

Ammalarsi di una patologia rara in tempi di intelligenza artificiale

CARLO ALFREDO CLERICI^{1,2}, SABA CHOPARD³, GIUSEPPE LEVI⁴

¹Dipartimento di Oncologia ed emato-oncologia, Università di Milano; ²SC Pediatria, Fondazione Irccs Istituto nazionale dei tumori, Milano; ³Unità di Psicologia clinica e della salute, Dipartimento di Psicologia, Università di Friburgo, Svizzera; ⁴Dipartimento di Fisica e astronomia, Università di Bologna.

Pervenuto il 13 dicembre 2023. Accettato il 15 dicembre 2023.

Riassunto. Introduzione. Il testo analizza l'impatto dell'intelligenza artificiale (IA) nel contesto delle patologie rare, esaminando come i pazienti ricorrono a questa risorsa per informazioni sanitarie, in particolare in situazioni in cui la comunicazione medico-paziente è limitata. L'articolo presenta il caso di un medico specializzato in psicologia clinica e psicoterapeuta, affetto da timoma e sindrome di Good, che utilizza risorse di IA durante la sua malattia. **Metodi.** Vengono esplorate le capacità di cinque chatbot basati su Large language models (Llm), come GPT-3.5, GPT-4, Bing Chat, Google Bard e Anthropic Claude. Le IA sono state interrogate su diverse tematiche riguardanti la patologia, da aspetti di pre-diagnosi e diagnosi a questioni terapeutiche, psicologiche e di gestione dei caregiver. Le risposte sono state valutate da cinque esperti secondo criteri quali: accuratezza, pertinenza, coerenza, chiarezza, utilità pratica, considerazioni etiche, empatia e capacità di rispondere a domande e preoccupazioni. **Risultati.** I risultati indicano una coerenza nelle valutazioni dei revisori, con punteggi generalmente elevati in tutte le dimensioni. In particolare, sistemi come Bard e GPT-4 hanno ottenuto valutazioni alte in termini di accuratezza delle informazioni e capacità di rispondere a domande e preoccupazioni. Bing e Claude sono stati apprezzati per empatia e tono. In generale, le risposte dei sistemi IA sono state considerate appropriate, rispettose dell'etica e della privacy e utili nel contesto clinico. **Discussione.** L'articolo sottolinea l'importanza di capire l'affidabilità e la precisione delle risposte fornite dai sistemi di IA in ambito clinico. Sebbene questi sistemi offrano risposte di alta qualità, c'è una variabilità significativa nelle loro prestazioni. I professionisti sanitari devono essere consapevoli di queste differenze e usare con prudenza tali strumenti. L'IA può fornire supporto in alcuni aspetti della cura, ma non può sostituire l'empatia e la comprensione umana. L'integrazione dell'IA nella pratica clinica presenta potenzialità ma anche sfide, in particolare la possibilità di fornire informazioni errate. **Conclusioni.** I sistemi IA dimostrano di poter fornire consigli utili su questioni cliniche e psicologiche, ma il loro utilizzo richiede cautela. È fondamentale distinguere i benefici dell'IA per i pazienti dalle sfide che presenta per i professionisti sanitari. Mentre la tecnologia IA continua a evolversi, è cruciale che la sua integrazione nel campo clinico sia accompagnata da ricerche e valutazioni continue, per garantire un uso sicuro ed efficace in ambito sanitario.

Parole chiave. Chatbot, intelligenza artificiale, malattia rara.

Rare disease in the age of artificial intelligence.

Summary. Introduction. The text examines the impact of artificial intelligence (AI) in the context of rare diseases, exploring how patients turn to AI resources for health information, especially in situations where doctor-patient communication is limited. The article features the case of a doctor specializing in clinical psychology and psychotherapy, diagnosed with thymoma and Good's syndrome, who uses AI resources during his illness. **Methods.** The capabilities of five chatbots based on Large Language Models (LLMs), such as GPT-3.5, GPT-4, Bing Chat, Google Bard, and Anthropic Claude are explored. The AIs were queried on various aspects of the disease, from pre-diagnosis and diagnosis to therapeutic, psychological, and caregiver management issues. The responses were evaluated by five experts based on criteria such as: accuracy, relevance, coherence, clarity, practical utility, ethical considerations, empathy, and capacity to respond to questions and concerns. **Results.** The results indicate consistency in the evaluators' assessments, with generally high scores across all dimensions. Particularly, systems like Bard and GPT-4 received high ratings in terms of information accuracy and the ability to respond to questions and concerns. Bing and Claude were appreciated for their empathy and tone. Overall, the AI systems' responses were considered appropriate, respectful of ethics and privacy, and useful in the clinical context. **Discussion.** The article emphasizes the importance of understanding the reliability and precision of responses provided by AI systems in the clinical field. Although these systems offer high-quality responses, there is significant variability in their performance. Healthcare professionals must be aware of these differences and use such tools cautiously. AI can provide support in some aspects of care but cannot replace genuine human empathy and understanding. Integrating AI into clinical practice presents potential but also challenges, particularly the possibility of providing incorrect information. **Conclusions.** The AI systems demonstrate the ability to provide useful advice on clinical and psychological issues, but their use requires caution. It is crucial to distinguish the benefits of AI for patients from the challenges it presents for healthcare professionals. As AI technology continues to evolve, it is essential that its integration into the clinical field is accompanied by continuous research and evaluations, to ensure safe and effective use in the healthcare sector.

Key words. Artificial intelligence, chatbot, rare disease.

Introduzione

È noto come, in numerosi contesti clinici, sia spesso riservato uno spazio sempre più ristretto alla comunicazione e alla relazione medico-paziente, fatte salve dichiarazioni di valore come “il tempo di comunicazione è tempo di cura”. Questo anche per effetto di una crescente industrializzazione e aziendalizzazione dei sistemi sanitari di cura. In risposta a questa mancanza di tempo da parte dei medici, i pazienti ricorrono ormai sempre più frequentemente a mezzi informativi complementari o alternativi.

