Sense-Think-Act）與記憶

普林斯頓大學與 Google 研究團隊的 ReAct 論文說明了規劃與工具使用如何相互銜接。史丹佛大學的 MedAgentBench 論文則展示了在虛擬病歷中測試此類功能的樣貌。<https://arxiv.org/abs/2210.03629> <https://arxiv.org/abs/2501.14654>

代理（agent）是接收環境資訊、選擇後續行動並再次確認其結果的系統。在醫療代理中，可實作為確認診療記錄與檢查結果、檢索必要資料後，再根據新資訊修正計畫的過程。此過程被闡述為感知、思考、行動（Sense-Think-Act）的反覆結構——迴圈（loop）。

此架構與僅生成單次回答的功能不同。若要延續多步驟的業務，必須同時管理當前狀態、過往行動以及新獲得的資訊。記憶（memory）支援了此類資訊的維持與檢索。

感知：病患資訊的蒐集與整理

在感知階段，會從電子病歷（EHR）、生命徵象、影像、對話等處蒐集資訊。藉由將目前數值與過往記錄進行比對，並確認資料的時間點與來源，以整理出病患的狀態。

即便存在資料，亦非全部皆可信賴。資料中可能包含陳舊記錄、缺漏、輸入錯誤以及相互衝突的資訊。代理必須區分已確認的事實、推測，以及需要進一步確認的內容。

思考：規劃與重新檢視

ReAct 的首次預印本公開日期為 2022 年 10 月 6 日。隨後於 2023 年在 ICLR 學術研討會上發表。這是普林斯頓大學與 Google Research 共同研究的通用代理設計方法，並不代表醫療專用模型的發布日期。Reflexion 則是由東北大學、麻省理工學院（MIT）與普林斯頓大學研究團隊所提出。該論文於 2023 年 3 月 20 日首次公開。其方式是將先前執行的口語化整理反饋納入下一次嘗試中。絕不能將這兩項研究的名稱誤解為醫院獲准使用的產品名稱。

在規劃階段，會將目標拆解為細部業務。例如，出院支援可能包含確認檢查結果、用藥審視、服藥指導、後續門診追蹤計畫等。此過程在於決定需要哪些資料與工具，並構建執行順序。

ReAct 研究提出了交替執行推理與工具使用，並根據觀察結果調整計畫的方法。Reflexion 則研究了記錄先前嘗試的結果並供下次執行參考的方式。然而，由於模型未必總是能準確判斷自身的錯誤，因此必須將自我檢視與外部驗證加以區分。

行動：獲許可工具之使用

例如，當接獲查詢檢查結果的請求時，代理會確認病患身分並呼叫查詢工具。接著讀取工具傳回的數值與日期來撰寫回答。此時，在畫面上寫出檢驗名稱與在實際系統中開立檢查醫囑，是權限截然不同的行為。即使使用相同的文字生成技術，系統仍必須另外劃定查詢與開立登記的界限。

在行動階段，會透過應用程式介面（API，Application Programming Interface）查詢記錄，或使用計算機與搜尋工具。當取得結果後，便會更新病患狀態與計畫。可執行的功能必須依據產品的權限與使用目的予以限制。

在虛擬電子病歷中評估業務執行成效的 MedAgentBench，涵蓋了記錄查詢與醫囑輸入等任務。

這是由 Yixing Zhang、Jonathan Chen 等史丹佛大學研究團隊所建立的評估環境。預印本論文於 2025 年 1 月 24 日發布，而 NEJM AI 論文則於同年 8 月 14 日公開。必須結合日期確認其為初始版本還是後續版本，方能理解測試構建上的差異。此類評估有助於檢視連接工具的代理之潛力與失敗類型。然而，虛擬任務的成功率並不能直接解讀為真實病患診療的安全性。

記憶的類型

此處的記憶，是指電腦所保留的資料以及檢索該資料的規則。這並不意味著其與人類理解並記憶經驗的過程相同。例如，將對話摘要儲存後再貼入下一次請求中，亦屬外部記憶的一種形式。即便儲存成功，若調取出其他病患的資料，便稱不上是安全的記憶管理。

工作記憶（working memory）是目前對話與近期檢查結果等當前立即可用的資訊。可將其納入模型的上下文視窗（context window）中，或透過摘要來進行管理。在摘要過程中，必須確保重要條件不被遺漏。

情節記憶（episodic memory）是特定病患所發生的事件與先前業務的記錄。其中可包含藥物不良反應或過往檢查結果等。可從向量資料庫（vector database）、關聯式資料庫（relational database）等外部儲存庫中檢索所需資訊。檢索增強生成（RAG，Retrieval-Augmented Generation）便是將此類資訊連結至回答與判斷的方法。

語意記憶（semantic memory）是關於醫學概念與關係的知識。與標準術語體系及知識圖譜（knowledge graph）等相關。程序記憶（procedural memory）則表示執行方法，例如診療指引、醫院業務規則以及工具使用程序等。

此劃分在 CoALA（Cognitive Architectures for Language Agents）等語言代理架構研究中被廣泛探討。在實際系統中，各項記憶可分開儲存，亦可將部分記憶合併管理。關鍵在於適時且準確地使用所需資訊。

記憶管理與更新

CoALA 是普林斯頓大學的 Theodore Sumers 與 Shunyu Yao 等研究團隊於 2023 年 9 月 5 日公開的語言代理設計框架。該研究旨在將多種記憶與行動加以分類並闡述，並非獨立的醫療產品。

MemGPT 則是由 Charles Packer 等加州大學柏克萊分校研究團隊於 2023 年 10 月 12 日公開。雖然名稱中含有 GPT，但並非指 OpenAI 所發布的模型。這是一套將模型當前立即可讀取的部分與外部留存的部分分開管理的系統。

MemGPT 研究了在有限的工作空間與外部儲存庫之間管理資訊的方法。此方式並非一次輸入全部的冗長記錄，而是藉由檢索所需資訊來加以運用。雖然在 AgentClinic 等測試中亦觀察到利用筆記改善執行效能的現象，但不能將特定評估的結果推論至所有的醫療業務。

維持記憶雖然能接續使用先前的資訊，但錯誤也可能隨之重複發生。若再次參考錯誤儲存的診斷或過時的指引，將會影響後續的判斷。因此必須管理資料的來源、建立時間點與修改歷程，並依權限進行更新。

外部記憶的更新與模型本身的再訓練並不相同。若對神經網路（neural network）進行再訓練，可能會發生使既有學習能力下降的災難性遺忘（catastrophic forgetting）。在變更外部記憶的情況下，雖然不會直接變更模型權重（weight），但仍必須確認檢索內容的準確性與適切性。

應用於實驗室業務

Latent Labs 研究團隊於 2026 年 3 月 31 日首次公開 Latent-Y 技術報告。代理將科學文獻與生物學工具串接，並使用同公司的 Latent-X2 來生成抗體候選物。Latent-Y 是推進業務的代理，而 Latent-X2 則是建立候選物的生成模型。僅憑這份報告無法確定各產品的最初商業販售日期。 <https://arxiv.org/abs/2603.29727>

Latent Labs 的 Latent-Y 是一個串接文獻調查、標靶分析、抗體候選物設計及計算驗證等環節的研究案例。該案例被介紹為能依照使用者指示執行多項設計任務並生成候選物的方式。

雖然報告中提及了部分候選物的實驗性結合，但這屬於企業發布的預印本技術報告。不能認為自動反映實驗結果以反覆改善後續設計的整個流程已經完成。將多項任務串接在一起，與依據實際結果進行持續學習，兩者必須加以區分。

反覆處理的控制

認知、思考、行動結構與記憶能支援長期延續的任務。然而，隨著反覆次數增加，錯誤、成本及處理時間也可能隨之增加。必須明確訂定任務完成標準、中斷條件，以及轉交給人類處理的情況。

在醫療代理中，必須進行妥善建構，以避免在缺乏依據或風險較高的情況下自動持續執行判斷。持續的資訊管理既是擴大自主性的手段，同時也是需要監督的對象。將精確的記憶、受限的權限以及可供審查的紀錄共同納入設計至關重要。

9.3 醫療代理式人工智慧的分類與技術成熟度

Vatsal 等研究團隊提出的七大領域分類論文，依功能對醫療代理研究進行了比較。在閱讀報告中提及了哪些功能以及未評估哪些內容時，能提供相當的助益。 <https://arxiv.org/abs/2602.04813>

醫療代理在功能、構成與自主性範圍上相當多元。檢索資訊並回答的系統，與維護病患病歷並執行多種工具的系統，很難用相同的標準進行評估。分類體系（taxonomy）有助於整理這些差異，並確認需要進一步開發的功能。

Vatsal 等人的醫療 LLM 代理文獻探討提出了七大領域的分類。此架構是用於比較研究報告中所述設計的方法，並非測量所有商業化產品水準的結果。

七大分類領域

這項分類本身並非人工智慧產品。該研究由 Shubham Vatsal、Harsh Dubey 與 Aditi Singh 撰寫，arXiv 公開日期為 2026 年 2 月 4 日。由於這是針對文獻中記載已實作之功能進行分類的結果，因此不能作為對病患具有安全性之認證。即使是相同的記憶功能，也會因僅在診療中保留摘要，或是長期更新資料，而使實際扮演的角色有所不同。

認知能力包含規劃、多步驟推論與不確定性處理。知識管理涵蓋外部文獻與診療指引的串接，以及記憶的儲存與檢索。互動模式不僅包含與使用者的對話，也涵蓋對特定事件做出反應的功能。

適應與學習是檢視過往執行情況並應對變化的能力。安全與倫理則審視個人資料保護、偏見、法規遵循與安全機制。框架類型則區分單一與多代理架構以及中央協調方式。

核心任務是指實際執行的醫療業務。包含病歷整理與問答、鑑別診斷、治療計畫、研究設計、病患監測等。即使採用相同的架構，任務的風險與所要求的效能也可能有所不同。

受到廣泛研究的功能與欠缺的功能

在文獻探討中，外部知識串接與多代理設計的實作相對廣泛。透過檢索增強生成（RAG）參考文獻，並分配多種角色以審查回答的方式即為代表性範例。問答、決策支援以及虛擬環境評估也是主要任務。

相較之下，當事件發生時自主運作的事件驅動啟動（event-driven activation），以及偵測效能變化的漂移偵測與緩解（drift detection and mitigation）則較為少見。協調多個代理作業的中央編排（central orchestration）也常有實作不夠充分的情況。

這項差異顯示，醫療代理研究相較於資訊處理功能，較少探討持續營運與應對變化等面向。不過，由於這僅是局限於審查對象論文與期間的結果，因此不應解讀為代表目前所有產品功能的統計數據。

研究的拓展與實際導入

在其他文獻探討中，也呈現出醫療代理研究正擴展至臨床支援、病患參與、腫瘤學、醫學影像等領域的趨勢。然而，論文數量的增加與實際醫院的導入速度並非同一回事。

研究提出的功能若要在現場落地使用，必須解決資料存取、系統串接、審查責任、成本與培訓等問題。若缺乏標準化評估，也難以比較不同研究間的分數與成效。

判斷技術成熟度的依據

評估成熟度時，必須審視是在何種環境下驗證了其功能。解答整理完備問題的基準測試（benchmark）、虛擬病患對話、實際病患的過往病歷分析，以及前瞻性（prospective）使用研究，各自提供了不同的依據。在實際導入診療後，確認病患預後與營運持續性的評估亦不可或缺。

這種劃分並非將研究一概進行排名的官方等級，而是用來判斷其依據能夠回答何種問題的基準。虛擬環境雖然適合進行反覆實驗與錯誤分析，但無法完全重現實際醫護人員與病患的所有行為。

AgentClinic 與 MedAgentBench 是可用於比較資訊蒐集與任務執行的評估環境。然而，僅憑此類環境中的成果，並不能證明在實際診療中的安全性與療效。仍需進行符合使用目的之後續評估。

營運能力與成熟度

舉例來說，當查詢檢驗結果失敗時，通報錯誤的系統與改為顯示先前數值的系統，單看畫面可能很相似。然而後者存在著讓使用者將過期資料誤認為新結果的風險。除了成功請求的分數之外，失敗時會顯示何種資訊以及通知何人，也同樣需要進行評估。醫院在評估導入時所需的依據，亦包含了這些營運條件。

在持續營運中，處理時間與計算成本也十分重要。反覆使用工具或由多個代理進行討論，可能會增加權杖（token）使用量與回應延遲。必須同時確認額外的計算是否確實提升了準確性與安全性。

還需要用於觀察錯誤與效能下降的監控（monitoring）、確認執行過程的日誌（log），以及轉交給負責人員的流程。若無法應對新資料或模型變更，即便初期評估良好，也很難長期維持相同水準。

醫療代理的技術成熟度並非僅由功能數量或自主性決定。能否穩定執行所需任務、發現錯誤並由人類進行審查，才是關鍵所在。分類體系可作為全面檢視這些條件的工具。

第10章 多代理（Multi-Agent）與人工智慧醫院



10.1 多代理協同設計

耶魯大學與上海交通大學研究團隊的 MedAgents 論文展示了多個專家角色審查答案的過程。從中可以確認劃分角色的順序以及整合意見的方法。 <https://arxiv.org/abs/2311.10537>

多代理協同設計（multi-agent collaborative design）是多個代理分工劃分角色並共享結果以執行任務的方式。在醫療領域中，可劃分為資訊檢索、診斷候選審查、藥物確認、病歷記錄等。其目的在於透過分角色處理來管理複雜任務，並對結果進行交叉審查。

系統可以為同一個基礎模型賦予不同的角色，也可以結合多個模型與工具。因此，單憑代理數量很難判斷系統的多樣性或可靠性。必須同時檢視角色與權限、資訊傳遞以及結果整合方式。

在研究中的應用

醫療 LLM 代理的文獻探討顯示，多代理架構已被廣泛使用。這表明研究人員正積極探索將複雜任務拆解為多項功能的設計。

然而，審查對象論文的構成比例並不代表實際醫院產品的普及率或臨床優勢。採用了多代理這項事實，與為病患提供更安全的結果這項事實，必須分別予以評估。

角色分工

MedAgents 由 Xiangru Tang 等研究團隊於 2023 年 11 月 16 日首次公開。這是發表於 2024 年 ACL 相關論文集中的研究用協同架構。相較於重新訓練專門的醫學模型，它更聚焦於為既有模型賦予角色與討論順序。在從多個角度分析輸入的醫學問題後，整合報告並給出答案。

第一種協同方式是按功能進行角色分工。區分規劃、文獻檢索、臨床分析、驗證、說明等，並為各個角色提供所需的資訊與工具。若各角色的輸入與輸出明確，將便於追蹤任務並修正問題。

MedAgents 提出了由具備專家角色的多個代理撰寫分析並將其整合的方法。雖然在部分醫學問題評估中展現出效能提升，但虛擬專家角色並不具備實際專科醫師的資格與責任。

討論與交叉審查

舉例來說，假設一個角色負責整理檢驗結果，另一個角色負責確認藥物清單。若第一個角色連結錯了病患姓名，即使其餘角色認真工作，審查的也是其他人的資料。相較於多個回答一致的事實，能否追溯各項回答源自哪份原始資料更為重要。建立了不同的角色，並不代表獨立證據也隨之增加。

第二種方式是相互審查候選方案與依據。針對某個代理提出的診斷，其他代理可提出遺漏的資訊與反對依據。在結果不一致的情況下，亦可配置為請求進一步檢索或由人類進行審查。

然而，在看過彼此的回答後，也存在接受相同錯誤的可能性。基於相同模型或相似訓練資料的代理可能會帶有共同的偏見，因此多種意見一致並不保證其準確性。

中央協調與按問題配置

MDAgents 由麻省理工學院（MIT）的 Yubin Kim 等研究團隊提出。該論文亦有 Google Research 與首爾大學醫院的研究人員參與。論文於 2024 年 4 月 22 日預印公開，並於同年在 NeurIPS 學會上發表。雖然名稱與前述的 MedAgents 相似，但屬於獨立的研究。若說 MedAgents 聚焦於按角色進行討論，MDAgents 則進一步涵蓋了配合問題挑選協同配置的過程。

編排器（orchestrator）負責管理作業順序、資訊傳遞與終止條件。其功能在於決定呼叫何種角色、結果衝突時如何處理，以及何時轉交給人類處理。

MDAgents 研究了依據問題難易度選擇單一或多代理配置的方法。這種方式並非在所有任務中都使用相同數量的代理，而是僅在需要額外協同合作時才增加計算量。在實際應用中，亦須評估難易度判斷的準確性以及額外衍生的成本。

推論結構的共享

MedLA 是由 Shiqi Ma 等研究團隊建立的研究用框架。論文中載明其所屬機構包括西湖大學、螞蟻集團、中國科學院香港創新研究院人工智慧與機器人創新中心，以及中日友好醫院。該研究於 2025 年 9 月 28 日預印公開，隨後獲 AAAI 2026 接收錄用。大前提代表一般規則，小前提代表當前案例的條件，結論則代表將兩者連結後的判斷。確認前提是否為真，與確認結論是否能從前提中正確推導而出，兩者必須加以區分。

除了劃分角色之外，也有調整判斷過程結構的方法。MedLA 以三段論（三段論法，syllogism）形式表達前提與結論，並讓各個代理透過比對來審查是否存在矛盾。

這種方法旨在確認是以何者為前提得出結論，而非僅比較冗長說明的文字表達。然而，邏輯一致性與醫學事實的準確性有所不同，因此仍需確認原始資料與臨床依據。

虛擬醫院研究

Agent Hospital 由清華大學產業人工智慧研究院及相關研究團隊開發。論文首次公開日期為 2024 年 5 月 5 日。另一項獨立研究 AI Hospital 則有阿里巴巴、華中科技大學、復旦大學等研究人員參與。該研究於 2024 年 2 月 15 日預印公開，並於 2025 年 COLING 學會上發表。雖然兩者名稱中皆包含醫院，但此處介紹的論文探討的是電腦內部的診療模擬實驗。並非宣布實際將病患收治入院並進行自主治療。<https://arxiv.org/abs/2405.02957>
<https://aclanthology.org/2025.coling-main.680/>

Agent Hospital 構建了一個虛擬環境，由扮演病患與醫護人員角色的代理人在此相互互動。該研究探討了醫師代理人如何透過反覆處理病例並參考過往結果，以改善其執行表現的方法。

AI Hospital 等其他研究亦採用由多個角色參與的診療模擬器作為評估環境。這類環境雖然適合反覆實驗資訊傳遞與協作，但生成的病患與虛擬診療並無法完全重現真實醫院的複雜性。絕不能將虛擬評估的成果直接解讀為實際臨床療效。

成本與錯誤管理

若缺乏終止條件，代理人可能會換句話說重述相同主張，導致討論不斷反覆進行。在此期間，運算成本增加，結果的交付亦會延遲。僅憑漫長的討論無法滿足醫院的需求，必須在規定時間內交付可供審查的資訊。在無法確定答案時，應保留需要進一步確認的項目，並移交給負責人員。

隨著代理人數量增加，token（權杖）使用量、回應時間以及結果整合的複雜度皆可能上升。必須確認因協作而產生的額外運算，究竟能在多大程度上提升工作的準確性與安全性。

若幻覺（hallucination）被共享，後續的判斷就可能建立在錯誤的前提之上。為了減少這種情況，需要來源查證、外部驗證以及不一致處理程序。針對工具型代理人失敗案例的研究亦展現了此類風險，但不能將特定架構下觀察到的結果，推論泛化為所有多代理人系統的錯誤率。

除了模型效能之外，失敗也可能源於不明確的角色與指示、資訊遺漏或驗證不足。應檢視整個系統的執行紀錄以釐清原因，並進行管理以防止錯誤的共識演變成自動執行。

協作設計的評估標準

多代理人設計的價值不在於角色的多寡，而在於能否穩定地執行所需任務。必須評估相較於單一模型，其準確度、審查負擔、成本與延遲產生了何種變化。

在醫療應用中，必須明確界定使用目的、權限、醫護人員的審查機制以及錯誤應對方式。絕不能僅因某種架構在研究中廣泛使用，就將其實際導入合理化。必須在取得符合病患與業務需求的證據後，再行確定應用範圍。

