Intelligenza artificiale in ospedale: i cinque rischi per privacy e sicurezza dei dati che corrono i pazienti di Ruggiero Corcella

Una revisione sistematica individua le principali minacce dell'Ai sanitaria. Alcuni studiosi prevedono sistemi autonomi superiori ai medici entro il 2030. FDA ed esperti di sicurezza cercano nuove regole per tecnologie che cambiano anche dopo essere entrate in ospedale
(Fonte: https://www.corriere.it/ 22 agosto 2026)

Rendere anonimi i dati sanitari potrebbe non bastare più. **Un'intelligenza artificiale può riconoscere un paziente** partendo da un elettrocardiogramma, da una radiografia del torace o da una scansione della retina. **Può stabilire se i dati di una determinata persona siano stati utilizzati per addestrare un modello**. E un aggressore può alterare impercettibilmente un'immagine medica fino a indurre un algoritmo diagnostico a sbagliare.
Sono alcuni dei rischi che emergono dalla revisione sistematica di Elvin Khanjahani e colleghi, appena accettata dall'*International Journal of Medical Informatics*. Gli autori hanno passato al setaccio **7.285 record** provenienti da sei database scientifici, più 205 individuati attraverso le citazioni, **selezionando alla fine 22 studi empirici**. La maggioranza riguarda l'imaging medico, nove studi su 22.

