Generazione di dati sintetici in farmacoepidemiologia: nuove opportunità per l'uso secondario dei dati *real-world*

Maria Cutillo, Flavia Mayer, Stefania Spila Alegiani, Marco Massari, Centro Nazionale per la Ricerca e la Valutazione preclinica e clinica dei Farmaci, Istituto Superiore di Sanità, Roma
Giorgia Stendardo, Centro Nazionale per le Tecnologie Innovative in Sanità Pubblica, Istituto Superiore di Sanità, Roma

Domanda

I dati sintetici possono rappresentare una nuova opportunità per favorire l'uso e la condivisione dei dati real-world *nella ricerca farmacoepidemiologica, garantendo una riproduzione accurata dei dati reali e, al contempo, un'adeguata tutela della privacy?*

Contesto della domanda

La ricerca farmaco-epidemiologica utilizza sempre più frequentemente dati *real-world* (RW), ovvero dati generati e raccolti nell'ambito della pratica clinico-assistenziale e dei processi sanitari. Tra le principali fonti rientrano i registri clinici di patologia e di monitoraggio dei farmaci, le banche dati amministrative quali le Schede di Dimissione Ospedaliera, la Farmaceutica territoriale, l'Anagrafe degli assistiti e i dati del Pronto Soccorso; infine, le cartelle cliniche elettroniche. L'analisi di questi dati consente di studiare l'uso e gli effetti dei farmaci, in termini di benefici e rischi, nelle popolazioni e nelle condizioni della pratica clinica reale e rappresenta una risorsa importante per produrre evidenze a supporto delle decisioni regolatorie e delle politiche del farmaco [1-3].

La crescente disponibilità e rilevanza dei dati RW ha favorito, negli ultimi due decenni, lo sviluppo di numerose reti nazionali e internazionali che ne facilitano la raccolta, l'integrazione e l'utilizzo per la ricerca, la tutela della salute pubblica e la governance del farmaco. Tra queste rientrano importanti iniziative internazionali, come *Sentinel* negli Stati Uniti [4], CNODES in Canada [5], DARWIN EU [6] e ConcePTION [7] in Europa, mentre a livello nazionale possono essere citate BIFAP in Spagna [8], e TheShinISS in Italia [9-13].

La possibilità di utilizzare e condividere questi dati è tuttavia condizionata dalla necessità di proteggere la riservatezza delle persone cui i dati si riferiscono. Il problema è particolarmente rilevante quando i dati individuali vengono utilizzati per finalità diverse da quelle per cui sono stati inizialmente raccolti (c.d. uso secondario). La condivisione

dei dati può essere infatti limitata da vincoli legali ed etici, anche quando i dati sono stati sottoposti a procedure di anonimizzazione o pseudonimizzazione. Tali vincoli possono limitare l'accesso ai dati e ostacolare la conduzione di nuovi studi, la validazione indipendente dei risultati, lo sviluppo e il test di metodi analitici e la collaborazione tra ricercatori e istituzioni [14-16].

Il quadro normativo relativo all'utilizzo secondario dei dati sanitari è inoltre in continua evoluzione, anche in relazione allo sviluppo e alla diffusione dell'intelligenza artificiale. Il recente quadro legislativo nazionale in materia di intelligenza artificiale, in coerenza con l'*Artificial Intelligence Act* europeo [17-18], introduce specifiche disposizioni relative al trattamento dei dati sanitari per finalità di ricerca e sperimentazione scientifica. In questo contesto, assume particolare interesse la ricerca di modalità che consentano di valorizzare il patrimonio informativo dei dati RW riducendo, al contempo, la necessità di condividere dati individuali riferibili a persone reali.

Una possibile soluzione è rappresentata dalla generazione di dati sintetici [19]: dati generati artificialmente mediante metodi statistici o di intelligenza artificiale, con l'obiettivo di riprodurre la struttura, le proprietà statistiche e le relazioni osservate nei dati originali senza contenere informazioni riferibili direttamente a persone reali.

La qualità di un dataset sintetico, quindi, non può essere valutata sulla sola base della sua capacità di riprodurre adeguatamente le principali caratteristiche dei dati originali. Un dataset sintetico deve anche permettere di ottenere risultati analitici comparabili a quelli che si otterrebbero sui dati reali e, contemporaneamente, presentare un rischio molto basso di identificare un individuo o dedurre informazioni sensibili. I dati sintetici devono, cioè, bilanciare utilità statistica e protezione della privacy.

In questo contesto, un confronto sistematico tra diversi metodi di generazione di dati sintetici può fornire indicazioni utili a ricercatori, istituzioni sanitarie e autorità regolatorie interessati al loro utilizzo e alla loro condivisione in ambito farmaco-epidemiologico. La letteratura ha finora dedicato particolare attenzione alla valutazione dell'utilità statistica dei dati sintetici, mentre la valutazione del rischio di re-identificazione (*disclosure risk*) è stata meno frequentemente integrata nei confronti tra differenti metodi.

I metodi di generazione di dati sintetici tabulari, come quelli comunemente utilizzati negli studi epidemiologici e nei trial clinici, sono numerosi e in continua evoluzione. In generale, possono essere distinti in approcci basati su modelli statistici tradizionali e approcci che utilizzano tecniche di *machine learning* [20].

Alla luce di queste considerazioni, nel nostro studio, pubblicato su Epidemiologia & Prevenzione [21] sono stati confrontati due dei più utilizzati metodi di generazione di dati sintetici, rappresentativi delle due categorie sopra descritte: synthpop [22], basato su modelli statistici tradizionali, e *Conditional Tabular-Generative Adversarial Networks* (CT-GAN) [23], basato su tecniche di *deep learning*.

Il confronto ha preso in esame tre dimensioni principali: (i) la capacità dei dati sintetici di riprodurre le distribuzioni delle variabili reali e le relazioni tra esse; (ii) la

capacità di ottenere risultati analitici comparabili a quelli osservati sui dati reali; e (iii) la capacità di ridurre il rischio di re-identificazione o di divulgazione di informazioni sensibili.