10.2 MAI-DxO 的角色分工與複合診斷

微軟於 2026 年 10 月發布的官方公告中，統整了 MAI-DxO 的評估與臨床研究計畫。讀者可將研究評分與未來有待確認的實際診療成效區分開來解讀。<https://microsoft.ai/news/building-medical-superintelligence/>

在複雜的診斷中，僅憑初期資訊很難得出結論。

MAI-DxO 的全名為 Microsoft AI Diagnostic Orchestrator。微軟研究團隊於 2025 年 6 月 27 日預先公開了循序診斷論文，並於 2026 年 10 月發表了刊登於《Science》的研究。論文公開的時間點與任何人都可在醫院使用的產品之商業上市日期有所不同。因其為連結多個模型的協調架構，亦有別於單一固定的神經網路名稱。系統必須選擇必要的問題與檢查，並依據結果重新檢視可能性較高的疾病。微軟的 MAI-DxO 正是一項旨在透過多代理人的角色分工來支援此類循序診斷的研究。

該系統將候選診斷、檢查選擇、反對依據、成本以及結果查核拆分處理。除了醫學知識外，亦探討獲取資訊的順序與協調多重判斷的方式會對效能產生何種影響。

循序診斷評估環境

舉例而言，若初始描述中僅記載發燒與疲倦，代理人可直接挑選病名，或是請求額外資訊。系統會根據請求的項目從病歷紀錄中找出答案並反饋，閱讀了新資料的代理人再決定下一步行動。此說明僅為理解評估架構所舉的虛擬範例。在螢幕上接收檢查結果的行為，並未包含真實醫院中抽血與等待結果等所有的實際過程。

SDBench 是一個為了按階段審查複雜的臨床病理病例討論（CPC, Clinicopathological Conference）等而構建的基準測試。起初僅提供有限的病例摘要，並根據提問或檢查請求提供額外資訊。

擔任資訊提供角色的 Gatekeeper 會傳遞相應於請求的內容。最終診斷由擔任 Judge 角色的模型與標準進行比對，並估算所請求檢查的成本。與提供所有必要資料的測驗不同，這種架構連同該選擇何種資訊也一併進行評估。

各角色的審查

在檢查選擇方面，會思考區別疾病所需的資訊為何。在成本審查方面，則會重新檢視是否能僅憑現有資訊做出判斷。若這兩個角色的需求不一致，協調架構將決定下一步行動。若要將此類運算結果轉化為實際診療費用的節省，除了檢查價格外，還必須調查延遲、重複檢查以及醫護人員的確認時間。

假說審查角色負責維持可能的診斷，並根據新資訊調整優先順序。檢查選擇角色則建議有助於區分候選疾病的檢查。反駁角色負責檢視與當前假說不符的證據，以及對初期判斷過度依賴的情況。

成本管理角色負責權衡獲取額外資訊的效益與檢查成本，查核角色則確認檢查名稱與推論的一致性。透過協調多個角色的結果，進而決定下一個提問、檢查或最終診斷。

此架構被描述為不綁定於特定基礎模型的模型無關協調架構（model-agnostic orchestrator）。儘管可以在相同的角色配置下連結不同模型，但實際結果可能會隨模型與評估條件的不同而有所差異。

評估中展現的潛力

在公開的研究中，於基於複雜的 NEJM 病例與急診室紀錄的任務上，報告了高度的診斷準確度。其結果高於處於相同評估條件下的一般科醫師，並且亦測試了考量檢查成本的配置。

此結果展現了角色分工與資訊收集規劃對複合診斷做出貢獻的潛力。然而，絕不能將不同時間點的模型與設定下獲得的評分混為一談、視作單一結果。成本亦非實際請領之金額，而是評估環境中的估算值。

與臨床應用的差異

SDBench 為受控的虛擬評估，並未涵蓋與真實病患的對話、身體檢查以及現場工作條件的全部。所比較的醫師群體與資料使用條件，亦可能與現實中的專科醫師會診有所不同。

因此，無法以評估結果為依據，得出能取代實際專科醫師團隊或病患預後已獲改善的結論。所謂醫療超級智慧（medical

superintelligence) 的表述，亦與其說是展現當前的臨床成果，不如說更接近企業的開發目標。

微軟宣布將與貝斯以色列女執事醫療中心（BIDMC）攜手推動真實臨床環境下的研究。在前瞻性（前向的，prospective）研究中，可以比較與病患的互動以及醫護人員的獨立判斷。然而，研究計畫與已完成的成果必須予以區分。

與醫護人員的協作設計

在實際導入時，除了人工智慧的單獨準確度之外，與醫護人員協同使用時的成效亦至關重要。目前亦有研究探討由醫護人員先做出獨立判斷，再由人工智慧整理一致處與差異處的運作方式。

在基於病例的研究中，此類架構展現了提高診斷準確度的潛力。然而，成效可能因醫護人員與使用者的背景以及資訊呈現方式而有所不同。絕不應引導使用者盲目遵循人工智慧的建議，而應使其能夠比對必要的依據。

MAI-DxO 的意義在於協調整合多個角色，將資訊收集、診斷與成本審查相結合。未來必須在實際診療中確認其可用性、安全性以及審查負擔。若要將研究成果轉化為對病患有益的支援，就必須將醫護人員的角色與系統的權限一併納入設計。

10.3 ClinVerse 與人工智慧醫院的診療流程

復旦大學等共同研究團隊的 ClinVerse 論文中，提出了六個角色的互動與評估標準。讀者可將取自實際紀錄的資訊與為模擬實驗補充的資訊區分開來解讀。 <https://arxiv.org/abs/2610.34878>

評估醫療代理人時，不僅要關注診斷的正確答案，還必須審視收集資訊與選擇檢查的整個過程。

ClinVerse 為復旦大學、上海創新研究院、浙江大學及華中科技大學的共同研究。Sirry Chen 等研究團隊於 2026 年 10 月 26 日發表了預印本論文。ClinPatient 為用於評估的病例集，ClinLab 則是預測檢查資訊缺失部分的組件。這些名稱均不代表個別獲得核准的醫療產品。ClinVerse 是一個以真實電子病歷（EHR）為基礎來模擬此類互動的研究環境。它並非營運真實醫院的產品，而是用於代理人學習與評估的模擬器（simulator）。

在該環境中，代理人從初始症狀出發，向病患提問並請求檢查，隨後提出診斷。與從一開始就提供所有必要資訊的選擇題式評估不同，它檢視的是選擇應獲取何種資訊的能力。

以實際紀錄為基礎的虛擬病例

ClinVerse 的 ClinPatient 資料集（dataset）包含了從公開臨床紀錄中構建的約 15 萬件診斷病例。它是以 MIMIC-IV、MIMIC-IV-Note 及 SICdb 等去識別化（非識別化，de-identified）資料為基礎。由於主要為美國與奧地利的住院及加護重症病歷紀錄，因此並不能代表所有國家與門診病患。

病歷紀錄中欠缺的生活習慣與社會史等，則利用虛擬病患生成工具 Synthea 予以補充。

Synthea 是由美國非營利研究機構 MITRE 主導開發的虛擬病患生成軟體。Jason Walonoski 等研究團隊關於該工具的論文於 2017 年 8 月 30 日在線上公開，並於 2018 年刊登於學術期刊卷期中。它透過規則模擬病患的生活與疾病演變過程來建立病歷紀錄。這是一種與語言模型本身不同類型的工具。因此，所謂「以實際紀錄為基礎」的說法，並不意味著所有資訊皆為實際觀察所得。評估時必須將真實資料與合成資訊加以區分。

診療流程的角色分工

醫師代理人負責問診、請求檢查以及進行診斷。病患模擬器則將病患知曉且能說明的資訊，與需透過檢查等方式確認的資訊分開管理。臨床病理角色負責傳遞檢查資訊，環境控制器則協調整體進程。診斷裁判不介入對話，僅對結果與過程進行評估。

這種角色劃分是為了僅提供符合提問或檢查的資訊而設計的。它限制了可能提前洩漏診斷的資訊被不當傳遞。然而，並無法保證能完全防止資訊洩漏。

對於缺失的檢查，ClinLab 模組（module）可依據其他資料預測數值並予以提供。該數值並非實際測量結果，而是模擬的組成元素。由於預測錯誤可能會對後續診斷產生影響，因此必須另行審查。

最終診斷與過程的評估

評估採用結構化的臨床依據與國際疾病分類（ICD-10）診斷名稱。綜合檢視收集了多少必要的病史、確認了多少相關檢查與異常發現，以及診斷的類別與具體細節是否相符。

即便最終正確率相近，收集資訊的方式也可能有所不同。可能會請求許多檢查卻未能將結果與診斷連結，也可能主要依賴初始資訊便得出結論。針對各過程的指標有助於掌握此類差異。

評估分數亦可用作學習的獎勵訊號（reward signal）。然而，提高既定分數的行為是否真正與優質的診療相符，仍需持續檢視。

反覆學習與記憶修正

Dynamic Cheatsheet 是由史丹佛大學與 Together AI 的 Mirac Suzgun 等研究團隊所提出。首篇論文於 2025 年 4 月 10 日公開。該方法將先前解決過的問題中具實用性的手法整理成文字，並在下一次輸入時一同納入。不應將其理解為 ClinVerse 自行建立的模型名稱，而是引進現有的記憶管理方法並在診斷模擬環境中進行測試。

ClinVerse 研究比較了從既有對話中擷取技能與經驗教訓的方式、在互動過程中更新記憶的方式，以及調適更新程序本身的方式等。該研究將此描述為自我演化（self-evolution）或自我改進的方法。

許多方法並非直接改變模型權重（weight），而是修改外部記憶與執行規則。如同 Dynamic Cheatsheet 的方式，整理過往經驗以供

下次執行時參考。在部分評估中，資訊收集策略與診斷效能確實獲得了改善。

然而，此類結果是在虛擬環境中呈現的變化。檢查資訊有所缺失的事實本身就可能成為診斷的線索，預測的檢查數值亦可能扭曲評估。除了模型效能外，還必須一併分析模擬器的假設與局限。

可與實際業務相連結的功能

ClinVerse 雖非導入真實醫院的系統，但它使研究得以探討將問診、調閱資料、建議候選檢查及準備鑑別診斷相串聯的架構。若未來要應用於臨床支援，必須確認各項功能究竟能為醫護人員的業務帶來何種助益。

例如，即使人工智慧整理了病歷，也必須能夠進一步確認重要症狀。在提出候選檢查項目時，必須同時檢視既有結果與病患狀態。候選診斷與根據必須以醫療團隊能夠確認與修正的形式提供。這是關於應用方向的說明，並非 ClinVerse 在實際醫院中所證實的成果。

虛擬環境評估的角色與限制

例如，如果扮演病患角色的回應始終採用相同格式，模型可能會因為習慣該格式而非基於醫學理解而提高分數。為了辨別這一點，必須在變更句子或使用其他資料建立的案例中也檢視其效能。由於受評估的代理與提供資訊的模擬器會共同影響結果，因此也有必要分別記錄兩者的變更歷程。

ClinVerse 在相同條件下重複評估多種代理，並分析資訊收集問題方面具有意義。不過本稿所探討的研究是同儕審查前的預印本論文，並未評估實際病患的結果。

在實際診療中，存在著噪音、不明確的發言、記錄錯誤、系統連線問題、醫療團隊的工作中斷等各種變數。虛擬病患與預測檢查數值無法完全重現這些情況。

人工智慧醫院研究的意義，並不在於立即將醫院的所有判斷自動化，而是在於測試連結多種角色與資訊的方法，並分析發生失敗的過程。實際應用必須經過各項功能的驗證、醫療團隊的監督，以及權限與責任的釐清。

「醫療超級智慧」並非當前已達成的技術水準，而是代表未來目標的詞彙。

Part 6. 未來醫療與AI

第11章. 醫師與人工智慧代理的協同合作



11.1 人工智慧篩檢・人類審核・綜合診斷

三亞人民醫院與四川大學相關研究團隊的遠距心電圖論文，說明了基層機構與判讀中心的角色。從中可以確認人工智慧標記的結果經由人工審查後傳遞的過程。<https://www.frontiersin.org/journals/cardiovascular-medicine/articles/10.3389/fcvm.2026.1869716/full>

將醫療人工智慧應用於診療時，必須決定賦予其何種功能，以及由誰來確認結果。即使是擁有高精確度的模型，如果警報過多或未能及時傳達給負責人員，其效果也會受到限制。因此，在關注模型效能的同時，也必須設計工作流程（workflow）。

將人工智慧篩檢、受過訓練的專責人員審查，以及醫療團隊的綜合判斷相結合的架構，是一種應用方式。人工智慧先分類需要審查的資訊，專責人員確認原始資料與輸出的妥適性，醫師則綜合考量病患狀態以決定診斷與後續處置。具體角色會依產品與機構而有所不同。

區域心電圖協作體系

此案例探討的並非單一特定人工智慧模型，而是連結區域醫療機構的運作體系。2026年發表的研究分析了2025年下半年引進後的運作狀況。論文將核心人工智慧描述為醫療技術企業開發的商用解決方案。然而，論文並未透露企業名稱與產品最初上市日期。不應將醫院發表平台研究的日期當作該人工智慧的開發日期。

中國三亞（三亞）的遠距心電圖（tele-ECG）雲端平台，是連結基層醫療機構與上級醫院判讀體系的案例。基層機構記錄12導程心電圖（12-lead ECG）後共享資料，並由人工智慧、判讀專責人員及專科醫師依序逐步檢視。

雲端（cloud）是多個機構傳遞與處理資料的基礎。當上級機構確認基層取得的心電圖時，便能支援必要的判斷與轉診後送準備。此架構的成效，不應僅理解為基層醫療人員個人的判讀能力，更應視為連結資訊與專業知識的成果。

第1階段：人工智慧篩檢與品質確認

12導程心電圖是從多個方向記錄心臟電氣訊號的檢查。輸入的是此波形，人工智慧則與學習到的特徵進行比對並輸出潛在風險。手部或身體晃動產生的雜訊也可能混入波形中。這就是為何在查看危險標示之前，需要有確認訊號是否值得分析的過程。

人工智慧會在波形中標示潛在風險，並確認電極貼附與訊號品質。在急性心肌梗塞等緊急狀況下，設定篩檢標準以避免遺漏危險至關重要。ST段上升型心肌梗塞（STEMI, ST-segment elevation myocardial infarction）也是主要的審查對象。

然而，在提高敏感度（sensitivity）的設定下，正常或重要性較低的發現可能會被標記為異常。

敏感度是指在實際有異常的人群中成功檢測出的比例。偽陰性是指明明有異常卻被遺漏的情況，偽陽性則是指對沒有異常的人發出警報的情況。陽性預測值是指在接獲警報的人當中實際有異常的比例。

即使是相同的模型，其警報的意義也會因受檢族群的疾病盛行率而有所不同。必須同時檢視偽陰性率（false negative rate）、偽陽性率（false positive rate）與陽性預測值（PPV, positive predictive value），以判斷警報的意義。

第2階段：原始資料審查

例如，當人工智慧警報送達時，若部分波形中斷，專責人員可以確認是否為測量誤差，並傳達重新測量的必要性。若缺少這個過程而持續只將警報傳送給專科醫師，即使需要審查的資訊增加，判斷也未必會加快。記錄由誰接收並確認了通知的功能，也是協同合作的一部分。

中央判讀中心的專責人員會確認原始心電圖波形，並檢視人工智慧標記的內容。其職責在於減少訊號錯誤與過多警報，並整理供專科醫師審查的資訊。

若重複出現不必要的通知，可能會產生即使出現重要警告反應也會變得遲鈍的警報疲勞（alarm fatigue）。中間審查階段是管理此類負擔的方法，但前提是必須具備專責人員的培訓、工作量控管以及結果傳遞標準。

第3階段：專科醫師的綜合判斷

專科醫師不僅檢視心電圖，還會同時審查胸痛（胸痛）與其他症狀，以及肌鈣蛋白（troponin）等檢查結果。因為不能單憑單一訊號的異常來確診所有病症。

呈現人工智慧輸出的時間點也可能影響判斷。為了減少過度依賴先前提供答案的錨定效應（anchoring）與自動化偏誤（automation bias），必須能夠進行獨立審查與依據比對。研究中的雙盲評估方式與實際醫院的日常流程應有所區隔地進行設計。

運作中所報告的轉變

三亞的研究報告了在基層機構的診斷流程、決策時間以及判讀結果傳遞等方面的改善。共享檢查資料有助於減輕上級醫院重複進行相同檢查的負擔，並有可能支援病患抵達前的準備工作。

然而，由於這是引進前後的觀察結果，因此不能將所有轉變皆歸功於人工智慧的成效。同時也必須考量樣本與記錄的限制、各季節病患數量的變化，以及網路連線狀態。絕不能將少數病患的存活結果過度引申為該平台改善存活率的效果。

其他臨床領域的協同合作

臺灣的案例評估了由國防醫學院與三軍總醫院研究團隊建立的人工智慧心電圖警報體系。臨床試驗論文於2024年4月29日公開。該日期並不代表相關技術整體的最初上市日期。該架構是在心電圖中計算風險度後，若超過設定標準即通知醫療團隊，再由醫療團隊確認狀況。從醫院公開的說明中，也能確認實際臨床研究與後續驗證的必要性。<https://www.tsgh.ndmutsgh.edu.tw/news/191/10010/24887/10864>

臺灣的人工智慧心電圖警報研究，設計為將危險訊號傳遞給醫療團隊並促使其應對。此案例的意義在於評估了篩檢結果是否能與實際臨床行動相互連結。

在瑞典的 MASAI 乳癌篩檢研究中，依據人工智慧的風險分類調整了判讀方式。

MASAI 是隆德大學等機構的研究名稱，並非人工智慧產品名稱。研究中使用的 Transpara 是荷蘭 ScreenPoint Medical 的影像分析軟體。它接收乳房攝影影像作為輸入，並提供疑似部位與評分。早期產品的美國 FDA 510(k) 審查核准日期為2018年11月21日。這是該美國審查記錄的日期，並不等同於全球最初公開日期或 MASAI 研究的開始日期。其架構是由人工智慧支援部分審查順序與工作分配，並由醫療團隊負責判讀。該研究亦評估了間歇癌（interval cancer），即定期篩檢間隔期間發現的癌症等與病患相關的預後結果。

在結合人類與人工智慧診斷的病例研究中，也展現了相輔相成的潛力。然而，以文字呈現之案例的準確度與實際診療成效有所不同。必須確認符合各領域的臨床實證依據。

自主功能與監督範圍

IDx-DR 是由美國 IDx 開發的醫療器材，後來以 Digital Diagnostics 的 LumineticsCore 之名為人所知。FDA 的 De Novo 許可決定日期為2018年4月11日。它透過分析眼底視網膜照片，產出指定範圍內的糖尿病視網膜病變篩檢結果。解讀結果時不需要眼科專科醫師另外進行影像判讀，這與病患不需要後續診治是兩回事。FDA 文件中明確規定了根據結果進行轉診以及檢查限制的條件。

https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN180001.pdf

在狹窄且明確的任務中，也有獲准無需醫療團隊額外判讀即可提供結果的器材。糖尿病視網膜病變篩檢器材 IDx-DR，即現今的 LumineticsCore 便是相關案例。因此，不應以偏概全地認為所有人工智慧輸出都需要相同形式的最終審查。

在受限條件下處理既有處方續開的 Doctronic 試辦計畫亦曾被介紹。

這是由美國 Doctronic 營運之處方更新工具的試辦計畫。猶他州主管機關公開的現況中記載，在第1階段，所有處方申請均需持有執照的醫事人員核准。因此，在介紹該試辦計畫時，也必須同時敘明當時的階段與監督條件。單憑這份說明無法確定該公司模型的最初公開日期。<https://commerce.utah.gov/ai/regulatory-relief-4/authorized-pilots/doctronic/> 然而，重複既有處方與開立新處方、調整劑量是截然不同的業務。絕不能將特定區域的試辦許可誤解為通用自主診療的許可。