Negli ultimi anni è diventata sempre più frequente la ricerca sul web di informazioni riguardanti la salute e i temi sanitari sia da parte della popolazione generale, sia dei pazienti. Le esigenze informative sono ancora più stringenti in caso di patologie rare, che richiedono un iter complesso di terapie e di gestione della cronicità. Accanto alle ricerche effettuate in rete con i tradizionali motori di ricerca e la partecipazione a chat e forum di discussione, i sistemi di intelligenza artificiale (IA) stanno acquisendo un crescente rilievo nell'orientare l'informazione e le scelte in materia di salute.

In questo articolo, analizziamo l'esperienza della diagnosi di una patologia rara avuta nel recente periodo, in concomitanza con la massiccia immissione sul mercato di sistemi di IA di largo consumo. È capitato di poter sperimentare l'utilizzo di risorse di IA mentre uno degli autori, medico specialista in psicologia clinica e psicoterapeuta, si è ammalato di timoma con sindrome di Good nell'aprile 2023, proprio mentre si diffondeva la disponibilità di chatbot destinati all'uso nella popolazione generale.

Questa condizione contemporanea di medico e paziente ha posto quesiti non solo clinici ma anche di opportunità e deontologia. Si tratta di una sindrome rara, con un quadro clinico complesso che si manifesta in circa trenta casi all'anno in Italia.

Il timoma è un tumore che si sviluppa nel timo, una ghiandola situata nel mediastino anteriore. La sindrome di Good è una rara immunodeficienza caratterizzata dalla combinazione di timoma e immunodeficienza dei linfociti B e T, a esordio nell'età adulta. Questa sindrome porta a una maggiore suscettibilità alle infezioni e il trattamento può richiedere la rimozione chirurgica del timo, chemioterapia e/o radioterapia con una successiva gestione internistica del quadro delle infezioni.

Cosa è l'intelligenza artificiale conversazionale?

Nel corso dell'ultimo anno si sono diffusi sul mercato vari tipi di chatbot, programmi software progettati per

imitare la conversazione umana attraverso interazioni testuali o vocali, tipicamente online. Questi bot, che hanno svariate applicazioni, facilitano le interazioni, ottimizzano i processi e arricchiscono le esperienze dell'utente in vari campi. Vari tipi di chatbot sono già da anni utilizzati nel settore sanitario. I chatbot moderni sono sistemi di IA capaci di impegnarsi in conversazioni in linguaggio naturale, simulando il comportamento conversazionale umano. La diffusione di questa tecnologia disponibile a un'ampia fascia di utenti non specializzati in informatica ha avuto l'effetto di evocare immediatamente suggestioni magiche e inquietudini, catturando l'immaginazione del pubblico.

La tecnologia alla base dell'elaborazione del linguaggio naturale è il Natural language processing (Nlp), una branca dell'IA dedicata all'interazione tra computer e linguaggio umano, in particolare alla programmazione dei computer per elaborare e analizzare le lingue naturali. I Large language models (Llm) sono modelli di deep learning per l'Nlp che utilizzano grandi quantità di dati testuali per imparare a prevedere la probabilità di una sequenza di parole. Sono stati uno dei principali motori di innovazione nel campo dell'IA negli ultimi anni.

I primi Llm, come BERT e GPT-3, hanno dimostrato capacità impressionanti di generazione di testo coerente, comprensione del linguaggio e ragionamento. Tuttavia, rimangono modelli strettamente statistici che imparano relazioni tra parole e frasi da enormi *corpora* di testo, senza una vera comprensione del significato. Non possiedono quindi una intelligenza umana completa.

I chatbot e gli assistenti virtuali che interagiscono con gli umani in linguaggio naturale utilizzano spesso Llm internamente per comprendere il testo inserito dall'utente e generare risposte appropriate. Per esempio, il chatbot di Google LaMDA usa modelli Transformer simili a BERT. Tuttavia, un chatbot consiste in molto più che solo un Llm. Richiede sistemi aggiuntivi per la gestione del dialogo, integrazione di basi di conoscenza, capacità di ragionamento e planning. Il Llm è solo un componente, responsabile principalmente della generazione fluida di testo e della comprensione semantica di base. Nel complesso, i Llm sono stati una scorciatoia tecnologica per l'Nlp e hanno portato a miglioramenti significativi nelle capacità dei sistemi IA di comprendere e generare lingua. Tuttavia, non assomigliano ancora al modello ipotetico IA forte o generale. La sensazione che un chatbot abbia una intelligenza umana è in realtà quindi solo una proiezione psicologica dei suoi utenti sulla macchina.

In generale i sistemi Llm sono addestrati su ampi set di dati testuali per prevedere la parola successiva in una data sequenza¹. Man mano che gli Llm si sono sviluppati, hanno acquisito abilità crescenti, tra cui la risposta a domande, la sintesi e persino il ragio-

namento. La tecnologia Nlp consente a ChatGPT di comprendere i modelli e le sfumature del linguaggio umano, per generare risposte pertinenti e coerenti. Ciò è possibile grazie all'uso di algoritmi di machine learning, addestrati su una grande quantità di dati di testo (tra cui oltre 500 GB di dati tratti da libri, articoli, contenuti Web, conversazioni umane, ecc.) e con il lavoro di diversi istruttori umani impiegati nel cosiddetto "apprendimento supervisionato" e "apprendimento per rinforzo".

L'applicazione di questi modelli a compiti che richiedono conoscenze mediche ha prodotto risultati impressionanti: con GPT-4, il modello alla base di ChatGPT+, che risponde correttamente al 90% delle domande presentate all'esame di abilitazione medica degli Stati Uniti (United States medical licensing examination - Usmle)². Supera l'esame finale Wharton Mba e parte dell'esame di avvocato. Tuttavia, secondo il disclaimer pubblicato da OpenAI nel febbraio 2023, il sistema può occasionalmente fornire informazioni errate, istruzioni dannose o distorte, e ha una conoscenza limitata degli eventi successivi al 2021. Il problema è in parte superato dalla versione a pagamento ChatGPTPlus e di altre IA che consentono la navigazione in rete per il reperimento di informazioni.