Come siamo arrivati alla risposta

Dal dataset originale ai dataset sintetici

Il dataset di partenza è quello utilizzato nello studio osservazionale *Survival of Hospitalized COVID-19 Patients in Northern Italy: A Population-Based Cohort Study by the ITA-COVID-19 Network*, che aveva l'obiettivo di analizzare la sopravvivenza di pazienti adulti ospedalizzati per COVID-19, valutando differenze in base all'età, al sesso, alla regione/provincia di provenienza (centro) e al livello di comorbidità (*Charlson Index*). La coorte comprendeva 42.926 pazienti di età ≥18 anni, ricoverati in Lombardia, Veneto e provincia di Reggio Emilia tra il 21 febbraio e il 21 aprile 2020, durante la prima fase della pandemia [24].

Si trattava di un ampio dataset anonimizzato di dati RW generato, a livello locale, attraverso procedure di trasformazione in *Common Data Model* (CDM) e record linkage di banche dati amministrative e registro COVID-19 nell'ambito del sistema di analisi distribuite TheShinISS, sviluppato dall'Istituto Superiore di Sanità [9]. Per ciascun individuo erano disponibili 40 variabili relative a caratteristiche demografiche e cliniche al momento del ricovero, numero di ospedalizzazioni e di prescrizioni di farmaci nei 12 mesi precedenti l'ospedalizzazione per COVID-19, uso di differenti classi di farmaci, *Charlson Index*, durata del follow-up e tempo al decesso per qualunque causa durante il follow-up.

A partire dal dataset originale sono stati generati due dataset sintetici, della stessa dimensione di quello originale, utilizzando rispettivamente synthpop e CT-GAN. I due approcci si basano su principi diversi per imparare a riprodurre le caratteristiche dei dati reali:

- synthpop, implementato in R, utilizza modelli statistici per generare, in modo sequenziale, una variabile alla volta. La prima variabile viene generata campionando in modo casuale valori dalla variabile reale; le variabili successive vengono generate condizionatamente alle variabili già sintetizzate, utilizzando modelli predittivi parametrici o non parametrici. In questo studio è stato utilizzato un metodo basato sugli alberi decisionali (CART).
- CT-GAN è un metodo di *deep learning* basato sull'addestramento di due reti neurali avversarie: un generatore, che genera i dati sintetici, e un discriminatore, che cerca di distinguere i dati sintetici da quelli reali. Il generatore viene progressivamente migliorato sulla base degli errori individuati dal discriminatore, fino a produrre dati sempre più difficili da distinguere da quelli reali. In questo studio il modello è stato addestrato per 300 cicli completi sui dati (epoche),

utilizzando gruppi di 500 osservazioni alla volta (batch size). Può essere implementato in R e Python, in questo studio è stato usato Python.

Nel contesto specifico dello studio, synthpop ha mostrato un minore tempo computazionale di generazione del dataset sintetico: circa 1 minuto contro 19 minuti necessari per CT-GAN.

I due dataset sintetici ottenuti sono stati poi confrontati con il dataset originale sotto tre aspetti:

1. utilità generale: quanto i dati sintetici riescono a riprodurre le distribuzioni delle variabili reali;
2. utilità specifica: quanto i dati sintetici permettono di ottenere risultati analitici simili a quelli ottenuti a partire dai dati originali;
3. privacy: quanto i dati sintetici permettono di ridurre il rischio di identificare individui o di dedurre informazioni sensibili.

Utilità generale: similarità delle distribuzioni delle variabili tra dati reali e sintetici

La prima valutazione ha riguardato quanto i dati sintetici riuscissero a riprodurre le principali caratteristiche del dataset originale.

Per ogni variabile è stata quindi confrontata la distribuzione osservata nei dati reali con quella ottenuta nei due dataset sintetici, sia in termini di statistiche descrittive (**Tabella 1**) e rappresentazioni grafiche attraverso la valutazione della sovrapposizione degli istogrammi e dei diagrammi a barre (**Figura 1**), sia quantitativamente utilizzando la *Standardized Mean Difference* (SMD) e il *Propensity Score Mean Squared Error* standardizzato (S_pMSE) bivariato costruito considerando tutte le possibili coppie di variabili [25]. Per la SMD, valori prossimi a zero indicano una maggiore similarità tra i due dataset, valori compresi tra -0,1 e +0,1 sono stati considerati indicativi di una buona sovrapposizione. Per il S_pMSE, valori più bassi indicano una maggiore similarità tra dati reali e sintetici; valori fino a 3 indicano una buona sovrapposizione e valori fino a 10 sono generalmente considerati accettabili.

La Figura 2 mostra il confronto, in termini di SMD, tra i dati originali e quelli generati con i due metodi per le singole variabili. Nel dataset generato con synthpop, le variabili considerate hanno mantenuto una distribuzione sufficientemente vicina a quella osservata nei dati reali: infatti, tutte le differenze sono rimaste entro la soglia considerata accettabile ($|SMD| \leq 0{,}1$), indicando una buona riproduzione delle caratteristiche dei dati reali. Nel dataset generato con CT-GAN, invece, diverse variabili si sono discostate in maniera importante dai valori originali. Le differenze più evidenti riguardavano, tra le variabili continue, il numero di prescrizioni nei 12 mesi precedenti il ricovero (SMD=−0,324) e nei tre mesi precedenti (SMD=−0,577), mentre tra le variabili binarie si osservavano differenze per il sesso (SMD=0,376), l'utilizzo di farmaci antidepressivi (SMD=0,453) e la presenza di cardiopatia ischemica (SMD=0,339). Nel complesso,

la Figura 2 mostra quindi una maggiore similarità tra i dati reali e quelli generati con synthpop rispetto a CT-GAN per quanto riguarda il confronto tra singole variabili.

Il S_pMSE bivariato mostra, per synthpop, valori della mediana e del range interquartile (IQR) ampiamente inferiori a 10 (5,539; IQR 2,520–9,721), indicando una buona similarità con i dati reali; al contrario, i valori sono enormemente più alti per CT-GAN (757,936; IQR 325,111–1224,009), indicando una maggiore differenza con i dati reali.