監督層級應根據使用目的、風險及產品的許可條件來決定。同時也需要制定將需要專業判斷的案例轉介給醫療團隊的標準，以及系統無法運作時的替代處置程序。

協同合作體系的設計

人工智慧篩檢・人類審核・綜合診斷的架構，是明確區分各功能角色的方法。其目的在於傳遞資訊以避免遺漏，在各個階段發現錯誤，並確保最終判斷所需的臨床脈絡。

在實際引進時，不應僅將檢查與警報的增加視為成效，而必須確認是否進行了妥善的審查與應對。應同時評估處理時間、重複檢查、專責人員的負擔以及病患的預後結果，並且為連線障礙做好準備。醫療人工智慧的成效不僅取決於單一模型，更取決於這種整體的運作體系。

11.2 醫師的角色與綜合臨床判斷

比較醫療團隊與人工智慧協同合作方式的隨機對照試驗論文，探討了獨立判斷與意見比對之間的差異。應將病例評估所得出的結果與實際病患的治療成效區別看待。 <https://www.nature.com/articles/s41746-026-02545-1>

當人工智慧支援記錄與資訊檢索時，醫師的工作結構也可能隨之改變。整理檢查結果與藥物清單、輸入診療內容的負擔減輕，取而代之的是檢視所呈現的資訊並為病患做出適當決策的角色變得更加重要。

然而，即使資訊整理實現了自動化，醫師的角色也不僅限於做判斷。問診、身體檢查、溝通交流以及治療調整是相互關聯的業務。實際產生的轉變也會因將人工智慧應用於何處而有所不同。

記錄工作與診療時間

環境人工智慧書寫員（ambient AI scribe）是指在診療過程中聆聽周遭對話並擬定病歷草稿的產品系列。由於這並非單一公司製造的單一模型名稱，因此無法設定共同的上市日期。它將語音轉換為文字後，按項目整理症狀與診療計畫等。由於可能會增添對話中未提及的內容或遺漏否定詞句，因此需要醫療團隊確認草稿。

環境人工智慧書寫員（ambient AI scribe）可透過擬定記錄草稿來減輕醫療團隊的輸入負擔。雖然多項研究報告了縮短文書作業時間或改善職業倦怠（burnout）的可能性，但成效因產品與使用環境而異。也有研究顯示非上班時間的工作時數並未顯著減少。

因此，人工智慧所省下的時間能否用於與病患對話或進行審查，取決於醫院的業務運作模式。若時程安排、人力與其他行政工作維持不變，僅靠減少記錄時間並無法解決整體的負擔。

如同區域心電圖協作那樣結合人工智慧篩檢與專責人員審查的體系中，專科醫師的角色依然不可或缺。專科醫師將整理好的判讀意見與症狀及其他檢查結果相互連結並進行判讀詮釋。必須同時規劃資訊處理的分工與臨床判斷的責任。

掌握病患脈絡的角色

例如，即使提出了建議定期回診的計畫，若病患居住在偏遠地區或無人協助接送，該計畫也難以執行。若在病歷中找不到此類條件，就必須親自詢問。不能僅止於確認人工智慧所擬定之計畫的文字是否通順合理，還必須審視病患是否能夠實際落實遵從。

電子病歷（EHR, Electronic Health Record）包含許多診療所需的資訊，但無法呈現患者的整體狀況。能否持續接受治療的生活條件、照護者的支持、患者的理解與擔憂等，往往需要透過對話來確認。

醫師必須確認人工智慧參考的資訊是否充分且最新，並檢視病歷中未記載的條件。在審視可能診斷的優先順序時，解釋不確定性並評估進一步確認的必要性，也是一項重要的角色。

與患者共同選擇治療的角色

治療的適切性不能僅憑平均療效來決定。必須同時考量副作用與治療負擔、生活方式，以及患者重視的目標。

醫病共享決策（shared decision-making）是說明各種選項的優缺點並納入患者偏好以做出決定的過程。人工智慧可以整理資料並協助說明，但患者理解了什麼、希望做出何種選擇，必須透過與醫護人員的互動來確認。

審查人工智慧結果的角色

為減少自動化偏誤（automation bias），必須對照自身知識與患者狀態來審查各項建議。這是確認證據是否與實際資料相符、重要條件是否遺漏，以及是否有其他選項的過程。

審查方式與責任因產品的功能、核准條件及業務風險而有所不同。這並不代表醫師必須重複所有計算，但必須確認判斷所需的資訊，並具備修改或拒絕建議的能力。

當多個專科與人工智慧工具的意見不一致時，也需要扮演協調的角色。必須考量患者的狀態、治療目標、資源與時程等以排定優先順序。多代理人系統提出的共識，亦無法自動取代此類臨床審查。

協作的成效與使用方式

在分析人機協作的研究中，成效並不一致。即使在整理過的病例中表現出助益，在複雜的實際病歷中其優勢可能較小。即便人工智慧單獨評分很高，使用者也未必能獲得同等水準的成果。

亦有研究探討醫護人員先獨立做出判斷，再與人工智慧比較異同的方式。雖然在部分病例評估中顯示出改善的可能性，但不能視為已在實際患者預後中獲得證實。

提出兩種模型的意見或替代假說的設計，同樣有助於支援審查。另一方面，資訊過量會增加負擔，因此必須同時評估使用者體驗（UX, user experience）與實際工作流程。僅僅賦予工具使用權限是不足夠的。

專業能力的維持

當人工智慧停擺時的因應措施亦包含在培訓之中。必須具備閱讀原始資料的能力，才能發現自動分析的錯誤並透過其他程序接續工作。在培訓中，與其只展示答對的案例，讓學員說明為何必須修正錯誤建議的方式可能更有助益。此為培訓設計的建議，並非特定研究所證實的成效數據。

當人工智慧承擔部分偵測與分析工作時，醫護人員親自執行該業務的頻率可能會減少。內視鏡領域的觀察性研究報告了使用人工智慧輔助後，在未採用人工智慧的情況下執行檢查時的表現變化。然而，由於無法斷定其原因，因此不能將自動化必然導致能力下降予以概括化。

醫療機構應維持培訓與檢核，確保在沒有人工智慧的情況下也能進行必要的判斷。在受訓過程中，不僅要教授確認結果的方法，還必須培養解讀原始資料並發現錯誤的能力。

角色轉變的評估

醫師角色的轉變，應以實際工作如何重構來判斷，而非職稱的改變。有必要同時檢視整理與審查資訊所花費的時間、與患者的對話、決策的適切性、培訓以及疲勞程度等。

人工智慧若能減少重複性工作並提供必要的證據，便能支援醫護人員的判斷。要將其轉化為對患者有益的改變，需要充分的審查環境、責任的明確化以及持續的培訓。確認導入技術後的業務運作與患者預後，是評估角色轉變價值的標準。

第12章 未來醫療人工智慧的技術與倫理課題



美國 FDA 的變更管理指引說明了如何將人工智慧醫療器材的修改計畫與驗證相互銜接。這是一份從取得初次許可後仍須確認安全性與效能的角度來研讀的資料。<https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence>

醫療人工智慧正將應用範圍從病歷撰寫與影像分析，擴展至診療流程支援、新藥候選物探勘以及臨床試驗設計。將多種工具與代理人連結雖能處理複雜業務，但錯誤造成的影響與管理負擔也會隨之增加。

醫療超級智能（medical superintelligence）是表達此一發展目標的詞彙。在特定測試中展現出高效能，並不同於在整體實際醫療環境中能安全運作。必須區分當前的研究成果與未來展望，並取得符合使用目的的證據。

幻覺與連鎖錯誤

舉例來說，假設病歷摘要誤將已停用的藥物記錄為目前正在服用的藥物。後續的代理人若僅閱讀該摘要，便會直接承襲錯誤的前提。即使重複審查相同的說明多次，若不重新對照原始資料，仍可能無法發現錯誤。這就是為何在各個階段都必須一併傳遞來源與日期。

大型語言模型能以流暢自然的方式生成不實內容。幻覺（hallucination）會表現為錯誤的藥物資訊、毫無根據的診斷說明或

不存在的文獻等。在多步驟的工作中，先前的錯誤可能會反映在後續的搜尋與判斷之中。

僅憑提供來源的功能並不能解決問題。必須確認引用內容是否與實際原文相符，且能否適用於當前患者。在代理人使用工具與整合結果的過程中，同樣需要驗證程序。

效能變化與資料的代表性

即使整體平均維持不變，部分族群的錯誤率仍可能增加。例如更換攝影設備後，可能僅有特定機構的影像效能下降。在修改模型之前，必須先檢視輸入資料是否發生變化，或者實際疾病的分佈是否改變，才能針對原因進行因應。

漂移（drift）是指因患者組成、設備、資料生成方式等發生變化，導致模型效能改變的問題。即使在最初評估的環境中獲得了良好結果，也無法保證日後能持續維持相同的效能。

偏向特定地區與人口族群的訓練資料，可能會造成對其他患者的效能落差。預測目標本身也必須進行審查。因為像醫療費用支出這類間接反映健康需求的指標，可能會反映出不同族群間醫療可近性的差異。

評估效能時，不僅要檢視整體平均，還必須依患者族群、機構與疾病類別分別審視。一旦發現變化，必須能夠透過重新評估、修改以及調整使用範圍來因應。

生物安全

蛋白質與 DNA 設計人工智慧雖能支援藥物開發與生命科學研究，但也有可能被濫用於生成具有危險生物學功能的序列。生物安全（biosecurity）不僅在模型開發階段，在實際訂購合成與執行研究的過程中也必須一併考量。

相關研究顯示，現有的序列審查工具可能無法充分辨識新生成的變異體。因此，必須持續改善偵測體系，並同時落實確認合成訂單目的與客戶身分的程序。若要降低風險，模型提供者、合成服務商與研究機構的查核程序必須相互串聯。

責任與個人資料保護

美國衛生與公眾服務部的 HIPAA 指南說明了適用對象與健康資訊的保護原則。由於並非所有健康 App 或所有國家都適用相同法律，因此必須先確認服務的營運主體與管轄範疇。<https://www.hhs.gov/hipaa/for-professionals/privacy/index.html>

發生錯誤時的責任，取決於模型開發商、產品製造商、醫療機構、使用者等各方的角色與實際成因。很難以將所有責任歸咎於最終核准者的方式來解決。在合約與工作規範中，必須明確界定權限、審查義務、通報與因應程序。

在資料蒐集、訓練、搜尋及紀錄保存的全過程中，皆需要保護個人資料。必須確認各管轄區域的法律規範（如美國 HIPAA 與歐洲 GDPR 等），並在必要範圍內限制存取與使用。

聯邦學習（federated learning）是一種不將資料匯集於一處，而是共同訓練模型的方法。

各機構傳送以自身資料訓練後的結果，系統匯總這些結果來更新模型。雖然具有無須移動原始檔案的優點，但傳送的訓練結果中仍可能留下資訊的痕跡。聯邦學習是一項技術方式的名稱，無法以單一公司或發表日期來解釋。然而，它並不能完全消除資訊外洩的風險。必須評估模型及訓練過程遭受攻擊的可能性，並同步建立技術防護與制度規範。

醫護人員的能力與監督

對人工智慧的依賴加深時，親自執行部分業務的經驗可能會減少。觀察性研究中顯示的能力變化，其原因雖需謹慎解讀，但也展現了維持培訓與檢核的必要性。

為減少自動化偏誤（automation bias）與警報疲勞（alarm fatigue），必須提供確認結果的時間、原始資料以及修改權限。讓醫護人員能獨立判斷並與人工智慧的建議進行比較的設計可能有所助益。然而，適切的方式應根據業務風險與使用環境來評估。

自主性也必須依功能加以區分。在狹窄的篩檢任務中獨立產出結果的器材，與能診斷多種疾病並決定新治療方案的通用代理人截然不同。絕不能將其解釋為已超越核准範圍與各地試辦計畫條件而獲得了自主性認可。

變更與生命週期管理

FDA 目前公布的相關最終指引為 2025 年 8 月版。建議在計畫中納入變更內容、開發、驗證與應用方法，以及變更帶來的影響。

PCCP 指的是法規審查中所處理的變更管理計畫，應與人工智慧模型名稱有所區別。這並不代表超出預定範圍變更功能的情況也會被自動核准。

人工智慧在模型、檢索資料或提示詞發生改變時，輸出結果也可能隨之不同。預先變更控制計畫（PCCP, Predetermined Change Control Plan）是一種預先界定既定變更範圍與驗證方法的方式。這並非一項允許隨意進行任何變更的制度。

必須配合產品的使用目的與風險，將變更前後的評估、使用者指南與上市後效能監測相互串聯。相關法規與指引的定案與否及適用時間點各不相同，因此在實際導入時必須確認最新的法規要求。

醫院的管理體系

舉例來說，可以不將人工智慧的風險標示顯示在實際診療畫面上，而是分開儲存，日後再與醫護人員的判斷進行比較。在此類試驗中由於同樣會處理患者資料，因此必須訂定存取權限與保存期限。不能僅僅因為隱藏了結果，就免除所有的研究與安全程序。

醫療機構很難將引進人工智慧的工作僅交由技術部門的採購業務負責。臨床醫師、護理師、資訊安全負責人、法律與倫理專家等必須共同審查功能、風險與運作條件。亦須考量患者的觀點與醫療可近性。

在對實際診療產生影響之前，可以考慮採用以現場資料檢視效能的靜默試驗（silent trial）。這是一種不將結果用於診療決策，而在實際工作條件下確認運作情況的方式。必須在訂定資料處理、個人資料保護與評估範圍後再行實施。

上線運作後，必須蒐集錯誤與異常案例，並管理負責人員的處置與修改歷程。一旦發現安全問題，亦需具備限制使用或中止使用的標準。

以患者預後為中心的評估

即使有節省時間的結果，也必須檢視是衡量了誰的時間以及衡量的範圍為何。因為即使輸入時間減少，審查與修改的時間仍可能增加。若將診療結束後的工作一併納入比較，便能更準確地理解患者與醫護人員所經歷的改變。

基準測試（benchmark）雖有助於進行可重複的比較，但無法全面展現對實際患者產生的影響。使用者傳遞資訊與理解回覆的過程、醫護人員的因應，以及醫院系統的銜接，都會對結果造成影響。

在人工智慧心電圖警報與乳癌篩檢等方面，已有研究評估了實際診療流程與患者相關預後。如同這些案例，必須設定符合各項使用目的的預後指標。在風險較高的功能中，相應的臨床證據至關重要。

必須同時檢視時間節省、成本、診斷準確度、患者的可近性與安全性。在提高研究資料的多樣性並擴大機構間合作的同時，亦須維持個人資料保護與評估的獨立性。

醫療人工智慧發展的方向

醫療人工智慧的價值不能僅憑自主性或模型規模來判斷。能否準確處理必要資訊、支援醫護人員的判斷，並確保患者獲得適切的診療，才是關鍵所在。

隨著技術的進步，必須同步改善資料品質、驗證方法、責任與補償機制、培訓與運作體系。若要將研究階段的潛力轉化為實際的醫療服務，需要持續確認這些條件的過程。以患者所獲得的效益與承受的風險為基準來界定應用範圍，是醫療人工智慧發展的基本方向。

參考文獻

TXP Medical (2025). 生成AIによる音声入力カルテアプリ「SpeechER」が2025年グッドデザイン賞を受賞. 公司新聞稿, 2025-10-15. <https://txpmedical.jp/en/news/2jmiSPoM4bpIdI9GUZFOBI> (企業發布: "最多縮減70% (自家調査) ")

SIND (年份不詳, 2026-10-28瀏覽). Caretomo 看護記録作成AI支援システム 產品網站. <https://caretomo.jp/> (企業發布: 各醫院試用結果)

Xu, J., et al. (2026). AI-Enabled regional tele-ECG cloud platform and improving access to cardiovascular diagnosis: real-world evidence from southern China. *Frontiers in Cardiovascular Medicine*, 13, 1869716. <https://doi.org/10.3389/fcvm.2026.1869716> (觀察性研究, 實際運作)

Lukac, P. J., Turner, W., Vangala, S., et al. (2025). Ambient AI scribes in clinical practice: a randomized trial. *NEJM AI*, 2(12). 線上 2025-11-26. <https://doi.org/10.1056/Aloa2501000> (隨機分配臨床試驗, 單一機構)

Rotenstein, L. S., Holmgren, A. J., Thombley, R., et al. (2026). Changes in clinician time expenditure and visit quantity with adoption of artificial intelligence-powered scribes: a multisite study. *JAMA*, 335(16), 1408-1417. <https://doi.org/10.1001/jama.2026.2253> (觀察性研究, 5家機構)

Cleveland Clinic Consult QD (2025). Less Typing, More Talking: How Ambient AI Is Reshaping Clinical Workflow at Cleveland Clinic. 2025-08-14. <https://consultqd.clevelandclinic.org/less-typing-more-talking-how-ambient-ai-is-reshaping-clinical-workflow-at-cleveland-clinic> (機構內部報告)

Booth, S. E., Gould, R., et al. (2026). Impact of unrestricted artificial intelligence scribe deployment in a single health system on clinician workload and productivity: a quality improvement study. *Annals of Internal Medicine*, 線上預先發布 2026-10-27. <https://doi.org/10.7326/ANNALS-25-05294> (觀察性研究, 單一體系)

Frieden, J. (2026). AI scribes improved note-taking time but not productivity, study finds. *MedPage Today*, 2026-10-26.