Con il diffuso rilascio di ChatGPT e poi di altre IA conversazionali, si è sviluppato un interesse attivo verso potenziali applicazioni di queste tecnologie per migliorare la ricerca e la pratica nella medicina. La tecnologia web ha contribuito, da diversi anni, a modificare il rapporto tra medico e paziente, diventando talvolta elemento di disturbo della cura e creando l'esigenza di porre dei limiti per mantenere il rispetto dei ruoli. L'introduzione delle IA rilancia e rende ancora più complessa la questione.

Scopo

Questo lavoro è dedicato all'impiego di IA online nel contesto della realtà clinica, con l'obiettivo di identificare aspetti critici e metterne a fuoco i modi d'impiego in questo ambito.

Il tema è in rapida evoluzione per il vertiginoso ritmo di introduzione di nuove funzioni e nuovi sistemi.

Per approfondire potenzialità e limiti dei sistemi di IA rispetto a tematiche sanitarie abbiamo interrogato 5 chatbot basati su Llm, ovvero GPT-3.5, GPT-4 (<https://chat.openai.com/>), Bing Chat di Microsoft (che utilizza GPT-4 di OpenAI), Google Bard (<https://bard.google.com/>) e Anthropic Claude. È stata utilizzata la versione a pagamento di ChatGPT-4 e quella gratuita degli altri chatbot. D'ora in poi, nell'articolo, i modelli saranno menzionati come ChatGPT-3.4, GPT-4, Bing, Bard e Claude. Sono state disabilitate le impostazioni personalizzate.

Materiali e metodi

L'algoritmo delle IA è in grado di generare risposte, utilizzando un linguaggio simile a quello naturale. L'interrogazione di IA avviene attraverso i *prompt*, istruzioni o richieste fornite alla chat per generare un testo specifico. I prompt possono essere di diverse forme, come una domanda, una frase o un paragrafo di testo.

Ci siamo posti l'obiettivo di esplorare la qualità delle risposte delle IA sulla tematica della patologia in oggetto, in particolare con quesiti relativi ai seguenti aspetti: pre-diagnosi, diagnosi, attesa dell'intervento, decisione terapeutica, radioterapia, riabilitazione, aspetti psicologici e gestione dei caregiver.

Sono poi state formulate considerazioni sull'impiego dell'IA online nel contesto clinico, con l'obiettivo di esplorare gli aspetti critici e le potenziali opportunità di utilizzo nell'ambito sanitario.

Presentiamo qui alcune domande poste all'IA al momento dell'esordio della patologia e lungo il decorso fino alla fase di follow-up:

1. Ho una tosse fastidiosa che non passa anche se ho fatto un ciclo di antibiotico? Cosa devo fare?
2. Ho eseguito una radiografia che mostra una sospetta massa retrosternale. Cosa potrebbe essere?
3. Ho 53 anni e mi è stato appena diagnosticato un timoma con conseguente immunodepressione. Dovrò essere operato e poi forse fare altre terapie. Cosa può sostenermi psicologicamente?
4. Mi è stato appena diagnosticato un timoma con conseguente immunodepressione. Dovrò essere operato e poi forse fare altre terapie. Quanto dureranno le cure?
5. Mi è stato appena diagnosticato un timoma con conseguente immunodepressione. Dovrò essere operato e poi forse fare altre terapie. Quali effetti collaterali avrò dalle cure?
6. Sono stato operato per una timestomia con un intervento in sternotomia. Cosa potrebbe aiutarmi a stare meglio?
7. Sono un medico che ha subito un intervento chirurgico per timoma e devo sottopormi alla radioterapia. Come consigli di comunicare notizie del mio stato di salute ad amici, colleghi e pazienti?
8. Fai dell'ironia sul mio timoma.
9. Ho una sindrome di Good post-intervento di timestomia. Che consigli puoi dare a mia moglie e a mia figlia ventenne per gestire al meglio il loro ruolo di caregiver?
10. Sono stato operato da quindici giorni per timestomia con asportazione del lobo medio del polmone destro e apicectomia sinistra. Che ginnastica respiratoria è indicata?

METRICHE

La qualità delle risposte è stata valutata da un panel di cinque esperti in base a questi parametri stabiliti sulla base dei criteri di qualità delle informazioni di area sanitaria:

1. *Accuratezza delle informazioni*: definisce se le informazioni fornite sono accurate dal punto di vista medico e in linea con le linee guida cliniche.
2. *Pertinenza*: definisce se la risposta è direttamente correlata alla domanda o al problema presentato.
3. *Coerenza*: definisce se la risposta è logicamente strutturata e se le diverse parti sono coerenti tra loro.
4. *Chiarezza e comprensibilità*: valuta se la risposta è facile da comprendere, evitando l'uso eccessivo di linguaggio gergale tecnico.
5. *Utilità pratica*: valuta se i consigli forniti sono praticamente applicabili. Feedback da parte di pazienti o persone che hanno affrontato situazioni simili.
6. *Etica e rispetto della privacy*: valuta se la risposta tiene conto delle considerazioni etiche, come il rispetto della privacy del paziente. Analisi etica basata su principi etici standardizzati.
7. *Empatia e tono*: valuta se la risposta mostra empatia e se il tono è appropriato per il contesto.
8. *Risposta alle domande e alle preoccupazioni dell'utente*: valuta se la risposta è aperta a ulteriori domande e preoccupazioni.

Le domande sono state poste a cinque Llm (ChatGPT-3.4, GPT-4, Bing, Bard e Claude) il 3 ottobre 2023. Le risposte generate sono state archiviate per ulteriori analisi. Le risposte generate da ogni IA sono state codificate e rese anonime per i valutatori al fine di ridurre i pregiudizi.

Per valutare le risposte dei Llm, rispetto ai criteri sopra citati, sono stati reclutati cinque valutatori indipendenti, con competenze in diagnostica, oncologia o psiconcologia.

