DIAGNOSI SEQUENZIALE CON LA INTELLIGENZA ARTIFICIALE:PERCHE' NO???

Data 06 novembre 2027
Categoria Medicinadigitale

L'articolo **"Sequential diagnosis with language models"** , di Harsha Nori e collaboratori, pubblicato su Science il 22 ottobre 2026, studia una capacità dell'intelligenza artificiale particolarmente rilevante per la medicina: costruire progressivamente una diagnosi, scegliendo quali informazioni raccogliere e quali esami richiedere. I risultati sono notevoli, ma riguardano una simulazione controllata e richiedono attenzione prima di essere trasferiti alla pratica clinica.

L'idea centrale: valutare il percorso diagnostico
Molti studi sull'IA medica presentano al modello un caso già completo: sintomi, anamnesi, esame obiettivo e risultati degli accertamenti. Il sistema deve individuare la diagnosi corretta disponendo di tutte le informazioni rilevanti.
Nella realtà, il medico parte da dati limitati. Deve decidere quali domande porre, quali ipotesi considerare, quali esami prescrivere e quando le informazioni raccolte siano sufficienti per prendere una decisione. Ogni passaggio modifica le probabilità diagnostiche e comporta tempi, costi e possibili conseguenze per il paziente.
Gli autori propongono quindi di valutare la diagnosi sequenziale: un processo nel quale formulazione delle ipotesi e acquisizione delle informazioni si alternano. È un avanzamento metodologico importante, perché una diagnosi corretta ottenuta dopo una batteria indiscriminata di esami non equivale a una diagnosi raggiunta mediante un percorso appropriato.

Come è stato condotto lo studio
Gli autori hanno costruito un ambiente di valutazione, denominato Sequential Diagnosis Benchmark — SDBench — utilizzando due gruppi di casi:
• 304 casi clinico-patologici del New England Journal of Medicine, spesso complessi e caratterizzati da malattie rare o presentazioniatipiche;
• 325 accessi in pronto soccorso seguiti da ricovero, ricavati dal database MIMIC-IV-ED, più vicini alla casistica ospedalieraordinaria.
Ogni caso comincia con una breve presentazione. Il partecipante, medico oppure sistema di IA, può chiedere informazioni anamnestiche, reperti obiettivi, risultati di esami oppure formulare la diagnosi definitiva.
Un agente informatico, chiamato Gatekeeper, fornisce le informazioni richieste sulla base della documentazione originale. Un dettaglio metodologico rilevante è che, quando il dato richiesto non è disponibile, il sistema genera una risposta sintetica concepita per essere diagnosticamente neutra. Lo scopo è evitare che la semplice assenza di un'informazione suggerisca indirettamente la soluzione.
Un secondo agente, il Judge, valuta la diagnosi rispetto a quella di riferimento mediante una scala clinica a cinque livelli. Sono considerate corrette le risposte dei due livelli superiori, comprese quelle la cui differenza rispetto alla diagnosi di riferimento non modificherebbe la gestione clinica.
Gli esiti principali sono accuratezza diagnostica e costo stimato degli accertamenti. I costi derivano da un tariffario ospedaliero statunitense; le domande anamnestiche e sull'esame obiettivo sono comprese in un costo fisso della visita.

Che cosa fa il Diagnostic Orchestrator
Il sistema sperimentale, denominato Diagnostic Orchestrator — DxO — organizza il lavoro del modello linguistico simulando un gruppo di esperti con funzioni complementari.
Il procedimento comprende la revisione delle ipotesi diagnostiche, la ricerca di elementi che possano contraddirle, la selezione degli esami informativi e la valutazione dei relativi costi. Il sistema combina inoltre diagnosi prodotte in esecuzioniindipendenti.
Il contributo dell'"orchestrazione" consiste dunque nel dare una struttura al ragionamento: mantenere esplicita la diagnosi differenziale, sottoporla a critica e scegliere il passo successivo. Un gruppo di ruoli simulati, tuttavia, non equivale a un'équipe di specialisti indipendenti, con esperienze cliniche e responsabilità differenti.