<https://www.medpagetoday.com/practicemanagement/informationtechnology/123591> (媒體報導，輔助確認數值用)

Tian, J., Song, Y., Tang, L., An, L., & Zhou, J. (2026). Clinical artificial intelligence implementation in routine care: real-world operational outcomes from a provincial health system in China. *Risk Management and Healthcare Policy*, 19, 1-11. <https://doi.org/10.2147/RMHP.S604289> (觀察性研究，單一機構)

Rabotin, A., Aguilar, E., Sandoval Gonzalez, S., Iniguez, A., Jung, C., Lee, E., Bell, D. S., & Arroyo, J. (2026). Propagation of interpreter errors by ambient AI scribes: study using simulated clinical encounters. *JMIR Medical Informatics*, 14, e88734. <https://doi.org/10.2196/88734> (模擬診療實驗)

Olson, K. D., Meeker, D., Troup, M., et al. (2025). Use of ambient AI scribes to reduce administrative burden and professional burnout. *JAMA Network Open*, 8(10), e2534976. <https://doi.org/10.1001/jamanetworkopen.2025.34976> (觀察性研究，6個醫療體系，事前・事後問卷調查)

Rotenstein, L. S., Holmgren, A. J., Thombley, R., et al. (2026). Changes in clinician time expenditure and visit quantity with adoption of artificial intelligence-powered scribes: a multisite study. *JAMA*, 335(16), 1408-1417. <https://doi.org/10.1001/jama.2026.2253> (觀察性研究，5所學術醫療機構)

Samraik, M. (2025). AI scribes reduce physician burnout and return focus to the patient. *Yale School of Medicine News*, 2025-10-17. <https://medicine.yale.edu/news-article/ai-scribes-reduce-physician-burnout-return-focus-to-the-patient/> (機構內部報告)

Merlino, A., Blue, A., Boose, E., Meese, B., Owens, J., Vickers, J., Angeloff, A., Arnold, M., Chen, R., Grade, M., Pham-Viernickel, P. A., & Rosner, B. (2026). Accelerating ambient AI scribe enterprise-scale deployment: Cleveland Clinic's novel approach to health system-industry partnership. *npj Health Systems*, 3, 74. <https://doi.org/10.1038/s44401-026-00144-6> (實際臨床運作報告，無對照組，部分共同作者隸屬公司)

Fierce Healthcare (2025). Cleveland Clinic taps Ambience for ambient AI tech. 2025-02-19. <https://www.fiercehealthcare.com/ai-and-machine-learning/cleveland-clinic-chooses-ambience-ai-scribe> (媒體報導，引用機構發布內容)

ASCO AI in Oncology (2026). Measuring Ambient AI Adoption at Scale. 2026-10. <https://ascoai.org/articles/2026/10/measuring-ambient-ai-adoption-at-scale> (專業學會媒體的解說)

Abridge (2026). The Evolution of Enterprise AI at Sutter: A Timeline. 2026-08-27. <https://www.abridge.com/blog/sutter-health-partnership-timeline> (企業發布)

Olson, K. D., Meeker, D., Troup, M., et al. (2025). Use of ambient AI scribes to reduce administrative burden and professional burnout. JAMA Network Open, 8(10), e2534976. <https://doi.org/10.1001/jamanetworkopen.2025.34976> (<https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2839542>) (觀察性研究，無對照組的前後比較，自我通報。確認表2・3：分析參與者263名，職業倦怠模型186名，校正多變量模型184名)

Afshar, M., et al. (2025). A pragmatic randomized controlled trial of ambient artificial intelligence to improve health practitioner well-being. NEJM AI, 2(12). <https://doi.org/10.1056/Aloa2500945> (ClinicalTrials.gov NCT06517082) (隨機分配臨床試驗，威斯康辛大學門診，66名，24週)

Rotenstein, L. S., Holmgren, A. J., Thombley, R., et al. (2026). Changes in clinician time expenditure and visit quantity with adoption of artificial intelligence-powered scribes: a multisite study. JAMA, 335(16), 1408-1417. <https://doi.org/10.1001/jama.2026.2253> (觀察性研究)

Booth, S. E., Gould, R., et al. (2026). Impact of unrestricted artificial intelligence scribe deployment in a single health system on clinician workload and productivity: a quality improvement study. Annals of Internal Medicine, 線上預先發布 2026-10-27. <https://doi.org/10.7326/ANNALS-25-05294> (觀察性研究)

Frieden, J. (2026). AI scribes improved note-taking time but not productivity, study finds. MedPage Today, 2026-10-26. <https://www.medpagetoday.com/practicemanagement/informationtechnology/123591> (媒體報導)

Wojda, T., Berliner, J., Fischer, G. S., Curren, M., Arnold, J., & Giugale, L. (2026). From burden to balance: trends in documentation workload with ambient artificial intelligence scribe adoption. *Applied Clinical Informatics*, 17(3), 484-492. <https://doi.org/10.1055/a-2892-7988> (觀察性研究，重複橫斷問卷調查，基準期1,600名，6個月101名)

De Vito, A., Geremia, N., Bavaro, D. F., et al. (2025). Comparing large language models for antibiotic prescribing in different clinical scenarios: which performs better? *Clinical Microbiology and Infection*, 31(8), 1336-1342. <https://doi.org/10.1016/j.cmi.2025.03.002> (基準評測，基於案例的比較)

Russo, A., Geremia, N., Bavaro, D. F., De Vito, A., et al. (2026). Comparing large language models for antibiotic prescribing: an updated analysis using the same clinical benchmark. *Clinical Microbiology and Infection*, 32(10), 1733-1735. 線上 2026-04-16. <https://doi.org/10.1016/j.cmi.2026.04.012>

De Vito, A., Geremia, N., Russo, A., Colpani, A., Vaira, L. A., & Madeddu, G. (2026). Reproducibility of AI-generated antibiotic recommendations in standardised clinical scenarios: a proof-of-concept experimental study. *International Journal of Antimicrobial Agents*, 67(10), 107881. <https://doi.org/10.1016/j.ijantimicag.2026.107881> (基準評測，重複查詢實驗)

Zhang, Y., Li, L., Tan, C., Ji, M., Tian, X., Mu, X., Li, J., Gu, Y., & Liu, H. (2026). Large language model-based clinical decision support for antibiotic selection and dose recommendation in hospitalized patients with pneumonia: multicenter retrospective study. *JMIR Medical Informatics*, 14, e98207. <https://doi.org/10.2196/98207> (觀察性研究，回溯性2家機構，實際臨床應用前)

(背景知識) Shortliffe, E. H. (1976). *Computer-Based Medical Consultations: MYCIN*. New York: Elsevier. (歷史背景，1976年基礎文獻來源)

雲知聲 Unisound (2026). 登頂 OpenAI HealthBench 雙榜！雲知聲醫療模型性能全面超越 GPT-6 Astra、GPT-6.1 Sol，構建高水準 AI 醫生問診能力. 官方發布，2026-10-21. <https://www.unisound.com/news/165> (英文版) <https://www.unisound.com/en/news/165> (企業發布，基準評測自行報告)

Arora, R. K., et al. (2025). HealthBench: Evaluating large language models towards improved human health. arXiv:2505.08775.

<https://doi.org/10.48550/arXiv.2505.08775> (基準評測設計，2025年基礎文獻來源)

Bean, A. M., Payne, R. E., Parsons, G., et al. (2026). Reliability of LLMs as medical assistants for the general public: a randomized preregistered study. *Nature Medicine*, 32, 609-615. <https://doi.org/10.1038/s41591-025-04074-y> (隨機分配研究，經追加調查確認) (2026年4月17日出版社更正公告：Nature Medicine 32(5):1937，<https://doi.org/10.1038/s41591-026-04404-8>. 未能確認更正內容是否影響正文數值)

Nori, H., Daswani, M., Kelly, C., et al. (2026). Sequential diagnosis with language models. *Science*, 394(6822), 423-428. <https://doi.org/10.1126/science.aeh7930> (基準測試，經追加調查確認書目・摘要)

Microsoft AI (2026). Building Medical Superintelligence. 2026-10-22.

<https://microsoft.ai/news/building-medical-superintelligence> (企業發布)

Chen, S., Chen, F., Ding, X., Chen, W., & Wei, Z. (2026). ClinVerse: a realistic multi-agent diagnostic environment for training and evaluating medical agents. arXiv:2610.34878. <https://doi.org/10.48550/arXiv.2610.34878> (模擬研究，預先公開)

Liévin, V., Palepu, A., Weng, W.-H., Saab, K., Stutz, D., Cheng, Y., Kulkarni, K., Mahdavi, S. S., et al., Karthikesalingam, A., & Schaeckermann, M. (2026). Towards conversational artificial intelligence for disease management. *Nature*, 655, 1292-1299. <https://doi.org/10.1038/s41586-026-10764-5> (模擬研究，病患演員，隨機・盲法)

Tu, T., Schaeckermann, M., Palepu, A., et al. (2025). Towards conversational diagnostic artificial intelligence. *Nature*, 642, 442-450.

<https://doi.org/10.1038/s41586-025-08866-7> (模擬研究，2025年基礎出處)

Google (2026). Google advances its AMIE research medical AI from diagnosis to treatment. Google Blog, 2026-06-17. <https://blog.google/innovation-and-ai/models-and-research/google-research/amie-for-disease-management-in-nature> (企業發布)

Ferber, D., Hilgers, L., Höper, C., Kinny-Köster, B., Eckardt, J.-N., et al., Truhn, D., Jäger, D., & Kather, J. N. (2026). Towards autonomous medical artificial intelligence agents. *Nature*, 655(8125), 1282-1291.

<https://doi.org/10.1038/s41586-026-10675-5> (模擬研究，基於 MIMIC-IV 實際病患病史)

Brodeur, P., Koshy, J. M., Palepu, A., Saab, K., et al., Schaekermann, M., Karthikesalingam, A., & Rodman, A. (2026). A prospective clinical feasibility study of a conversational diagnostic AI in an ambulatory primary care clinic. *The Lancet*, 2026-10-08 線上. [https://www.thelancet.com/journals/lancet/article/PIIS0140-6736\(26\)01535-7/fulltext](https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(26)01535-7/fulltext) (預印本：A prospective clinical feasibility study of a conversational diagnostic AI in an ambulatory primary care clinic. arXiv:2603.08448) (前瞻性研究，實際診療，監督下)

ARISE Network (2026). ARISE, BIDMC and Google AMIE publish prospective clinical AI feasibility study in *The Lancet*. 2026-10-08.

<https://www.arise-ai.org/blog/amie-lancet> (研究參與機構解說)

Nature Medicine (2026). Agents AMIE and MIRA advance medical AI capabilities. 2026-06-30. <https://www.nature.com/articles/d41591-026-00034-2> (學術期刊新聞，僅確認標題與存在)

Beth Israel Deaconess Medical Center (2026). BIDMC researchers conduct first real-world study of safety and quality of patient-facing AI in primary care. 2026-10-09. <https://bidmc.org/news-stories/all-news-stories/news/2026/10/bidmc-research-shows-safety-quality-of-ai-in-primary-care> (機構新聞稿)

Brix, G., Durrant, M. G., Ku, J., et al. (2026). Genome modelling and design across all domains of life with Evo 2. *Nature*, 652(8112), 1349-1361.

<https://doi.org/10.1038/s41586-026-10176-5> (同行評審論文)

Avsec, Ž., Latysheva, N., Cheng, J., Novati, G., et al. (2026). Advancing regulatory variant effect prediction with AlphaGenome. *Nature*. 649(8099), 1206-1218.

<https://doi.org/10.1038/s41586-025-10014-0> (同行評審論文)

Wang, D., Zhang, J., Chen, C., Zhang, W., Wang, S., Meng, Y., Sonpavde, G., Horbinski, C., & Tian, Y. (2026). A multimodal foundation model linking

histopathology and DNA methylation. bioRxiv.

<https://doi.org/10.64898/2026.06.11.731518> (預印本，2026年6月14日發布)

Li, Q., Sang, J., Xiao, Y., Zhang, P., Chen, W., Tran, T. N., Li, Z., Li, S., Zhang, T., & Wang, G. (2026). Genome-scale molecular profiling from routine histopathology with a multimodal pathology foundation model. Research Square.

<https://doi.org/10.21203/rs.3.rs-9949149/v1> (預印本，PMID 42620258)

Chen, W., et al. (2025). A visual-omics foundation model to bridge histopathology with spatial transcriptomics. Nature Methods, 22, 1568-1582.

<https://doi.org/10.1038/s41592-025-02707-1> (同行評審論文)

Chaves, J. M. Z., et al. (2024). Tx-LLM: A large language model for therapeutics.

arXiv:2406.06316. <https://arxiv.org/abs/2406.06316> (預印本)

Wang, E., Schmidgall, S., et al. (2025). TxGemma: Efficient and agentic LLMs for therapeutics.

arXiv:2504.06196. <https://arxiv.org/abs/2504.06196> (預印本)

Sun, Q., Wu, X., Chen, J., Zhang, Y., Zhou, C., & Zhu, Z. (2026). Cross-attention guided multi-modal network for breast ultrasound diagnosis incorporating objective clinical semantics. Frontiers in Medicine, 13, 1802570.

<https://doi.org/10.3389/fmed.2026.1802570> (同行評審論文)

Gu, Z., Wang, C., Zhu, Q., Wang, W., et al. (2026). Multimodal large language models for prognostic prediction in cervical cancer treated with definitive chemoradiotherapy: An exploratory study of systematic multimodal data integration. Cancers, 18(19), 3136.

<https://doi.org/10.3390/cancers18193136> (同行評審論文，詳細 AUC 數值未核對)

Souza, H., & Mehta, P. (2026). Parameter-free representations outperform single-cell foundation models on downstream benchmarks.

arXiv:2602.16696.

<https://arxiv.org/abs/2602.16696> (預印本)

Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., & Wang, B. (2024). scGPT: Toward building a foundation model for single-cell multi-omics using generative AI. Nature Methods, 21(8), 1470-1480.

<https://doi.org/10.1038/s41592-024-02201-0> (同行評審論文)

Ahlmann-Eltze, C., Huber, W., & Anders, S. (2025). Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. *Nature Methods*, 22(8), 1657-1661. <https://doi.org/10.1038/s41592-025-02772-6> (同行評審論文)

Liang, W., Zhang, Y., Kwon, Y., Yeung, S., & Zou, J. (2022). Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems (NeurIPS 2022)*. <https://arxiv.org/abs/2203.02053> (同行評審研討會論文，早期基礎文獻)

Baker, C., Ren, T., Rafferty, K., Wang, H., & McDade, S. (2026). Autonomous computational prioritisation of colorectal cancer vulnerabilities via multi-scale AI swarms. *bioRxiv*. <https://doi.org/10.64898/2026.07.05.736565>

Baker, C., Ren, T., Rafferty, K., & Wang, H. (2026). Contextual invertible world models: A neuro-symbolic agentic framework for colorectal cancer drug response. *arXiv:2603.02274*. <https://arxiv.org/abs/2603.02274> (預印本，v3 2026年6月10日)

Cui, H., Cruz-Correa, M., Giardiello, F. M., et al. (2003). Loss of IGF2 imprinting: A potential marker of colorectal cancer risk. *Science*, 299(5613), 1753-1755. <https://doi.org/10.1126/science.1080902> (同行評審論文，早期基礎文獻)

Waddington, C. H. (1957). *The strategy of the genes: A discussion of some aspects of theoretical biology*. George Allen & Unwin. (單行本，早期基礎文獻)

Qiu, M., Zheng, G., Xu, Y., Zhang, R., Ding, Y., Long, Q., & Chen, T. (2026). Chreode: A cell world model for one-step temporal dynamics and perturbation prediction. *arXiv:2605.28111*. <https://doi.org/10.48550/arXiv.2605.28111> (預印本，2026年5月27日)

Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F. D., & Klein, A. M. (2020). Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science*, 367(6479), eaaw3381. <https://doi.org/10.1126/science.aaw3381> (同行評審論文，基礎資料集)

Norman, T. M., Horlbeck, M. A., Replogle, J. M., et al. (2019). Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365(6455), 786-793. <https://doi.org/10.1126/science.aax4438> (同行評審論文，基礎資料集)

Roohani, Y., Huang, K., & Leskovec, J. (2024). Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology*, 42, 927-935. <https://doi.org/10.1038/s41587-023-01905-6> (同行評審論文)

Chen, S., Yu, L., Lin, X., Mao, X., Zhang, S., Gu, X., Wu, H., Xu, S., Jin, K., Bai, L., Qian, Q., Chen, Q., Gao, Q., Sun, S., & Gao, Z. (2026). SCALE: Scalable conditional atlas-level endpoint transport for virtual cell perturbation prediction. *arXiv:2603.17380*. <https://doi.org/10.48550/arXiv.2603.17380> (預印本，v3 2026年8月31日)

Zhang, J., et al. (2025). Tahoe-100M: A giga-scale single-cell perturbation atlas for context-dependent gene function and cellular modeling. *bioRxiv*. <https://doi.org/10.1101/2025.02.20.639398> (預印本)

Di Francesco, A. G., et al. (2026). Differentiable retrieval-augmented generation for predicting cellular responses to gene perturbation (PT-RAG). *arXiv:2603.07233*. <https://doi.org/10.48550/arXiv.2603.07233> (預印本，v3 2026年10月9日。全體共同作者於本書調查中未予核對)

Roohani, Y. H., et al. (2025). Virtual Cell Challenge: Toward a Turing test for the virtual cell. *Cell*. [https://www.cell.com/cell/fulltext/S0092-8674\(25\)00675-0](https://www.cell.com/cell/fulltext/S0092-8674(25)00675-0) (同儕審查評論論文，卷期與頁碼未核對)

Arc Institute. (2025, December). Virtual Cell Challenge 2025 wrap-up: Winners and reflections. <https://arcinstitute.org/news/virtual-cell-challenge-2025-wrap-up> (機構發布)

Arc Institute. (2026, August 20). The 2026 Virtual Cell Challenge: Predicting perturbation responses in cell contexts a model has never seen. <https://arcinstitute.org/news/virtual-cell-challenge-2026> (機構發布)。相關論文：Virtual Cell Challenge 2026: Benchmarking zero-shot generalization across cellular contexts. *Cell* (2026), PMID 42648290 (作者資訊未核對)

Virtual Cell Challenge. (2026). Virtual Cell Challenge 2026: Timeline and rules. <https://virtualcellchallenge.org/> (2026年10月22日公布最終測試集，確認11月5日截止)

Lamb, J., Crawford, E. D., Peck, D., et al. (2006). The Connectivity Map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313(5795), 1929-1935. <https://doi.org/10.1126/science.1132939> (同儕審查論文，早期基礎文獻)

Subramanian, A., Narayan, R., Corsello, S. M., et al. (2017). A next generation Connectivity Map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6), 1437-1452. <https://doi.org/10.1016/j.cell.2017.10.049> (同儕審查論文，早期基礎文獻)

Xu, Z., Zhang, Z., Wang, L., Liu, Z., Liu, Q., Wu, S., & Wang, L. (2026). Reading the cell, designing the cure: Perturbation-conditioned molecular diffusion for function-oriented drug design. arXiv:2605.15243. <https://doi.org/10.48550/arXiv.2605.15243> (預印本，2026年5月14日)

Yuan, H., Li, C., Ma, W., Murata, T., & Jiang, Y. (2026). Gene expression-informed jointly controlled generative modeling for precision molecular design. arXiv:2607.11978. <https://arxiv.org/abs/2607.11978> (預印本，2026年7月13日)

Miller, H. E., et al. (2026). Deep learning perturbation models can outperform baselines on calibrated metrics. *Nature Biotechnology*. <https://doi.org/10.1038/s41587-026-03307-w>

Sadée, C., Testa, S., Barba, T., Hartmann, K., Schuessler, M., Thieme, A., Church, G. M., Okoye, I., Hernandez-Boussard, T., Hood, L., Shmulevich, I., Kuhl, E., & Gevaert, O. (2025). Medical digital twins: Enabling precision medicine and medical artificial intelligence. *The Lancet Digital Health*, 7, 100864. <https://doi.org/10.1016/j.landig.2025.02.004> (同儕審查政策論文)

National Academies of Sciences, Engineering, and Medicine. (2024). Foundational research gaps and future directions for digital twins. The National Academies Press. <https://www.ncbi.nlm.nih.gov/books/NBK605512> (機構報告，詳細文句未核對)

Sel, K., Hawkins-Daarud, A., Chaudhuri, A., Osman, D., Bahai, A., Paydarfar, D., Willcox, K., Chung, C., & Jafari, R. (2025). Survey and perspective on verification, validation, and uncertainty quantification of digital twins for precision medicine. *npj Digital Medicine*, 8, 40. <https://doi.org/10.1038/s41746-025-01447-y> (同儕審查論文)

Garcia Requesens, M., Mankan, A. K., Cantini, L., Calabresi, M., Vidal, L., Brown, J., Lang, S., Kaura, B., Veselkov, K., Xenarios, I., & Saini, K. S. (2026). Role of health digital twins in oncology drug development: A primer. *Frontiers in Oncology*, 16, 1885797. <https://doi.org/10.3389/fonc.2026.1885797> (同儕審查文獻回顧，2026年8月24日)

Blasiak, A., Truong, A. T. L., Foo, N., Tan, L. W. J., Kumar, K. S., Tan, S. B., Teo, C. B., Tan, B. K. J., Tadeo, X., Tan, H. L., Chee, C. E., Yong, W. P., Ho, D., & Sundar, R. (2025). Personalized dose selection platform for patients with solid tumors in the PRECISE CURATE.AI feasibility trial. *npj Precision Oncology*, 9, 49. <https://doi.org/10.1038/s41698-025-00835-7> (同儕審查論文，可行性試驗)

The Cleveland Clinic. (2022-2026). Randomized controlled trial of digital twin precision treatment: A novel whole body digital twin enabled precision treatment for type 2 diabetes (NCT05181449). *ClinicalTrials.gov*. <https://clinicaltrials.gov/study/NCT05181449> (臨床試驗註冊資訊)

U.S. Food and Drug Administration. (2023). Assessing the credibility of computational modeling and simulation in medical device submissions: Guidance for industry and FDA staff. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents> (監管文件，2023年11月最終指引)

Lim Il-gon. (2026年7月23日). OPRIMED於失智症學術大會展示數位分身平台. *Business Watch*. <https://news.bizwatch.co.kr/article/healthcare/2026/07/23/0014> (報導)

Kwon Sang-hee. (2026年7月28日). OPRIMED於春川企業創新園區推動數位分身臨床試驗據點. 電子新聞. <https://www.etnews.com/20260728000025> (報導)

European Medicines Agency. (2022). Qualification opinion for Prognostic Covariate Adjustment (PROCOVA™). <https://www.ema.europa.eu/en/documents/>

[regulatory-procedural-guideline/qualification-opinion-prognostic-covariate-adjustment-procovatm_en.pdf](#) (監管文件)

U.S. Food and Drug Administration & European Medicines Agency. (2026, January). Guiding principles of good AI practice in drug development. <https://www.fda.gov/media/189581/download> (監管文件)

U.S. Food and Drug Administration. (2025, January). Considerations for the use of artificial intelligence to support regulatory decision-making for drug and biological products: Draft guidance for industry. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents> (監管文件，草案。截至2026年8月仍為草案狀態之解說：CASRAI Editorial Board. (2026). Digital twins in clinical trials: Synthetic control arms and FDA's evolving posture. <https://casrai.org/guides/digital-twins-clinical-trials>)

Franklin, J. M., Patorno, E., Desai, R. J., Glynn, R. J., Martin, D., Quinto, K., Pawar, A., Bessette, L. G., Lee, H., Garry, E. M., Gautam, N., & Schneeweiss, S. (2021). Emulating randomized clinical trials with nonrandomized real-world evidence studies: First results from the RCT DUPLICATE initiative. *Circulation*, 143(10), 1002-1013. <https://doi.org/10.1161/CIRCULATIONAHA.120.051718> (同儕審查論文，早期核心文獻)

Biotechnology Innovation Organization, Informa Pharma Intelligence, & QLS Advisors. (2021). Clinical development success rates and contributing factors 2011-2020. <https://www.bio.org/clinical-development-success-rates-and-contributing-factors-2011-2020> (產業報告，2021年2月，早期基礎資料。12,728件階段轉換分析，確認以第1期為基準之獲准可能性為7.9%)

Olmos, L. (2026, October 8). Built on regulatory rigor: Advancing digital twins in clinical trials. Unlearn. <https://www.unlearn.ai/blog/built-on-regulatory-rigor-advancing-digital-twins-in-clinical-trials> (企業發布)

U.S. Food and Drug Administration. (2023, May). Adjusting for covariates in randomized clinical trials for drugs and biological products: Guidance for industry. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents> (監管文件)

CASRAI Editorial Board. (2026). Digital twins in clinical trials: Synthetic control arms and FDA's evolving posture (last updated 2026-08-25).