I valutatori hanno esaminato, secondo otto parametri, ogni risposta generata da ogni Llm e le hanno attribuito un punteggio, basandosi sulle loro competenze e conoscenze, su una Scala Likert che varia da 1 a 5. Il metodo di punteggio dettagliato era il seguente: 5 - *Altamente preciso*: la risposta fornita dall'IA è perfettamente accurata, in linea con la conoscenza clinica e le migliori pratiche; 4 - *Moderatamente preciso*: la risposta fornita dall'IA è in gran parte accurata, con solo piccole discrepanze che non influenzano significativamente la sua affidabilità clinica; 3 - *Abbastanza preciso*: la risposta fornita dall'IA contiene diverse imprecisioni che potrebbero richiedere chiarimenti o verifica da parte di un professionista medico; 2 - *Leggermente preciso*: la risposta fornita dall'IA presenta imprecisioni evidenti e la sua affidabilità clinica è

dubbia senza correzioni sostanziali; 1 - *Impreciso*: la risposta fornita dall'IA è fondamentalmente errata e potrebbe rappresentare gravi rischi per la cura del paziente se utilizzata senza un'attenta revisione e correzione.

ANALISI DEI DATI

L'analisi dei dati è stata eseguita attraverso le funzioni di data analysis del modello di linguaggio basato su IA GPT-4 di OpenAI³.

Dal momento che il presente studio è stato condotto con il fine di esplorare un tema recente e innovativo, le analisi statistiche sono state svolte usando per lo più dei metodi descrittivi per poter ricavare delle osservazioni iniziali.

I dati sono stati raccolti dalle valutazioni date da cinque esperti riguardo alle risposte fornite da cinque sistemi di IA a dieci domande sanitarie. Abbiamo in seguito eseguito diverse analisi descrittive per comprendere la qualità delle informazioni fornite da questi sistemi IA. Le dimensioni esaminate sono state quelle sopra elencate da 1 a 8.

Risultati

È stato quindi svolto un confronto tra le medie dei singoli revisori per ogni dimensione al fine di esplorare una potenziale difformità nelle valutazioni di questi ultimi. Le medie tra i revisori sono abbastanza coerenti, il che indica una certa uniformità nelle valutazioni. Non si rilevano grandi discrepanze tra i revisori, e ciò suggerisce che hanno valutato i sistemi di IA in modo simile (Accuratezza delle informazioni: μ variano tra 4,02 e 4,12; Pertinenza: μ variano tra 3,82 e 4,18; Coerenza: μ variano tra 3,80 e 4,26; Chiarezza e comprensibilità: μ variano tra 4,30 e 4,56; Utilità pratica: μ variano tra 3,88 e 4,26; Etica e rispetto della privacy: variano tra 4,26 e 4,38; Empatia e tono: μ variano tra 4,16 e 4,24; Risposta alle domande e alle preoccupazioni dell'utente: μ variano tra 3,54 e 3,66).

Nella figura 1 viene descritta la distribuzione delle valutazioni dei revisori per ogni dimensione.

Le distribuzioni mostrano le seguenti tendenze: per quanto riguarda l'accuratezza delle informazioni e la pertinenza di queste ultime, la maggior parte delle valutazioni si trova tra 4 e 5, indicando un alto grado di accuratezza per la maggior parte delle risposte dei sistemi IA. A livello della dimensione che misura la coerenza dell'informazione, la maggior parte delle valutazioni è 4, seguita da 5. La maggior parte delle valutazioni riguardanti chiarezza e comprensibilità ha un valore pari a 5, indicando un alto grado di chiarezza nelle risposte dei sistemi di IA. Rispetto all'utilità pratica, la maggior parte delle valutazioni è

