Come sapere se l'innovazione funziona?

Servono studi sperimentali rigorosi che valutino il ricorso all'intelligenza artificiale. Ne abbiamo parlato con una ricercatrice che sta lavorando all'integrazione delle checklist Consort e Spirit

Quali sono gli attuali limiti degli algoritmi di intelligenza artificiale applicati alle cure dei pazienti?

L'intelligenza artificiale (IA) ha una vasta gamma di applicazioni per l'assistenza sanitaria, dall'identificazione del paziente fino alla diagnosi e alla prescrizione del trattamento. Potenzialmente questi algoritmi potrebbero trasformare l'assistenza sanitaria in una miriade di modi diversi (dal fornire una diagnosi più precoce o più accurata al consentire un'erogazione del servizio più rapida ed efficiente e facilitare l'accesso alle cure), tuttavia per ora il limite principale è la scarsità di prove che il ricorso all'intelligenza artificiale faccia più bene che male ai pazienti. Questo è uno dei principali motivi della lenta diffusione delle tecnologie per l'assistenza sanitaria in tutto il mondo.

D'altro canto, ad oggi gli studi di interventi con IA sono per la maggior parte studi di validazione (per esempio studi di accuratezza diagnostica) e inoltre pochi di essi presentano risultati validati esternamente oppure confrontano le prestazioni di un'applicazione di IA con quelle ottenute da operatori sanitari senza un supporto di IA nella stessa popolazione di pazienti[1]. Per tradurre il potenziale dell'IA nella pratica clinica sono quindi necessari studi che valutino gli esiti per i pazienti e i servizi sanitari raggiunti a seguito dell'utilizzo di interventi di IA a confronto con la pratica corrente. Una rendicontazione ottimale di questi studi è fondamentale al fine di garantire che i risultati della ricerca possano essere utilizzati per informare le decisioni politiche e le valutazioni delle tecnologie sanitarie.

Gli studi clinici hanno ancora un ruolo nella valutazione degli interventi sanitari come gli algoritmi di intelligenza artificiale?

Tutti gli interventi sanitari devono essere valutati rigorosamente prima di una loro introduzione nella pratica clinica al fine di dimostrare che il loro impiego comporti più vantaggi che danni per la salute del paziente. Gli studi controllati randomizzati forniscono le evidenze scientifiche di più elevata qualità sull'efficacia degli interventi sanitari e non vediamo perché gli interventi sanitari di IA dovrebbero rappresentare un'eccezione. L'esigenza di un tale livello di valutazione diventa ancora più critica con gli algoritmi "black-box", in cui le conseguenze intenzionali e non intenzionali dell'implementazione possono essere imprevedibili.

Ci sono delle criticità degli algoritmi di intelligenza artificiale che non sono state prese in considerazione nei documenti di indirizzo Consort e Spirit?

Le guidance originali sono state pensate per la valutazione di trattamenti terapeutici (come un farmaco o un intervento chirurgico), pertanto le estensioni degli statement di Consort e Spirit per l'IA sono state concepite per identificare e considerare le sfide ulteriori o nuove nella valutazione degli interventi sanitari con l'IA. Discutendo con tutte le parti interessate, stiamo procedendo a identificare ogni potenziale criticità che è aggiunta. Tuttavia ipotizziamo che gli aspetti che richiederanno report dettagliati e specifici comprenderanno l'impostazione dello studio e la sua capacità di gestire un intervento di machine learning in tempo reale, i criteri per l'inclusione a livello di input-data nonché a livello dei partecipanti, le interazioni tra uomo e algoritmi e di questi ultimi le potenziali ripercussioni a valle, gli effetti delle tecnologie di machine learning adattivo (che hanno il potenziale per migliorare continuamente in performance)[2].

Intervista a

Lavinia Ferrante di Ruffano

Test evaluation research group

Institute of applied health research

University of Birmingham

Il limite principale è la scarsità di prove che il ricorso all'intelligenza artificiale faccia più bene che male ai pazienti.

Come verrà affrontato questo problema dal gruppo direttivo Consort-AI e Spirit-AI?

Il gruppo direttivo ha disegnato un progetto internazionale per lo sviluppo delle estensioni per l'IA nelle checklist e nei documenti di indirizzo di Consort e Spirit già esistenti. Tali estensioni si concentreranno in particolare sugli studi clinici in cui l'intervento sanitario comprende applicazioni di machine learning o altri sottoinsiemi dell'IA. Applicando il quadro metodologico del network Equator (*Enhancing the quality and transparency of health research* – miglioramento della qualità e trasparenza della ricerca sanitaria) per lo sviluppo di linee guida[3], saranno prodotte in quattro fasi: una prima fase per definire cosa deve essere aggiunto, due fasi che prevedono l'interazione tra esperti secondo il metodo Delphi, a seguire un incontro per il consenso finale in cui votare gli item aggiuntivi che hanno ricevuto un maggior numero di segnalazioni. La nostra iniziativa è complementare agli sforzi condotti da altri gruppi che lavorano su standard di reporting, come il Tripod-Ml (*Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis* – reporting trasparente di un modello di previsione multivariabile per la prognosi o la diagnosi individuale) di Collins e Moon che si propone di migliorare il reporting per lo sviluppo e la validazione di modelli predittivi basati sull'apprendimento[4].

