

DATI SINTETICI IN SANITA'

Utilizzo e prospettive per la produzione di evidenze
cliniche

Gabriele D'Amico - Senior Associate LS CUBE Società Tra Avvocati

Cosa sono i dati sintetici

I dati sintetici sono dei dati originati attraverso varie metodologie (incluse quelle di Intelligenza Artificiale) che utilizzano come input dei dati reali e producono come output dei dati che sono “simili” a quelli reali senza essere identici

QUATTRO PROPRIETÀ CHIAVE

Origine generativa

Generati da modelli matematici, statistici o algoritmi, non da misurazioni reali

Fedeltà statistica

Capacità di preservare le stesse proprietà statistiche dei dati reali, mantenendo intatti i raggruppamenti naturali (*cluster*) e i modelli semantici presenti nei dati originali, senza tuttavia contenere alcuna informazione reale o specifica sugli individui cui tali dati si riferiscono

Varietà e tipologie

Immagini mediche ed istopatologiche, dati tabulari, dati monodimensionali, dati multidimensionali, dati longitudinali

Compliance GDPR & privacy

Se generati correttamente, rispettano il GDPR e integrano garanzie come differential privacy

Metriche di qualità

Cinque dimensioni per valutare la qualità dei dati sintetici nelle applicazioni cliniche e regolatorie

FIDELITY

Riproduzione della stessa struttura statistica dei dati reali: distribuzioni, correlazioni, momenti

UTILITY

Capacità di sostituire o integrare i dati reali nell'addestramento di modelli predittivi

PRIVACY

Preservazione della riservatezza dei soggetti reali

DIVERSITY

Riflesso dell'intera distribuzione dei dati reali, evitando *overfitting* e perdita di variabili eccezionali (cc.dd. *outlier*)

CLINICAL UTILITY

Rilevanza clinica e interpretabilità dei dati sintetici nei contesti d'uso reali

Potenzialità dei dati sintetici

L'impiego di bracci di controllo sintetici e più in generale di dati sintetici nel disegno e conduzione di sperimentazioni cliniche offre molteplici vantaggi, già documentate in letteratura, che possono essere articolati in diverse dimensioni strategiche

Espansione disponibilità dati

Nelle malattie rare, la generazione di dati sintetici aumenta significativamente la dimensione campionaria, superando l'ostacolo della scarsità intrinseca di informazioni e permettendo di soddisfare requisiti regolatori

Applicazioni sistemiche

Ottimizzazione dei percorsi diagnostico-terapeutici e previsione dei fabbisogni di risorse in scenari critici, modelli predittivi dell'evoluzione dei bisogni di salute e l'impatto futuro delle patologie croniche

Ottimizzazione degli studi clinici e riduzione dei dilemmi etici

Simulazione dell'effetto placebo, riducendo la necessità di includere pazienti reali in gruppi di controllo e l'esposizione a trattamenti inattivi

Personalizzazione terapeutica

Identificazione di sottogruppi di *responders* e stratificazione di popolazioni in base a caratteristiche predittive di efficacia, nonché modellando, virtualmente, le variabilità genetiche, fenotipiche e ambientali che influenzano la risposta ai trattamenti

I *case studies* del progetto NET – HEALTH - Sanità in Rete

Tre scenari per esplorare le applicazioni dei dati sintetici in contesti clinici e regolatori diversi

01 Patologie a maggiore prevalenza — *Alzheimer (AD) e Diabete mellito*

02 Patologie rare e ultra-rare — *Leucemia Mieloide Acuta (AML) e White-Sutton (WSS)*

03 RWE post-approval — *Bracci di confronto sintetici ex-post*

Patologie a maggiore prevalenza

CASE STUDY 1

Performance dei dati sintetici su due dataset paradigmatici di patologie a maggiore prevalenza

Diabete Mellito

Dati sintetici bilanciati, generati per compensare lo sbilanciamento tra classi nel *dataset* originale, hanno migliorato le prestazioni del modello predittivo. L'applicazione di tecniche statistiche ha confermato la similarità statistica con i dati reali

Alzheimer

I dati sintetici hanno riprodotto fedelmente l'andamento temporale dei biomarcatori per la previsione dell'età biologica, con allineamento quasi perfetto rispetto ai dati reali. Risultato che ne conferma l'affidabilità per bracci sintetici da database longitudinali consolidati

Patologie rare e ultra-rare

CASE STUDY 2

Braccio di controllo costruito a partire da modelli predittivi con dati sintetici per superare i limiti etici e la scarsità di dati nelle malattie rare e ultra-rare

Leucemia Mieloide Acuta (AML)

Framework predittivo unificato che collega la classificazione diagnostica alla stratificazione del rischio. La generazione di dati sintetici ha consentito di riprodurre la complessità dei dati genomici, clinici e longitudinali, preservando le distribuzioni delle frequenze mutazionali e la struttura prognostica appresa dai dati reali

Sindrome di White-Sutton

La scarsità di dati reali è stata superata tramite generazione sintetica, ampliando artificialmente il campione e preservando le relazioni multivariate e i pattern delle caratteristiche fenotipiche della popolazione reale di pazienti

I profili regolatori

Il quadro normativo europeo e nazionale e gli spazi per l'impiego dei dati sintetici nei percorsi autorizzativi e di HTA dei medicinali.

Assenza di un quadro normativo specifico e dedicato

Nessuna disposizione normativa, europea o nazionale, disciplina puntualmente l'uso dei dati sintetici nei percorsi autorizzativi e di HTA dei medicinali. Adottata nel febbraio 2026, la linea guida ICH M15 stabilisce i criteri di valutazione delle evidenze generate da modelli computazionali in ambito regolatorio.