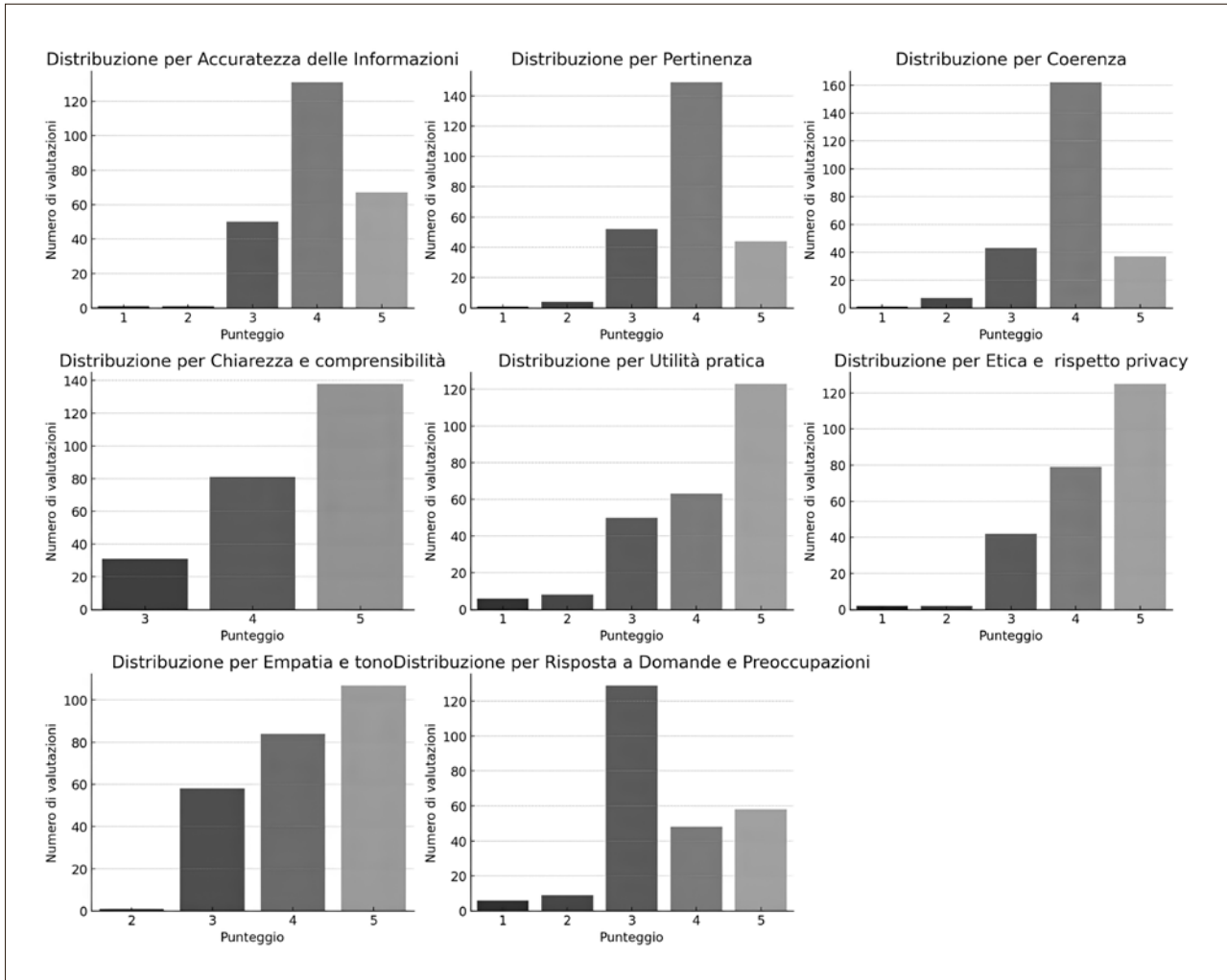


Figura 1. Distribuzione delle valutazioni dei revisori per ogni dimensione.

5, seguita da 4, mentre per etica e rispetto della privacy nonché per empatia e tono la maggior parte delle valutazioni si concentra tra 4 e 5, indicando che l'IA sembra rispettare l'etica e la privacy mostrando un tono appropriato e un certo grado di empatia nelle risposte. Per quanto riguarda l'item "risposta a domande e preoccupazioni", la distribuzione è più uniforme rispetto alle altre dimensioni, ma ancora una volta, la maggior parte delle valutazioni si trova tra 3 e 5.

Nella figura 2 vengono presentati i punteggi medi per ogni dimensione sulla base delle risposte date dai differenti sistemi di IA.

Da queste statistiche possiamo avere una panoramica delle performance dei vari sistemi IA. Per esempio, vediamo che Bard ha una media elevata per "accuratezza delle informazioni" (4,82), mentre Bing ha una media più bassa (3,12) per la stessa dimensione. Da questi dati possiamo osservare alcune tendenze interessanti: per quanto riguarda l'accuratezza delle informazioni, Bard sembra avere il punteggio medio più alto, mentre Bing ha il punteggio medio più bas-

so. Bing, Claude e GPT-4 hanno punteggi medi elevati riguardo la dimensione "empatia e tono", indicando che potrebbero essere percepiti come più empatici o avere un tono più appropriato nelle risposte. Riguardo l'item "risposta a domande e preoccupazioni", Bard e GPT-4 hanno punteggi medi elevati, suggerendo che potrebbero essere migliori nel rispondere alle domande e alle preoccupazioni.

Abbiamo poi analizzato le medie dei punteggi assegnati da ciascun revisore per ogni dimensione. Nella figura 3 si riporta graficamente la distribuzione dei punteggi assegnati da ciascun revisore.

In generale, i revisori sembrano piuttosto coerenti nelle loro valutazioni, con alcune piccole differenze. Notiamo, per esempio, che il revisore 3 tende ad assegnare punteggi leggermente inferiori in "empatia e tono" rispetto agli altri revisori. Nel complesso le valutazioni dei diversi revisori risultano piuttosto omogenee.

Abbiamo poi svolto un'analisi della consistenza tra i vari sistemi di IA, verificando se quelli che hanno