Tabella 1
Confronto delle statistiche descrittive tra dataset reale e dataset sintetici generati con metodo synthpop e CT-GAN

	Dataset originale	synthpop	synthpop	CT-GAN	CT-GAN
		Dataset sintetico	SMD	Dataset sintetico	SMD
N. record	42.926	42.926		42.926	
Età, anni, mediana [IQR]	69,0 [57,0-79,0]	69,0 [57,0-79,0]	0,006	75,0 [58,0-86,0]	-0,261
Classe d'età, anni (%)					
18-49	5.561 (13,0)	5.500 (12,8)	0,004	3.955 (9,2)	0,119
50-59	7.172 (16,7)	7.292 (17,0)	-0,007	6.911 (16,1)	0,016
60-69	8.754 (20,4)	8.791 (20,5)	-0,002	9.439 (22,0)	-0,039
70-79	10.953 (25,5)	11.021 (25,7)	-0,004	9.159 (21,3)	0,099
80-89	8.880 (20,7)	8.774 (20,4)	0,006	8.491 (19,8)	0,023
90+	1.606 (3,7)	1.548 (3,6)	0,007	4.971 (11,6)	-0,298
Sesso, M (%)	26.873 (62,6)	26.886 (62,6)	-0,001	18.951 (44,1)	0,376
N. ospedalizzazioni (%)					
0	31.458 (73,3)	31.431 (73,2)	0,001	29.017 (67,6)	0,125
1	6.452 (15,0)	6.422 (15,0)	0,002	6.410 (14,9)	0,003
2	2.600 (6,1)	2.598 (6,1)	<0,001	1.628 (3,8)	0,105
3+	2.416 (5,6)	2.475 (5,8)	-0,006	5.871 (13,7)	-0,275
Charlson index (%)					
0	29.775 (69,4)	29.898 (69,7)	-0,006	30.401 (70,8)	-0,032
1-2	10.575 (24,6)	10.693 (24,9)	-0,006	6.968 (16,2)	0,210
3+	2.576 (6,0)	2.335 (5,4)	0,024	5.557 (12,9)	-0,239
Tempo di follow-up, mediana [IQR]	24,00 [10,00-35,00]	24,00 [10,00-35,00]	-0.008	22,00 [10,00-35,00]	-0,024
Decesso (%)	11.205 (26,1)	11.121 (25,9)	-0,004	8.598 (20,0)	-0,145
Centro (%)					
Lombardia	38.715 (90,2)	38.776 (90,3)	-0,005	40.234 (93,7)	-0,130
Reggio Emilia	982 (2,3)	1.006 (2,3)	-0,004	741 (1,7)	0,040
Veneto	3.229 (7,5)	3.144 (7,3)	0,008	1.951 (4,5)	0,125
Terapia intensiva (%)	4.392 (10,2)	4.340 (10,1)	-0,004	5.572 (13,0)	0,086
N. prescrizioni, mediana [IQR]					
12 mesi precedenti	15,00 [3,00-36,00]	16,00 [3,00-36,00]	-0,004	27,00 [4,00-54,00]	-0,325
3 mesi precedenti	3,00 [0,00-9,00]	4,00 [0,00-9,00]	-0,004	8,00 [1,00-16,00]	-0,577

Tabella 1 (*segue*)
Confronto delle statistiche descrittive tra dataset reale e dataset sintetici generati con metodo synthpop e CT-GAN

	Dataset originale	synthpop	synthpop	CT-GAN	CT-GAN
		Dataset sintetico	SMD	Dataset sintetico	SMD
Farmaci (%)					
Anticoagulanti	6.743 (15,7)	6.733 (15,7)	-0,001	8.004 (18,6)	0,078
Inibitori dell'aggregazione piastrinica	9.025 (21,0)	8.923 (20,8)	-0,006	11.525 (26,8)	0,137
Agenti modificatori dei lipidi	12.550 (29,2)	12.584 (29,3)	0,002	7.487 (17,4)	-0,282
Antiaritmici	1.924 (4,5)	1.919 (4,5)	-0,001	1.821 (4,2)	-0,012
Farmaci antiulcera	16.192 (37,7)	16.078 (37,5)	-0,005	17.700 (41,2)	0,072
Farmaci per BPCO	6.356 (14,8)	6.486 (15,1)	0,008	6.735 (15,7)	0,025
Antibiotici	1.7513 (40,8)	17.802 (41,5)	0,014	17.591 (41,0)	0,004
Farmaci anti-HIV	300 (0,7)	314 (0,7)	0,004	866 (2,0)	0,114
Farmaci anti Parkinson	754 (1,8)	734 (1,7)	-0,004	595 (1,4)	-0,030
Farmaci antiepilettici	2.936 (6,8)	2.969 (6,9)	0,003	6.131 (14,3)	0,244
Antipsicotici	1.737 (4,0)	1.836 (4,3)	0,012	3.711 (8,6)	0,189
Antidepressivi	5.339 (12,4)	5.393 (12,6)	0,004	13.138 (30,6)	0,453
FANS	2.680 (6,2)	2.784 (6,5)	0,010	4.436 (10,3)	0,149
Corticosteroidi per uso sistemico	2.358 (5,5)	2.437 (5,7)	0,008	5.431 (12,7)	0,251
Clorochina/idrossiclorochina	222 (0,5)	219 (0,5)	-0,001	862 (2,0)	0,134
Immunosoppressori	695 (1,6)	766 (1,8)	0,013	1.672 (3,9)	0,139
Comorbidità (%)					
Cardiopatia ischemica	4.436 (10,3)	4.162 (9,7)	-0,021	9.761 (22,7)	0,339
Fibrillazione atriale	2.899 (6,8)	2.657 (6,2)	-0,023	3.556 (8,3)	0,058
Insufficienza cardiaca	2.684 (6,3)	2.443 (5,7)	-0,024	2.171 (5,1)	-0,052
Ipertensione	5.610 (13,1)	5.506 (12,8)	-0,007	4.750 (11,1)	-0,062
Malattie cerebrovascolari	3.441 (8,0)	3.187 (7,4)	-0,022	6.669 (15,5)	0,235
Diabete mellito	7.710 (18,0)	7.910 (18,4)	0,012	8.949 (20,8)	0,073
Epatopatia	701 (1,6)	804 (1,9)	0,018	664 (1,5)	-0,007
Demenza	1.014 (2,4)	1.006 (2,3)	-0,001	2.139 (5,0)	0,140
Polmonite/influenza	2.737 (6,4)	2.671 (6,2)	-0,006	1.683 (3,9)	-0,111
Insufficienza renale cronica	1.046 (2,4)	1.057 (2,5)	0,002	1.549 (3,6)	0,068
BPCO	1.521 (3,5)	1.447 (3,4)	-0,009	1.330 (3,1)	-0,025
Tumori	6.446 (15,0)	6.642 (15,5)	0,013	3.530 (8,2)	-0,213
Malattie reumatiche	375 (0,9)	403 (0,9)	0,007	2.234 (5,2)	0,254