<https://casrai.org/guides/digital-twins-clinical-trials> (解說資料)

International Council for Harmonisation. (2000). ICH E10: Choice of control group and related issues in clinical trials. <https://www.ich.org/page/efficacy-guidelines> (監管文件，早期基礎文獻)

Jahanshahi, M., Gregg, K., Davis, G., Ndu, A., Franolic, T., et al. (2021). The use of external controls in FDA regulatory decision making. *Therapeutic Innovation & Regulatory Science*, 55, 1019-1035. <https://doi.org/10.1007/s43441-021-00302-y> (同儕審查文獻回顧，2000-2019年FDA核准決定分析)

U.S. Food and Drug Administration. (2023, February). Considerations for the design and conduct of externally controlled trials for drug and biological products: Draft guidance for industry. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents> (監管文件，草案)

European Medicines Agency. (2024). Concept paper on the development of a reflection paper on the use of external controls for evidence generation in regulatory decision-making. <https://www.ema.europa.eu/en/development-reflection-paper-use-external-controls-evidence-generation-regulatory-decision-making-scientific-guideline> (監管文件)

Franklin, J. M., Patorno, E., Desai, R. J., Glynn, R. J., Martin, D., Quinto, K., Pawar, A., Bessette, L. G., Lee, H., Garry, E. M., Gautam, N., & Schneeweiss, S. (2021). Emulating randomized clinical trials with nonrandomized real-world evidence studies: First results from the RCT DUPLICATE initiative. *Circulation*, 143(10), 1002-1013. <https://doi.org/10.1161/CIRCULATIONAHA.120.051718> (同儕審查論文)

Wang, S. V., Schneeweiss, S., & RCT-DUPLICATE Initiative. (2023). Emulation of randomized clinical trials with nonrandomized database analyses: Results of 32 clinical trials. *JAMA*, 329(16), 1376-1385. <https://doi.org/10.1001/jama.2023.4221> (同儕審查論文)

Garcia Requesens, M., Mankan, A. K., Cantini, L., et al. (2026). Role of health digital twins in oncology drug development: A primer. *Frontiers in Oncology*, 16, 1885797. <https://doi.org/10.3389/fonc.2026.1885797> (同儕審查回顧)

Scannell, J. W., Blanckley, A., Boldon, H., & Warrington, B. (2012). Diagnosing the decline in pharmaceutical R&D efficiency. *Nature Reviews Drug Discovery*, 11(3), 191-200. <https://doi.org/10.1038/nrd3681> (同儕審查論文，早期基礎文獻)

U.S. Food and Drug Administration. (2026). New Approach Methodologies (NAMs) (content current as of 2026-09-21); General considerations for the use of new approach methodologies in drug development: Draft guidance (CDER, 2026). <https://www.fda.gov/science-research/science-and-research-special-topics/new-approach-methodologies-nams> (監管機構資料)

U.S. Congress. (2022). FDA Modernization Act 2.0, enacted as part of the Consolidated Appropriations Act, 2023, Public Law 117-328. <https://www.congress.gov/> (法律，條款編號在本書調查中未予核對)

U.S. Congress. (2025-2026). H.R. 2821 - FDA Modernization Act 3.0, 119th Congress (passed House 2026-07-20; received in Senate 2026-07-21). <https://www.congress.gov/bill/119th-congress/house-bill/2821/all-info> (國會紀錄)

U.S. Food and Drug Administration. (2025, April). Roadmap to reducing animal testing in preclinical safety studies. <https://www.fda.gov/media/186092/download> (監管文件)

U.S. Food and Drug Administration. (2026, April). Reducing animal testing in nonclinical studies: Year one progress and the path forward; FDA achieves year 1 goals in reducing animal testing in drug development (press release, 2026-04-20). <https://www.fda.gov/media/191986/download> (監管文件)

U.S. Food and Drug Administration. (2026, September 21). FDA updates regulations to advance innovative alternatives to animal testing (press release); Nonclinical testing terminology: Direct final rule, Federal Register document 2026-19350. <https://www.fda.gov/news-events/press-announcements/fda-updates-regulations-advance-innovative-alternatives-animal-testing> (監管文件)

U.S. Food and Drug Administration. (2026, June). M15 general principles for model-informed drug development: Guidance for industry (ICH, docket FDA-2024-D-5580). <https://www.fda.gov/media/184747/download> (監管文件，最終指引)

U.S. Food and Drug Administration. (2018-). Model-Informed Drug Development Paired Meeting Program. <https://www.fda.gov/> (FDA 監管計畫說明。準確的頁面網址與詳細文字在本書調查中未予核對)

International Council for Harmonisation. (2026, February 9). Harmonised ICH M15 guideline, general principles for model-informed drug development, adopted. <https://www.ich.org/news/harmonised-ich-m15-guideline-general-principles-model-informed-drug-development-adopted> (監管機構公告。第4階段最終版：https://database.ich.org/sites/default/files/ICH_M15_Step4_Final_Guideline_2026_0129.pdf)

U.S. Food and Drug Administration. (2026). Webinar for stakeholder engagement on the new final guideline on general principles for model-informed drug development (ICH M15), 10/29/2026. <https://www.fda.gov/drugs/news-events-human-drugs/webinar-stakeholder-engagement-new-final-guideline-general-principles-model-informed-drug> (監管機構公告)

Ren, F., Aliper, A., Chen, J., et al. (2024). A small-molecule TNIK inhibitor targets fibrosis in preclinical and clinical models. *Nature Biotechnology*, 43(1), 63-75. <https://doi.org/10.1038/s41587-024-02143-0> (同儕審查論文，線上 2024-03-08)

Xu, Z., Ren, F., Wang, P., et al. (2025). A generative AI-discovered TNIK inhibitor for idiopathic pulmonary fibrosis: A randomized phase 2a trial. *Nature Medicine*, 31(8), 2602-2610. <https://doi.org/10.1038/s41591-025-03743-2> (同儕審查論文。部分二次資料中出現的 DOI 10.1038/s41591-025-03120-w 並非本論文的 DOI)

Insilico Medicine. (2026, July 7). Insilico initiates Phase III clinical trial for rentosertib, its AI-empowered TNIK inhibitor for idiopathic pulmonary fibrosis. <https://insilico.com/news/xmjns4l091-insilico-initiates-phase-iii-clinical-tr> (企業發布)

InSilico Medicine Hong Kong Limited. (2026). A prospective phase 3, randomized, double-blind, placebo-controlled, parallel-group study evaluating the efficacy and safety of rentosertib (INS018_055) administered orally over 52 weeks in patients with idiopathic pulmonary fibrosis (NCT07687459). ClinicalTrials.gov. <https://clinicaltrials.gov/study/NCT07687459> (臨床試驗註冊資訊，已確認 2026-10-13 更新版本)

Jayatunga, M. K. P., Ayers, M., Bruens, L., Jayanth, D., & Meier, C. (2024). How successful are AI-discovered drugs in clinical trials? A first analysis and emerging lessons. *Drug Discovery Today*, 29(6), 104009. <https://doi.org/10.1016/j.drudis.2024.104009> (同儕審查論文)

Sahni, N. R., & Carrus, B. (2023). Artificial Intelligence in U.S. Health Care Delivery. *New England Journal of Medicine*, 389(4), 348-358. DOI: 10.1056/NEJMr2204673

VoiceCare AI (2025, February 14). Agentic AI Startup VoiceCare AI Launches to Automate Healthcare Back Office and Super-Staff Workforce. PR Newswire. <https://www.prnewswire.com/news-releases/agentic-ai-startup-voicecare-ai-launches-to-automate-healthcare-back-office-and-super-staff-workforce-302376944.html>

Halamka, J., & Cerrato, P. (2026, August 7). An Expanding Role for Agentic AI in Healthcare. Digital Health Frontier Column, Mayo Clinic Platform. <https://www.mayoclinicplatform.org/2026/08/07/ai-agents-in-healthcare/>

Hao, Y., Holmes, J., Waddle, M. R., Davis, B. J., Yu, N. Y., et al. (2025). Personalizing prostate cancer education for patients using an EHR-Integrated LLM agent. *npj Digital Medicine*, 8, 770. DOI: 10.1038/s41746-025-02166-0

Bromelkamp, R. / Mayo Clinic News Network (2026, January 7). New Mayo Clinic study advances personalized prostate cancer education with an EHR-integrated AI agent. <https://newsnetwork.mayoclinic.org/discussion/new-mayo-clinic-study-advances-personalized-prostate-cancer-education-with-an-ehr-integrated-ai-agent>

NVIDIA Corporation (2025, January 13). NVIDIA Partners With Industry Leaders to Advance Genomics, Drug Discovery and Healthcare. Press Release.

<https://investor.nvidia.com/news/press-release-details/2025/NVIDIA-Partners-With-Industry-Leaders-to-Advance-Genomics-Drug-Discovery-and-Healthcare/default.aspx>

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1), 121-127. DOI: 10.1136/amiajnl-2011-000089

※ 未收錄資料：部分二次資料中的 Collaco et al.(2026, npj Digital Medicine) 與 Alhazba et al.(2026) 未能確認原文，且後者的期刊名稱與 DOI 互不相符，故未予引用。

Tempus AI, Inc. (2026, May 28). Tempus Expands Next Platform to Deliver Real-Time Clinical Intelligence and Close Critical Care Gaps Across Oncology. Business Wire / Tempus Investor Relations.
<https://investors.tempus.com/news-releases/news-release-details/tempus-expands-next-platform-deliver-real-time-clinical>

Tempus AI, Inc. (2026, May 31). Tempus Unveils the Next-Generation of Lens, Expanding its Agentic AI Platform for Oncology Drug Development. Press Release.
<https://www.tempus.com/news/pr/tempus-unveils-the-next-generation-of-lens-expanding-its-agentic-ai-platform-for-oncology-drug-development>

Tempus AI, Inc. (2026, May 29). Tempus Announces Initial Results from its Multimodal Foundation Model Efforts for Novel and Scalable Insight Generation in Oncology. Press Release. <https://investors.tempus.com/news-releases/news-release-details/tempus-announces-initial-results-its-multimodal-foundation-model>

Wong, A., Otles, E., Donnelly, J. P., Krumm, A., et al. (2021). External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. *JAMA Internal Medicine*, 181(8), 1065-1070. DOI: 10.1001/jamainternmed.2021.2626

Tempus AI, Inc. (2026, July 20). Tempus to Acquire Personalis, More Tightly Integrating Molecular Residual Disease (MRD) into Its AI-Enabled Precision Oncology Platform. Press Release. <https://investors.tempus.com/news-releases/news-release-details/tempus-acquire-personalis-more-tightly-integrating-molecular-residual-disease-mrd-into-its-ai-enabled-precision-oncology-platform>

ar / 輔助：Santhosh, C. (2026, July 20). Tempus to buy cancer test maker Personalis for \$1.5 billion. Reuters.

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44-56. DOI: 10.1038/s41591-018-0300-7

Marmarelis, M. E., Dow, J., Penney, J., et al., & Zimmerman, N. (2026). Multi-center prospective study evaluating an AI-enabled clinical decision support tool to improve biomarker testing in early-stage NSCLC. *Journal of Clinical Oncology*, 44(16_suppl), 8044. ASCO 2026 Annual Meeting.

https://doi.org/10.1200/JCO.2026.44.16_suppl.8044 (學會摘要，技術性前後比較。

Tempus 公開發布資料：<https://tempus.com/publications/multi-center-prospective-study-evaluating-an-ai-enabled-clinical-decision-support-tool-to-improve-biomarker-testing-in-early-stage-nsclc>)

Personalis, Inc. & Tempus AI, Inc. (2026, October 21). Definitive proxy statement/prospectus (Schedule 14A, DEFM14A). U.S. Securities and Exchange Commission. <https://www.sec.gov/Archives/edgar/data/1527753/000119312526427872/d141436ddefm14a.htm> / 輔助：Personalis, Inc. Form 8-K (報告日 2026-10-02). <https://www.sec.gov/Archives/edgar/data/1527753/000119312526415547/d155320d8k.htm>

※ 未能確認原文的資料：NVIDIA・Tempus 共同技術白皮書(2025)、PRECISION Challenge 相關一手資料。

Greenhalgh, T., Wherton, J., Papoutsis, C., Lynch, J., et al. (2017). Beyond Adoption: A New Framework for Theorizing and Evaluating Nonadoption, Abandonment, and Challenges to the Scale-Up, Spread, and Sustainability of Health and Care Technologies. *Journal of Medical Internet Research*, 19(11), e367. DOI: 10.2196/jmir.8775

Sendak, M. P., Ratliff, W., Sarro, D., Alderton, E., et al. (2020). Real-World Integration of a Sepsis Deep Learning Technology Into Routine Clinical Care: Implementation Study. *JMIR Medical Informatics*, 8(7), e15182. DOI: 10.2196/15182

Tian, J., Song, Y., Tang, L., An, L., & Zhou, J. (2026). Clinical Artificial Intelligence Implementation in Routine Care: Real-World Operational Outcomes from a Provincial Health System in China. *Risk Management and Healthcare Policy* (Dove Medical Press), 19, 1-11. 2026年6月刊登。DOI: 10.2147/RMHP.S604289. PMCID: PMC13245460. <https://pmc.ncbi.nlm.nih.gov/articles/PMC13245460/>

Tian, J., Zhao, Z., Tang, L., Song, Y., Li, Y., & Jiang, N. (2026). From Pilot Trap to Institutional Capacity: A Governance Framework for Sustainable Clinical AI Implementation in Health Systems. *Journal of Medical Internet Research*, 28, e92680. DOI: 10.2196/92680

Yip, W., Fu, H., Chen, A. T., Zhai, T., Jian, W., Xu, R., et al. (2019). 10 years of health-care reform in China: progress and gaps in Universal Health Coverage. *The Lancet*, 394(10204), 1192-1204. DOI: 10.1016/S0140-6736(19)32136-1

Zhuang, B., Song, C., Lu, H., et al. (2025). WiNGPT-3.0 Technical Report. arXiv preprint arXiv:2505.17387. DOI: 10.48550/arXiv.2505.17387. <https://arxiv.org/abs/2505.17387> (預印本，同儕審查前)

衛寧健康科技集團股份有限公司 (Winning Health Technology Group) (2025, September 21). 投資者關係活動記錄表 編號 2025-003 (業績說明會投資人關係活動記錄). 巨潮資訊網(cninfo) 公告. <http://static.cninfo.com.cn/finalpage/2025-09-21/1224672897.PDF>

Unisound (雲知聲) (2026, October 21). 登頂 OpenAI HealthBench 雙榜！雲知聲醫療模型性能全面超越 GPT-6 Astra、GPT-6.1 Sol，構建高水平 AI 醫生問診能力. Official News. <https://www.unisound.com/news/165> (英文版：<https://www.unisound.com/en/news/165>)

衛寧健康科技集團股份有限公司 (2026, April 27). 投資者關係活動記錄表 編號 2026-001. 巨潮資訊網(cninfo) 公告. <http://static.cninfo.com.cn/finalpage/2026-04-27/1225212083.PDF> / 輔助：證券日報 (2026, February 3). 衛寧健康：WiNEX Copilot 已在全國近150家醫療機構部署應用. 新浪財經 轉載. <https://finance.sina.com.cn/stock/relnews/cn/2026-02-03/doc-inhkpvfi7241242.shtml>

Arora, R. K., Wei, J., Hicks, R. S., et al. (2025). HealthBench: Evaluating Large Language Models Towards Improved Human Health. arXiv preprint arXiv:2505.08775. DOI: 10.48550/arXiv.2505.08775 (OpenAI, 預印本)

Ant Group (2026, September 10). Ant Group's AI health app AQ reaches 150 million users as it expands AI-powered day-to-day health use cases. Press Release. <https://www.antgroup.com/en/news-media/press-releases/1789024200000>

李保金 (Li, Baojin) (2026, October 26). AI赋能 基层医疗数字化提速. 经济参考报 (Economic Information Daily, 新华社). <http://jjckb.xinhuanet.com/20261026/53401657b61d443cad7ad4d015ae003a/c.html>

Xu, J., Liu, J., Lao, J., et al. (2026). PulseMind: A Multi-Modal Medical Model for Real-World Clinical Diagnosis. arXiv preprint arXiv:2601.07344. <https://arxiv.org/abs/2601.07344>

国家卫生健康委员会 规划发展与信息化司 (2025, November 4). 《关于促进和规范"人工智能+医疗卫生"应用发展的实施意见》政策解读. <https://www.nhc.gov.cn/guihuaxs/c100132/202511/25d0e47241b04a6c85214aa1351e7ff7.shtml>

Ant Group (2025, December 15). AQ 升級及命名為「螞蟻阿福」新聞稿. <https://www.antgroup.com/en/news-media/press-releases/1765779300000> (額外調查資料中所記錄的第一手資料。於本書調查中未能重新確認頁面內容)

國家衛生健康委辦公廳, 國家發展改革委辦公廳, 工業和信息化部辦公廳, 國家中醫藥局綜合司, 國家疾控局綜合司 (2025, October 20). 關於促進和規範「人工智慧+醫療衛生」應用發展的實施意見 (國衛辦規劃發〔2025〕30號). 中國政府網 發布. https://www.gov.cn/zhengce/zhengceku/202511/content_7047018.htm

北京市卫生健康委员会 (2025, December 30). 北京市支持医疗健康领域人工智能应用发展行动计划 (2026-2027年) (京卫信息〔2025〕53号). https://wjw.beijing.gov.cn/zwgk_20040/zcwj2024/202512/t20251229_4375683.html

深圳邁瑞生物醫療電子股份有限公司 (Shenzhen Mindray Bio-Medical Electronics) (2026, August 28). 2026年半年度報告 (2026 Semi-annual Report). 巨潮資訊網 (cninfo) 公告. <https://static.cninfo.com.cn/finalpage/2026-08-29/1225529856.PDF>

Zheng, L., Lin, Z., Liu, X., Fan, Z., He, Z., Xu, J., & Yin, L. (2026). Critical Care-Specific vs General-Purpose Large Language Models in Emergency Intensive Care Unit Diagnosis: Single-Center Retrospective Paired Comparative Study. *Journal of Medical Internet Research*, 28, e98026. DOI: 10.2196/98026. PMID: 42580688

邁瑞醫療 (2026). 2026CMEF | 走進邁瑞AI的「集群時代」.