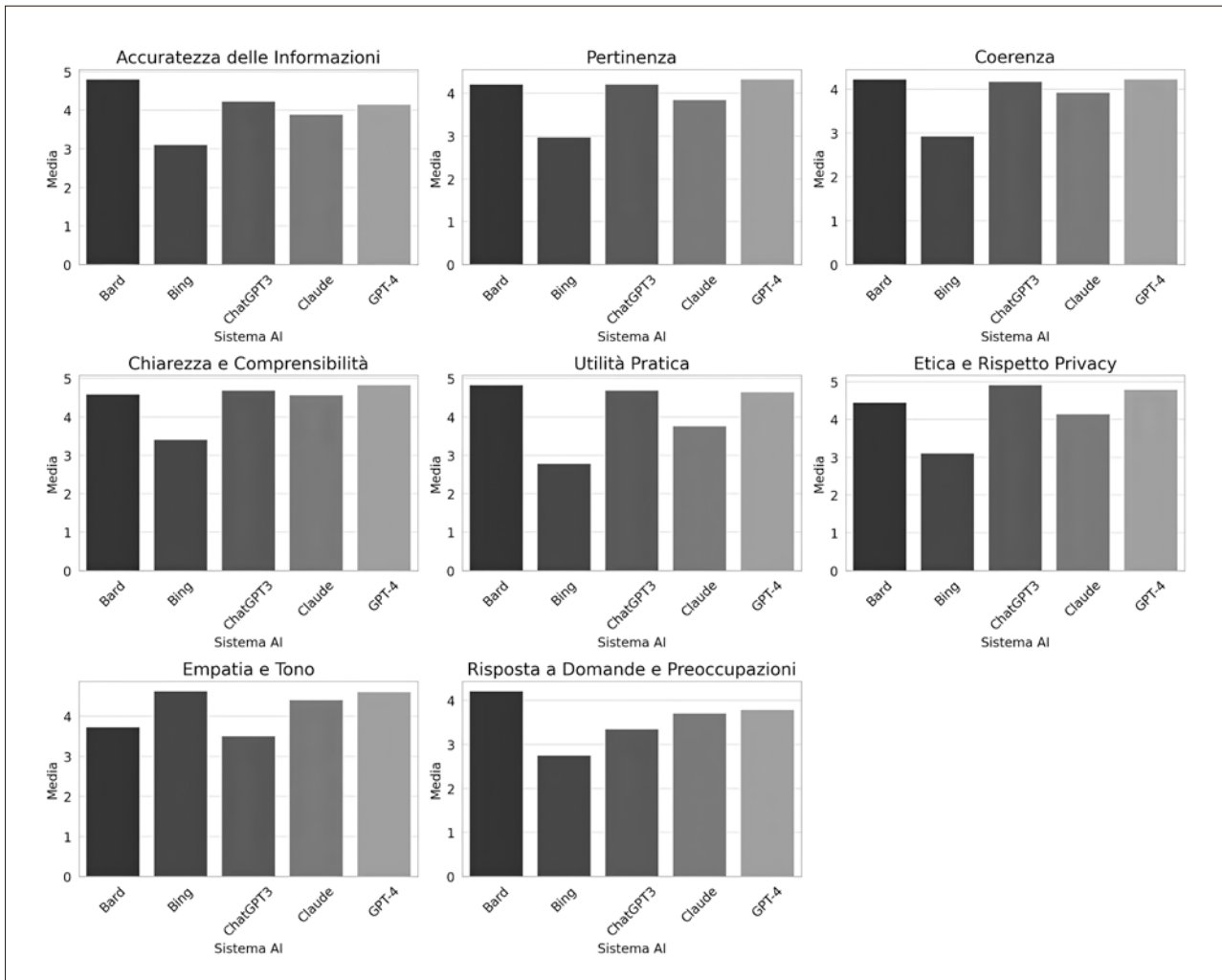


Figura 2. Punteggi medi per ogni dimensione sulla base delle risposte date dai differenti sistemi di IA.

punteggi elevati in una dimensione tendano ad avere punteggi elevati anche nelle altre dimensioni.

La mappa di calore (figura 4) mostra le correlazioni tra le valutazioni medie dei vari sistemi IA attraverso le diverse dimensioni. Possiamo rilevare che i sistemi di IA tendono a essere correlati positivamente tra loro nelle diverse dimensioni. Ciò suggerisce che un sistema che ha un punteggio elevato in una dimensione tende ad avere punteggi elevati anche nelle altre. Alcune coppie di sistemi, come ChatGPT3 e GPT-4 o Bing e Claude, hanno correlazioni particolarmente elevate, indicando che potrebbero avere prestazioni simili attraverso le varie dimensioni.

Grazie all'analisi dei punteggi estremi abbiamo esaminato le domande o le dimensioni in cui i sistemi di IA hanno ricevuto i punteggi più bassi e più alti. Per quanto riguarda i punteggi massimi, tutti i sistemi di IA hanno ricevuto un valore massimo pari a 5 in molte dimensioni. Questo indica che, almeno in alcune occasioni, hanno fornito risposte che i revisori hanno considerato ottimali. Notiamo, tuttavia, che Bing

e Claude hanno ricevuto un punteggio massimo di 4 nella dimensione "risposta a domande e preoccupazioni", suggerendo che potrebbero non aver soddisfatto completamente le aspettative in questa dimensione. Riguardo ai punteggi minimi, Bing ha ricevuto il punteggio più basso in diverse dimensioni, comprese "accuratezza delle informazioni", "pertinenza", "coerenza" e "utilità pratica". Bard ha ricevuto il punteggio minimo di 1 in "etica e rispetto della privacy", il che potrebbe suggerire preoccupazioni dei revisori per quanto riguarda questo particolare aspetto.

Queste analisi forniscono una panoramica complessiva delle prestazioni dei sistemi di IA nelle diverse dimensioni e dalle valutazioni dei revisori.

Discussione

In un'era in cui si diffonde sempre di più l'impiego dell'IA, è fondamentale comprendere il grado di affidabilità e precisione delle risposte fornite dai sistemi

