Dati sintetici per accelerare gli studi clinici: il ruolo dell'AI

Saverio D'Amico - Humanitas/Train
Daniele Di Ianni - Roche

laboratorio SANITÀ 2030
27-28 GIUGNO 2024 FRIULI VENEZIA GIULIA
VILLA MANIN CODROIPO - UDINE
L'INTELLIGENZA ARTIFICIALE IN SANITÀ
TRA ETICA E APPLICAZIONE

About Humanitas AI Center for Health

#sanita2030

www.sanita2030.it

laboratorio SANITÀ 2030
27-28 GIUGNO 2024 FRIULI VENEZIA GIULIA
VILLA MANIN CODROIPO - UDINE
L'INTELLIGENZA ARTIFICIALE IN SANITÀ
TRA ETICA E APPLICAZIONE

About Train

Train is a spin-out from Humanitas Research Hospital founded in 2023.

Our competitive advantage is the **clinical validation** and the **scientific evidence** achieved through our collaborations with **domain experts**.

Wired Health 2023, Co-founders Saverio D'Amico and Prof. Matteo Della Porta

laboratorio SANITÀ 2030
27-28 GIUGNO 2024 FRIULI VENEZIA GIULIA
VILLA MANIN CODROIPO - UDINE
L'INTELLIGENZA ARTIFICIALE IN SANITÀ
TRA ETICA E APPLICAZIONE

Low complexity vs high complexity in medicine

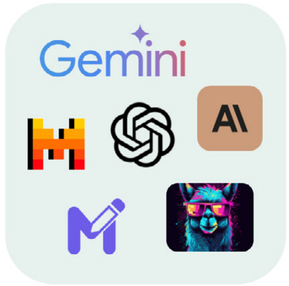
General doctor
Low medical complexity

Specialized doctor, MD, PhD...
High medical complexity

Unmet needs will increase due to the aging population

laboratorio SANITÀ 2030
27-28 GIUGNO 2024 FRIULI VENEZIA GIULIA
VILLA MANIN CODROIPO - UDINE
L'INTELLIGENZA ARTIFICIALE IN SANITÀ
TRA ETICA E APPLICAZIONE

AI in medicine



Generalistic model
Low medical complexity

Domain-trained models
High medical complexity

#sanita2030

www.sanita2030.it

laboratorio SANITÀ 2030
27-28 GIUGNO 2024 FRIULI VENEZIA GIULIA
VILLA MANIN CODROIPO - UDINE
L'INTELLIGENZA ARTIFICIALE IN SANITÀ
TRA ETICA E APPLICAZIONE

2021 WHO guidance on ethics and governance of AI for health

We have to address three important topics for a right deployment of AI in healthcare:

Transparency of models
We have to provide a good understanding of the models (interpretability and explainability).

Reliability of models
The main vulnerabilities of AI models are related to the lack of generalizability. Therefore, extensive, independent validation of generated AI models is required.

Protection of data and data sharing
Innovative technologies such as federated learning procedures for data collection and analysis (without moving sensitive medical data from their original locations) are required to facilitate clinical implementability of AI solutions.

#sanita2030

www.sanita2030.it

laboratorio SANITÀ 2030
27-28 GIUGNO 2024 FRIULI VENEZIA GIULIA
VILLA MANIN CODROIPO - UDINE
L'INTELLIGENZA ARTIFICIALE IN SANITÀ
TRA ETICA E APPLICAZIONE

The opportunity of synthetic data

Generative AI in healthcare holds promise by generating synthetic information to improve basic research, diagnosis and drug discovery/repurposing.

Synthetic data are artificially generated data that mimics real data by capturing the statistical distribution and dependencies of the original datasets.