<https://www.mindray.com/cn/media-center/news/mindray-qiyuan-ai-cmef-2026/> / 深圳市衛生健康委員會 發布文章(2025-03-18) https://wjw.sz.gov.cn/wzx/content/post_12075823.html (B~C) / 新華網 (2025-12-11) <http://www.news.cn/tech/20251211/5daf09a9b5034f1eb1eeb6688293cb32/c.html>

※ 未採納資料：部分二手資料提及之「國民健康十五五規劃（2026年7月，預期壽命80歲目標）」、2024年84個場景指引（國衛辦規劃函〔2024〕420號）原文，以及 Sidebench(2026) 報告，因於本書調查中未能確認原文，故未列入書目。

U.S. Code. 21 U.S.C. § 360e-4, Predetermined change control plans for devices (FD&C Act § 515C), added by Pub. L. 117-328, div. FF, title III, § 3308(a), 136 Stat. 5835 (2022). Legal Information Institute, Cornell Law School. <https://www.law.cornell.edu/uscode/text/21/360e-4>

U.S. Food and Drug Administration (最初最終版本 2024年12月4日；現行發布版本 2025年8月18日發行). Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions: Guidance for Industry and FDA Staff (Final). CDRH • CBER • CDER. Docket FDA-2022-D-2628. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence> (PDF: <https://www.fda.gov/media/166704/download>)

Federal Register (2024-12-04). Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device

Software Functions; Availability. Document 2024-28361.

<https://www.federalregister.gov/documents/2024/12/04/2024-28361>

U.S. FDA, CDRH (2025-01-14). Webinar transcript: Final Guidance: Marketing Submission Recommendations for a PCCP for AI-Enabled Device Software Functions. <https://www.fda.gov/media/187905/download>

U.S. FDA, Health Canada, MHRA (2023; FDA 網頁更新 2025-08-18). Predetermined Change Control Plans for Machine Learning-Enabled Medical Devices: Guiding Principles. <https://www.fda.gov/medical-devices/artificial-intelligence-enabled-medical-devices/predetermined-change-control-plans-machine-learning-enabled-medical-devices-guiding-principles> ; 併參：U.S. FDA, Health Canada, MHRA (2021). Good Machine Learning Practice for Medical Device Development: Guiding Principles. <https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles>

U.S. FDA (截至 2026年9月更新). Artificial Intelligence-Enabled Medical Devices. Digital Health Center of Excellence. <https://www.fda.gov/medical-devices/digital-health-center-excellence/artificial-intelligence-enabled-medical-devices>

U.S. FDA (2025-01). Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations (Draft Guidance). <https://www.fda.gov/media/184856/download>

U.S. FDA, CDRH Digital Health Center of Excellence (2026-08-18). Considerations for the Regulation of Generative AI-Enabled Medical Devices: Discussion Paper and Request for Feedback. Docket FDA-2026-N-7874. <https://www.fda.gov/media/194242/download>

U.S. Food and Drug Administration, CDRH Digital Health Center of Excellence (2026-08-18). Considerations for the Regulation of Generative AI-Enabled Medical Devices: Discussion Paper and Request for Feedback. Docket FDA-2026-N-7874. <https://www.fda.gov/media/194242/download>

U.S. FDA (2026-08-18). FDA Seeks Public Feedback to Inform Regulatory Approach for Generative AI-Enabled Medical Devices (Press Announcement).

<https://www.fda.gov/news-events/press-announcements/fda-seeks-public-feedback-inform-regulatory-approach-generative-ai-enabled-medical-devices>

U.S. FDA, CDER (2025-01). Considerations for the Use of Artificial Intelligence To Support Regulatory Decision-Making for Drug and Biological Products (Draft Guidance). <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-use-artificial-intelligence-support-regulatory-decision-making-drug-and-biological>

Patel B, Blumenthal D (2026). A Novel Approach to Overseeing the Clinical Application of Generative AI. JAMA Health Forum, 7(3), e256947. <https://doi.org/10.1001/jamahealthforum.2025.6947>

Bergman A, Wachter RM, Emanuel EJ (2026). A Licensure Framework for Autonomous Clinical AI. JAMA, 335(20), 1751-1754. <https://doi.org/10.1001/jama.2026.5483>

Freyer O, Jayabalan S, Kather JN, Gilbert S (2025). Overcoming regulatory barriers to the implementation of AI agents in healthcare. Nature Medicine, 31, 3239-3243. <https://doi.org/10.1038/s41591-025-03841-1>

U.S. FDA (2025-10-20 發布). Digital Health Advisory Committee, November 6, 2025: Executive Summary (Generative AI-enabled devices). <https://www.fda.gov/media/189391/download>

American Hospital Association (2026-10-19). AHA Provides Feedback to FDA on Regulatory Considerations for Generative AI-Enabled Medical Devices. <https://www.aha.org/2026-10-19-aha-provides-feedback-fda-regulatory-considerations-generative-ai-enabled-medical-devices>

Bedi S, Liu Y, Orr-Ewing L, et al. (2025). Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review. JAMA, 333(4), 319-328. <https://doi.org/10.1001/jama.2024.21700>

Bean AM, Payne RE, Parsons G, et al. (2026). Reliability of LLMs as medical assistants for the general public: a randomized preregistered study. Nature Medicine, 32, 609-615. <https://doi.org/10.1038/s41591-025-04074-y>

Medicines and Healthcare products Regulatory Agency (2026-06-09, updated 2026-07-27). AI Airlock Sandbox Phase 2 Programme Report. GOV.UK. <https://www.gov.uk/government/publications/ai-airlock-sandbox-phase-2-programme-report>

MHRA. AI Airlock: a regulatory sandbox for AI enabled Medical Devices (Collection page). GOV.UK. <https://www.gov.uk/government/collections/ai-airlock-a-regulatory-sandbox-for-ai-enabled-medical-devices>

MHRA (2026-04-08). MHRA expands AI Airlock programme with a £3.6 million funding boost over three years. GOV.UK. <https://www.gov.uk/government/news/mhra-expands-ai-airlock-programme-with-a-36-million-funding-boost-over-three-years>

UK Government (2026-10-06). Government backs recommendations of NHS doctors-led AI Commission. GOV.UK. <https://www.gov.uk/government/news/government-backs-recommendations-of-nhs-doctors-led-ai-commission>

European Parliament and Council (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L, 2024/1689, 12.7.2024. <http://data.europa.eu/eli/reg/2024/1689/oj> ；整合版本(2026-07-27)：
<https://eur-lex.europa.eu/eli/reg/2024/1689/2026-07-27/eng>

European Commission, AI Act Service Desk. Article 113: Entry into force and application. <https://ai-act-service-desk.ec.europa.eu/en/ai-act/article-113>

Medical Device Coordination Group & AI Board (2025-06). MDCG 2025-6 / AIB 2025-1: Interplay between the Medical Devices Regulation (MDR) & In vitro Diagnostic Medical Devices Regulation (IVDR) and the Artificial Intelligence Act (AIA). European Commission. https://health.ec.europa.eu/document/download/b78a17d7-e3cd-4943-851d-e02a2f22bbb4_en (注意：敘述修訂前之適用日期)

European Parliament and Council (2026). Regulation (EU) 2026/1744 of 8 July 2026 amending Regulations (EU) 2024/1689, (EU) 2018/1139 and (EU) 2023/1230 as regards the simplification of the implementation of harmonised rules on

artificial intelligence (Digital Omnibus on AI). Official Journal of the European Union, L, 2026/1744, 24.7.2026. <http://data.europa.eu/eli/reg/2026/1744/oj>

European Commission (2026-07-27). AI Omnibus enters into force. Shaping Europe's digital future. <https://digital-strategy.ec.europa.eu/en/news/ai-omnibus-enters-force>

Taylor Wessing (2026-07-02). Update: AI-Enabled Medical Devices and IVDs Confirmed as High-Risk. Life Sciences Legal Lens (二手資料). <https://www.taylorwessing.com/en/international-life-sciences-newsletter/life-sciences-legal-lens-vol-2/update-ai-enabled-medical-devices-and-ivds-confirmed-as-high-risk>

White & Case (2026-08-04). EU AI Omnibus enters into force, amending the AI Act (二手資料). <https://www.whitecase.com/insight-alert/eu-ai-omnibus-enters-force-amending-ai-act>

Rabotin A, Aguilar E, Sandoval Gonzalez S, Iniguez A, Jung C, Lee E, Bell DS, Arroyo J (2026). Propagation of Interpreter Errors by Ambient AI Scribes: Study Using Simulated Clinical Encounters. JMIR Medical Informatics, 14, e88734. <https://doi.org/10.2196/88734> (PMID 42520279)

Flores G, Laws MB, Mayo SJ, Zuckerman B, et al. (2003). Errors in Medical Interpretation and Their Potential Clinical Consequences in Pediatric Encounters. Pediatrics, 111(1), 6-14. <https://doi.org/10.1542/peds.111.1.6>

Koenecke A, Nam A, Lake E, Nudell J, Quartey M, Mengesha Z, Toups C, Rickford JR, Jurafsky D, Goel S (2020). Racial disparities in automated speech recognition. Proceedings of the National Academy of Sciences, 117(14), 7684-7689. <https://doi.org/10.1073/pnas.1915768117>

Zolnoori M, Vergez S, Xu Z, et al. (2024). Decoding disparities: evaluating automatic speech recognition system performance in transcribing Black and White patient verbal communication with nurses in home healthcare. JAMIA Open, 7(4), ooae130. <https://doi.org/10.1093/jamiaopen/ooae130>

Topaz M, Peltonen LM, Zhang Z (2025). Beyond human ears: navigating the uncharted risks of AI scribes in clinical practice. *npj Digital Medicine*, 8, 569. <https://doi.org/10.1038/s41746-025-01895-6>

Goddard K, Roudsari A, Wyatt JC (2012). Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1), 121-127. <https://doi.org/10.1136/amiajnl-2011-000089>

Lyell D, Coiera E (2017). Automation bias and verification complexity: a systematic review. *Journal of the American Medical Informatics Association*, 24(2), 423-431. <https://doi.org/10.1093/jamia/ocw105>

Dratsch T, Chen X, Rezazade Mehrizi M, et al. (2023). Automation Bias in Mammography: The Impact of Artificial Intelligence BI-RADS Suggestions on Reader Performance. *Radiology*, 307(4), e222176. <https://doi.org/10.1148/radiol.222176>

Taib, A. G., Partridge, G. J. W., Phillips, P., Maxwell-Armstrong, C., Chen, X., Hofvind, S. S. H., James, J. J., & Chen, Y. (2026). Automation Bias in Action: Eye Tracking of Humans Reading Screening Mammograms with and without AI Prompts. *Radiology*, 320(1), e252590. 線上 2026-07-14. <https://doi.org/10.1148/radiol.252590>

Qazi IA, Ali A, Khawaja AU, Akhtar MJ, Sheikh AZ, Alizai MH (2026). Automation Bias in Large Language Model-Assisted Diagnostic Reasoning among Physicians Trained in AI Literacy: A Randomized Clinical Trial. *NEJM AI*, 3(5). <https://doi.org/10.1056/Aloa2501001>

Wu L, Xu Z, Wang H, et al. (2026). Dual- versus Single-Suggestion AI Support for Radiographic Interpretation in Residents: Randomized Multireader Study. *arXiv preprint arXiv:2610.09589* (同儕審查前). <https://doi.org/10.48550/arXiv.2610.09589>

European Parliament and Council (2024). Regulation (EU) 2024/1689 (Artificial Intelligence Act), Article 14 Human oversight. <http://data.europa.eu/eli/reg/2024/1689/oj>

Jiang YH, Black KC, Geng G, et al. (2025). MedAgentBench: A Virtual EHR Environment to Benchmark Medical LLM Agents. NEJM AI, 2(9).
<https://doi.org/10.1056/Aldbp2500144>

ECRI (2026-01). Top 10 Health Technology Hazards for 2026: Misuse of AI chatbots tops annual list. <https://home.ecri.org/blogs/ecri-news/misuse-of-ai-chatbots-to-ps-annual-list-of-health-technology-hazards>

Grand View Research (2026). Global Agentic AI in Healthcare Market Size & Outlook, 2033 (Horizon Outlook). <https://www.grandviewresearch.com/horizon/outlook/agentic-ai-in-healthcare-market-size/global> ；報告：Agentic AI In Healthcare Market (2026-2033) Size, Share & Trends Analysis Report. <https://www.grandviewresearch.com/industry-analysis/agentic-ai-healthcare-market-report>

Grand View Research. North America Agentic AI in Healthcare Market Size & Outlook. <https://www.grandviewresearch.com/horizon/outlook/agentic-ai-in-healthcare-market/north-america>

Towards Healthcare (2026-08-17). Agentic AI in Healthcare Market Sizing (比較用二手市場資料). <https://www.towardshealthcare.com/insights/agentic-ai-in-healthcare-market-sizing>

VoiceCare AI (2025-02-14). Agentic AI Startup VoiceCare AI Launches to Automate Healthcare Back Office and Super Staff Workforce. PR Newswire. <https://www.prnewswire.com/news-releases/agentic-ai-startup-voicecare-ai-launches-to-automate-healthcare-back-office-and-super-staff-workforce-302376944.html>

NVIDIA (2025-01-13). NVIDIA Partners With Industry Leaders to Advance Genomics, Drug Discovery and Healthcare. <https://investor.nvidia.com/news/press-release-details/2025/NVIDIA-Partners-With-Industry-Leaders-to-Advance-Genomics-Drug-Discovery-and-Healthcare/default.aspx>

Bipartisan Policy Center (2026-02-18). Paying for AI in U.S. Health Care (Issue Brief). <https://bipartisanpolicy.org/issue-brief/paying-for-ai-in-u-s-health-care>

American Medical Association (2026). CPT Appendix S: Taxonomy for Artificial Intelligence in Medical Services and Procedures (CPT 2027) 及 Appendix S Table. <https://www.ama-assn.org/practice-management/cpt/cpt-appendix-s-taxonomy-artificial-intelligence-medical-services-procedures> ; <https://www.ama-assn.org/system/files/appendix-s-table.pdf>

American Medical Association (2026-06-10). CPT Editorial Panel strengthens AI taxonomy to keep pace with tech. <https://www.ama-assn.org/practice-management/cpt/cpt-editorial-panel-strengthens-ai-taxonomy-keep-pace-tech>

Centers for Medicare & Medicaid Services (2020). FY 2021 Hospital Inpatient Prospective Payment System (IPPS) and LTCH PPS Final Rule (CMS-1735-F), New Technology Add-On Payment for ContaCT (Viz.ai). Federal Register. <https://www.federalregister.gov/>

Bipartisan Policy Center (2026-10-26). Paying for Clinical AI: Principles for Policymakers. <https://bipartisanpolicy.org/article/paying-for-clinical-ai-principles-for-policymakers/>

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the Middle: How Language Models Use Long Contexts. Transactions of the Association for Computational Linguistics, 12, 157–173. https://doi.org/10.1162/tacl_a_00638 (2024年核心先行研究)

Bedi, S., Liu, Y., Orr-Ewing, L., et al. (2025). Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review. JAMA, 333(4), 319–328. <https://doi.org/10.1001/jama.2024.21700>

Ma, S., Huang, J., Zhang, F., Wu, J., Shen, Y., Fan, G., et al. (2026). MedLA: A Logic-Driven Multi-Agent Framework for Complex Medical Reasoning with Large Language Models. Proceedings of the AAAI Conference on Artificial Intelligence, 40(2), 845–853. <https://doi.org/10.1609/aaai.v40i2.37052> (擴充版 arXiv:2509.23725)

Schmidgall, S., et al. (2024, v5 2025). AgentClinic: a multimodal agent benchmark to evaluate AI in simulated clinical environments. arXiv preprint arXiv:2405.07960. <https://doi.org/10.48550/arXiv.2405.07960> (預印本論文)

Liu, Y., Carrero, Z. I., Jiang, X., Ferber, D., Wölflein, G., Zhang, L., Jayabalan, S., Lenz, T., Hui, Z., & Kather, J. N. (2026). Benchmarking large language model-based agent systems for clinical decision tasks. *npj Digital Medicine*, 9, 259. <https://doi.org/10.1038/s41746-026-02443-6>

Chen, M., Wu, Y., Ma, J., Jia, X., Gao, C., Zhao, F., & Qiao, Y. (2026). Independent and collaborative performance of large language models and healthcare professionals in diagnosis and triage. *npj Digital Medicine*, 9, 222. <https://doi.org/10.1038/s41746-026-02409-8>

Bean, A. M., Payne, R. E., Parsons, G., et al. (2026). Reliability of LLMs as medical assistants for the general public: a randomized preregistered study. *Nature Medicine*, 32, 609–615. <https://doi.org/10.1038/s41591-025-04074-y> (2026年4月17日出版社更正公告：Nature Medicine 32(5):1937，<https://doi.org/10.1038/s41591-026-04404-8>. 未能確認更正內容是否影響正文數值)

Alhazba, S. A., Ali, M. H., Alawi, D. A., Al-Betar, M. A., & Elaziz, M. A. (2026). From Reactive AI to Agentic Systems: A Review of Autonomous Medical AI Agents in Healthcare. *Archives of Computational Methods in Engineering*, 33(8), 11693–11730. <https://doi.org/10.1007/s11831-026-10655-y>

※ 參考文獻註：部分次級資料中所列的 "AgentClinic, *npj Digital Medicine* 8:112(DOI 10.1038/s41746-025-01452-1)" 與 "統合分析 8:245(DOI 10.1038/s41746-025-01588-0)"，經查詢 DOI 確認為其他論文，故未予採用。
1.4%・1.0% 數值的實際出處為 Liu 等人的 2026 年論文「Benchmarking large language model-based agent systems for clinical decision tasks」。

Russell, S., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson. (代理人的古典定義，教科書)

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing Reasoning and Acting in Language Models. *International Conference on Learning Representations (ICLR 2023)*. <https://arxiv.org/abs/2210.03629>

Shinn, N., Cassano, F., Berman, E., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language Agents with Verbal Reinforcement Learning. *Advances in*

Neural Information Processing Systems 36 (NeurIPS 2023).

<https://arxiv.org/abs/2303.11366>

Jiang, Y. H., Black, K. C., Geng, G., et al. (2025). MedAgentBench: A Virtual EHR Environment to Benchmark Medical LLM Agents. NEJM AI, 2(9).

<https://doi.org/10.1056/Aldbp2500144>

Sumers, T. R., Yao, S., Narasimhan, K., & Griffiths, T. L. (2024). Cognitive Architectures for Language Agents. Transactions on Machine Learning Research.

