Dalla letteratura

Dati sintetici per accelerare la ricerca clinica

Roma, 8 ottobre 2024, Sala Convegni AIL

Le pagine del numero di gennaio della rubrica "Dalla letteratura" di *Recenti Progressi in Medicina* sono dedicate a un rapido resoconto degli interventi dei relatori che, da punti di vista diversi, hanno approfondito il tema dei dati sintetici nella ricerca clinica in un evento organizzato lo scorso ottobre dal progetto Forward insieme alla Fondazione Gimema. Se da un lato i dati sintetici, che mimano le caratteristiche statistiche dei dati reali senza contenere informazioni sensibili, sembrano poter accelerare e in qualche modo migliorare il processo decisionale della ricerca clinica, la loro adozione solleva nuovi e numerosi interrogativi dal punto di vista scientifico e regolatorio, come il pericolo di generare dati affetti da bias, il rispetto della proprietà intellettuale dei dati di addestramento, il nascere di nuovi e in parte inattesi problemi di privacy.
www.forward.recentiprogressi.it

Shalini Kurapati

Il punto di vista della data scientist

Cosa sono i dati sintetici? Come si generano? Perché vengono generati in un modo piuttosto che in un altro? Shalini Kurapati, co-fondatrice e Ceo di Clearbox AI, si sofferma sui fondamentali, per offrire, con il suo contributo, una panoramica il più generale possibile dei dati sintetici. Di fatto i dati sintetici sono dati fittizi, generati artificialmente tramite diverse metodologie, con l'obiettivo di rappresentare una cosa vera dal mondo reale. Secondo alcune previsioni, entro il 2030 quasi tutti i dati utilizzati dall'intelligenza artificiale come big data saranno generati in modo sintetico e le ragioni sono ovvie, sottolinea Kurapati: costi minimi, vantaggio dal punto di vista della privacy, possibilità di sfruttarli per coprire i dati mancanti in caso di scarsità o di sbilanciamento.

Se generiamo da zero simulazioni e modelli, facciamo un uso dei dati sintetici secondo un approccio diverso da quello che prevede la partenza da dati reali. Si tratta di metodologie che si rivelano utili, ad esempio, nella renderizzazione con motori di gaming o *unreal engine*, permettendo di realizzare elementi virtuali. Esiste anche una metodologia di questo tipo, l'*agent based modeling*, che consente di simulare un comportamento in modo controllato. In questo modo, spiega Kurapati, è possibile creare una simulazione di come si propaga una pandemia o di come si comportano le persone in una città durante tutte le loro transazioni.

Shalini Kurapati.

Se passiamo all'altro approccio, «i modelli che possiamo fare a partire da dati reali sono molti di più», prosegue Kurapati. Questi modelli generativi che sfruttano il dato originale riescono a rilevare le proprietà statistiche del dataset originale e ne ricreano uno sintetico, simile ma non uguale. Esistono diversi modelli generativi di questo tipo, ciascuno con pregi e difetti. Kurapati cita i quattro più usati nel settore dei dati sintetici. Uno di questi è il *generative adversial network*. È semplice e molto diffuso, perché di fatto ci sono due modelli che competono tra loro. Uno è il generatore, l'altro è il discriminatore. Quando vengono forniti tantissimi dati, il generatore e il discriminatore devono convergere dove il discriminatore non riesce più a riconoscere se un'immagine è vera o fittizia. Quando si giunge a questa convergenza, possono essere creati dati sintetici. Il problema è che la convergenza tra discriminatore e generatore non è sempre facile e anche il training è molto instabile. Altro modello generativo di cui parla Kurapati è il *variation auto encoder*. Si tratta di un modello che apprende dal dataset originale e mappa in uno spazio latente la distribuzione normalmente gaussiana, che poi ricostruisce in un dataset sintetico. Questi encoder sono molto stabili e controllabili, consentono di generare dati in modo estremamente snello. Utili per le serie temporali, non ottengono buoni risultati con le immagini. Tutti i modelli di intelligenza artificiale che generano immagini (Midjourney, Dall-e, ecc.) sono basati invece sui *diffusion models*. Attraverso i *diffusion models* si generano dati partendo da quelli originali (un'immagine di un certo soggetto, ad esempio), aggiungendo "rumore", per poi ricostruire un'immagine sintetica. Si tratta, com'è evidente, dello stato dell'arte proprio per la creazione di immagini.

La domanda è "perché non generiamo dati sintetici con i *large language models*?", chiede al pubblico Kurapati. Innanzitutto questi sistemi (il più noto dei quali è ChatGPT) operano tramite transformer: milioni/miliardi di parametri che trasformano un input in un output, usando sempre il metodo probabilistico. Funzionano molto bene per la generazione dei testi, ma non sono affatto adatti per le serie temporali e altri tipi di dati.

La scelta di metodo di generazione non è banale perché dipende dal contesto, dal tipo di dato e dalla complessità. In ambito clinico, ad esempio, abbiamo diverse multimodalità e l'idea è quella di rigenerare le stesse proprietà anche mantenendo relazioni tra dataset. Capire quale può essere il metodo più adatto si rivela quindi un'operazione molto complicata che richiede sia capacità tecniche sia expertise di dominio. La scelta, continua

Kurapati, dipende anche dallo scopo. I dati sintetici possono aiutare a migliorare i modelli di machine learning specialmente in presenza di problemi molto sbilanciati come la predizione di malattie rare, il rilevamento di anomalie nei dati clinici e la previsione degli esiti. Si possono, ad esempio, creare dataset utili alla sperimentazione (seguendo lo schema *what if…*).