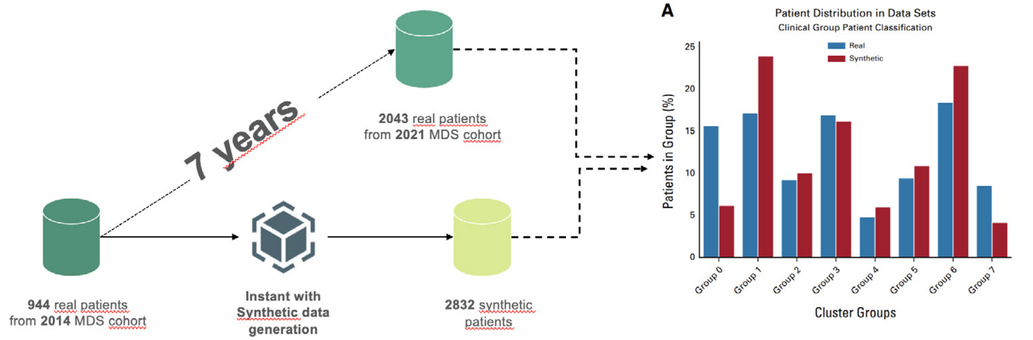
The adoption of synthetic data in healthcare is driven by the scarcity of openly available data, concerns about patient privacy preservation, and the potential risk of re-identification associated with traditional methods, as well as the need to address limited availability of data for training AI models.

What if the control arm of clinical trials currently conducted using standard treatments or placebos were replaced by a synthetic cohort based on historical data?

#sanita2030
www.sanita2030.it

L'INTELLIGENZA ARTIFICIALE IN SANITÀ
TRA ETICA E APPLICAZIONE

We provided evidence that Generative AI accelerate clinical research.
By generating synthetic data from a small cohort of patients available in 2014, we were able to recapitulate the definition of a molecular prognostic score as described in real cohorts 7 years later (2021).



Source: D'Amico et al, J Clin Oncol CCI, 30 June 2023

#sanita2030

www.sanita2030.it

Benefits of synthetic data in clinical trials

01 **Less patients to recruit**

Synthetic data simulates patient data and outcomes, minimizing the need for recruiting actual patients, thus halving recruitment numbers in clinical trials.

02 **Saving time and cost of clinical trials**

Synthetic data expedites trial design, reduces recruitment time and costs, potentially enhancing success rates by optimizing resource allocation and efficiency.

03 **Reduce the patient burden of participation**

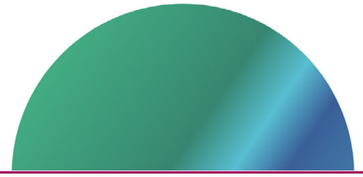
Synthetic data lessens the need for real patient involvement, easing burdens by minimizing recruitment demands and participation requirements.

04 **Mitigates the long-standing ethical concern**

Synthetic data addresses ethical concerns by reducing reliance on real patient data, safeguarding privacy and minimizing potential risks.

05 **Minimizing the need for placebo patient enrollment**

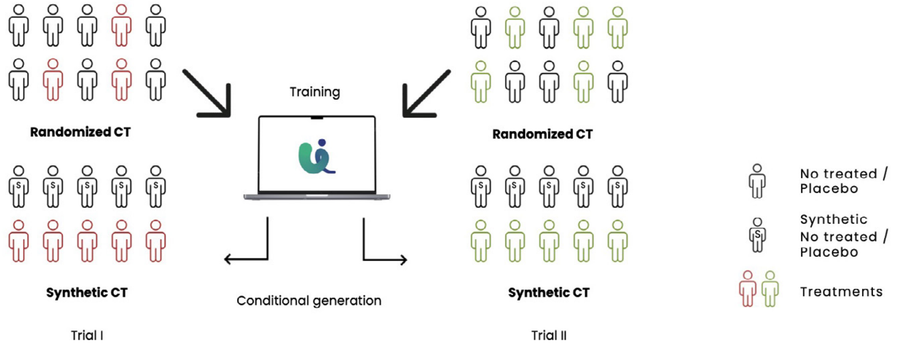
Synthetic data replicates control group characteristics, reducing the necessity for placebo patients, streamlining trials while maintaining scientific validity.



laboratorio SANITÀ 2030
27-28 GIUGNO 2024 FRIULI VENEZIA GIULIA
VILLA MANIN CODROIPO - UDINE
L'INTELLIGENZA ARTIFICIALE IN SANITÀ
TRA ETICA E APPLICAZIONE

How to build synthetic control arms

To construct a synthetic control arm, it is necessary to aggregate historical data from conducted studies of patients with similar characteristics.