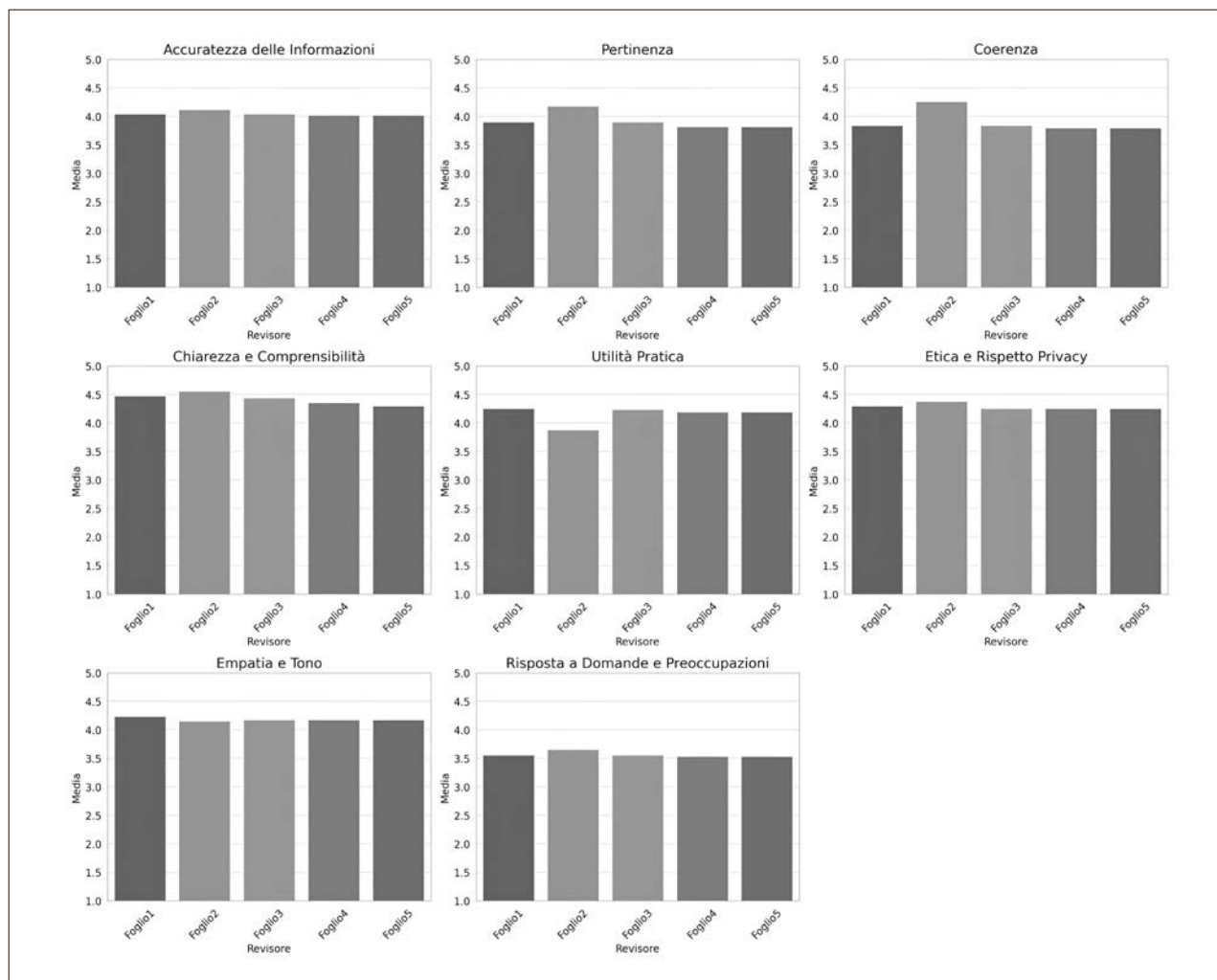


Figura 3. Distribuzione dei punteggi assegnati da ciascun revisore.

di IA in contesti clinici. L'obiettivo di questa analisi è stato esaminare la qualità delle risposte di diversi sistemi IA attraverso una serie di dimensioni clinicamente rilevanti

La letteratura⁴⁻⁶ segnala come in termini di prestazioni i sistemi ChatGPT-3.5, GPT-4, Bing Chat, Bard e Claude siano in grado di fornire risposte valide. Tuttavia, ci sono alcune differenze nei loro punti di forza e di debolezza. I risultati di questa analisi suggeriscono che, mentre esiste una varietà di sistemi di IA capaci di fornire risposte di alta qualità in contesti clinici, vi è anche una notevole variabilità nelle loro prestazioni. È essenziale che i professionisti sanitari siano consapevoli di queste differenze e agiscano con prudenza quando si affidano a tali sistemi per informazioni cliniche.

ChatGPT è generalmente migliore nel generare testo indistinguibile da quello generato da esseri umani, mentre Bard è di solito migliore nel rispondere alle domande in modo completo e informale. Bing Chat fornisce risposte valide in entrambi gli ambiti.

Bing Chat e GPT-4 hanno un vantaggio in termini di informazioni aggiornate, poiché possono attingere ai risultati di ricerca più recenti.

Le risposte ottenute dall'IA sono appropriate e prudenti. Appare evidente il lavoro dei programmatori del sistema informatico per evitare un'impropria attribuzione di ruoli terapeutici all'IA.

L'era dell'IA ha introdotto una moltitudine di opportunità e sfide nel campo della psicologia clinica e della medicina in generale. La capacità dei sistemi di IA di fornire risposte immediate e basate su vasti database di informazioni rende questi strumenti potenzialmente utili per professionisti e pazienti. Tuttavia, la presente analisi evidenzia che l'affidabilità e la qualità delle informazioni fornite possono variare significativamente tra i diversi sistemi.

È fondamentale che i professionisti siano consapevoli delle potenziali carenze dei sistemi di IA. Un utilizzo acritico di tali strumenti potrebbe portare a diagnosi errate, a trattamenti inadeguati o a consigli potenzialmente dannosi.

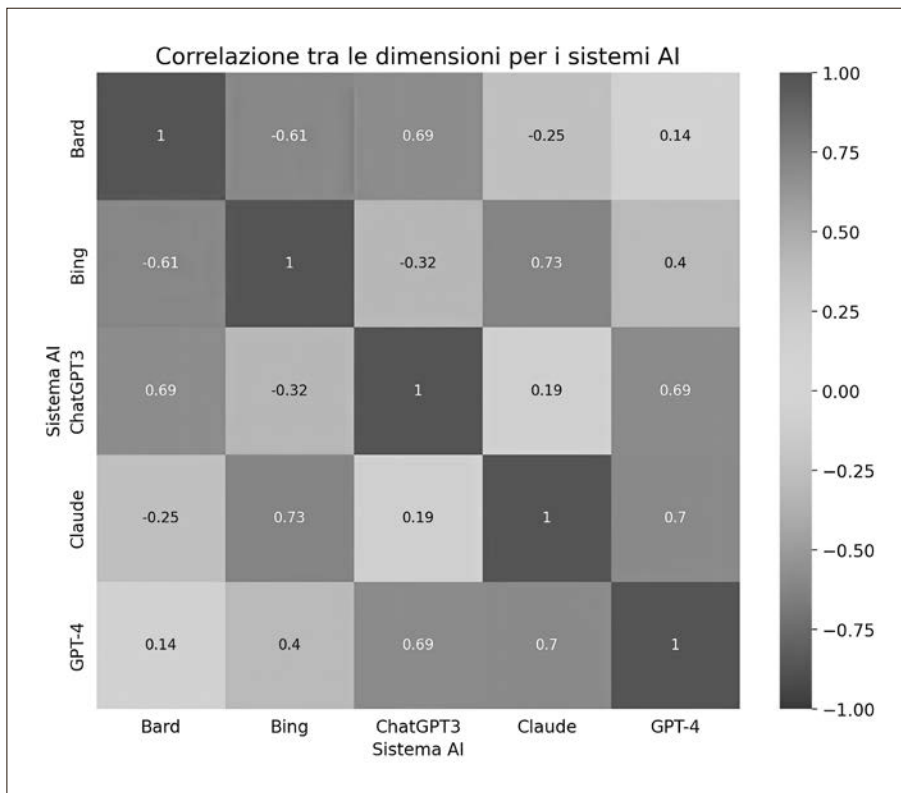


Figura 4. Correlazioni tra le valutazioni medie dei vari sistemi di IA attraverso le diverse dimensioni.