Tornando a un tema soltanto accennato all'inizio, se i dati sintetici sono utilizzati per superare problemi di privacy (uno dei principali motivi del loro utilizzo), allora dobbiamo essere ragionevolmente sicuri di stare valutando i dataset in modo robusto. Esiste infatti il rischio che possa spuntare fuori qualcosa di molto più simile al dato originale rispetto a quanto atteso, cioè un output virtualmente uguale al training dataset. Esistono fortunatamente diverse tecniche per evitare risultati del genere, per esempio la simulazione di veri e propri attacchi alla privacy, i cosiddetti *shadow modeling-based membership inference attacks*. Nel tentativo di garantire la privacy non bisogna comunque perdere di vista il criterio dell'utilità: se si finisce per togliere tutte le informazioni nel dataset sintetico, poi lo si rende di fatto inutile. È necessario ricorrere a un sistema in grado di bilanciare privacy e utilità, conclude Kurapati.

A cura di Alessio Malta
Il Pensiero Scientifico Editore

Guido Scorza
Il punto di vista giuridico

«Quando si dice che la privacy è un ostacolo per la ricerca medica siamo di fronte a un paradosso». Inizia così l'intervento di Guido Scorza, componente Garante per la protezione dei dati personali (Gdpr). Se ci rifacciamo alla definizione di salute dell'Organizzazione mondiale della sanità – che la identifica come uno stato completo di benessere fisico, mentale e sociale –, il benessere è legato a doppio filo alla dignità della persona, e la dignità della persona è esattamente il bene della vita tutelato dal diritto alla privacy. Secondo questa definizione, non è possibile garantire la salute sacrificando la privacy o la ricerca medico-scientifica. Il rapporto tra le due non è di rivalità: la "parola magica" è, infatti, bilanciamento. Nella Carta dei diritti fondamentali dell'Unione europea, all'articolo 54, troviamo l'algoritmo di bilanciamento, nel quale si afferma che, in un contesto di rivalità tra due diritti fondamentali, quali salute e privacy appunto, l'unica regola di riferimento prevede di comprimere un diritto nella misura minima necessaria a garantire l'esercizio, l'esistenza dell'altro.

Giudo Scorza.

Il Gdpr è sempre citato come Regolamento generale sulla privacy in Europa, dimenticando che contiene invece qualcosa in più: non solo protezione dei dati, ma libera circolazione degli stessi. Le parole sono importanti: il legislatore europeo ha promosso lo stesso valore per protezione e circolazione. Non solo: la libera circolazione dei dati personali è proprio l'obiettivo, raggiungibile con il mezzo di una disciplina uniforme, come abbiamo constatato durante la pandemia, quando solo grazie a un'unica disciplina europea siamo riusciti a ottenere un passaporto sanitario comune in due settimane. Proseguendo la lettura del Gdpr, nel Considerando 4 troviamo scritto: «Il trattamento dei dati personali dovrebbe essere al servizio dell'uomo. Il diritto alla protezione dei dati di carattere personale non è una prerogativa assoluta, ma va considerato alla luce della sua funzione sociale e va contemperato con altri diritti fondamentali in ossequio al principio di proporzionalità». Questa indicazione, spesso soggetta a cattive interpretazioni, è il contesto dei dati sintetici, tra bilanciamento e compromesso. Ricordando, inoltre, che oggi il dato artificiale è figlio di quello personale, intimo, la disciplina rimane quella del trattamento. L'operazione è la stessa, così come le regole, nel percorso di generazione del dato sintetico: partendo dal dato personale, che sia sanitario o no, occorre rispettare tutte le regole caratteristiche della disciplina sulla privacy. E far sì che, una volta arrivati a destinazione, nessun dato personale sia rintracciabile.

Eppure, sottolinea Scorza, in partenza i dati personali esistono e necessitano di una base giuridica, del consenso, di una legge che permetta di operare, cioè di un presupposto giuridico che legittimi l'operazione del trattamento nel contesto della generazione dei dati di sintesi. Per raggiungere questo risultato non sono sufficienti le basi giuridiche attualmente sul tavolo. Nel Parlamento italiano c'è una questione aperta intorno al disegno di legge sull'intelligenza artificiale. Tra gli emendamenti proposti, ce n'è uno particolarmente prezioso che muove dal principio per il quale la ricerca medico-scientifica è un interesse pubblico. In nome di questo principio, il legislatore ha la possibilità di "far firmare la pace" tra quanti si occupano di privacy e mondo della ricerca, stabilendo che il processo di sintesi possa essere svolto nell'universo della ricerca medico-scientifica in quanto risponde a un interesse pubblico. «Dobbiamo, insomma, liberare il processo di sintesi», conclude Scorza.