<https://arxiv.org/abs/2309.02427>

Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems 33 (NeurIPS 2020). <https://arxiv.org/abs/2005.11401>

Packer, C., Wooders, S., Lin, K., Fang, V., Patil, S. G., Stoica, I., & Gonzalez, J. E. (2023). MemGPT: Towards LLMs as Operating Systems. arXiv preprint arXiv:2310.08560. <https://arxiv.org/abs/2310.08560>

Schmidgall, S., et al. (2024, v5 2025). AgentClinic: a multimodal agent benchmark to evaluate AI in simulated clinical environments. arXiv preprint arXiv:2405.07960. <https://doi.org/10.48550/arXiv.2405.07960>

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative Agents: Interactive Simulacra of Human Behavior. Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23). <https://doi.org/10.1145/3586183.3606763>

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., et al. (2017). Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences, 114(13), 3521–3526. <https://doi.org/10.1073/pnas.1611835114>

Latent Labs Team (2026). Latent-Y: A Lab-Validated Autonomous Agent for De Novo Drug Design. arXiv preprint arXiv:2603.29727. <https://arxiv.org/abs/2603.29727> (企業技術報告，預印本)

[輔助文獻] Alhazba, S. A., et al. (2026). From Reactive AI to Agentic Systems: A Review of Autonomous Medical AI Agents in Healthcare. Archives of

Computational Methods in Engineering, 33(8), 11693–11730.

<https://doi.org/10.1007/s11831-026-10655-y> (將 Agentic AI 定義為「感知-推理-行動循環、持續性記憶、工具使用」)

※ 參考文獻註：部分次級資料中的 "Latent-Y / DBTL (Nature 2026)" 未能確認是否刊登於 Nature，實際出處為 Latent Labs Team 的 arXiv 技術報告。"

ProtoPilot"(arXiv:2606.31763，生物學實驗協定自動生成代理)雖為真實存在的預印本論文，但因與本節主題未直接相關，故未於正文中探討。

Vatsal, S., Dubey, H., & Singh, A. (2026). Agentic AI in Healthcare and Medicine: A Seven-Dimensional Taxonomy for Empirical Evaluation of LLM-Based Agents. IEEE Access, 14, 4840–4863. <https://doi.org/10.1109/ACCESS.2026.3651218> (預印本 arXiv:2602.04813, <https://doi.org/10.48550/arXiv.2602.04813>)

Liu, Y., Carrero, Z. I., Jiang, X., et al., & Kather, J. N. (2026). Benchmarking large language model-based agent systems for clinical decision tasks. npj Digital Medicine, 9, 259. <https://doi.org/10.1038/s41746-026-02443-6>

※ 參考文獻註：部分二手資料提及之 Sidebench 產業報告、McKinsey 約 6%、MIT Project NANDA 95% 等數值，因於本書調查中未能直接確認原始出處，故未予採用。部分二手資料將 Vatsal 研究範圍記為「2021～2026年初」之處，經確認論文正文後已更正為「2023年10月～2025年6月」。

Vatsal, S., Dubey, H., & Singh, A. (2026). Agentic AI in Healthcare and Medicine: A Seven-Dimensional Taxonomy for Empirical Evaluation of LLM-Based Agents. IEEE Access, 14, 4840–4863. <https://doi.org/10.1109/ACCESS.2026.3651218> (arXiv:2602.04813)

Tang, X., Zou, A., Zhang, Z., Li, Z., Zhao, Y., Zhang, X., Cohan, A., & Gerstein, M. (2024). MedAgents: Large Language Models as Collaborators for Zero-shot Medical Reasoning. Findings of the Association for Computational Linguistics: ACL 2024. <https://arxiv.org/abs/2311.10537>

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2024). Improving Factuality and Reasoning in Language Models through Multiagent Debate. Proceedings of the 41st International Conference on Machine Learning (ICML 2024). <https://arxiv.org/abs/2305.14325>

Kim, Y., Park, C., Jeong, H., et al. (2024). MDAgents: An Adaptive Collaboration of LLMs for Medical Decision-Making. *Advances in Neural Information Processing Systems* 37 (NeurIPS 2024). <https://arxiv.org/abs/2404.15155>

Ma, S., Huang, J., Zhang, F., Wu, J., Shen, Y., Fan, G., et al. (2026). MedLA: A Logic-Driven Multi-Agent Framework for Complex Medical Reasoning with Large Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(2), 845–853. <https://doi.org/10.1609/aaai.v40i2.37052>

Li, J., Lai, Y., Li, W., Ren, J., Zhang, M., Kang, X., Wang, S., Li, P., Zhang, Y.-Q., Ma, W., & Liu, Y. (2024, v3 2025). Agent Hospital: A Simulacrum of Hospital with Evolvable Medical Agents. *arXiv preprint arXiv:2405.02957*. <https://doi.org/10.48550/arXiv.2405.02957>

Chen, S., Chen, F., Ding, X., Chen, W., & Wei, Z. (2026). ClinVerse: A Realistic Multi-Agent Diagnostic Environment for Training and Evaluating Medical Agents. *arXiv preprint arXiv:2610.34878*. <https://doi.org/10.48550/arXiv.2610.34878>

Fan, Z., Wei, L., Tang, J., Chen, W., Wang, S., Wei, Z., & Huang, F. (2025). AI Hospital: Benchmarking Large Language Models in a Multi-agent Medical Interaction Simulator. *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*, 10183–10213. (書目已透過 [7] 之參考文獻重新確認)

Yao, Z., & Yu, H. (2026). LLM-Based Multi-Agent Systems for Clinical Workflows: A Survey of AI Hospitals. *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 45772–45793. <https://doi.org/10.18653/v1/2026.acl-long.2123> (書目確認，未閱覽內文詳細內容)

Cemri, M., Pan, M. Z., Yang, S., et al. (2025). Why Do Multi-Agent LLM Systems Fail? *arXiv preprint arXiv:2503.13657*. <https://arxiv.org/abs/2503.13657>

※ 參考文獻註：部分二手資料中的「MAGDA」案例因未能確認原始出處而未予探討。「ProtoPilot」（arXiv:2606.31763）雖為實際存在的預印本論文，但因超出本節範圍而未予探討。

Nori, H., Daswani, M., Kelly, C., Lundberg, S., Ribeiro, M. T., Wilson, M., Liu, X., Sounderajah, V., Carlson, J., Lungren, M. P., Gross, B., Hames, P., Suleyman, M., King, D., & Horvitz, E. (2026). Sequential diagnosis with language models. *Science*, 394(6822), 423–428. <https://doi.org/10.1126/science.aeh7930> (2026-10-22刊登，全體作者均隸屬Microsoft)

Nori, H., et al. (2025). Sequential Diagnosis with Language Models. arXiv preprint arXiv:2506.22405 (v2, 2025-07-02). <https://doi.org/10.48550/arXiv.2506.22405> (預印本，數值與Science版不同)

Microsoft AI (2026, October 22). Building Medical Superintelligence. <https://microsoft.ai/news/building-medical-superintelligence> (企業官方公告)

Microsoft (2026, June 2). Mayo Clinic and Microsoft collaborate to develop a frontier AI model for healthcare. <https://news.microsoft.com/source/2026/06/02/mayo-clinic-and-microsoft-collaborate-to-develop-a-frontier-ai-model-for-health-care/> (企業新聞稿，於[3]確認連結)

Microsoft AI研究團隊 (2026). The Building Blocks of Medical Superintelligence. *BMJ Digital Health and AI*, 2(1), e000395. <https://bmjdigitalhealth.bmj.com/content/2/1/e000395>

Everett, S. S., Bunning, B. J., Jain, P., Lopez, I., Agarwal, A., Desai, M., Gallo, R., Goh, E., et al. (2026). From tool to teammate in a randomized controlled trial of clinician-AI collaborative workflows for diagnosis. *npj Digital Medicine*, 9, 409. <https://doi.org/10.1038/s41746-026-02545-1> (NCT06911645)

※ 參考文獻註：部分二手資料中Science的卷期號「394(6821), eah7930」，經Crossref查詢後更正為「394(6822), 423–428」。部分二手資料中提及的「至2027年EHR的程式化革新」發言（引用自中國媒體），因未能確認原始出處而未使用。

Chen, S., Chen, F., Ding, X., Chen, W., & Wei, Z. (2026). ClinVerse: A Realistic Multi-Agent Diagnostic Environment for Training and Evaluating Medical Agents. arXiv preprint arXiv:2610.34878 (v1, 2026-10-26). Fudan University, Shanghai Innovation Institute, Zhejiang University, Huazhong University of Science and Technology. <https://doi.org/10.48550/arXiv.2610.34878> (預印本，同儕審查前)

Johnson, A. E. W., Bulgarelli, L., Shen, L., et al. (2023). MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10, 1.

<https://doi.org/10.1038/s41597-022-01899-x>（資料原始出處，2023年核心資料） / SICdb: Rodemund, N., et al. (2023). Salzburg Intensive Care database (SICdb). 轉引自[1]之參考文獻，由PhysioNet提供。

Walonoski, J., Kramer, M., Nichols, J., et al. (2018). Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, 25(3), 230–238. <https://doi.org/10.1093/jamia/ocx079>（收錄於2018年卷期；線上公開2017-08-30）

Li, J., Lai, Y., Li, W., et al. (2024, v3 2025). Agent Hospital: A Simulacrum of Hospital with Evolvable Medical Agents. arXiv preprint arXiv:2405.02957. <https://doi.org/10.48550/arXiv.2405.02957>（MedAgent-Zero方法的出處）

Nori, H., et al. (2026). Sequential diagnosis with language models. *Science*, 394(6822), 423–428. <https://doi.org/10.1126/science.aeh7930>（比較對象SDBench）

※ 參考文獻註：部分二手資料中的「ClinPatient主要針對中國特定醫院」、「照護者Agent」、「DPO/PPO等9種技術」、「病歷書寫時間縮短50%」，因與論文原文不符或未經證實，已予以更正或刪除。

Xu, J., Pan, M., Chen, L., Fu, J., Shi, R., & Xie, J. (2026). AI-Enabled regional tele-ECG cloud platform and improving access to cardiovascular diagnosis: real-world evidence from southern China. *Frontiers in Cardiovascular Medicine*, 13, 1869716. <https://doi.org/10.3389/fcvm.2026.1869716>（2026年7月16日更正公告：<https://doi.org/10.3389/fcvm.2026.1927426>）

Thygesen, K., Alpert, J. S., Jaffe, A. S., Chaitman, B. R., et al. (2019). Fourth universal definition of myocardial infarction (2018). *European Heart Journal*, 40(3), 237–269. <https://doi.org/10.1093/eurheartj/ehy462>

Dratsch, T., Chen, X., Rezazade Mehrizi, M., Kloeckner, R., et al. (2023). Automation bias in mammography: The impact of artificial intelligence BI-RADS suggestions on reader performance. *Radiology*, 307(4), e222176. <https://doi.org/10.1148/radiol.222176>

Lin, C.-S., Liu, W.-T., Tsai, D.-J., Lou, Y.-S., et al. (2024). AI-enabled electrocardiography alert intervention and all-cause mortality: a pragmatic randomized clinical trial. *Nature Medicine*, 30(5), 1461–1470.

<https://doi.org/10.1038/s41591-024-02961-4>

Lång, K., Josefsson, V., Larsson, A.-M., Larsson, S., et al. (2023). Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. *The Lancet Oncology*, 24(8), 936–944.

[https://doi.org/10.1016/S1470-2045\(23\)00298-X](https://doi.org/10.1016/S1470-2045(23)00298-X)

Gommers, J., Hernström, V., Josefsson, V., Sartor, H., et al. (2026). Interval cancer, sensitivity, and specificity comparing AI-supported mammography screening with standard double reading without AI in the MASAI study: a randomised, controlled, non-inferiority, single-blinded, population-based, screening-accuracy trial. *The Lancet*, 407(10527), 505–514. [https://doi.org/10.1016/S0140-6736\(25\)02464-X](https://doi.org/10.1016/S0140-6736(25)02464-X)

Zöller, N., Berger, J., Lin, I., Fu, N., et al. (2025). Human–AI collectives most accurately diagnose clinical vignettes. *Proceedings of the National Academy of Sciences*, 122(24), e2426153122. <https://doi.org/10.1073/pnas.2426153122>

Abràmoff, M. D., Lavin, P. T., Birch, M., Shah, N., & Folk, J. C. (2018). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *npj Digital Medicine*, 1, 39.

<https://doi.org/10.1038/s41746-018-0040-6>

Utah Department of Commerce, Office of Artificial Intelligence Policy. (2026). Doctrionic regulatory mitigation agreement: AI prescription renewal pilot (發布於 2026-01；合約期間註明為2025年10月~2026年10月).

<https://commerce.utah.gov/ai/regulatory-relief/authorized-ai-pilots/doctrionic/>

Utah Office of Artificial Intelligence Policy. (2026, May 19). Key statistics on the Doctrionic pilot program (Public Report 1).

<https://commerce.utah.gov/wp-content/uploads/2026/05/Doctrionic-Outcomes-May-2026.pdf>

Olson, K. D., Meeker, D., Troup, M., Barker, T. D., et al. (2025). Use of ambient AI scribes to reduce administrative burden and professional burnout. *JAMA Network Open*, 8(10), e2534976. <https://doi.org/10.1001/jamanetworkopen.2025.34976>

Lukac, P. J., Turner, W., Vangala, S., Chin, A. T., et al. (2025). Ambient AI scribes in clinical practice: A randomized trial. *NEJM AI*, 2(12). <https://doi.org/10.1056/Aloa2501000>

Afshar, M., Ryan Baumann, M., Resnik, F., Hintzke, J., et al. (2025). A pragmatic randomized controlled trial of ambient artificial intelligence to improve health practitioner well-being. *NEJM AI*, 2(12). <https://doi.org/10.1056/Aloa2500945>

Rotenstein, L. S., Holmgren, A. J., Thombley, R., Sriram, A., et al. (2026). Changes in clinician time expenditure and visit quantity with adoption of artificial intelligence–powered scribes. *JAMA*, 335(16), 1408-1417. <https://doi.org/10.1001/jama.2026.2253>

Xu, J., Pan, M., Chen, L., Fu, J., Shi, R., & Xie, J. (2026). AI-Enabled regional tele-ECG cloud platform and improving access to cardiovascular diagnosis: real-world evidence from southern China. *Frontiers in Cardiovascular Medicine*, 13, 1869716. <https://doi.org/10.3389/fcvm.2026.1869716>

Microsoft AI. (2026, October 22). Building medical superintelligence [企業發布聲明]. <https://microsoft.ai/news/building-medical-superintelligence>

Nori, H., Daswani, M., Kelly, C., Lundberg, S., et al. (2026). Sequential diagnosis with language models. *Science*, 394(6822), 423–428. <https://doi.org/10.1126/science.aeh7930>

Goh, E., Gallo, R., Hom, J., Strong, E., et al. (2024). Large language model influence on diagnostic reasoning: A randomized clinical trial. *JAMA Network Open*, 7(10), e2440969. <https://doi.org/10.1001/jamanetworkopen.2024.40969>

Everett, S. S., Bunning, B. J., Jain, P., et al. (2026). From tool to teammate in a randomized controlled trial of clinician-AI collaborative workflows for diagnosis. *npj Digital Medicine*, 9, 409. <https://doi.org/10.1038/s41746-026-02545-1>

Wu, L., Xu, Z., Wang, H., Zhou, F., et al. (2026). Dual- versus single-suggestion AI support for radiographic interpretation in residents: Randomized multireader study [Preprint]. arXiv:2610.09589. <https://doi.org/10.48550/arXiv.2610.09589>

Budzyń, K., et al. (2025). Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study. *The Lancet Gastroenterology & Hepatology*, 10(10), 896–903. [https://doi.org/10.1016/S2468-1253\(25\)00133-5](https://doi.org/10.1016/S2468-1253(25)00133-5)

Suleyman, M. (2025, November 6). Towards humanist superintelligence [企業發布聲明]. Microsoft AI. <https://microsoft.ai/news/towards-humanist-superintelligence/>

Nori, H., Daswani, M., Kelly, C., Lundberg, S., Ribeiro, M. T., Wilson, M., Liu, X., Sounderajah, V., Carlson, J., Lungren, M. P., Gross, B., Hames, P., Suleyman, M., King, D., & Horvitz, E. (2026). Sequential diagnosis with language models. *Science*, 394(6822), 423–428. <https://doi.org/10.1126/science.aeh7930>

Microsoft AI. (2026, October 22). Building medical superintelligence [企業發布聲明]. <https://microsoft.ai/news/building-medical-superintelligence> (同一文章所介紹之相關論文：The building blocks of medical superintelligence. *BMJ Digital Health & AI*, 2(1), e000395. <https://bmjdigitalhealth.bmj.com/content/2/1/e000395> — 作者與DOI於本書調查中未予確認)

ECRI. (2026). Top 10 health technology hazards for 2026: Misuse of AI chatbots tops annual list [新聞稿]. <https://home.ecri.org/blogs/ecri-news/misuse-of-ai-chatbots-tops-annual-list-of-health-technology-hazards>

Vatsal, S., Dubey, H., & Singh, A. (2026). Agentic AI in healthcare & medicine: A seven-dimensional taxonomy for empirical evaluation of LLM-based agents. *IEEE Access*, 14, 4840–4863. <https://doi.org/10.1109/ACCESS.2026.3651218> (arXiv:2602.04813)

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453. <https://doi.org/10.1126/science.aax2342>

Wittmann, B. J., Alexanian, T., Bartling, C., Beal, J., Clore, A., Diggans, J., et al. (2025). Strengthening nucleic acid biosecurity screening against generative protein design tools. *Science*, 390(6768), 82–87.

<https://doi.org/10.1126/science.adu8578>

European Commission. (2025, October 6). Withdrawal of Commission proposals (C/2025/5423) [包含COM(2022)496 › AI Liability Directive]. Official Journal of the European Union, C series. <http://data.europa.eu/eli/C/2025/5423/oj>

Utah Office of Artificial Intelligence Policy. (2026). Doctrionic regulatory mitigation agreement and Public Report 1 (May 19, 2026).

<https://commerce.utah.gov/ai/regulatory-relief/authorized-ai-pilots/doctrionic/> ;
<https://commerce.utah.gov/wp-content/uploads/2026/05/Doctrionic-Outcomes-May-2026.pdf>

U.S. Food and Drug Administration. (最初最終版本 2024-12-04 ; 現行發布版本 2025-08-18). Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions: Final guidance. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence>

U.S. Food and Drug Administration & European Medicines Agency. (2026, January). Guiding principles of good AI practice in drug development. <https://www.fda.gov/media/189581/download>

Regulation (EU) 2024/1689 (Artificial Intelligence Act), consolidated version 2026-07-27; Regulation (EU) 2026/1744 (Digital Omnibus on AI).
<https://eur-lex.europa.eu/eli/reg/2024/1689/2026-07-27/eng> ;
<https://eur-lex.europa.eu/eli/reg/2026/1744/oj/eng>

Office of Science and Technology Policy (Executive Office of the President). (2024). Framework for nucleic acid synthesis screening.