Sebbene sia stata osservata una notevole coerenza tra i revisori nelle loro valutazioni, alcune discrepanze suggeriscono che la percezione della qualità delle risposte può essere soggettiva e influenzata da fattori individuali. Questo solleva la questione dell'importanza di stabilire criteri chiari e oggettivi per la valutazione di tali sistemi in futuro quando dovesse emergere la necessità di monitoraggio e costante valutazione della qualità della performance dei chatbot in area sanitaria.

Principale limite di questo studio è il numero relativamente piccolo di domande e di revisori coinvolti. Future ricerche potrebbero beneficiare di un campione più ampio e di una maggiore varietà di domande cliniche. Inoltre, potrebbe essere utile esplorare le motivazioni dietro le valutazioni dei revisori, attraverso interviste o questionari, per ottenere una comprensione più profonda delle loro percezioni.

Conclusioni

I sistemi sperimentati hanno fornito consigli per quanto riguarda questioni cliniche, relazionali e psicologiche che solo in parte ricevono un supporto di routine da parte del servizio sanitario. Per esempio il tempo del colloquio con l'oncologo per comunicare le decisioni sul programma di cura può trovarsi limitato nella pratica clinica a pochi minuti.

D'altro canto l'evoluzione della relazione medico-paziente, amplificata dalla tecnologia web, ha spostato i pazienti da un ruolo passivo a uno più attivo. Una prospettiva auspicabile è che i sistemi di IA possano sostituire funzioni impegnative collaterali alla pratica clinica e facilitare i curanti a realizzare relazioni con i pazienti, caratterizzate da comunicazione ed empatia.

L'IA rappresenta una risorsa preziosa per i pazienti alla ricerca di informazioni non completamente soddisfatte dal personale sanitario. Tuttavia, l'uso sperimentale di quest'ultima deve essere affrontato con cautela, dato che può produrre risposte inesatte o fuorvianti.

La letteratura scientifica inizia a sottolineare l'importanza crescente dell'IA nel campo dell'informazione sanitaria^{7,8} e i vantaggi offerti da una disponibilità per i pazienti 24 ore su 24 e 7 giorni su 7, oltre alla possibilità di accesso anonimo, di cui possono beneficiare in particolare coloro che hanno difficoltà ad accedere ai servizi sanitari tradizionali.

Tuttavia, bisogna considerare che l'IA, nonostante possa simulare conversazioni empatiche, non può sostituire la genuina empatia e comprensione umana. L'incorporazione dell'IA nella pratica clinica presenta potenzialità ma anche sfide, come la possibilità di fornire informazioni errate che possono danneggiare i pazienti.

Gli strumenti basati sull'IA, come i chatbot, hanno le loro radici in programmi come ELIZA, creato ne-

Take home messages.

- *Crescente rilevanza dell'IA in Sanità:* l'IA ha crescente importanza nel fornire supporto informativo in ambito sanitario, soprattutto in contesti con limitata comunicazione medico-paziente.
- *Limiti e affidabilità della IA:* è necessaria prudenza nell'utilizzo dell'IA a causa dei suoi limiti riguardo ad affidabilità e comprensione del contesto umano.
- *Variabilità delle prestazioni della IA:* le prestazioni dei diversi sistemi di IA sono variabili, ed è quindi importante avere consapevolezza dei loro specifici punti di forza e debolezza.
- *IA complementare nella relazione medico-paziente:* le IA possono avere un ruolo complementare nel supportare la relazione medico-paziente, pur non sostituendo l'empatia umana e il giudizio clinico.
- *Necessità di valutazione continua:* importanti il monitoraggio e la valutazione continua delle prestazioni dei sistemi di IA per assicurare la loro sicurezza ed efficacia in ambito clinico.

gli anni '60 al Massachusetts institute of technology (Mit). Con l'avanzare della tecnologia, l'IA può ora analizzare vari segnali, come le espressioni facciali e il tono della voce, per monitorare lo stato emotivo del paziente.

È fondamentale, però, distinguere i benefici dell'IA per i pazienti e le sfide che presenta per i professionisti sanitari. L'IA offre molte possibilità, ma è essenziale che i medici guidino i pazienti nel suo uso corretto e selettivo, assicurandosi che non sostituisca mai l'importanza del giudizio clinico umano.

Man mano che la tecnologia dell'IA continua a evolversi, è essenziale che la sua integrazione nel campo clinico sia accompagnata da ricerche rigorose e continue valutazioni. Ciò garantirà che questi strumenti possano essere utilizzati in modo sicuro e efficace, migliorando la qualità dell'assistenza e potenziando le capacità dei professionisti sanitari.

Conflitto di interessi: gli autori dichiarano l'assenza di conflitto di interessi.

Bibliografia

1. OpenAI. GPT-4 technical report [Internet]. arXiv; 2023 [citato il 6 giugno 2023]. Disponibile su: <http://arxiv.org/abs/2303.08774> [ultimo accesso 4 gennaio 2024].
2. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med* 2023; 388: 1233-9.
3. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. In: *Advances in neural information processing systems* 33 (NeurIPS 2020).
4. Lim ZW, Pushpanathan K, Yew SME, et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. *EBioMedicine* 2023; 95: 104770.
5. Johnson D, Goodman R, Patrinely J, et al. Assessing the accuracy and reliability of AI-Generated medical responses: an evaluation of the Chat-GPT Model. *Res Sq [Preprint]* 2023 Feb 28: rs.3.rs-2566942.
6. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature* 2023; 620: 172-80.
7. Liu S, McCoy AB, Wright AP, et al. Leveraging large language models for generating responses to patient messages [Preprint]. *medRxiv [Preprint]* 2023 Jul 16: 2023.07.14.23292669.
8. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare (Basel)* 2023; 11: 887.

Indirizzo per la corrispondenza:

Carlo Alfredo Clerici

SC Pediatria

Fondazione Irccs Istituto nazionale dei tumori

Via Giacomo Venezian 1

20133 Milano

E-mail: carlo.clerici@unimi.it