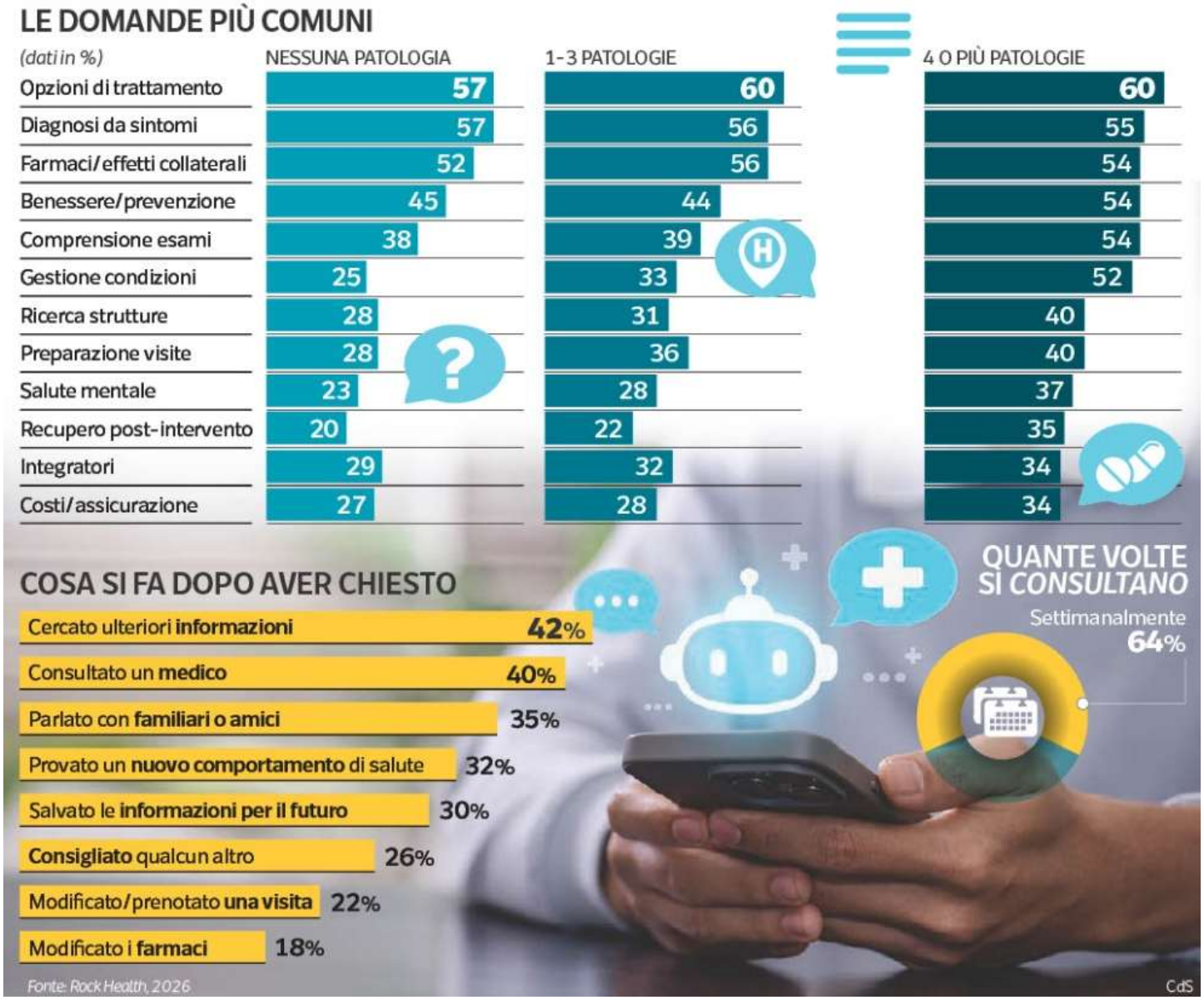
Le principali minacce

Dall'analisi emergono cinque categorie di minacce: **re-identificazione dei pazienti, membership inference (che permettono di scoprire se i dati di una determinata persona sono stati utilizzati per addestrare un sistema di intelligenza artificiale), accesso non autorizzato e attacchi ostili**

dei sistemi, alterazione dei dati forniti all'Ai, uso improprio o sovrainterpretazione dell'Ai.
La più studiata è la re-identificazione: **12 dei 22 studi**. Altri tre analizzano i membership inference attacks, tre l'accesso non autorizzato o gli attacchi avversariali, tre la manipolazione degli input e due l'uso improprio, la sovrainterpretazione o una progettazione inadeguata dei modelli. Alcuni studi ricadono in più di una categoria.

Il paziente nascosto dentro i dati
Il primo risultato mette in discussione uno dei presupposti tradizionali della protezione dei dati sanitari: eliminare nome, cognome e altri identificatori potrebbe non rendere davvero anonimo un dataset. Gli algoritmi possono infatti individuare **segnali biometrici latenti** nei dati clinici. I dodici studi sulla re-identificazione spaziano dalla Cardiologia alla Radiologia, dall'Oftalmologia alla Patologia digitale e utilizzano ECG, risonanze magnetiche, Tac, radiografie, immagini del **fondo oculare** e OCT (Tomografia a Coerenza Ottica) e «vetrini» istopatologici.
Gli esempi rendono concreto il problema. **Sistemi deep learning sono riusciti a riconoscere persone attraverso il loro ECG.** Le risonanze della testa possono permettere di ricostruire caratteristiche del volto e confrontarle con fotografie pubblicamente disponibili, anche se tecniche di *de-facing* riducono sensibilmente il rischio. Metodi di riconoscimento facciale e deep metric learning hanno identificato persone **anche partendo da immagini tridimensionali derivate da Tac,** persino confrontando esami realizzati in istituzioni diverse e con apparecchiature di produttori differenti. E nelle immagini istopatologiche uno studio è arrivato fino all'**80% di accuratezza nell'identificazione del paziente di origine**.
Il fenomeno non scompare con i modelli più recenti. La revisione riporta che immagini del fondo oculare, OCT e radiografie toraciche analizzate attraverso *foundation model* (modelli-base, cioè grandi sistemi di Ai addestrati su enormi quantità di dati e adattabili a molti compiti diversi) hanno prodotto tassi di re-identificazione del **26-46%**, mentre modelli specifici per determinati compiti sono arrivati fino al **94% nelle immagini OCT**. Un altro studio ha ottenuto una re-identificazione quasi perfetta da radiografie toraciche anonimizzate, anche confrontando immagini raccolte a più di dieci anni di distanza.