I risultati principali
I risultati riportati per DxO con GPT-5 mostrano un vantaggio ampio nel contesto sperimentale:
Casistica e modalità Accuratezza diagnostica Costo medio stimato per caso
Casi NEJM: medici con accesso a fonti informative 36,1% 4.725 dollari
Casi NEJM: DxO senza vincolo di budget 83,2% 5.431 dollari
Casi NEJM: DxO con una configurazione a budget limitato 79,6% 1.303 dollari
Pronto soccorso: medici con accesso a fonti informative 46,8% 3.400 dollari
Pronto soccorso: DxO senza vincolo di budget 80,3% 4.979 dollari
Pronto soccorso: DxO con una configurazione a budget limitato 73,2% 2.634 dollari
Una precisazione è necessaria: il confronto umano sui casi NEJM è stato condotto su un sottoinsieme di 56 casi, esclusi dalle attività di sviluppo del sistema. Le percentuali riassuntive non vanno quindi interpretate come risultati di tutti i partecipanti sull'intera raccolta di 304 casi.
Un secondo punto riguarda i costi. La configurazione di DxO senza limiti di budget è più accurata, ma anche più costosa rispetto alla media dei medici. Il vantaggio economico emerge nelle configurazioni che impongono attenzione alla spesa:

in quelle condizioni, l'accuratezza diminuisce moderatamente e rimane superiore a quella dei partecipanti umani.
Il risultato interessante è quindi la possibilità di ottenere un rapporto migliore tra accuratezza e costo, scegliendo differenti configurazioni del sistema.
Anche l'organizzazione del ragionamento contribuisce al risultato: rispetto agli stessi modelli utilizzati con istruzioni minime, DxO aumenta mediamente l'accuratezza di 7,9 punti percentuali nei casi NEJM e di 4,4 punti nei casi di pronto soccorso.

Un dato particolarmente significativo: l'IA fa più domande
Prima di richiedere il primo esame, i medici formulavano mediamente 5,6 domande, mentre DxO con GPT-5 ne formulava 25,5 sugli stessi casi.
Il sistema tendeva quindi ad approfondire l'anamnesi prima di procedere agli accertamenti, soprattutto quando doveva contenere i costi. Gli autori osservano che una parte del vantaggio riguardava diagnosi nelle quali un'interrogazione più accurata poteva far emergere indizi decisivi: esposizioni insolite, viaggi o infezioni associate a circostanze particolari.
Per un medico di famiglia è un risultato di particolare interesse: una tecnologia avanzata sembra valorizzare una competenza tradizionale della medicina, saper porre domande pertinenti prima di prescrivere esami.
Va però ricordato che nel modello economico utilizzato le domande aggiuntive non comportano un costo marginale di tempo. Nella visita reale, venticinque domande richiedono ascolto, comprensione, chiarimenti e collaborazione del paziente. Il vantaggio osservato deve quindi essere verificato anche rispetto alla durata e alla qualità dell'incontro clinico.