Coinvolgerete tutti gli stakeholder interessati nel processo di consenso?

L'integrità dell'output di un progetto di consenso, quale l'estensione all'IA delle checklist Consort e Spirit, è direttamente correlata all'ampiezza delle parti interessate che vi possono contribuire. Il gruppo direttivo Consort-AI e Spirit-AI ha preso seriamente in considerazione come garantire il coinvolgimento dei rappresentanti di tutti i gruppi di portatori di interesse identificati, e di più nazioni, nella identificazione iniziale degli item nonché nelle fasi Delphi e nella riunione finale di consenso. Possiamo confermare che hanno già partecipato o accettato di partecipare i membri dei seguenti gruppi di stakeholder (elencati senza nessun ordine particolare): rappresentanti dei pazienti, responsabili politici (enti governativi, enti medici e istituti di ricerca), enti regolatori, riviste mediche, industria e sviluppatori di sistemi

VEDI ANCHE

Consort e **Spirit** per migliorare il valore delle sperimentazioni cliniche

All'inizio degli anni novanta, due gruppi di direttori di riviste scientifiche, ricercatori e metodologi pubblicarono autonomamente delle raccomandazioni sui metodi per la rendicontazione delle evidenze della ricerca. In un editoriale successivo, Drummond Rennie invitò i due gruppi a incontrarsi per sviluppare una serie di raccomandazioni condivise; il risultato è stato la dichiarazione Consort (*Consolidated standards of reporting trials*)[1]. Comprende una checklist degli elementi essenziali che dovrebbero essere inclusi nei report degli studi controllati randomizzati e una *flow-chart* che illustri il percorso dei partecipanti lungo una sperimentazione. Si rivolgeva ad articoli che riportavano obiettivi, metodi e risultati di studi controllati randomizzati che arruolavano due popolazioni con disegni paralleli. Gran parte delle indicazioni Consort era rilevante anche per una più ampia tipologia di sperimentazioni. Nel corso degli anni sono state pubblicate estensioni della checklist Consort per la rendicontazione delle evidenze derivanti da diversi disegni di ricerca, tra cui quelli per la segnalazione di danni o di risultati di trattamenti non farmacologici come gli interventi di fitoterapia. Dunque la finalità del gruppo Consort è fornire delle guide agli autori per migliorare il reporting delle prove mirando alla chiarezza, alla completezza e alla trasparenza.

Parallelamente, il gruppo Spirit (*Standard protocol items: recommendations for interventional trials*) lavora invece al miglioramento dei protocolli di ricerca[2]. Ha prodotto una checklist di 33 item per valutare i protocolli di qualsiasi tipo di studio clinico, concentrandosi sulla sostanza piuttosto che sulla forma. La checklist raccomanda la descrizione completa di ciò che è pianificato che avvenga nel corso della sperimentazione includendo nei protocolli le informazioni necessarie per la valutazione critica e l'interpretazione del trial. L'idea di fondo non è quella di prescrivere come progettare o condurre uno studio ma piuttosto quella di facilitare l'elaborazione di protocolli di alta qualità. Spirit incentiva la trasparenza e la completezza dei protocolli a beneficio dei ricercatori, dei partecipanti alla sperimentazione, di pazienti, sponsor, finanziatori, comitati etici di ricerca o commissioni istituzionali di valutazione, riviste, registri di sperimentazione, decisori sanitari, autorità regolatorie. •

1. www.consort-statement.org
2. www.spirit-statement.org

di intelligenza artificiale, metodologi, statistici, sperimentatori, gruppi di standardizzazione dell'IA, clinici di diverse specialità, istituti che fanno ricerca sull'IA applicata alla salute, scienziati computazionali, ricercatori nel campo del machine learning, specialisti in informatica clinica/sanitaria, studiosi di etica e organismi che si occupano di finanziamento della ricerca.

Perché considera così importante il ruolo dei direttori delle riviste mediche?

Qualsiasi documento orientativo avrà successo solo se è visibile e facilmente applicabile a tutte le valutazioni pertinenti. Le riviste mediche, rappresentate dai loro direttori, svolgono quindi un ruolo cruciale nel successo delle linee guida relative ai metodi e al reporting. Raggiungono questo obiettivo in due modi. Primo, partecipando alla produzione e discussione di nuovi item di Consort e Spirit, i direttori delle riviste ci consentono di tenere conto del punto di vista esclusivo di chi è abituato a vedere in prospettiva la portata delle ricerche sull'IA sottoposte alle riviste e quelle pubblicate, e in più ha una lunga esperienza nell'implementare le attuali istruzioni per gli autori. Secondo, le riviste mediche svolgono un ruolo sostanziale nel diffondere e nel far conoscere linee guida e checklist di reporting, garantendone così la visibilità tra gli autori di tutto il mondo e allo stesso tempo l'utilizzo da parte dei ricercatori che hanno sottoposto il proprio articolo alla rivista.