<https://aspr.hhs.gov/S3/Documents/OSTP-Nucleic-Acid-Synthesis-Screening-Framework-Sep2024.pdf>

Bedi, S., Liu, Y., Orr-Ewing, L., et al. (2025). Testing and evaluation of health care applications of large language models: A systematic review. JAMA, 333(4), 319–328. <https://doi.org/10.1001/jama.2024.21700>

增補版追加參考資料

TXP Medical. SpeechER 獲獎・產品發布。 <https://txpmedical.jp/en/news/2jmiSPoM4bpIdI9GUZFOBI/>

SIND. Caretomo 產品說明。 <https://caretomo.jp/ja>

Cleveland Clinic. Less Typing, More Talking. <https://consultqd.clevelandclinic.org/less-typing-more-talking-how-ambient-ai-is-reshaping-clinical-workflow-at-cleveland-clinic>

Yale 等. Use of Ambient AI Scribes to Reduce Administrative Burden and Professional Burnout. <https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2839542>

Abridge. Abridge Launches. <https://www.abridge.com/press-release/abridge-launches-to-help-people-stay-on-top-of-their-health>

首都醫科大學・清華大學等. Antibiotic selection CDSS. <https://medinform.jmir.org/2026/1/e98207>

Unisound. U2-Med 推出。 <https://unisound.com/en/research/126>

Unisound. U2-Med World. <https://unisound.com/en/news/165>

Microsoft AI. Sequential Diagnosis with Language Models. <https://arxiv.org/abs/2506.22405>

Fudan University, Shanghai Innovation Institute, Zhejiang University, Huazhong University of Science and Technology. ClinVerse: A Realistic Multi-Agent Diagnostic Environment for Training and Evaluating Medical Agents. <https://arxiv.org/abs/2610.34878>

Google Research. AMIE announcement. <https://research.google/blog/amie-a-research-ai-system-for-diagnostic-medical-reasoning-and-conversations/>

Google. AMIE disease management. <https://blog.google/innovation-and-ai/models-and-research/google-research/amie-for-disease-management-in-nature/>

Heidelberg • TU Dresden 等. Towards autonomous medical artificial intelligence agents. <https://www.nature.com/articles/s41586-026-10675-5>

BIDMC. AMIE clinical study. <https://bidmc.org/news-stories/all-news-stories/news/2026/10/bidmc-research-shows-safety-quality-of-ai-in-primary-care>

Arc • NVIDIA. Evo 2 announcement. <https://arcinstitute.org/news/evo2>

Google DeepMind. AlphaGenome announcement.
<https://deepmind.google/blog/alphagenome-ai-for-better-understanding-the-genome/>

Google 研究團隊. Tx-LLM. <https://arxiv.org/abs/2406.06316>

Google 研究團隊. TxGemma. <https://arxiv.org/abs/2504.06196>

Wang • Tian 等. HistoMethyl. <https://www.biorxiv.org/content/10.64898/2026.06.11.731518v1>

Houston Methodist Research Institute • Weill Cornell Medicine • NCI. Fuji.
<https://pmc.ncbi.nlm.nih.gov/articles/PMC13484847/>

Q. Sun 等. CGA-Net. <https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2026.1802570/full>

Christopher Baker 等. Octopus. <https://arxiv.org/abs/2607.16262>

Christopher Baker 等. Octopus 修改稿。 <https://www.biorxiv.org/content/10.64898/2026.07.05.736565v2>

Christopher Baker 等. CIWM. <https://arxiv.org/abs/2603.02274>

UNC • UT Austin • UPenn. Chreode. <https://arxiv.org/html/2605.28111v1>

Shanghai AI Lab 等. SCALE. <https://arxiv.org/html/2603.17380v3>

Di Francesco 等. PT-RAG. <https://arxiv.org/abs/2603.07233>

CAS • UCAS • NUS. CURE. <https://arxiv.org/html/2605.15243v1>

South China Normal • Osaka. JoPMol. <https://arxiv.org/html/2607.11978v1>

Tianqi Chen • Carlos Guestrin. XGBoost. <https://arxiv.org/abs/1603.02754>

Scott Lundberg • Su-In Lee. SHAP. <https://arxiv.org/abs/1705.07874>

Romain Lopez • Nir Yosef 等. scVI. <https://arxiv.org/abs/1709.02082>

William Peebles • Saining Xie. DiT. <https://arxiv.org/abs/2212.09748>

Stanford University. GEARS. <https://www.nature.com/articles/s41587-023-01905-6>

bioRxiv / University of Central Florida. HistoMethyl bioRxiv official metadata.
<https://api.biorxiv.org/details/biorxiv/10.64898/2026.06.11.731518>

The First People' s Hospital of Taicang. CGA-Net first-page affiliations.
<https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2026.1802570/pdf>

Sapienza • ISTI-CNR • Cambridge • Wellcome Sanger. PT-RAG initial preprint.
<https://arxiv.org/html/2603.07233v1>

NUS 等. PRECISE CURATE.AI. <https://www.nature.com/articles/s41698-025-00835-7>

NUS. CURATE.AI 2018 research. <https://news.nus.edu.sg/nus-researchers-use-ai-to-successfully-treat-metastatic-cancer-patient/>

Cleveland Clinic. Twin Precision Treatment.
<https://newsroom.clevelandclinic.org/2025/08/20/cleveland-clinic-led-research-shows-how-ai-supported-precision-health-and-lifestyle-coaching-program-can-improve-outcomes-in-patients-with-type-2-diabetes>

Dassault Systemes. Living Heart launch.

<https://www.3ds.com/newsroom/press-releases/dassault-systemes-unveils-3d-simulated-heart-breaking-new-ground-study-heart-disease-and-personalized-treatment>

OPRIMED. OPTIVIS 產品說明。 <https://www.oprimed.com/>

EMA. PROCOVA qualification. https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/qualification-opinion-prognostic-covariate-adjustment-procovatm_en.pdf

EMA. EMA qualification opinion 原文副本。 https://www.aindo.com/docs/regulatory-hub/37_Qualification_opinion_for_Prognostic_Covariate_Adjustment.pdf

Unlearn. PROCOVA regulatory rigor. <https://www.unlearn.ai/blog/built-on-regulatory-rigor-advancing-digital-twins-in-clinical-trials>

FDA. NAMs. <https://www.fda.gov/science-research/science-and-research-special-topics/new-approach-methodologies-nams>

FDA. MIDD program. <https://www.fda.gov/drugs/development-resources/model-informed-drug-development-paired-meeting-program>

Insilico Medicine. Rentosertib Phase III. <https://insilico.com/news/xmjns4l091-insilico-initiates-phase-iii-clinical-tr>

K-MELLODDY. 事業團介紹。 <https://www.kmelloddy.org/aboutus>

Insilico Medicine. Insilico official Technology introduction.
<https://insilico.com/mediakit>

VoiceCare AI. VoiceCare AI 成立・梅約試驗計畫。
<https://www.prnewswire.com/news-releases/agentice-ai-startup-voicecare-ai-launches-to-automate-healthcare-back-office-and-super-staff-workforce-302376944.html>

Mayo Clinic Platform. An Expanding Role for Agentice AI in Healthcare.
<https://www.mayoclinicplatform.org/2026/08/07/ai-agents-in-healthcare/>

Mayo Clinic/Cornell/MIT. Personalizing prostate cancer education for patients using an EHR-Integrated LLM agent. <https://pubmed.ncbi.nlm.nih.gov/41413170/>

Mayo Clinic. MedEduChat 機構解說。 <https://newsnetwork.mayoclinic.org/discussion/new-mayo-clinic-study-advances-personalized-prostate-cancer-education-with-an-ehr-integrated-ai-agent/>

Tempus. Tempus Next 官方產品介紹。 <https://www.tempus.com/solutions/next/>

Tempus. Lens 首次公開。 <https://www.tempus.com/news/pr/tempus-unveils-its-lens-platform-offering-unparalleled-access-to-one-of-the-worlds-largest-de-identified-clinical-and-molecular-datasets/>

Tempus. Lens 次世代公開。 <https://www.tempus.com/news/pr/tempus-unveils-the-next-generation-of-lens-expanding-its-agentic-ai-platform-for-oncology-drug-development/>

Mayo Clinic/Cornell/MIT. MedEduChat 論文 PMC 全文。
<https://pmc.ncbi.nlm.nih.gov/articles/PMC12714722/>

河北醫科大學第一醫院研究團隊. Clinical Artificial Intelligence Implementation in Routine Care. <https://pmc.ncbi.nlm.nih.gov/articles/PMC13245460/>

Winning Health 研究團隊. WinGPT-3.0 Technical Report.
<https://arxiv.org/abs/2505.17387>

Unisound. Medical World Models. <https://www.unisound.com/en/news/167>

Ant Group. AQ 推出。 <https://www.antgroup.com/news-media/press-releases/1750924800000>

Ant Group. How Does AI Shape the Next Decade of Healthcare.
<https://www.antgroup.com/en/news-media/press-releases/1751436000000>

深圳市衛生健康委員會. ICU 重症模型引進深圳。 https://wjw.sz.gov.cn/wzx/content/post_12075823.html

中國國家衛生健康委員會. 人工智慧+醫療衛生政策解讀。

<https://www.nhc.gov.cn/guihuaxxs/c100132/202511/25d0e47241b04a6c85214aa1351e7ff7.shtml>

U.S. FDA. PCCP 最終指引。 <https://www.fda.gov/media/166704/download>

FDA/Health Canada/MHRA. PCCP 國際原則。 <https://www.fda.gov/medical-devices/artificial-intelligence-enabled-medical-devices/predetermined-change-control-plans-machine-learning-enabled-medical-devices-guiding-principles>

U.S. FDA. Generative AI Discussion Paper.
<https://www.fda.gov/media/194242/download>

UK MHRA. AI Airlock Sandbox Phase 2 Programme Report.
<https://www.gov.uk/government/publications/ai-airlock-sandbox-phase-2-programme-report>

European Commission. AI Act Article 6. <https://ai-act-service-desk.ec.europa.eu/en/ai-act/article-6>

European Commission. AI Act Article 113. <https://ai-act-service-desk.ec.europa.eu/en/ai-act/article-113>

EU EUR-Lex. AI Act 整合法令。 <https://eur-lex.europa.eu/eli/reg/2024/1689/2026-07-27/eng>

UCLA • AltaMed 等研究團隊/JMIR. Propagation of Interpreter Errors by Ambient AI Scribes. <https://medinform.jmir.org/2026/1/e88734>

RSNA. AI Bias May Impair Radiologist Accuracy on Mammogram.
<https://www.rsna.org/media/press/i/2432>

研究團隊/Radiology. Automation Bias in Mammography.
<https://pubs.rsna.org/doi/10.1148/radiol.222176>

Stanford 研究團隊. MedAgentBench 專案。 <https://stanfordmlgroup.github.io/projects/medagentbench/>

Stanford 研究團隊. MedAgentBench 預印本公開。

<https://arxiv.org/abs/2501.14654>

Stanford University, Yixing Jiang • Jonathan H. Chen 等. MedAgentBench: A Virtual EHR Environment to Benchmark Medical LLM Agents.

<https://doi.org/10.1056/Aldb2500144>

U.S. FDA. IDx-DR De Novo 判定。 https://www.accessdata.fda.gov/cdrh_docs/pdf/18/DEN180001.pdf

AMA. CPT Appendix S. <https://www.ama-assn.org/practice-management/cpt/cpt-appendix-s-taxonomy-artificial-intelligence-medical-services-procedures>

Bipartisan Policy Center. Paying for AI in U.S. Health Care.

<https://bipartisanpolicy.org/issue-brief/paying-for-ai-in-u-s-health-care/>

Grand View Research. Agentic AI in Healthcare Market.

<https://www.grandviewresearch.com/horizon/outlook/agentic-ai-in-healthcare-market-size/global>

European Commission. Article 14 Human oversight.

<https://ai-act-service-desk.ec.europa.eu/en/ai-act/article-14>

Johns Hopkins University, Stanford University, ETH Zurich 等共同研究團隊.

AgentClinic: a multimodal agent benchmark to evaluate AI in simulated clinical environments. <https://arxiv.org/abs/2405.07960>

Johns Hopkins University, Stanford University, ETH Zurich 等共同研究團隊.

AgentClinic: a multimodal agent benchmark to evaluate AI in simulated clinical environments. <https://arxiv.org/html/2405.07960v5>

Westlake University, Ant Group, CAIR HKISI CAS, China-Japan Friendship Hospital. MedLA: A Logic-Driven Multi-Agent Framework for Complex Medical Reasoning with Large Language Models. <https://arxiv.org/abs/2509.23725>

Westlake University, Ant Group, CAIR HKISI CAS, China-Japan Friendship Hospital. MedLA: A Logic-Driven Multi-Agent Framework for Complex Medical Reasoning with Large Language Models. <https://arxiv.org/html/2509.23725v3>

Princeton University, Google Research. ReAct: Synergizing Reasoning and Acting in Language Models. <https://arxiv.org/abs/2210.03629>

Princeton University, Google Research. ReAct: Synergizing Reasoning and Acting in Language Models. <https://arxiv.org/html/2210.03629v3>

Northeastern University, MIT, Princeton University. Reflexion: Language Agents with Verbal Reinforcement Learning. <https://arxiv.org/abs/2303.11366>

Northeastern University, MIT, Princeton University. Reflexion: Language Agents with Verbal Reinforcement Learning. <https://arxiv.org/html/2303.11366v4>

Stanford University, Yixing Jiang, Jonathan H. Chen 等人。MedAgentBench: Dataset for Benchmarking LLMs as Agents in Medical Applications。 <https://arxiv.org/html/2501.14654v1>

Princeton University, Theodore R. Sumers, Shunyu Yao 等人。Cognitive Architectures for Language Agents。 <https://arxiv.org/abs/2309.02427>

Princeton University, Theodore R. Sumers, Shunyu Yao 等人。Cognitive Architectures for Language Agents。 <https://arxiv.org/html/2309.02427v3>

University of California, Berkeley. MemGPT: Towards LLMs as Operating Systems. <https://arxiv.org/abs/2310.08560>

University of California, Berkeley. MemGPT: Towards LLMs as Operating Systems. <https://arxiv.org/html/2310.08560v2>

Latent Labs Team. Latent-Y: A Lab-Validated Autonomous Agent for De Novo Drug Design. <https://arxiv.org/abs/2603.29727>

Shubham Vatsal, Harsh Dubey, Aditi Singh. Agentic AI in Healthcare & Medicine: A Seven-Dimensional Taxonomy for Empirical Evaluation of LLM-based Agents. <https://arxiv.org/abs/2602.04813>

Yale University, Shanghai Jiao Tong University. MedAgents: Large Language Models as Collaborators for Zero-shot Medical Reasoning. <https://arxiv.org/abs/2311.10537>

Yale University, Shanghai Jiao Tong University. MedAgents: Large Language Models as Collaborators for Zero-shot Medical Reasoning.

<https://arxiv.org/html/2311.10537v4>

MIT、Google Research、Seoul National University Hospital 共同研究團隊。MDAgents: An Adaptive Collaboration of LLMs for Medical Decision-Making。

<https://arxiv.org/abs/2404.15155>

MIT、Google Research、Seoul National University Hospital 共同研究團隊。MDAgents: An Adaptive Collaboration of LLMs for Medical Decision-Making。

<https://arxiv.org/html/2404.15155v3>

Tsinghua University AIR 等。Agent Hospital: A Simulacrum of Hospital with Evolvable Medical Agents。

<https://arxiv.org/abs/2405.02957>

Tsinghua University AIR 等。Agent Hospital: A Simulacrum of Hospital with Evolvable Medical Agents。

<https://arxiv.org/html/2405.02957v3>

Alibaba, Huazhong University of Science and Technology, University of Southern California, Fudan University. AI Hospital: Benchmarking Large Language Models in a Multi-agent Medical Interaction Simulator. <https://arxiv.org/abs/2402.09742>

Alibaba, Huazhong University of Science and Technology, University of Southern California, Fudan University. AI Hospital: Benchmarking Large Language Models in a Multi-agent Medical Interaction Simulator.

<https://aclanthology.org/2025.coling-main.680/>

無單一開發主體；由 Microsoft 等表達為目標。Building Medical Superintelligence。

<https://microsoft.ai/news/building-medical-superintelligence/>

Fudan University, Shanghai Innovation Institute, Zhejiang University, Huazhong University of Science and Technology. ClinVerse: A Realistic Multi-Agent Diagnostic Environment for Training and Evaluating Medical Agents.

<https://arxiv.org/html/2610.34878v1>

MITRE 主導及 HIKER Group 等共同研究團隊。Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic

electronic health care record。 <https://academic.oup.com/jamia/article/25/3/230/4098271>

Stanford University, Together AI. Dynamic Cheatsheet: Test-Time Learning with Adaptive Memory. <https://arxiv.org/abs/2504.07952>

Stanford University, Together AI. Dynamic Cheatsheet: Test-Time Learning with Adaptive Memory. <https://arxiv.org/html/2504.07952v1>

論文研究團隊 / npj Digital Medicine。 From tool to teammate in a randomized controlled trial of clinician-AI collaborative workflows for diagnosis。 <https://www.nature.com/articles/s41746-026-02545-1>

平台研究：三亞人民醫院、四川大學相關研究團隊；商用人工智慧製造商在論文中未公開。 AI-Enabled regional tele-ECG cloud platform and improving access to cardiovascular diagnosis: real-world evidence from southern China。 <https://www.frontiersin.org/journals/cardiovascular-medicine/articles/10.3389/fcvm.2026.1869716/full>

National Defense Medical Center, Tri-Service General Hospital 研究團隊。 AI-enabled electrocardiography alert intervention and all-cause mortality: a pragmatic randomized clinical trial。 <https://pubmed.ncbi.nlm.nih.gov/38684860/>

National Defense Medical Center, Tri-Service General Hospital 研究團隊。 人工智慧心電圖之落地運用-三軍總醫院經驗 (2024-05-31 Grand Round 公告)。 <https://www.tsgh.ndmutsgh.edu.tw/news/191/10010/24887/10864>

MASAI: Lund University 等；Transpara: ScreenPoint Medical BV。 Mammography Screening with Artificial Intelligence。 <https://portal.research.lu.se/en/projects/mammography-screening-with-artificial-intelligence/>

MASAI: Lund University 等；Transpara: ScreenPoint Medical BV。 Transpara, 510(k) K181704 decision and summary。 https://www.accessdata.fda.gov/cdrh_docs/pdf18/K181704.pdf

MASAI: Lund University 等；Transpara: ScreenPoint Medical BV。 Breast AI Research Portfolio。 <https://screenpoint-medical.com/evidence>

IDx LLC，現為 Digital Diagnostics。De Novo classification request for IDx-DR, DEN180001。 https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN180001.pdf

IDx LLC，現為 Digital Diagnostics。LumineticsCore。 <https://www.digitaldiagnostics.com/products/eye-disease/lumineticscore/>

Doctronic. Doctronic AI Regulatory Mitigation Agreement. <https://commerce.utah.gov/ai/regulatory-relief-4/authorized-pilots/doctronic/>

美國 FDA 監管指引。Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions。 <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence>

U.S. HHS. The HIPAA Privacy Rule. <https://www.hhs.gov/hipaa/for-professionals/privacy/index.html>

European Commission. Legal framework of EU data protection. https://commission.europa.eu/law/law-topic/data-protection/legal-framework-eu-data-protection_en

Wittmann et al. ; Microsoft、IBBIS 等研究團隊。Strengthening nucleic acid biosecurity screening against generative protein design tools。 <https://pubmed.ncbi.nlm.nih.gov/41037625/>

OSTP / HHS. Framework for nucleic acid synthesis screening. <https://aspr.hhs.gov/S3/Documents/OSTP-Nucleic-Acid-Synthesis-Screening-Framework-Sep2024.pdf>

書籍資訊

AI與醫療的大轉型

作者 | 金京鎮

發行人 | 金京鎮

出版處 | 金京鎮律師出版社

出版社登記 | 2025. 3. 10. (第2025-000015號)

地址 | 首爾特別市東大門區典農路 91，百日大樓 304 室

電話 | 02-6338-1905

電子郵件 | kimkj008@gmail.com

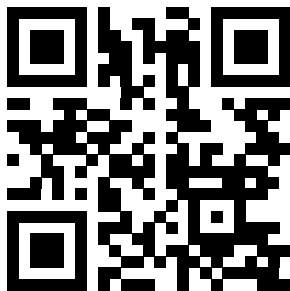
© 金京鎮 2026

本書為著作人之智慧財產，未經授權禁止擅自轉載與複製。

備註) 封面與各章插圖均為由人工智慧生成之概念意象，並非實際拍攝之臨床現場照片。本書之部分內容與編輯過程獲得了人工智慧工具的協助。

支持本人工智慧書庫書籍

若本書對您有所助益，您可以透過 PayPal 支持未來的翻譯、公開
版本發行以及開放人工智慧知識專案。



PayPal

<https://paypal.me/kimkjj>

請掃描 QR 碼或開啟上方連結。