Prima di chiudere, l'avvocato si concentra su una seconda questione: non tutti i dati sintetici sono uguali. Dunque, il processo di generazione deve essere governato da un approccio multidisciplinare, non esclusivamente orientato alla ricerca medico-scientifica. In poche parole: l'utilità perseguita dal ricercatore non può trascurare la privacy *by design* e *by default*. Se il processo di sintesi alleggerisce il potere informativo dei dati, comprendiamo come i ricercatori siano poco disponibili a cedere potenzialità identificativa, con il rischio che il dato sintetico continui a essere riconducibile a una persona. L'unica indicazione utile è costruire un processo di sintesi applicando il Gdpr e i suoi due principi della privacy *by default* e *by design*, con l'obiettivo dell'utilità per la ricerca ma anche rispettando l'anonimizzazione assoluta. Così, si opera al di fuori dell'ambito di applicazione del Gdpr. In caso contrario, gli investimenti, di tempo e di risorse economiche, sarebbero sprecati.

Insomma, se la privacy è salute, in questo percorso verso i dati sintetici il Garante è un interlocutore, un alleato.

A cura di Maria Frega
giornalista freelance

Matteo Della Porta
Il punto di vista del medico

Pensando al più grande ostacolo all'innovazione in ematologia, la risposta sembra banale: abbiamo bisogno di più dati per accelerare la ricerca traslazionale clinica e migliorare la cura e l'assistenza per i pazienti. È questa la premessa all'intervento di Matteo

Della Porta, ematologo all'Humanitas di Milano. «La stragrande maggioranza dei dati che noi produciamo durante l'attività clinica – spiega – non è utilizzabile perché la loro raccolta non è uniforme né strutturata, e per il rispetto della privacy».

«L'istituzione nella quale lavoro tre anni fa ha creato, all'interno dell'ospedale, un centro di intelligenza artificiale – l'Humanitas AI Center – a disposizione dei ricercatori, per utilizzare i dati prodotti nell'attività clinica, e ci siamo innamorati di questa tecnologia così utile dal punto di vista scientifico. Da clinico, invece, mi chiedo: se abbiamo già il processo di anonimizzazione classica del dato, perché cambiare paradigma?». L'Organizzazione mondiale della sanità, prosegue, ha indicato i requirement minimi per gli strumenti che utilizzano l'intelligenza artificiale nella pratica clinica[1]. Devono essere trasparenti; comprensibili a medico e paziente; validati, come i farmaci; rispettosi della privacy.

In questa direzione, all'Humanitas hanno creato *Train*, una spin-out che, senza voler competere con i *large language models*, sviluppa soluzioni di intelligenza artificiale validate su un'ampia base di dati clinici di settore. Come caso studio[2,3] è stata scelta la sindrome mielodisplastica, poiché particolarmente peculiare: è una malattia rara, quindi ci sono pochi dati; è estremamente complicata dal punto di vista clinico; ha una base genetica e biologica complessa. Dal punto di vista tecnologico, è dunque il migliore stress test per capire se i dati sintetici riescono a estrarre un valore clinico sufficiente per utilizzi scientifici. Da un repository pubblico di duemila pazienti con questa malattia rara hanno generato uno stesso numero di dati sintetici per poi valutarne il valore clinico traslabile, mantenendo le relazioni complesse di interazione tra dati dello stesso paziente: la rappresentazione delle co-occorrenze e mutue esclusività di lesioni geniche, cromosomiche e mutazioni della popolazione reale sono catturate con esattezza. E, infine, hanno creato un generatore di dati sintetici, pubblico, per dimostrare la duttilità di questa tecnologia: dai duemila pazienti del repository si possono generare fino a diecimila pazienti sintetici con 250 diverse caratteristiche cliniche, genomiche, di trattamento, sopravvivenza. Ogni dato sintetico, inoltre, è misurato in un framework di validazione che fornisce informazioni su performance e fitness in ambito clinico, genomico e di privacy.

Come utilizzare questi dati sintetici? «Da clinici – ha continuato Della Porta – la priorità è accelerare lo sviluppo dei trial che mediamente durano dai quindici ai venti anni. Abbiamo testato la possibilità sullo studio accademico della Fondazione italiana sindromi mielodisplastiche (Fisim) riproducendo i pazienti sintetici per verificare la riproducibilità dei valori di efficacia e di sicurezza di un farmaco per l'anemia[3]. La risposta è sì, è possibile ridurre i tempi. Con l'ultimo release della nostra startup, presentato all'ultimo congresso di ematologia Ash 2023 a San Diego, ai dati clinici e genomici abbiamo aggiunto la possibilità di sintetizzare dati che riguardano la biopsia midollare e le immagini radiologiche per una valutazione clinica multidisciplinare del paziente».

Secondo Della Porta, l'accelerazione dei clinical trial è importante per superare i limiti attuali, come quelli legati allo studio randomizzato, un generatore di evidenza prezioso ma non privo di vincoli, anche etici, soprattutto nel caso delle malattie rare. In ambito oncologico, un'iniziativa importante è diretta ai pazienti con carcinoma del colon retto metastatico e refrattari al trattamento convenzionale, per i quali l'aspettativa di vita è minore di sei mesi. Un'alleanza globale di accademie e istituti clinici negli Stati Uniti, in Europa e in Giappone, a stretto contatto con le agenzie regolatorie (statunitense, europea e italiana), ha costruito un impianto per gli studi randomizzati chiamato "No Placebo Initiative". Un repository di dati real world già esistenti e certificati verrà utilizzato per costruire i bracci di controllo di tutti gli studi clinici di fase III. La Comunità europea, tra l'altro, in stretta connessione con l'agenzia regolatoria, sta già investendo in quattro consorzi europei, di cui Humanitas fa parte, per validare nuove tecnologie, tra cui quelle di dati sintetici per uso clinico nell'ambito delle malattie rare ematologiche.