Gli studi randomizzati controllati sono il gold standard della ricerca clinica: non c'è ragione perché l'intelligenza artificiale rappresenti un'eccezione nel ricorso ai trial sperimentali.

Prevede che gli orientamenti di Consort-AI e Spirit-AI possano avere un impatto sul processo normativo della Food and drug administration?

In quanto stakeholder nella valutazione degli interventi sanitari, stiamo collaborando con diverse agenzie regolatorie internazionali nell'ambito del processo di consenso per la produzione delle checklist Consort-AI e Spirit-AI. Tuttavia non rientra nei fini del nostro progetto modificare o influenzare i processi normativi attuali. Al contrario, l'obiettivo principale è quello di migliorare la rendicontazione e la progettazione delle sperimentazioni condotte per valutare l'efficacia degli interventi sanitari con sistemi di IA, affinché il regolatorio e gli organismi di valutazione delle tecnologie sanitarie abbiano accesso a una base di prove di qualità sufficiente a facilitare l'introduzione di sistemi di IA efficaci nell'assistenza sanitaria.

1. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digital Health 2019;1:e271-97.
2. Consort-AI and Spirit-AI steering group. Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. Nat Med 2019;25:1467-8.
3. EQUATOR Network. Reporting guidelines under development.
4. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. Lancet 2019;393:1577-9.

Come dovremmo **leggere uno studio sul machine learning?**

Nella famosa serie delle Users' guides to the medical literature del *JAMA* è stato pubblicato il 12 novembre 2019 un articolo che propone in dettaglio il metodo per leggere criticamente una ricerca sul machine learning. Il riferimento di fondo degli autori è un'altra guida della stessa serie uscita quasi in apertura dell'avventura della evidence-based medicine[2]. I primi requisiti di qualità discussi nell'articolo riguardano i metodi statistici. Procedendo oltre, gli autori considerano un elemento essenziale – "I risultati di uno studio sono troppo belli per essere veri?" – arrivando alla conclusione che "i metodi di machine learning non dovrebbero essere in grado di superare le prestazioni di medici estremamente attenti ed esperti ai quali è stato concesso il tempo sufficiente per prendere una decisione".

Altro punto importante è quello della ripetibilità e della riproducibilità, aspetti critici della misurazione della coerenza delle prestazioni del modello di apprendimento automatico: "quando viene fornita (al sistema) la stessa immagine per due volte, i risultati di un determinato modello di machine learning dovrebbero essere identici". Ancora: "Come un test diagnostico può essere utilizzato (in linea di principio) per scopi di triaging, screening o diagnostica, un modello di machine learning sviluppato per eseguire un compito specifico deve poter essere utilizzato per diversi scopi. Per esempio, in un'applicazione diagnostica, l'apprendimento automatico può essere utile in tre fasi distinte: prediagnosi, peridiagnosi e postdiagnosi".

L'implementazione di modelli di machine learning non prevede particolari problemi dal punto di vista della tecnologia informatica, ma alcune peculiarità possono influire sulla riservatezza dei dati o sull'integrazione degli stessi con altri dati del malato archiviati nel sistema ospedaliero e, pertanto, questi aspetti vanno tenuti in debito conto. Sempre ai fini del trasferimento dei risultati di uno studio alla pratica clinica, è necessario che siano attentamente considerati l'efficacia clinica a partire dagli esiti sulla malattia e sul malato nonché i costi. Inoltre, va considerato il possibile aggravio sul carico di lavoro del personale sanitario (anche in termini di verifica attenta di possibili falsi positivi) e le conseguenze che possono derivare dall'eccessiva fiducia degli operatori nella risposta della tecnologia: per questo gli autori raccomandano che l'efficacia sia valutata da ampie sperimentazioni controllate randomizzate.

Va considerato anche che il machine learning permette l'aggiornamento costante del software così che in prospettiva ci può essere una ragionevole certezza di un miglioramento di qualità dei prodotti: un aspetto, questo, di cui dovrebbero tener conto non solo chi consulta un articolo su progetti di machine learning ma anche le autorità regolatorie.

1. Liu Y, Chen PC, Krause J, Peng L. How to read articles that use machine learning: Users' guides to the medical literature. JAMA 2019;322:1806-16.
2. Jaeschke R, Guyatt GH, Sackett DL. Evidence-based medicine working group. Users' guides to the medical literature, III: how to use an article about a diagnostic test, B: what are the results and will they help me in caring for my patients? JAMA 1994;271:703-7.