Commento critico: che cosa dimostra e che cosa rimane aperto
Il merito principale dello studio è aver spostato l'attenzione dalla risposta finale alla qualità del percorso che conduce alla diagnosi. Rimangono, tuttavia, limiti sostanziali.
1. La simulazione semplifica la raccolta delle informazioni.
Il medico reale incontra racconti incompleti, sintomi descritti ambiguamente, difficoltà linguistiche, problemi di memoria e informazioni contraddittorie. Deve inoltre distinguere ciò che il paziente riferisce da ciò che comprende, teme o preferisce non comunicare.
Nel benchmark, invece, le risposte arrivano da un intermediario informatico. Anche quando le domande sono appropriate, non viene misurata la capacità di ottenere informazioni attendibili attraverso una relazione clinica.
La generazione di dati sintetici mancanti aggiunge un'ulteriore cautela: la loro neutralità diagnostica è un obiettivo del metodo, ma non può essere considerata equivalente a un dato realmente osservato.
2. La popolazione studiata non rappresenta la medicina generale.
I casi NEJM sono selezionati per complessità e valore didattico. I casi MIMIC riguardano pazienti ricoverati dopo una valutazione in pronto soccorso. Mancano gli incontri con persone sane o con disturbi benigni, che costituiscono una parte importante dell'attività territoriale.
Rimane quindi da verificare come si comporti il sistema quando la decisione appropriata consiste nel rassicurare, osservare l'evoluzione o evitare ulteriori accertamenti. Un'elevata capacità di riconoscere patologie in una raccolta di casi patologici non dimostra, da sola, un basso rischio di sovradiagnosi.
3. Il confronto riguarda singoli medici in condizioni sperimentali.
Gli autori riconoscono che il lavoro non dimostra superiorità rispetto a specialisti, équipe multidisciplinari o percorsi clinici reali. Il medico utilizza consulenze, osservazione nel tempo, conoscenza pregressa del paziente e confronto con altri professionisti.
Le percentuali relativamente basse dei partecipanti non possono pertanto essere interpretate come il loro tasso di correttezza diagnostica nell'attività quotidiana.
4. Accuratezza diagnostica e sicurezza clinica sono esiti diversi.
Lo studio non valuta direttamente appropriatezza terapeutica, eventi avversi, mortalità o qualità della gestione nel tempo. Una diagnosi finale corretta può essere raggiunta attraverso passaggi rischiosi; una diagnosi inizialmente incompleta può accompagnarsi a una gestione prudente e sicura.
Anche il riscontro che gli errori dell'IA siano spesso "vicini" alla diagnosi corretta non prova che siano innocui: una differenza di sottotipo o di specificità può avere conseguenze terapeutiche importanti.
5. I costi sono stime, non risparmi dimostrati nei servizi sanitari.
Il tariffario statunitense consente un confronto interno, ma non rappresenta direttamente i costi del Servizio sanitario nazionale. Tempi di attesa, disponibilità degli esami, invasività, disagio del paziente e conseguenze degli accertamenti non sono pienamente rappresentati.
6. Sono necessarie verifiche indipendenti.
Tutti gli autori dichiarano di essere dipendenti Microsoft e di possedere partecipazioni nell'azienda; alcuni sono inventori di un brevetto collegato al metodo. Ciò non invalida i risultati, ma rende particolarmente importante una replicazione esterna, possibilmente prospettica e su pazienti reali.
Le implicazioni per il ragionamento clinico e la formazione
Il contributo più promettente riguarda l'uso dell'IA come supporto alla revisione critica delle ipotesi. Un sistema organizzato può aiutare a ricordare diagnosi alternative, individuare informazioni mancanti e chiedere quale accertamento abbia davvero probabilità di modificare una decisione.
Il metodo è progettato per contrastare ancoraggio e chiusura prematura. Lo studio, però, non dimostra direttamente che questi bias diminuiscano nei medici che utilizzano DxO: manca una valutazione della collaborazione medico–IA nella pratica.
Qui si apre anche il problema della dipendenza cognitiva. Se il medico si limita ad accettare la proposta del sistema, un supporto potenzialmente utile può diventare una fonte di automatismo. Come applicazione formativa — proposta che va oltre quanto verificato nell'articolo — sarebbe ragionevole far formulare prima al medico una propria ipotesi, poi confrontarla con l'analisi dell'IA e infine motivare le decisioni assunte.

La domanda decisiva per gli studi successivi sarà quindi: un medico che lavora con questo sistema ottiene diagnosi più tempestive, meno errori e migliori esiti per i pazienti, mantenendo capacità critica e autonomia? Questo articolo offre basi convincenti per affrontare tale domanda, ma non ne fornisce ancora la risposta.

Riccardo De Gobbi, Giampaolo Collecchia, Renato Rossi

Bibliografia: Nori H, Daswani M, Kelly C, et al. Sequential diagnosis with language models. Science. 2026;394(6822):423–428.doi:10.1126/science.aeh7930.