«Nel prossimo futuro dei dati sintetici, soprattutto in ematologia, nella mia visione personale sarà necessario creare una piattaforma per quelli che potrei definire *Next Generation Clinical Trial*, con quattro silos: un'iniziativa Ema per la qualificazione dei registri, per aumentare la disponibilità di dati real world per fini regolatori; un'alleanza con tutti i consorzi europei esistenti per definire le nuove tecnologie capaci di catturare efficientemente il valore clinico dei dati di qualità; un nuovo disegno per i trial clinici; una piattaforma che colleghi comunità clinica e agenzie regolatorie».

Bibliografia

1. World Health Organization. Ethics and governance of artificial intelligence for health. Disponibile su: https://lc.cx/QvuR9y [ultimo accesso 16 dicembre 2024].
2. Jacobs F, D'Amico S, Benvenuti C, et al. Opportunities and challenges of synthetic data generation in oncology. JCO Clin Cancer Inform 2023 Aug7; e2300045.
3. D'Amico S, Dall'Olio D, Sala C, et al. Synthetic Data generation by artificial intelligence to accelerate research and precision medicine in hematology. JCO Clin Cancer Inform 2023 Jun7: e2300021

A cura di Maria Frega
giornalista freelance

Matteo Della Porta.

MARTA CIPRIANI
Il punto di vista della ricercatrice

La ricerca della Fondazione Gimema da tempo utilizza i dati sintetici nei trial clinici per replicare la struttura e le proprietà statistiche dei dati osservati per ottenere analisi sul campione sintetico equivalenti a quelle condotte sul campione reale. In questa occasione ci ha pensato Marta Cipriani, ricer-

Marta Cipriani.

catrice della Fondazione, a raccontare le attività che stanno portando avanti negli ultimi anni.

Quando non è possibile condurre un trial clinico randomizzato – per limitazioni etiche, di risorse o nel caso delle malattie orfane di terapia – esistono diverse alternative, ha spiegato: trial adattativi, cioè flessibili, modificando il trattamento in base alle risposte dei pazienti; utilizzo di controlli esterni, rappresentati sia da dati storici sia da dati real world; coorti sintetiche, a oggi l'opzione più utilizzata, che prevede l'impiego di pazienti sintetici, generati emulando coorti già osservate in altri studi per simulare i gruppi di controllo, mentre i pazienti reali sono invece arruolati nel gruppo sperimentale del trial.

Esistono diversi metodi di generazione dei pazienti sintetici da utilizzare nei trial clinici, riassumibili in tre macro categorie. «Nella prima troviamo i metodi basati sul deep learning, che consentono di catturare relazioni anche molto complesse all'interno dei dati, ma necessitano di grandi dataset (e di molta cautela con la privacy) per addestrare i modelli generativi», ha spiegato Cipriani. «Gli altri due metodi sono quelli non parametrici (basati su alberi decisionali) e quelli parametrici: entrambi sono ideali con dataset di partenza più ridotti. Su questi ultimi due metodi ci siamo maggiormente concentrati al Gimema, in particolare con l'utilizzo dello strumento Synthpop del software statistico R. Partendo da un dataset di dati reali a livello individuale generiamo la coorte (anonimizzata) di pazienti fittizi che preserva la struttura e le proprietà statistiche della popolazione di partenza».

La prima esperienza di generazione di dati sintetici del Gimema è stata condotta nel contesto delle leucemie mieloidi acute. In particolare, partendo dai 445 pazienti dello studio Gimema AML1310[1] hanno generato una coorte di 890 pazienti sintetici: il tasso di remissione completa e l'analisi di sopravvivenza nei due gruppi sono risultati sovrapponibili. Ma come impiegare una coorte di pazienti sintetici? Uno dei possibili utilizzi è il gruppo di controllo in un trial virtuale, come fatto con lo studio AML 1718[2] che, come il precedente, coinvolge pazienti affetti da leucemia mieloide acuta. È uno studio multicentrico di fase II che valuta l'efficacia e la tollerabilità di venetoclax, aggiunto alla chemioterapia standard. I buoni risultati hanno permesso di estendere lo studio di fase III con pazienti reali nel gruppo sperimentale e pazienti sintetici (generatici dal gruppo sottoposto al trattamento chemioterapico standard) in quello di controllo.

Un ulteriore contesto in cui Gimema ha fatto esperienza di generazione di dati sintetici è quello delle leucemie linfoblastiche acute Philadelphia negative, con due studi. Il primo[3], con 203 pazienti, oggi rappresenta lo standard clinico di trattamento per questo tipo di leucemia. A quella coorte hanno aggiunto 421 pazienti, osservati in real life e trattati con lo stesso schema terapeutico dello studio 1913[4]. Dalla coorte complessiva di questi due studi hanno generato quella sintetica, cinque volte più grande. I risultati, anche in questo caso, confermano la quasi perfetta concordanza tra le due coorti. Recentemente è stato allora proposto uno studio di fase II per il trattamento delle leucemie linfoblastiche acute che prevede trattamenti analoghi ai precedenti con, in aggiunta, daratumumab. Se i risultati saranno promettenti, nella fase III sarà eseguito un trial virtuale, oltre alla generazione di un sottocampione fittizio.