Le forme di integrazione nei processi di produzione di evidenze

Il principale interrogativo attiene alla possibilità, per i dati sintetici, di essere validamente integrati nell'ambito della costruzione di protocolli di studio per la produzione di evidenze cliniche di valore regolatorio

Il nodo giuridico: uso complementare vs sostitutivo

A seconda che i dati sintetici affianchino o sostituiscano le evidenze randomizzate cambiano i requisiti metodologici e il grado di scrutinio regolatorio richiesto dalle Autorità

La questione definitoria

L'assenza di una definizione condivisa di «dato sintetico» genera incertezza giuridica sui confini del suo impiego nello sviluppo dei medicinali.

FDA — Stati Uniti

La FDA definisce i dati sintetici come dati creati artificialmente - ad es. tramite modellazione statistica o simulazione - che riproducono struttura, proprietà e relazioni dei dati reali dei pazienti, senza contenere alcuna informazione reale sui singoli individui.

EMA — Unione europea

- Reflection Paper sull'IA (9 settembre 2024): cita i dati sintetici per l'ampliamento dei dataset di addestramento, ma non ne fornisce una definizione.
- Draft Concept Paper (luglio 2025): riconosce l'assenza di linee guida specifiche sui controlli esterni.
- Reflection Paper dedicato atteso per il 2° trimestre 2027: definizione standardizzata di «controllo esterno».

SCA $\subset$ ECA — i bracci di controllo sintetici (SCA) sono una *species* del più ampio genere dei controlli esterni (ECA) e, per le loro peculiarità, meriterebbero una definizione propria e specifica.

Responsabilità e requisiti metodologici

Il Reflection Paper EMA sull'intelligenza artificiale nel ciclo di vita dei medicinali (9 settembre 2024) fissa i principi cardine per i dati generati con AI/ML

Responsabilità dello sponsor

Resta in capo allo sponsor, al richiedente o al titolare dell'AIC garantire che algoritmi, modelli, dataset e pipeline siano «fit for purpose» e conformi agli standard legali, etici, tecnici, scientifici

Mitigazione del bias

I modelli di AI/ML sono data-driven e vulnerabili all'integrazione di *bias*: occorre acquisire un training dataset bilanciato e sufficientemente ampio rispetto al contesto d'uso previsto

Documentazione e tracciabilità

Va documentato che i dataset reali alla base della generazione dei dati sintetici siano adeguati e rappresentativi del contesto clinico di riferimento

I domini principali di impiego

Fase autorizzativa — AIC e AIC condizionata

- RCT gold standard (Reg. 536/2014, Dir. 2001/83/CE); controlli esterni solo se giustificati.
- AIC condizionata (art. 14-bis Reg. 726/2004): accesso precoce per bisogni insoddisfatti.
- Dati sintetici per integrare o simulare evidenze limitate (es. malattie rare).

Fase post-autorizzativa

- PASS (sicurezza) e PAES (efficacia) in condizioni reali.
- Reflection paper RWD (2025): requisiti per gli studi non interventistici.
- Possibile supporto alla generazione di RWE pianificata.

HTA europeo

- Reg. 2021/2282 (dal 12 gen. 2025): valutazioni cliniche congiunte (JCA).
- Linee guida HTACG prive di riferimenti ai dati sintetici.
- JSC: sede per l'allineamento metodologico ex ante.

Prezzo e rimborso nazionale

- Competenza esclusiva degli Stati membri (art. 168(7) TFUE).
- AIFA - DM 2 agosto 2019: negoziazione di prezzo e rimborso.
- Usi possibili: cost-effectiveness e BIA; manca un framework condiviso.

Lo strumento delle *sandbox* regolatorie

Ambienti controllati di sperimentazione normativa: strumento naturale per testare modelli generativi di dati sintetici e bracci di controllo sintetici prima dell'adozione regolatoria stabile.

Livello europeo

Nuovo Pharma Regulation

- Sandbox interna al sistema autorizzativo: AIC con deroghe a tempo limitato e requisiti adattati, per medicinali con caratteristiche o metodi incompatibili con il regime ordinario.

Biotech Act

- Coordina la sandbox del Pharma Regulation con quella sulle sperimentazioni cliniche: ove servano nuovi approcci anche per i trial, la Commissione può istituirne una ai sensi del Reg. Clinical Trial.
- Deroghe procedurali nei trial: disegno dello studio, consenso informato (one-shot con follow-up pluriennale), farmacovigilanza per eventi avversi tardivi.

Livello italiano

L. 132/2025

Art. 9

- Introduce gli «spazi speciali di sperimentazione»: la sandbox del Ministero della Salute per i dati sanitari.

Art. 20

- AgID e ACN istituiscono e gestiscono congiuntamente gli spazi di sperimentazione.

Raccomandazioni operative

Alcune raccomandazioni operative per abilitare l'utilizzo dei dati sintetici nell'ambito della generazione delle evidenze cliniche.

01

Framework di validazione

Standardizzati, condivisi e context-specific, con metriche multidimensionali per corrispondenza statistica e utilità clinica.

02

Data quality assurance

Sistemi di data governance che documentino provenienza, trasformazioni e criteri di selezione lungo tutto il processo.

03

Valutazione rischio privacy

Fase di valutazione del rischio per verificare che i dati sintetici non siano riconducibili a dati personali reali.

04

Dialogo con autorità

Scientific advice, consultazioni precoci con le Agenzie Regolatorie e progetti pilota in sandbox regolatorie.

05

Formazione e infrastrutture

Sviluppo delle competenze e delle infrastrutture tecnologiche per sostenere l'adozione su larga scala.

06

Governance etica e legale

Definire le responsabilità di sviluppatori e ricercatori nella generazione e utilizzo dei dati sintetici.

Grazie per l'attenzione