Che cosa chiediamo al «dottor chatbot»



Anche sapere che eri nel database è una violazione

La seconda minaccia è più difficile da percepire, ma non necessariamente meno delicata. Con un **membership inference attack** l'attaccante non deve ricostruire nome e cognome direttamente dal dato: cerca di scoprire se i dati di una persona facevano parte del dataset, utilizzato per addestrare l'Ai. In medicina, la semplice appartenenza a un database può già rivelare informazioni sensibili, per esempio **la partecipazione a uno studio clinico o la presenza di una determinata condizione.**

Tre studi della revisione documentano questo rischio in Genomica, Neurologia e dataset sanitari. In uno, modelli addestrati su dati genomici ad alta dimensionalità sono risultati vulnerabili; una tecnica chiamata «**privacy differenziale**», che introduce alterazioni controllate nei dati per rendere più difficile ricavare informazioni sul singolo paziente, ha ridotto le prestazioni del modello senza eliminare adeguatamente l'attacco, mentre una tecnica per rendere i dati meno identificabili ha ridotto l'efficacia dell'attacco fino a portarla al livello del caso (AUC da **0,71 a 0,50**).

Ancora più significativo il risultato sui **dati sanitari parzialmente sintetici**: gli aggressori hanno

raggiunto una precisione superiore a **0,9 per una quota fino al 44% dei pazienti** nel dataset *All of Us*, mentre i dati completamente sintetici si sono dimostrati molto più resistenti. Anche sistemi brain-computer interface basati su EEG sono risultati vulnerabili.

Quando basta cambiare pochi pixel

Privacy e cybersecurity si intrecciano poi con la sicurezza clinica. Gli studi esaminati mostrano che attacchi di *data poisoning* (alterazione intenzionale dei dati usati per addestrare l'Ai) possono provocare classificazioni miratamente errate in diversi algoritmi e dataset sanitari. In un esperimento sui sistemi per diagnosticare Covid-19 dalle radiografie del torace, perturbazioni universali pari appena all'**1-2% della norma dell'immagine** hanno fatto classificare erroneamente oltre l'85% delle immagini negli attacchi non mirati; quelli mirati hanno ottenuto risultati prossimi al successo totale.

Un'altra categoria riguarda la **manipolazione dei dati in ingresso,** senza dover violare l'infrastruttura informatica. In un modello che prevedeva il parto in struttura sanitaria tra le donne di Zanzibar, modificare una sola variabile categoriale — il luogo del parto precedente — cambiava la classificazione nel **38-56% dei casi**. In modelli per riconoscere Covid-19 da radiografie e Tac, alterazioni impercettibili hanno fatto precipitare l'accuratezza, fino al **16% per VGG16**. Ancora più immediato l'esperimento con applicazioni mobili per individuare il melanoma: adesivi trasparenti con pattern di punti applicati alla fotocamera dello smartphone hanno provocato classificazioni errate nel **50-85% dei casi**, senza che la manipolazione risultasse evidente all'utente.

Gli autori invitano però a non trasformare questi risultati nella prova di un'emergenza già diffusa negli ospedali. Gran parte dei 22 studi è stata condotta in ambienti sperimentali, simulazioni o benchmark; la validazione esterna e la replica indipendente sono poco frequenti. Le vulnerabilità sono quindi **empiricamente dimostrate**, ma questo non significa che gli attacchi descritti siano già comuni nei sistemi clinici operativi.