Prima di chiudere, Cipriani ci tiene a sollevare quelli che per lei saranno i vantaggi e le sfide future dell'utilizzo di questa tecnologia. «Il principale vantaggio nell'utilizzo dei pazienti sintetici è sicuramente l'anonimizzazione che consente un accesso ampio ai dati e la riproducibilità dell'esperimento scientifico. Riusciamo così a fornire modelli predittivi potenzialmente più accurati e, infine, acceleriamo la ricerca e ottimizziamo i costi. L'utilizzo di questi trial virtuali pone poi sicuramente sfide sul fronte regolatorio perché, se un passo avanti è rappresentato dall'AI Act, mancano leggi nazionali specifiche. L'ultima sfida è la rappresentatività dei dati: le coorti sintetiche potrebbero non replicare adeguatamente dei confondenti ignoti, presenti invece nei dati reali».

Bibliografia

1. Piciocchi A, Cipriani M, Messina M, et al. Unlocking the potential of synthetic patients for accelerating clinical trials: results of the first GIMEMA experience on acute myeloid leukemia patients. EJHaem 2024; 5: 353-9.
2. Marconi G, Piciocchi A, Audisio E, et al. Gimema AML1718 Part 1: Planned interim analysis of a safety run-in and phase 2 open-label study of Venetoclax, Fludarabine, Idarubicin and Cytarabine (V-FLAI) in the induction therapy of non low-risk acute myeloid leukemia. Blood 2022; 140 (Suppl 1): 1705-7.
3. Bassan R, Chiaretti S, Della Starza I, et al. Pegaspargase-modified risk-oriented program for adult acute lymphoblastic leukemia: results of the GIMEMA LAL1913 trial. Blood Adv 2023; 7: 4448-61.
4. Lazzarotto D, Cerrano M, Papayannidis C. Outcome of 421 adult patients with Philadelphia-negative acute lymphoblastic leukemia treated under an intensive program inspired by the GIMEMA LAL1913 clinical trial: a Campus ALL study. Haematologica 2025; 110: 55-67.

A cura di Maria Frega
giornalista freelance

Annarita Vestri
Il punto di vista del comitato etico

Le stime affermano che entro il 2030 molti settori saranno governati dai dati sintetici e dall'intelligenza artificiale: da un certo punto di vista questi strumenti possono essere una buona alternativa ai dati reali, come nei casi di difficoltà di reperimento dei dati o quando si è di fronte alla possibilità di lavorare su grandi database. È questa la premessa da cui parte la relazione di Annarita Vestri, presidente del Comitato etico territoriale Lazio 4. Gli ambiti di applicazione dei dati sintetici possono essere molteplici e interessanti nel settore sanitario: dall'epidemiologia alla ricerca, dallo sviluppo dell'informatica sanitaria alla formazione; tuttavia, poiché il set di dati artificiali è creato da un modello (il generatore di dati sintetici) che può assumere diverse forme, tutto è condizionato fortemente dalla scelta del modello, dal quale non si può prescindere.

Vestri prosegue provando ad analizzare i diversi quesiti aperti rispetto ai dati sintetici applicati alla ricerca clinica: dai problemi che possono derivare dalla rapida crescita della generazione

Annarita Vestri.

di dati sintetici, fino a correre il rischio di arrivare all'obsolescenza, con la necessità di andare sempre più veloci altrimenti il modello trainante sarà già vecchio quando si va ad applicarlo.

Come ripetuto più volte nel corso della giornata, una prima sfida è legata alla protezione della privacy: se da un lato i modelli migliorano con l'input di dati da parte degli utenti, dall'altro i vincoli di privacy limitano la possibilità di aggiungere al modello dati clinici e personali di pazienti. Ancora, un vantaggio è l'aumento della dimensione del campione: soprattutto per un ricercatore o un biostatistico la grande mole dei dati è importante perché permette di avere modellistiche più affidabili. Non esiste mai un database ideale: quelli della pratica clinica sono spesso inficiati con errori e contraddizioni e, d'altro canto, c'è anche scarsità di dati, quindi poter aumentare le numerosità sembra offrire un buon vantaggio. Tuttavia, è necessario tener presente tutti i bias intrinseci che non si riesce a cogliere appieno: se quell'aumento dei dati significa anche aumentare i bias, gli errori si trascinano lungo il processo. In sintesi, se non ci si rende conto che ci sono degli errori che si perpetuano, non si può più correggere il modello una volta realizzato.

Un'altra sfida per il futuro da tenere presente è l'estrema complessità di fronte a cui manca formazione per i professionisti che affrontano le questioni che riguardano questi strumenti nuovi. Per poter valutare se il protocollo ha una sua validità scientifica è necessario studiare e formarsi. Ma chi forma i valutatori? In un contesto in cui i comitati etici non lavorano in modo armonizzato, c'è sempre qualche discrepanza e diversità di background. Come stabilire quando utilizzare i dati sintetici e quando invece quelli reali, per rispondere alle esigenze del paziente? Come decidere chi è il responsabile nel processo quando si utilizzano dati? Dal punto di vista statistico, qual è il divario tra il modello e la realtà? Qual è l'effettivo contributo che possono dare i dati sintetici alla ricerca? Vestri ritiene che sia importante capire in che modo lavorare sui gruppi sottorappresentati poiché rimane il problema dell'equità nella ricerca clinica.

