Intelligenza Artificiale e decisione clinica

Un percorso attraverso evidenze e concetti chiave della letteratura recente.

Human in the Loop, Automation Bias ed Explainability

Dr. Erik Lagolio
Gruppo di coordinamento AI in Medicina Generale SIMG
Contributing Physicians Progetto Health Bench OpenAI

Tavola rotonda AI 25 novembre 2025

AI come supporto alle decisioni cliniche

- L'Intelligenza Artificiale entra in **diagnostica, triage, gestione delle cronicità, refertazione**.

- Gli editoriali di *Nature Medicine*, *NEJM* e *JAMA* invitano a valutare l'AI **nel contesto reale**, non solo su metriche di su metriche di laboratorio: come vengono accettate, modificate o ignorate le raccomandazioni.

- Topol descrive una "**medicina ad alte prestazioni**" che nasce dalla **convergenza fra intelligenza umana e artificiale**, **umana e artificiale**, con l'AI in ruolo di potenziamento, non sostitutivo.

Rif. Topol EJ. Nat Med 2019; Kohane IS. N Engl J Med 2024; Nature Medicine 2025.

Il progetto HealthBench: validazione su scala globale

5K+

Scenari clinici

Scenari realistici costruiti su linee guida internazionali e casi clinici reali

262

Medici coinvolti

Professionisti sanitari distribuiti in 60 paesi per validazione cross-culturale

48.5K

Rubriche valutative

Metriche su accuratezza, sicurezza, empatia, chiarezza e gestione dell'incertezza

Benchmark open-access misura la *clinical alignment* dei Large Language Models, evidenziando gap di performance significativi tra diverse aree cliniche.

Contributo personale: sviluppo di rubriche specifiche per primary care (multimorbidità, incertezza, continuità assistenziale) e analisi comparativa per definire soglie minime di sicurezza in Medicina Generale.

Human in, on, out of the loop

Human in the loop (HITL)

- L'AI elabora dati e propone diagnosi, priorità, piani o testi.
- Il clinico **esamina, integra, modifica o rifiuta** la proposta.
- La **decisione finale** e la responsabilità restano umane.

Human on the loop

L'AI agisce quasi in autonomia, mentre il clinico mantiene un ruolo di ruolo di **sorveglianza** e intervento in intervento in caso di anomalie.

Human out of the loop

L'AI prende decisioni che vengono vengono eseguite senza un controllo controllo umano significativo.

Rif. Nature Medicine 2025 – "For trustworthy AI, keep the human in the loop".

Perché il "human in the loop" è centrale

- Gli editoriali di *Nature Medicine* sottolineano che la **fiducia** nell'AI dipende dalla capacità di **integrare il giudizio clinico** e non di sostituirlo.

- La prospettiva HITL considera il **binomio clinico+AI** come unità di analisi e misura l'effetto sui pazienti, non solo sulla metrica del modello.

Rif. Nat Med 2025 – "For trustworthy AI, keep the human in the loop".

loop".

Automation Bias: definizione operativa

Revisione sistematica di Goddard et al. su *JAMIA*: l'automation bias è la tendenza a **sovrastimare l'accuratezza** del sistema del sistema automatizzato.

Si manifesta come:

Errori di omissione

il clinico trascura informazioni cliniche non evidenziate evidenziate dall'AI.

Errori di commissione

il clinico segue raccomandazioni errate, ignorando dati discordanti.

Fattori mediatori: **fiducia nell'automazione, carico di lavoro, tempo limitato, stile cognitivo** del medico.

Rif. Goddard K et al. J Am Med Inform Assoc. 2012;19(1):121–127.

Automation Bias: dati da studi empirici

Goddard et al., JAMIA 2012

- 74 studi inclusi su oltre 13.000 record iniziali.
- In alcuni contesti clinici, errori associati ad automation bias nel **6–11% delle consultazioni**.

Studi recenti con assistenti AI in patologia computazionale

- L'integrazione di AI migliora la performance media.
- Si osserva comunque un tasso di circa **7%** di casi in cui un giudizio corretto viene modificato in uno errato per seguire il modello.

Rif. Goddard K et al., JAMIA 2012; Rosbach E et al., 2024 (computational pathology).

Assistive AI: quando il “supporto” può creare danno

Khera et al. su *JAMA*: discussione di uno studio con centinaia di clinici che gestiscono un caso di dispnea con supporto AI.

supporto AI.