GAN: *Generative Adversarial Network*; SMD: *Standardized Mean Difference*; N: Numero; IQR: *Interquartile Range*; BPCO: Broncopneumopatia cronica ostruttiva; HIV: *Human Immunodeficiency Virus*; FANS: Farmaci antinfiammatori non steroidei.

Figura 1
Confronto delle distribuzioni delle variabili tra dataset reale e dataset sintetici generati con synthpop e CT-GAN



a) numeriche; b) categoriche; c) binarie.
VE: Veneto; RE: Reggio Emilia; LO: Lombardia; M: Maschi; HIV: *Human Immunodeficiency Virus*; FANS: Farmaci antinfiammatori non steroidei; BPCO: Broncopneumopatia cronica ostruttiva; CT-GAN: *Conditional Tabular-Generative Adversarial Networks*.

Figura 2
***Standardized Mean Differences* tra dati originali e sintetici generati con CT-GAN e synthpop**



VE: Veneto; RE: Reggio Emilia; LO: Lombardia; M: Maschi; HIV: *Human Immunodeficiency Virus*; FANS: Farmaci antinfiammatori non steroidei; BPCO: Broncopneumopatia cronica ostruttiva; SMD: *Standardized Mean Difference*; CT-GAN: *Conditional Tabular-Generative Adversarial Networks.*

Utilità specifica: similarità dei risultati delle analisi statistiche tra dati reali e sintetici

La valutazione dell'utilità specifica permette di rispondere alla seguente domanda: se conducessimo sui dati sintetici la stessa analisi statistica effettuata sui dati reali, otterremmo risultati simili?

A questo scopo, è stata valutata la coerenza delle misure di associazione tra dati reali e sintetici confrontando i risultati dell'analisi di sopravvivenza ottenuti per il dataset originale e per i due dataset sintetici.

Sono state confrontate le curve di Kaplan-Meier (K-M) ottenute dai dati reali con quelle ottenute dai dati sintetici. È stata inoltre valutata la similarità dei risultati dei modelli di regressione di Cox, attraverso la sovrapposizione degli intervalli di confidenza al 95% (IC95%) degli *Hazard Ratio* (HR) ottenuti sui dati reali e sintetici.

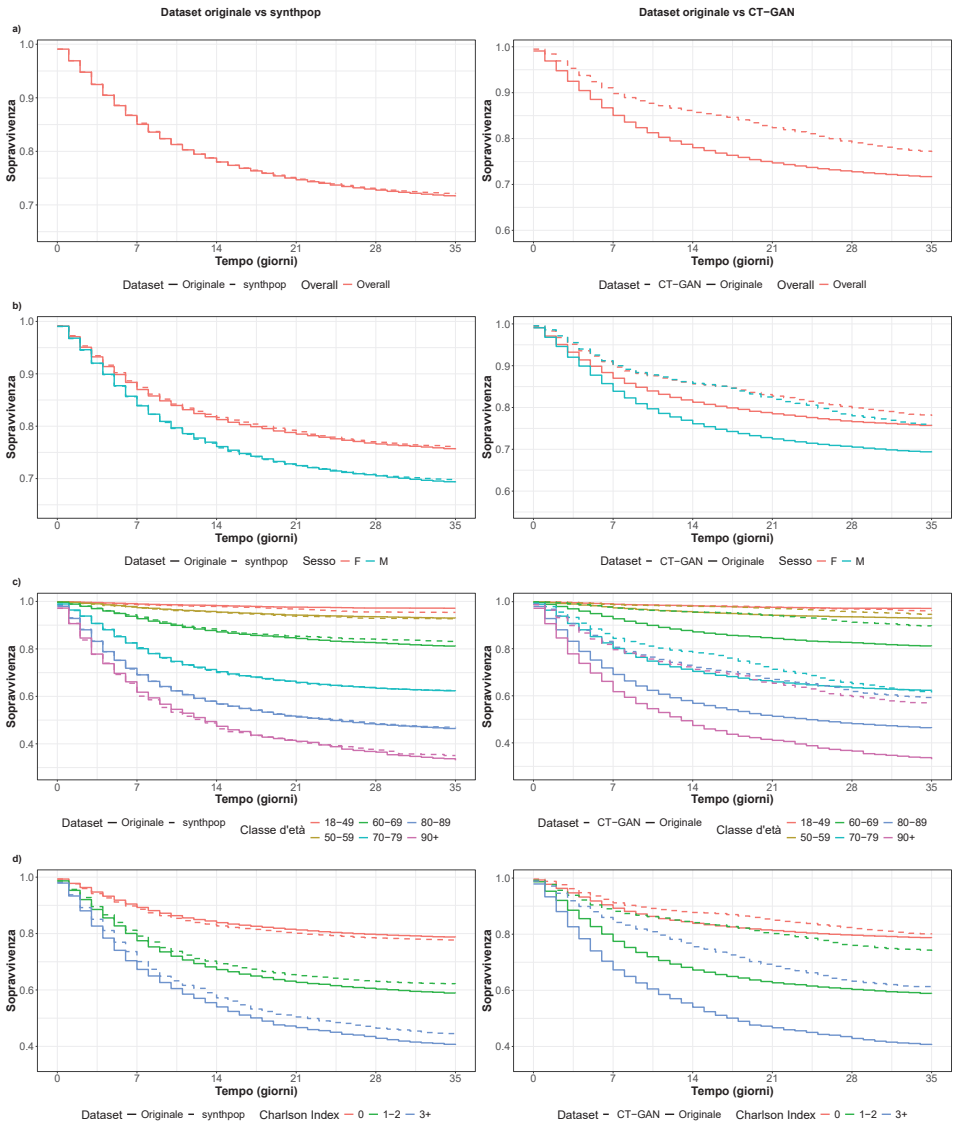
Il grado di sovrapposizione complessivo degli IC95% è stato valutato misurando la percentuale mediana e il range interquartile di sovrapposizione tra gli IC95%.