In conclusione, i dati sintetici possono essere strumenti molto utili, ma per un comitato etico nella valutazione è importante considerare i seguenti aspetti: la protezione della privacy, i bias, l'equità, la conformità alle normative, la responsabilità e il contesto applicativo che può variare a seconda degli scopi della ricerca. Il comitato etico ha bisogno di sapere di chi è la proprietà dei dati e dunque, in ambito regolatorio, si è di fronte a una sfida. Secondo Vestri esistono delle carenze regolatorie su tutto ciò che non è farmacologico attualmente in Italia, motivo per cui dovranno essere emesse nuove leggi. Tutto questo ci porrà sfide molteplici per il futuro.

A cura di Federica Ciavoni
Il Pensiero Scientifico Editore

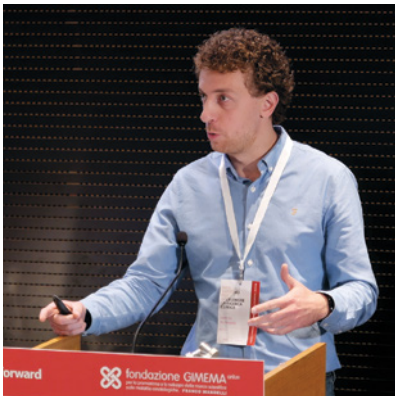
Lorenzo De Angelis
Il punto di vista dell'ente regolatorio

«Riflettendo sul tema dei dati sintetici, ho pensato a una citazione del rinomato fisico Richard Feynman: What I cannot create I don't understand. Se non posso creare qualcosa non la capisco veramente». Ha aperto così il suo intervento, ricordando la sua formazione da fisico, Lorenzo De Angelis, che lavora nel team di Information processing and analytics della European medicines agency (Ema). «Per creare dati sintetici occorre partire da un buon modello di linguaggio per poi verificare eventuali paradossi: è quello che faccio, tra le altre cose, all'Ema, dove stiamo investendo molto sul processare il linguaggio. Lavoriamo, infatti, con una grossa mole di documenti che continua a crescere. La sfida per noi come per altri professionisti in ambito clinico è, dunque, lavorare con dati non strutturati».

L'Ema, racconta De Angelis, riceve sempre più richieste per nuovi medicinali e per *scientific advice* che contengono un mix di dati strutturati e non strutturati. Per non rischiare la perdita di informazioni e per assicurare coerenza verso le diverse entità che si rivolgono all'Ema, hanno quindi creato il motore di ricerca *Scientific Explorer*, dotato di un'interfaccia semplificata che consente di accedere contestualmente all'informazione contenuta nei database strutturati, nei documenti non strutturati, oltre a informazioni aggiuntive estratte usando l'intelligenza artificiale. Per affrontare questa sfida hanno utilizzato l'approccio del machine learning, addestrando un modello di *natural language processing* in grado di estrarre informazioni da un alto numero di documenti. Non sempre i dati utili, per esempio per pianificare le strategie per il design di un clinical trial, sono evidenti, anzi: sono sparsi, presentati in maniera diversa e spesso richiedono input da esperti. «Per risolvere problemi di questo tipo, abbiamo pensato di creare dati sintetici: su dataset artificiali, cioè documenti finti, abbiamo inserito termini noti per addestrare il modello», continua. «Questa semplificazione, però, non funziona perché le parole non bastano: occorre il contesto per ottenere il giusto significato». Nel frattempo, sono arrivati i modelli di grandi dimensioni (*large language models*), come ChatGPT e Claude, risolvendo il problema in pochi minuti: dando le istruzioni appropriate, si genera il testo che si desidera produrre ottenendo risultati di alta qualità. «Si tratta di documenti che simulano quelli reali e, poiché sono disegnati da noi, non necessitano di essere annotati*. I modelli di grandi dimensioni sono in grado di processare il linguaggio, per questo possiamo utilizzarli per estrarre informazioni dai nostri documenti reali. E sanno generare testi di grande qualità».

Oggi, utilizzando i *large language models* per estrarre informazioni dalle lettere di *Scientifc Advice*, Ema è riuscita a completare il processo dello strumento *Scientific Explorer* che, da marzo 2024, è utilizzato dall'European medicine regulatory network. Stanno adesso lavorando per espandere l'uso in altri *use cases* per l'approvazione dei farmaci.

Resta comunque un paradosso: per generare buoni dati sintetici abbiamo bisogno di un buon modello, ma se abbiamo un buon modello ci serve generare i dati sintetici? Investiamo sui modelli o sulla generazione di dati sintetici? «Dal mio background in fisica – continua De Angelis – dove la simulazione di dati ha una lunga storia, arriva uno spunto. Le equazioni di Maxwell ci offrono un modello matematico che spiega nei dettagli che la luce si propaga come un'onda. Le simulazioni di questo fenomeno sono svariate e servono a spiegare molteplici fenomeni fisici a cui non pen-



Lorenzo De Angelis.

siamo osservando semplicemente le equazioni». Occorre, infine, sottolineare che generare dati usando modelli esistenti porta anche a scoprire i limiti di questi ultimi. Le simulazioni ci danno il vantaggio di cambiare parametri a nostro piacimento per testare ipotesi senza coinvolgere nuovi pazienti. Certo, sottolinea, in fisica non esiste il problema della privacy quando si "maltrattano" fotoni ed elettroni.