Il paradosso: un'Ai più autonoma, mentre emergono nuovi rischi

Il tema diventa ancora più rilevante perché una parte della ricerca spinge nella direzione di una maggiore autonomia delle macchine. In una «perspective» pubblicata su *JAMA*, Ezekiel Emanuel, Abe Baker-Butler, Neal Khosla e Vinod Khosla sostengono una tesi provocatoria: nelle attività mediche cognitive**, entro il 2030 l'Ai da sola potrebbe diventare migliore non soltanto del medico, ma anche della coppia medico-Ai**. Una puntualizzazione appare però doverosa: **qui l'«hype» sull'Ai ha anche un interesse specifico** (ma almeno, dichiarato). Neal Khosla è infatti CEO di Curai Health, un'azienda di assistenza sanitaria primaria basata sull'intelligenza artificiale. Suo padre, Vinod Khosla, è un miliardario del settore del venture capital che investe in tecnologie sperimentali come la biomedicina e la robotica.

Gli autori individuano **cinque funzioni fondamentali** nelle quali i sistemi generativi già competono con i professionisti: raccolta delle informazioni cliniche, diagnosi differenziale, scelta degli esami,

prescrizione di trattamenti conformi alle linee guida e gestione delle malattie croniche. Citano, fra gli altri, uno studio nel quale ChatGPT o3 ha collocato al primo posto la diagnosi finale nel **60% di 377 casi complessi reali**, contro il 15,9% ottenuto da 20 internisti su un sottoinsieme di 302 casi. In un altro confronto, un sistema di Ai che coordina le diverse fasi del percorso diagnostico, scegliendo anche quali esami richiedere, ha raggiunto la diagnosi corretta 4,02 volte più frequentemente dei medici, con costi inferiori del 19,1%.
È però appunto un articolo di «orizzonte», non una revisione sistematica equivalente a quella sui rischi, e **gli stessi autori riconoscono limiti rilevanti:** molti confronti sono simulazioni di singoli compiti cognitivi e non interazioni cliniche reali; esistono problemi di fragilità dei modelli, **allucinazioni, cyberattacchi e perdita di connessione**; molte attività mediche richiedono inoltre procedure fisiche che l'Ai non può svolgere autonomamente. Nonostante ciò, Emanuel e colleghi ritengono probabile che **entro i prossimi quattro anni sistemi autonomi superiori** possano essere pronti per alcuni, forse molti, workflow cognitivi reali.

L'American Medical Association prende posizione

L'American Medical Association (AMA) ha pubblicato un nuovo Quadro di riferimento sul ruolo dei medici nell'era dell'intelligenza artificiale, fornito in anteprima ad Axios, che ribadisce come i medici debbano rimanere centrali nell'assistenza ai pazienti.
Secondo quanto riportato da Axios, il Quadro di riferimento, pubblicato in collaborazione con la Digital Medicine Society, sostiene che, con l'evoluzione delle tecnologie, «anche la pratica medica deve evolversi di pari passo».

- «I singoli compiti si evolveranno con il progresso tecnologico, ma le responsabilità fondamentali del medico restano sostanzialmente invariate», aggiunge l'introduzione.
- Tra questi rientrano il contatto umano, il giudizio clinico e il ruolo di promotori di un uso responsabile della tecnologia.

«Allo stato attuale, i medici devono essere coinvolti. Stiamo parlando dell'erogazione dell'assistenza sanitaria alle persone», ha dichiarato ad Axios John Whyte, CEO e vicepresidente esecutivo dell'AMA. «Vorresti davvero andare al Pronto Soccorso ed essere curato da un medico senza licenza se hai dolore al petto? Non credo proprio».

Il medico può diventare parte del problema?