Le curve di sopravvivenza (Figura 3), sia per la popolazione overall che stratificata per sesso, classe d'età e *Charlson Index*, hanno mostrato che, per il dataset sintetico generato con synthpop, la maggior parte delle curve è risultata sovrapponibile a quelle ottenute sul dataset originale, contrariamente a quanto osservato nel dataset sintetico generato con CT-GAN.

La Figura 4a mostra che, nel caso di synthpop, le stime ottenute dai dati sintetici risultano generalmente vicine a quelle ottenute dai dati reali e i relativi intervalli di confidenza sono ampiamente sovrapposti. La percentuale mediana di sovrapposizione degli intervalli di confidenza è risultata pari al 75% (IQR 66%–95%). Al contrario, per CT-GAN (Figura 4b) le differenze nei risultati rispetto ai dati reali sono molto più marcate: la sovrapposizione mediana degli intervalli di confidenza è risultata pari a 0% (IQR 0%–6%).

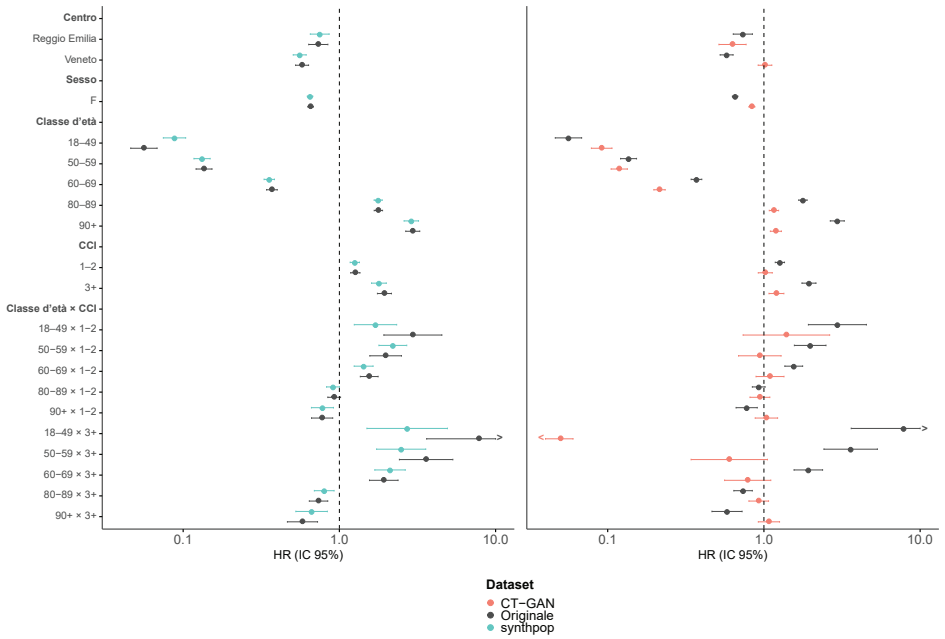
Questi risultati suggeriscono che, nel dataset analizzato, la maggiore similarità tra dati reali e sintetici osservata per synthpop nella valutazione dell'utilità generale si traduce anche in una maggiore capacità di preservare i risultati delle analisi statistiche rispetto a CT-GAN.

Figura 3
Confronto delle curve di Kaplan-Meier tra dati originali e dati sintetici generati con CT-GAN e synthpop



a) overall; b) sesso; c) classe d'età; d) *Charlson Index*
F: Femmine; M: Maschi; CT-GAN: *Conditional Tabular-Generative Adversarial Networks*

Figura 4
Forest plot del modello di regressione multipla a rischi proporzionali di Cox con interazione



a) dataset originale vs synthpop; b) dataset originale vs CT-GAN.
CT-GAN: *Conditional Tabular-Generative Adversarial Networks*; F: Femmine; CCI: *Charlson Comorbidity Index*; HR: *Hazard Ratio*; IC%: Intervallo di confidenza.

PRIVACY: RISCHIO DI RE-IDENTIFICAZIONE DI UN INDIVIDUO O DI INFORMAZIONI SENSIBILI

La terza valutazione ha riguardato la capacità dei dati sintetici di preservare la privacy. A tale scopo, è stato considerato il rischio che le informazioni contenute nei dati sintetici potessero essere utilizzate per riconoscere una persona reale o dedurre informazioni sensibili che la riguardano. Sono stati quindi valutati due tipi di rischio: *identity disclosure*, cioè il rischio di identificare uno specifico individuo, e *attribute disclosure*, cioè il rischio di dedurre correttamente un'informazione sensibile, note alcune caratteristiche dell'individuo.

La valutazione è stata effettuata ipotizzando uno scenario realistico in cui un potenziale intruso fosse a conoscenza di alcune informazioni di pubblico dominio o facilmente reperibili, chiamate *quasi-identifier*, e avesse accesso all'intero dataset sintetico. Le variabili *quasi-identifier* non contengono informazioni sensibili ma una combinazione univoca di queste varabili potrebbe permettere di identificare un individuo nel dataset. Questo rischio è noto come *identity disclosure*. Nel nostro studio

sono state identificate come *quasi-identifier* l'età, il sesso, il centro di provenienza e il ricovero in terapia intensiva. C'è poi un secondo tipo di variabili chiamate variabili target. Queste contengono informazioni sensibili, come la presenza di una comorbidità o l'assunzione di uno specifico farmaco. Anche senza identificare un individuo specifico nel dataset, si può inferire una caratteristica sensibile quando un particolare valore della variabile target è associato a una combinazione unica delle variabili *quasi-identifier*. Questo rischio è noto come *attribute disclosure*. Nel nostro studio abbiamo identificato 29 variabili target.