«L'esperienza sull'uso dei dati di linguaggio all'Ema ci ha reso consapevoli che produrre buoni dati sintetici necessita di una buona comprensione del sistema che si vuole simulare. Utilizzare un modello per generare dati o per capire come funziona il nostro sistema sono due opzioni in grado di creare benefici, come nuove conoscenze, limiti compresi».

* "Annotazione", in questo contesto, si riferisce alla pratica di *data annotation*, cioè a tutte quelle azioni (etichettare, contestualizzare e definire parole, collegare testo e immagini, per esempio) che si compiono nel processo di addestramento delle macchine.

A cura di Maria Frega
giornalista freelance

Noemi Porrello
Il punto di vista dell'industria

Con l'evoluzione tecnologica, il contesto della ricerca clinica è cambiato profondamente. Il settore sanitario è secondo solo all'automotive per investimento in Research & Development, come dimostrano i dati pubblicati ogni anno dalla Commissione europea[1]. L'innovazione è, infatti, un driver fondamentale in questo ambito: vuol dire migliori terapie e migliori possibilità per i pazienti.

«Ho il privilegio di lavorare in un'azienda che, secondo il ranking della Commissione europea, da sempre investe molto in ricerca e sviluppo», si presenta Noemi Porrello, Roche Italia. «Con il nostro portfolio abbastanza diversificato di farmaci, abbiamo avuto la possibilità di vedere le applicazioni specifiche dei dati sintetici, delle evidenze in generale, in un percorso che va dalla nascita dei farmaci stessi alla loro commercializzazione, con riguardo a molteplici variabili tra cui complessità, invecchiamento della popolazione, sostenibilità. Nel confronto sui benefici delle nuove tecnologie, ciò che vedo mancare è il senso di urgenza. A volte sottovalutiamo che i dati possono essere un acceleratore essenziale di certi processi. La disponibilità tardiva di un farmaco non comprime anche il benessere mentale e sociale di una persona? Sarebbe perciò necessaria una pressione sulle autorità regolatorie, sulla politica, per sostenere un contesto che favorisca l'accelerazione della ricerca».

Noemi Porello.

Le domande che ci poniamo nella ricerca clinica, tradizionale, sul generare evidenze e su come un intervento, in condizioni perfette, possa funzionare non sono nuove. Se ne parlava già venticinque anni fa nel noto articolo in cui si ricordavano i concetti definiti da Archie Cochrane per valutare la sperimentazione di interventi sanitari: «Può funzionare? Funziona? Ne vale la pena?»[2]. Nel trial clinico randomizzato controllato si possono prendere decisioni da trasferire nella pratica clinica; in ambito real world è necessario essere più severi rispetto all'errore legato alla tecnologia, alle macchine, perché la ricerca e l'innovazione implicano un rischio e una possibilità di errore. Inoltre, è diventato rilevante il tema della sostenibilità, cioè cosa ci possiamo permettere. Utilizzando il dato sintetico è possibile testare una quantità di ipotesi che la mente umana non riuscirebbe neanche a immaginare. L'urgenza riguarda dunque la possibilità della tecnologia di complementare, senza sostituire, le metodiche tradizionali.

«Roche in partnership con Nvidia – che si occupa di intelligenza artificiale – ha sviluppato un modello interno di collaborazione in ambito discovery che può trovare applicazione nella ricerca clinica. Dal metodo sperimentale, sempre valido, derivano dati che possono aiutare le macchine, gli algoritmi, a generare nuove ipotesi, da ritestare negli esperimenti. La dicotomia tra ricerca cliniche e nuove metodiche, quindi, deve trovare una sinergia anziché un contrasto o una sostituzione. Per questo è però necessario un nuovo mindset per affrontare le domande», continua Porrello. In sintesi, gli aspetti su cui concentrarsi sono tre. Innanzitutto, occorre trovare un punto di compromesso tra la massima accuratezza del dato e la tempestività o tra la protezione estrema della privacy e l'utilità. Questo è solo uno dei contrasti tra dimensioni importanti (qualità/quantità; individuo/salute pubblica; macchina/uomo) in cerca di bilanciamento. La seconda domanda: quali sono le potenzialità del dato sintetico nella ricerca clinica? Su tutte, l'accelerazione e un migliore e più efficace investimento delle risorse. E, di conseguenza, maggiore capacità nella prevenzione ed esiti migliori, ottenendo nel tempo risposte per le quali non sarà necessario effettuare studi specifici, grazie a dati strutturati e sottoposti a una governance più rigida, con investimenti rapidi e sostenibili. Infine: a che punto siamo? Come ogni innovazione, anche i dati sintetici possono essere spiegati con il Gartner Hype Cycle*. Si è partiti, quindi, con un picco di entusiasmo, pensando che la tecnologia potesse risolvere problemi, trovandosi poi a scontrarsi con difficoltà tecniche, come la privacy, che ci portano allo sconforto totale. Adesso ci troviamo in quella fase di risalita che porta al plateau della produttività, dal quale possiamo trarre il meglio dalla tecnologia. Ci si arriva con la fiducia, abbassando la resistenza al cambiamento.