#sanita2030
www.sanita2030.it

laboratorio SANITÀ 2030
27-28 GIUGNO 2024 FRIULI VENEZIA GIULIA
VILLA MANIN CODROIPO - UDINE
L'INTELLIGENZA ARTIFICIALE IN SANITÀ
TRA ETICA E APPLICAZIONE

Real case study

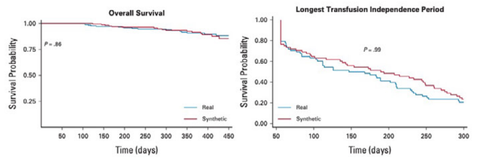
Synthetic patients control arms in luspatercept clinical trial for treatment of myelodysplastic syndrome (MDS)

We investigated the possibility to use a synthetic data set as a comparison group in a clinical trial. We therefore aimed to replicate a real patient cohort from a multicenter study including 187 patients with MDS who were treated with luspatercept.

Eligible patients were:
- Age 18 years or older
- Had an MDS with ring sideroblasts
- Were receiving regular red blood cells transfusions
- Were refractory to erythropoiesis-stimulating agent therapy.

Primary end point was transfusion independence (TI) for ≥8 weeks during weeks 1-24. Key secondary end point was TI for ≥12 weeks during both weeks 1-24 and 1-48.

We generated a synthetic cohort (N=187) from the patients included in the study using all data for training, and we compared the synthetic endpoints with the original study results.
All the characteristics and metrics of the synthetic cohort were comparable with respect to the original data set, with high efficient coefficient of privacy preservability.



Kaplan-Meier survival probability curves compared for real and synthetic patients' overall survival. (B) Kaplan-Meier curves of longest transfusion independence period for real and synthetic patients. The P values of the log-rank test are calculated, confirming the hypothesis of no difference in survival probabilities between real and synthetic cohorts.

Clinical endpoint	Real data	Synthetic data	P-value
RBC-TI ≥ 8 weeks 1-24	56 (31.5)	56 (31.5)	1.0
Longest transfusion indepence period, weeks, median (range)	195 (56-490)	280 (56-490)	< .05
RBC-TI ≥ 8 weeks 1-48	68 (38.2)	61 (34.3)	.50
RBC-TI ≥ 12 weeks 1-24	36 (20.2)	41 (23.0)	.60
RBC-TI ≥ 12 weeks 1-48	51 (28.7)	46 (25.8)	.63
Reduction ≥ 4 RBC	62 (34.8)	63 (35.4)	1.0
Reduction ≥ 50%	77 (43.3)	72 (40.4)	.66
AML evolution	4 (2.2)	6 (3.4)	.75
Discontinued patients	74 (41.6)	82 (46.1)	.64

Study end point comparison between real and synthetic cohorts. RBC-TI, rate of red blood cell transfusion independence.

Source: D'Amico et al, J Clin Oncol CCI, 2023

#sanita2030
www.sanita2030.it

laboratorio SANITÀ 2030
27-28 GIUGNO 2024 FRIULI VENEZIA GIULIA
VILLA MANIN CODROIPO - UDINE
L'INTELLIGENZA ARTIFICIALE IN SANITÀ
TRA ETICA E APPLICAZIONE

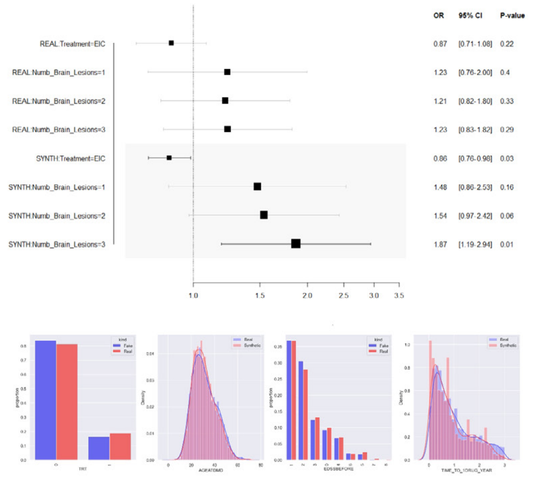
Real case study

Synthetic Data Generation by Artificial Intelligence in multiple sclerosis (MS) applied to the Italian MS Register

To test the capability of advanced SD generation methods to replicate results of a comparative effectiveness analysis between an early initiation treatment (EIT) with high-efficacy disease-modifying therapies (HE-DMTs) and escalation strategies (ESC) on the first progression independent of relapse activity (PIRA) event risk performed on real data from the Italian MS Register (IMSR).