Risultati chiave:

AI "corretta" migliora la probabilità di diagnosi corretta rispetto al gruppo senza AI.	AI "distorta" porta i clinici a **allontanarsi allontanarsi sistematicamente** dalla diagnosi appropriata.	Le spiegazioni attenuano ma non eliminano l'effetto negativo.

Conclusione: anche un sistema "assistivo" può introdurre **nuovi tipi di errore** se l'automation bias non viene gestito.

Rif. Khera R, Simon MA, Ross JS. JAMA. 2023;330(23):2255–2257.

"When the model trains you" (NEJM AI 2024)

Studio di Kwong et al. su modello AI per idronefrosi pediatrica.

Durante il "silent trial", senza modifica apparente del contesto:

- L’uso di scintigrafia renale scende da circa 80% a 58%.
- Non vengono rilevate variazioni di linee guida, team o casi.

Interpretazione:

- L'esposizione a dataset e predizioni del modello induce **revisione delle soglie decisionali** dei clinici.
- Il modello non solo supporta decisioni, ma **plasma il modo di decidere**.

Rif. Kwong JCC et al. NEJM AI. 2024;1(2):AIcs2300004.

Studio Multicentrico in Colonscopia: evidenze concrete di deskilling percettivo

Un trial osservazionale multicentrico condotto in quattro centri endoscopici ha rilevato che l'utilizzo routinario di un sistema AI per la rilevazione dei polipi è stato associato a una **riduzione significativa dell'adenoma detection rate (ADR)** nelle procedure effettuate senza AI dopo il periodo di esposizione.

Prima dell'esposizione all'AI

ADR senza AI: 28,4%

(226 su 795 pazienti)

Dopo l'uso routinario dell'AI

ADR senza AI: 22,4%

(145 su 648 pazienti)

-6%

Differenza assoluta

IC 95% −10,5 a −1,6; p = 0,0089

1,443

Pazienti coinvolti

Studio multicentrico su 4 centri

> I risultati sono compatibili con un effetto di **deskilling percettivo e cognitivo**, con endoscopisti meno abituati a un'esplorazione visiva attiva quando l'AI non è disponibile.

Explainability (XAI): a cosa serve realmente

Amann et al. (BMC 2020) analizzano l'explainability da prospettive: **tecnologica, clinica, legale, del paziente**.

Funzioni principali in sanità:

- Supportare la **valutazione critica** della raccomandazione da parte del clinico.
- Sostenere **responsabilità** e consenso informato.
- Favorire la **fiducia** nel sistema, se usata in modo credibile. credibile.

Limiti:

- Spiegazioni post-hoc generiche possono risultare poco informative o fuorvianti.
- Un'eccessiva complessità rischia di aumentare, non ridurre, il ridurre, il carico cognitivo.

Rif. Amann J et al. BMC Med Inform Decis Mak. 2020;20:310.

Explainability e "illusione di comprensione"

Hildt (Bioengineering 2025) discute come i bisogni di explainability varino con tipo di decisione (screening, diagnosi, triage, allocazione risorse).

Per decisioni ad alto impatto individuale:

- Serve una comprensione di massima di **perché** il sistema propone X.
- Occorre sapere quali **alternative** sono state considerate o escluse.

Il rischio:

Spiegazioni poco trasparenti o troppo sofisticate possono generare una **falsa sensazione di capire**, rafforzando l'automation bias.

Rif. Hildt E. Bioengineering. 2025;12(4):375.

XAI e progettazione centrata sull'utente

Studi recenti mostrano che le tecniche XAI risultano utili solo se:

1. si integrano nei **flussi di lavoro reali** dei clinici;
2. usano un **linguaggio e visualizzazioni visualizzazioni** comprensibili;
3. vengono co-progettate con gli utilizzatori finali.

Gli autori raccomandano:

- approccio di **user-centered design** e iterazione continua;
- valutazione prospettica dell'impatto su errori, bias e tempi di decisione.

Rif. Prince EW et al. Front Radiol. 2025; altri lavori recenti su XAI e CDSS.

Principi di design per ridurre l'automation bias

Favorire il giudizio attivo attivo del clinico

- Richiedere una valutazione valutazione preliminare prima di mostrare il suggerimento.
- Evitare flussi che permettono di accettare tutto con un solo click.

Esporre incertezza e fattori fattori chiave

- Mostrare probabilità, intervalli, feature principali che hanno guidato la previsione.

Monitorare l'uso nel tempo

- Analizzare pattern di "approvazione automatica".
- Osservare eventuali cambiamenti nelle soglie decisionali dei clinici.