«Sono cresciuta in un ambito di ricerca clinica nel quale la qualità è stata la prima cosa che mi hanno insegnato», conclude Porrello. «Adesso però la qualità va anche accompagnata alla quantità, che, a sua volta, diventa un ulteriore elemento di qualità: il fatto di avere molti dati è importante quanto averli qualitativamente affidabili. Dobbiamo cercare un'alleanza non solo fra le parti, ma anche fra le diverse metodologie a disposizione. Infine, la governance: in questi contesti complessi, un perimetro regolatorio top-down che aiuti a capire come procedere è l'elemento fondamentale per avere vera evoluzione e vero progresso».

*Modello esplicativo del ciclo di vita di una tecnologia, articolato in cinque fasi: innesco, picco delle aspettative esagerate, fossa della disillusione, salita dell'illuminazione, plateau della produttività.

Bibliografia

1. European Commission. Industrial R&D investment: EU's growth highest since 2015. Disponibile su: https://lc.cx/F79qaL [ultimo accesso 16 dicembre 2024].
2. Haynes B. Can it work? Does it work? Is it worth it? The testing of health care interventions is evolving. BMJ 1999; 319: 652-3.

A cura di Maria Frega
giornalista freelance

FRANCESCO NONINO
Il punto di vista del metodologo

«Mi troverò a ripetere alcune cose che sono già state dette nel corso della giornata. Non solo perché sono l'ultimo, ma soprattutto perché gli aspetti metodologici sono trasversali» afferma Francesco Nonino (Irccs Istituto di scienze neurologiche di Bologna e direttore della Cochrane Italia), per sottolineare come questi siano temi comuni a medici e ricercatori, data scientist e giuristi.

Prima di entrare nel merito degli aspetti metodologici, Nonino si sofferma sulle opportunità offerte dall'utilizzo dei dati sintetici in diversi ambiti. Il primo è lo sviluppo di coorti sintetiche, cioè gruppi di pazienti artificiali che potrebbero essere utilizzati per testare l'affidabilità di algoritmi e farmaci sperimentali nel trattamento delle malattie rare e ultra rare o delle malattie degenerative. Un altro campo nel quale i dati sintetici potrebbero avere un ruolo rilevante è quello del *record linkage* per l'integrazione di dati provenienti da fonti diverse, in particolare dati clinici e amministrativi. Ma quali sono i potenziali rischi? Dopo aver esposto le potenzialità di questa tecnologia, Nonino fa luce su alcuni problemi che derivano dall'utilizzo dell'intelligenza artificiale. Primo fra tutti il rischio di bias e allucinazioni: i dati con i quali vengono addestrati gli algoritmi devono essere di qualità, perché eventuali pregiudizi, polarizzazioni o distorsioni della realtà verranno poi amplificati. E allenare un sistema di machine learning o di generazione di dati sintetici su un dataset reale incompleto o con dei dati sbilanciati potrebbe condurre il sistema verso relazioni e associazioni non sempre sensate o appropriate.

Francesco Nonnino.

La domanda di ricerca per la quale i dati sintetici sono generati dovrebbe essere uno dei criteri metodologici per determinarne la qualità. In quest'ottica Nonino individua tre misure qualitative dalle quali partire per costruire uno strumento di valutazione di tali sistemi: la *fidelity* (o *utility*) per capire se i campioni sintetici possiedono caratteristiche simili a quelli reali; la *diversity*, la rappresentatività della popolazione reale, per identificare sottogruppi non rappresentati; la *generalization*, strettamente legata alla privacy, che quantifica in che misura l'algoritmo "copia" i dati reali.

In generale, per valutare da un punto di vista metodologico la qualità degli studi randomizzati controllati e quella degli studi diagnostici i medici hanno a disposizione una serie di strumenti riconosciuti e validati dalla comunità scientifica come RoB 2, ROBINS-I, QUADAS-2, AMSTAR 2. «Non abbiamo ancora – e secondo me questo è un terreno su cui bisognerebbe mettersi a lavorare – degli strumenti metodologici per valutare la qualità dei dati sintetici: a una domanda di ricerca deve corrispondere un sistema di generazione di dati sintetici adeguato».

Un altro problema, legato a quello della generazione di dati sintetici di qualità, è la trasparenza della ricerca: i difetti dei dati sintetici, infatti, sono subordinati alla qualità dei dati da cui derivano. Nonostante la comunità scientifica stia incoraggiando alcune buone pratiche come la condivisione dei dati, uno studio del 2023 pubblicato sul *BMJ*[1] ha mostrato che tra il 2016 e il 2021 veniva dichiarata disponibile una quantità di dati pari all'8% del totale, mentre quella effettiva era del 2%. Guardando invece i codici statistici e analitici la percentuale non arrivava all'1%.

«Ho aggiunto ottimisticamente un'altra parola al titolo di questo convegno: i dati sintetici possono servire non solo ad accelerare ma anche a migliorare la ricerca». Secondo il neurologo, nei tavoli di pianificazione e gestione della ricerca clinica metodologica dovrebbero trovare posto pazienti, ricercatori, metodologici, clinici e data scientist, che possono dare un contributo prezioso per la realizzazione di una ricerca utile e tempestiva che possa andare incontro ai reali bisogni dei pazienti.

Bibliografia

1. Hamilton DG, Hong K, Fraser H, et al. Prevalence and predictors of data and code sharing in the medical and health sciences: systematic review with meta-analysis of individual participant data. BMJ 2023; 382: e075767.

A cura di Andrea Calignano
Il Pensiero Scientifico Editore