Al fine di quantificare questo rischio sia sul dataset reale che su quello sintetico, tra le tante misure disponibili in letteratura [26, 27], si è scelto di utilizzare in questo studio quelle raccomandate da Raab et al. [28].

In particolare, per la valutazione dell'*identity disclosure*, rispettivamente nel dataset originale e in quello sintetico, sono stati utilizzati gli indicatori:

- UiO (*Unique in Original*), che misura quanti soggetti nel dataset originale presentano una combinazione unica dei *quasi-identifier*, e quindi sono potenzialmente identificabili nel dataset originale sulla base di queste informazioni;
- RepU (*Replicated Uniques*), che misura quante combinazioni uniche dei *quasi-identifier* presenti nel dataset originale sono replicate come uniche anche nel sintetico, e potrebbero quindi consentire il riconoscimento di una persona reale a partire dal dataset sintetico.

Quindi, valori di RepU minori di UiO indicano che la quota di soggetti potenzialmente riconoscibili nei dati sintetici è inferiore rispetto a quella osservata nei dati originali, e quindi che i dati sintetici riducono il rischio di identificare un soggetto rispetto ai dati originali.

Per la valutazione dell'*attribute disclosure* rispettivamente nel dataset originale e in quello sintetico, sono stati utilizzati gli indicatori:

- Dorig (*Disclosive in Original*), che misura quante combinazioni di *quasi-identifier* sono associate ad un solo valore della variabile target (informazione sensibile), e quindi potenzialmente sufficienti per dedurla correttamente già nei dati originali;
- DiSCO (*Disclosive in Synthetic Correct Original*), che misura quanti soggetti, nel dataset sintetico, presentano una combinazione di *quasi-identifier* associata ad un solo valore della variabile target (informazione sensibile), e tale valore corrisponde a quello effettivamente osservato nei dati originali. In altre parole, valuta il rischio di dedurre correttamente un'informazione sensibile di una persona reale a partire dal dataset sintetico.

Quindi, valori della DiSCO più bassi di Dorig indicano un minore rischio di divulgazione di informazioni sensibili nei dati sintetici rispetto ai dati originali.

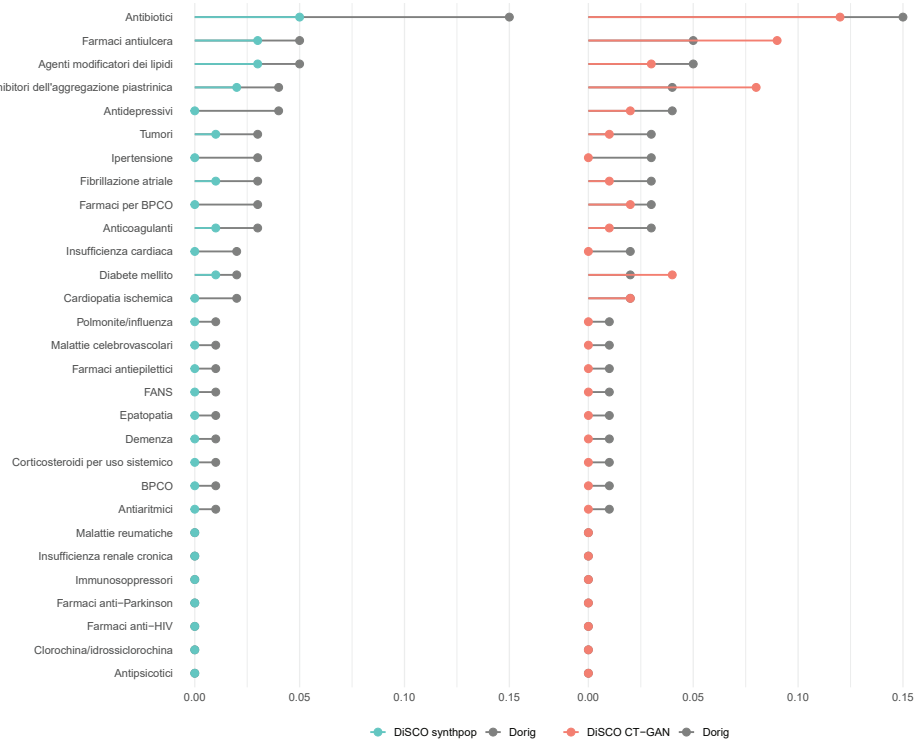
Per quanto riguarda la valutazione del rischio di *identity disclosure*, nel dataset originale, 99 persone su 42.926 (UiO=0,23%) presentavano una combinazione unica dei *quasi-identifier* (classe d'età, centro, sesso, ricovero in terapia intensiva). Questo risultato indica che nel dataset originale, costruito attraverso procedure di trasformazione e anonimizzazione dei dati, il possibile rischio di re-identificazione di soggetti specifici era già di per sé molto basso. Nei dataset sintetici, la percentuale di soggetti con combinazioni uniche dei *quasi-identifier* che rimangono uniche anche nei dati sintetici (RepU) è scesa a 0,06% (pari a 26 individui) con synthpop, e 0,05% (pari a 21 individui) con CT-GAN. La generazione di dati sintetici ha quindi ridotto il numero di soggetti potenzialmente riconoscibili rispetto ai dati originali, in maniera sostanzialmente equivalente con entrambi i metodi.

Per valutare il rischio di *attribute disclosure*, è stato verificato se la conoscenza dei *quasi-identifier* potesse consentire di dedurre una caratteristica sensibile, come l'uso di un farmaco o la presenza di una comorbidità, anche senza identificare un soggetto specifico.