Per restare nel tema della sicurezza, in un altro lavoro pubblicato su Diagnosis diciotto esperti di tre Paesi, coordinati da Emily Patterson e Michael Bruno, hanno individuato sei grandi lacune di conoscenza da affrontare nei prossimi cinque-sette anni per usare l'Ai in sicurezza. Circa il **75% dei dispositivi Ai è impiegato nella diagnosi clinica in Radiologia**, osservano gli autori, ma questi strumenti vengono inseriti in processi di lavoro spesso già frammentati, distribuiti tra diversi team e fortemente dipendenti dalla comunicazione.
Il gruppo indica tra le questioni aperte proprio il rapporto uomo-macchina: **come evitare**

il *deskilling* (la perdita progressiva di competenze) dei sanitari, **rendere comprensibili le conclusioni dell'Ai e la loro incertezza**, calibrare la fiducia senza produrre *over-trust* (eccesso di fiducia nelle indicazioni dell'Ai). Un modello multimodale citato nel lavoro riconosceva correttamente la modalità di imaging in tutti i casi, ma presentava un tasso di **allucinazioni diagnostiche superiore al 40%**. L'Ai, concludono gli esperti, non migliorerà sicurezza e qualità «di default»: tutto dipenderà da come verrà integrata nei sistemi sociotecnici della medicina.
Emanuel e colleghi arrivano quasi al problema opposto. Citando una metanalisi di 106 esperimenti, sostengono che quando l'Ai è già superiore all'uomo, **il sistema ibrido può avere risultati peggiori dell'Ai sola**; una revisione del 2025, su 52 studi clinici, avrebbe analogamente trovato nei sistemi medico-Ai un'affidabilità inferiore a quella del migliore tra uomo e macchina. Secondo la «perspective», **il medico può introdurre errori** quando corregge erroneamente una risposta corretta dell'algoritmo, anche a causa della cosiddetta *algorithm aversion* (diffidenza verso gli algoritmi).

Le regole devono seguire un algoritmo «mutante»
È anche la questione che sta arrivando sui tavoli dei regolatori. Negli Stati Uniti, osserva il panel di Patterson, negli ultimi cinque anni **17 Stati hanno approvato 29 leggi sull'Ai**, concentrate soprattutto su privacy, discriminazione e accountability. Gli esperti prevedono tuttavia che per l'Ai sanitaria il principale centro della regolazione resterà federale, sotto la FDA. Restano aperte domande sul **consenso informato** all'utilizzo e alla **vendita dei dati**, sulla possibilità per il paziente di scegliere un'alternativa «human-only», sulla **responsabilità di sviluppatori e strutture sanitarie** e sulla **ripartizione dei costi legali,** quando un sistema sbaglia.
La FDA sta intanto valutando un cambio di paradigma per i dispositivi medici basati su AI generativa. In un documento anticipato da *Axios*, l'agenzia ipotizza una valutazione **«competency-based»**, ispirata ad alto livello a quella utilizzata per verificare e accreditare le competenze dei clinici: definizione di benchmark e conferma delle prestazioni nell'ambiente clinico. La ragione è che **un dispositivo GenAI può fornire output variabili ed evolvere nel tempo**, caratteristiche difficili da incasellare nella regolazione tradizionale dei dispositivi. La FDA valuta persino se accettare una maggiore incertezza prima dell'immissione sul mercato, compensandola con una sorveglianza post-market più intensa.
Sul versante europeo, la revisione di Khanjahani richiama esplicitamente **GDPR e Ai Act**. Il GDPR riconosce i limiti di anonimizzazione e pseudonimizzazione quando la re-identificazione resta ragionevolmente possibile; l'Ai Act punta invece, per i sistemi ad alto rischio, su gestione del rischio lungo il ciclo di vita, trasparenza, documentazione e monitoraggio continuo. Secondo gli autori, le evidenze raccolte indicano che la governance non può fermarsi alla protezione del database: **anche modelli addestrati, rappresentazioni delle caratteristiche e dataset derivati devono essere considerati asset sensibili**, sottoposti a controllo degli accessi e monitoraggio.
Che fare, allora? Gli autori della revisione su 22 studi propongono una difesa a strati: **privacy e**

security by design, pipeline sicure, controllo degli accessi, test periodici, monitoraggio continuo, procedure di risposta agli incidenti e supervisione lungo l'intero ciclo di vita dell'Ai.