Our results demonstrate that:

- The use of advanced SD generation methods may confirm and improve the results obtained by using real data.
- An early start of HE-DMTs was confirmed to be more effective in reducing the risk of reaching a first PIRA event, by using synthetic augmented dataset.



Source: Iaffaldano et al, abstract for ECTRIMS, 2024

#sanita2030

www.sanita2030.it

laboratorio SANITÀ 2030
27-28 GIUGNO 2024 FRIULI VENEZIA GIULIA
VILLA MANIN CODROIPO - UDINE
L'INTELLIGENZA ARTIFICIALE IN SANITÀ
TRA ETICA E APPLICAZIONE

Real case study

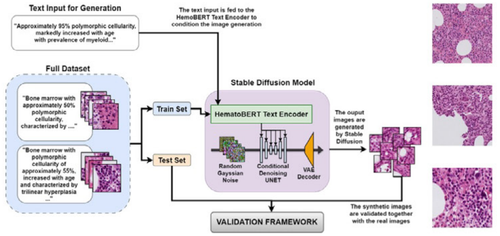
Synthetic histopathological images generation from medical text reports with AI to accelerate research and improve clinical outcomes in hematology

Hematological malignancies are rare and complex diseases and as a consequence, multimodal data (ranging from clinical and genomic information to images) are required to improve diagnosis, prognosis and personalized treatments.

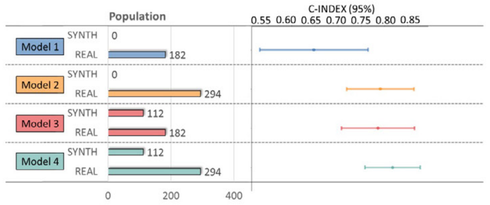
However, collecting all these layers of information is challenging, in particular when collecting cytological and histological images from the bone marrow (BM) reproducing disease morphologic features.

Synthetic data generation by Artificial Intelligence (AI) can circumvent these issues by generating images conditioned from textual inputs (i.e. reports from pathologists), which are widely available and contain many useful clinical information.

AI generated images preserve properties of real-world images, replicating cells morphological features relevant to identify hematological diseases and their clinical status. This approach based on widely available textual data allows effective data augmentation and effortless data sharing, thus accelerating and improving precision medicine research in hematology.



Overview of the Stable Diffusion fine-tuning process and the clinical and statistical validation process.



Cox's proportional hazards model to predict individual probability of overall survival in patients affected with myeloid neoplasms.

Source: Asti et al, Blood, 2023

Synthetic control arm for refractory metastatic colorectal cancer: the no placebo initiative

Since 2022, **FDA** and the **scientific community** are forming an alliance for the no placebo initiative to use external comparison arms to study new therapies for gastrointestinal stromal tumor and other rare cancers, facilitating drug trials and regulatory approvals.

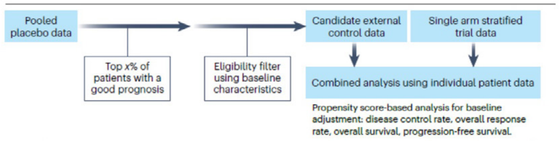


Fig. 1 | Three-step analysis for no-placebo initiative. First, participants enrolled in trials with placebo arms will be selected based on compatible patient demographics and key characteristics. These data will form the synthetic control arm. Second, patients in the top percentile for overall survival will be extracted from the synthetic control arm. Third, the synthetic control arm will be compared with patients in the trial, using propensity scored-based analysis.