La Figura 5 mostra il rischio di dedurre correttamente informazioni sensibili (diverse condizioni cliniche e uso di farmaci) a partire dalla conoscenza di alcune caratteristiche (*quasi-identifier*) nei dati originali (Dorig) e nei dati sintetici (DiSCO). Come si può osservare, per i due dataset sintetici generati con entrambi i metodi, tale rischio si è ridotto per quasi tutte le variabili target rispetto a quello osservato nei dati originali, in maniera sostanzialmente equivalente per entrambi i dataset sintetici. Infatti, per 22 variabili su 29 variabili target (76%), i valori di DiSCO sui dataset sintetici generati con i due metodi erano confrontabili.

Il valore di DiSCO più elevato è stato osservato nel dataset generato da CT-GAN relativamente all'uso di antibiotici (DiSCO=0,12%). Ciò significa che, nello scenario considerato, meno di 1 soggetto su 800 presentava una combinazione di *quasi-identifier* che avrebbe consentito di dedurre correttamente, a partire dai dati sintetici, il valore di una variabile sensibile osservato nei dati reali.

Figura 5
Misure di attribute disclosure (Dorig e DiSCO) per dataset originale e dataset sintetici generati con synthpop e CT-GAN



CT-GAN: Conditional Tabular-Generative Adversarial Network; HIV: *Human Immunodeficiency Virus*; BPCO: Broncopneumopatia cronica ostruttiva; FANS: Farmaci antinfiammatori non steroidei; DiSCO: *Disclosive in Synthetic Correct Original*; Dorig: Disclosive in Original
N.B. L'analisi è stata effettuata escludendo la modalità "No", relativa all'assenza di utilizzo del farmaco o della comorbidità, poiché la sua divulgazione non costituirebbe un rischio per la privacy.

RISPOSTA ALLA DOMANDA

I dati sintetici possono rappresentare una potenziale opportunità per favorire l'uso e la condivisione dei dati real-world in farmacoepidemiologia, ma il loro impiego deve essere valutato caso per caso e la loro utilità dipende dalla capacità di raggiungere un adeguato equilibrio tra fedeltà ai dati reali e tutela della privacy. I risultati del nostro studio mostrano infatti che è possibile generare dati sintetici che, nelle condizioni considerate, presentino un rischio molto basso di identificazione degli individui e delle loro informazioni sensibili e, allo stesso tempo, riproducano in buona misura le caratteristiche e le relazioni presenti nei dati originali. Tuttavia, i due metodi considerati hanno mostrato prestazioni differenti nelle diverse dimensioni valutate.

Il metodo CT-GAN ha mostrato una capacità di protezione della privacy comparabile a synthpop, ma una minore capacità di riprodurre le caratteristiche e le relazioni presenti nel dataset originale. Synthpop ha invece fornito, nelle condizioni dello studio, il miglior equilibrio complessivo tra utilità dei dati sintetici, tutela della privacy e tempo di elaborazione. Infatti, ha mostrato una maggiore capacità di riprodurre sia le distribuzioni delle singole variabili sia le relazioni tra esse e, soprattutto, di fornire risultati analitici comparabili a quelli ottenuti sui dati reali. È inoltre risultato più rapido e, basandosi su modelli statistici tradizionali, più facile da interpretare, documentare e riprodurre.

Dal punto di vista della privacy, entrambi i metodi hanno prodotto dataset con un rischio di identificazione degli individui e di divulgazione di informazioni sensibili inferiore rispetto al dataset originale. Tuttavia, un risultato particolarmente importante è che il dataset di partenza presentava già un livello di rischio contenuto, grazie ai processi di anonimizzazione e trasformazione applicati nella sua costruzione nell'ambito del sistema di analisi distribuite TheShinISS. Pertanto, prima di procedere alla generazione di dati sintetici, dovrebbe essere valutato il livello di protezione già garantito dal dataset di partenza e il beneficio aggiuntivo ottenuto attraverso la generazione di dati sintetici.

Riferimenti bibliografici

[1] International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). General Principles on Planning, Designing, Analysing, and Reporting of Non-interventional Studies That Utilise RealWorld Data for Safety Assessment of Medicines. 2025 [ultimo accesso 17/09/2026]. Disponibile su: https://database.ich.org/sites/default/files/ICH_M14_Step4_Final_Guideline_2025_0905.pdf

[2] Council for International Organizations of Medical Sciences (CIOMS) Working Group XIII. Real-world data and real-world evidence in regulatory decision making. CIOMS; 2024 [ultimo accesso 17/09/2026]. Disponibile su: https://cioms.ch/publications/product/real-world-data-and-real-world-evidence-in-regulatory-decision-making/

[3] Hernandez RK, Critchlow CW, Dreyer N, et al. Advancing Principled Pharmacoepidemiologic Research to Support Regulatory and Healthcare Decision Making: The Era of Real-World Evidence. Clin Pharmacol Ther. aprile 2025;117(4):927–37.

[4] Sentinel Initiative. 2025 [ultimo accesso 17/09/2026]. Disponibile su: https://www.sentinelinitiative.org/

[5] Canadian Network for Observational Drug Effects Studies (CNODES). CNODES. 2025 [ultimo accesso 17/09/2026]. Disponibile su: https://www.cnodes.ca/

[6] European Medicines Agency (EMA). Data Analysis and Real World Interrogation Network (DARWIN EU). DARWIN EU. 2021 [ultimo accesso 17/09/2026]. Disponibile su: https://www.ema.europa.eu/en/about-us/how-we-work/data-regulation-big-data-other-sources/real-world-evidence/data-analysis-real-world-interrogation-network-darwin-eu

[7] Thurin NH, Pajouheshnia R, Roberto G, et al. From Inception to ConcePTION: Genesis of a Network to Support Better Monitoring and Communication of Medication Safety During Pregnancy and Breastfeeding. Clin Pharmacol Ther. 2022;111(1):321-331. doi:10.1002/cpt.2476

[8] Base de Datos para la Investigación Farmacoepidemiológica en Atención Primaria. BIFAP. [ultimo accesso 17/09/2026]. Disponibile su: https://www.bifap.org/learn-more?lang=en

[9] Massari M, Spila Alegiani S, Da Cas R, Menniti Ippolito F. TheShinISS: un applicativo open-source per la conduzione di analisi distribuite in studi di farmacoepidemiologia di tipo multi-database. Boll Epidemiol Naz 2020; 1(2):39-45. DOI: https://doi.org/10.53225/BEN_006.