Source: A synthetic control arm for refractory metastatic colorectal cancer: the no placebo initiative, Nature, 2022

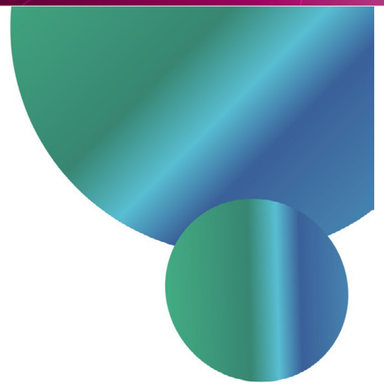
Synthema: synthetic hematological over federated learning frameworks

Synthema, as a European consortium founded in 2023 (www.synthema.eu) consisting of academic and industrial partners, is building a Synthetic Validation Framework (SVF) for privacy, utility and clinical evaluation of synthetic data in healthcare. SVF is being submitted to the EMA for approval and design improvement to expedite and extend the use of synthetic data as control arms in clinical trials.

Synthetic data in the AI Act

Data Governance obligations for High Risks Systems:
Art. 10(5) lett. a: "(...) the bias detection and correction cannot be effectively fulfilled by processing other data, including synthetic or anonymised data;"

Further Processing of Personal Data for Developing Certain AI Systems in the Public Interest in the AI Regulatory Sandbox:
Art. 59(1) lett. b: " the data processed are necessary for complying with one or more of the requirements referred to in Chapter III, Section 2 where those requirements cannot be effectively fulfilled by processing anonymised, synthetic or other non-personal data;"



SYNTHEMA

Delitti in materia di violazione del diritto d'autore (Art. 25-novies, D.Lgs. n. 231/2001) [articolo aggiunto dalla L. n. 99/2009]

- Messa a disposizione del pubblico, in un sistema di reti telematiche, mediante connessioni di qualsiasi genere, di un'opera dell'ingegno protetta, o di parte di essa (art. 171, legge n.633/1941 comma 1 lett. a) bis)
- Reati di cui al punto precedente commessi su opere altrui non destinate alla pubblicazione qualora ne risulti offeso l'onore o la reputazione (art. 171, legge n.633/1941 comma 3)
- Abusiva duplicazione, per trarne profitto, di programmi per elaboratore; importazione, distribuzione, vendita o detenzione a scopo commerciale o imprenditoriale o concessione in locazione di programmi contenuti in supporti non contrassegnati dalla SIAE; predisposizione di mezzi per rimuovere o eludere i dispositivi di protezione di programmi per elaboratori (art. 171-bis legge n.633/1941 comma 1)
- Riproduzione, trasferimento su altro supporto, distribuzione, comunicazione, presentazione o dimostrazione in pubblico, del contenuto di una banca dati; estrazione o reimpiego della banca dati; distribuzione, vendita o concessione in locazione di banche di dati (art. 171-bis legge n.633/1941 comma 2)
- Abusiva duplicazione, riproduzione, trasmissione o diffusione in pubblico con qualsiasi procedimento, in tutto o in parte, di opere dell'ingegno destinate al circuito televisivo, cinematografico, della vendita o del noleggio di dischi, nastri o supporti analoghi o ogni altro supporto contenente fonogrammi o videogrammi di opere musicali, cinematografiche o audiovisive assimilate o sequenze di immagini in movimento; opere letterarie, drammatiche, scientifiche o didattiche, musicali o drammatico musicali, multimediali, anche se inserite in opere collettive o composite o banche dati; riproduzione, duplicazione, trasmissione o diffusione abusiva, vendita o commercio, cessione a qualsiasi titolo o importazione abusiva di oltre cinquanta copie o esemplari di opere tutelate dal diritto d'autore e da diritti connessi; immissione in un sistema di reti telematiche, mediante connessioni di qualsiasi genere, di un'opera dell'ingegno protetta dal diritto d'autore, o parte di essa (art. 171-ter legge n.633/1941)
- Mancata comunicazione alla SIAE dei dati di identificazione dei supporti non soggetti al contrassegno o falsa dichiarazione (art. 171-septies legge n.633/1941)
- Fraudolenta produzione, vendita, importazione, promozione, installazione, modifica, utilizzo per uso pubblico e privato di apparati o parti di apparati atti alla decodificazione di trasmissioni audiovisive ad accesso condizionato effettuate via etere, via satellite, via cavo, in forma sia analogica sia digitale (art. 171-octies legge n.633/1941).

Torna all'inizio