[10] Spila Alegiani S, Massari M, Trifirò G, Da Cas R, Menniti Ippolito F. ITA-COVID19: rete multiregionale per la conduzione di studi osservazionali sull'associazione tra esposizione a farmaci e/o vaccini e COVID-19. Epidemiol Prev. 2020;44(5–6):426–7.

[11] Trifirò G, Isgrò V, Ingrasciotta Y, et al. Large-Scale Postmarketing Surveillance of Biological Drugs for Immune-Mediated Inflammatory Diseases Through an Italian Distributed Multi-Database Healthcare Network: The VALORE Project. BioDrugs. novembre 2021;35(6):749–64.

[12] Belleudi V, Rosa AC, Finocchietti M, et al. An Italian multicentre distributed data research network to study the use, effectiveness, and safety of immunosuppressive drugs in transplant patients: Framework and perspectives of the CESIT project. Front Pharmacol. 15 settembre 2022;13:959267.

[13] Spila Alegiani S, Morciano C, Menniti-Ippolito F, et al. Postmarketing observational study on the safety of 2021/2022 and 2022/2023 influenza vaccination campaigns in Italy: TheShinISS-Vax|Flu study protocol. BMJ Open. 1 agosto 2023;13(8):e069858.

[14] Scheibner J, Ienca M, Kechagia S, et al. Data protection and ethics requirements for multisite research with health data: a comparative examination of legislative governance frameworks and the role of data protection technologies†. Journal of Law and the Biosciences. 25 luglio 2020;7(1):lsaa010.

[15] Barone-Adesi F, Bisceglia L, Venturini F, Maglio G. Protezione dati personali e uso dei dati epidemiologici: novità legislative e questioni aperte - Quotidiano Sanità. quotidianosanita.it. [ultimo accesso 17/09/2026]. Disponibile su: https://www.quotidianosanita.it/lettere-al-direttore/articolo.php?articolo_id=122786

[16] Garante per la protezione dei dati personali. Codice Privacy e Normativa italiana. Codice Privacy e Normativa italiana. [ultimo accesso 17/09/2026]. Disponibile su: https://www.garanteprivacy.it/codice

[17] DDL 132/2025 Disposizioni e deleghe al Governo in materia di intelligenza artificiale. Gazzetta Ufficiale, G.U. n. 223 del 25 settembre 2025; rettifica in G.U. n. 242 del 17 ottobre 2025 set 23, 2025. [ultimo accesso 17/09/2026]. Disponibile su: https://www.gazzettaufficiale.it/eli/id/2025/09/25/25G00143/sg

[18] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). giu 13, 2024. [ultimo accesso 17/09/2026]. Disponibile su: http://data.europa.eu/eli/reg/2024/1689/oj

[19] FDA. FDA Glossary on Digital Health and Artificial Intelligence. Digital Health and Artificial Intelligence Glossary – Educational Resource. FDA; 2025 [ultimo accesso 17/09/2026]. Disponibile su: https://www.fda.gov/science-research/artificial-intelligence-and-medical-products/fda-digital-health-and-artificial-intelligence-glossary-educational-resource

[20] Wang, E.H., Myles, P., Foraker, R. et al. Crossing borders securely: synthetic data and federated networks for privacy-preserving access to real-world data and emerging use cases. npj Digit. Med. 8, 758 (2025). Disponibile su: https://doi.org/10.1038/s41746-025-02126-8

[21] Mayer F, Fazio ML, Cutillo M, Stendardo G, Spila Alegiani S, Trifirò G, Massari M. Confronto tra metodi di generazione di dati sintetici in ambito farmacoepidemiologico: conciliare la riproducibilità delle stime con la tutela della privacy. Epidemiol Prev 2026; 50 (3): 279-289. Disponibile su: https://doi.org/10.19191/EP26.3.A992.059

[22] Nowok B, Raab GM, Dibben C, Snoke J, Lissa C van, Pater L. synthpop: Generating Synthetic Versions of Sensitive Microdata for Statistical Disclosure Control. 2025 [ultimo accesso 17/09/2026]. Disponibile su: https://cran.r-project.org/web/packages/synthpop/index.html

[23] Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. Modeling Tabular data using Conditional GAN. In: Advances in Neural Information Processing Systems. Curran Associates, Inc.; 2019 [ultimo accesso 17/09/2026]. Disponibile su: https://proceedings.neurips.cc/paper/2019/hash/254ed7d2de3b23ab10936522dd547b78-Abstract.html

[24] Ferroni E, Giorgi Rossi P, Spila Alegiani S, et al. Survival of Hospitalized COVID-19 Patients in Northern Italy: A Population-Based Cohort Study by the ITA-COVID-19 Network. Clin Epidemiol. 8 dicembre 2020;12:1337–46.

[25] Raab GM, Nowok B, Dibben C. Assessing, visualizing and improving the utility of synthetic data. arXiv; 2021 [ultimo accesso 17/09/2026]. Disponibile su: http://arxiv.org/abs/2109.12717

[26] Raab GM. Privacy risk from synthetic data: practical proposals. arXiv; 2025 [ultimo accesso 17/09/2026]. Disponibile su: http://arxiv.org/abs/2409.04257

[27] Pilgram L, Dankar FK, Drechsler J, et al. A consensus privacy metrics framework for synthetic data. Patterns. ottobre 2025;6(10):101320.

[28] Raab GM, Nowok B, Dibben C. Practical privacy metrics for synthetic data. arXiv; 2025 [ultimo accesso 17/09/2026]. Disponibile su: http://arxiv.org/abs/2406.16826

[29] Miletic M, Sariyar M. Utility-based Analysis of Statistical Approaches and Deep Learning Models for Synthetic Data Generation With Focus on Correlation Structures: Algorithm Development and Validation. JMIR AI. 20 marzo 2025;4(1):e65729.

[30] Guillaudeux M, Rousseau O, Petot J, et al. Patient-centric synthetic data generation, no reason to risk re-identification in biomedical data analysis. npj Digit Med. 10 marzo 2023;6(1):37.