FACULTY OF MEDICINE AND HEALTH SCIENCES

Synthetic Data for Research Acceleration (SYNDARA) collaborative research unit

672P

Decruyenaere A, Dehaene H, Rabaey P, Polet C, Decruyenaere J, Demeester T, Vansteelandt S, Rottey S

INFERENCE FAILURE WITH SYNTHETIC ARMS: EMPIRICAL APPLICATION TO PHASE 3 ONCOLOGY TRIALS

BACKGROUND

Open sharing of synthetic trial data could facilitate secondary analyses and synthetic control arm development. This study investigates whether the conclusions of phase 3 oncology trials would have changed if synthetic data had been used in the same analysis instead of original data.

Synthetic data

Synthetic data are artificial data created by a generative model trained on the original data. They replicate the statistical properties of the original data without disclosing individual records. As such, synthetic data might be able to replace the original data in statistical analysis, while preserving the privacy of the individuals in the original data and thereby enabling data sharing.

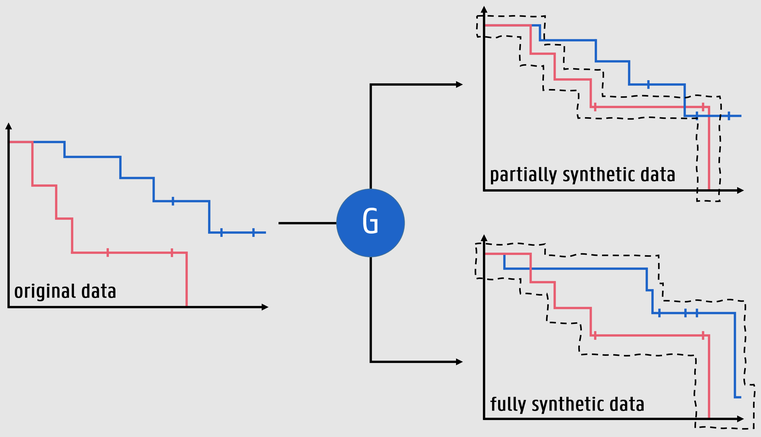
Unfortunately, synthetic data analysis may induce bias (if the generative model fails to capture the underlying data structure) and imprecision (as synthetic data are not real observations but predictions from a generative model that is estimated itself), which should ideally be acknowledged in the analysis. However, it is not yet entirely clear how this should be done.

METHODS

This study included 128 experimental vs. control treatments for solid tumors from 115 phase 3 randomized clinical trials (RCTs) published in 7 high-impact journals in 2023.

Original individual patient data were reconstructed from the published Kaplan-Meier (KM) curves of overall survival (OS) in the intention-to-treat population.

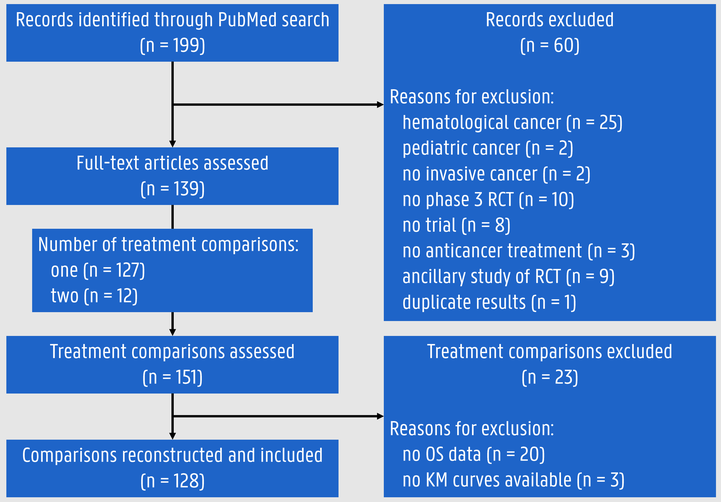
Either both arms (**fully synthetic**; retaining original sample size) or only the control arm (**partially synthetic**; retaining original control arm size) were synthesized using statistical models (**Weibull** and **Synthpop**) and a deep generative model (**SurvivalGAN**).



Generative models

Synthetic data can be created using statistical or deep generative models. The latter offer more flexibility in capturing the underlying data structure (potentially safeguarding against bias), albeit at the expense of more uncertainty (since more model parameters need to be estimated).

RESULTS



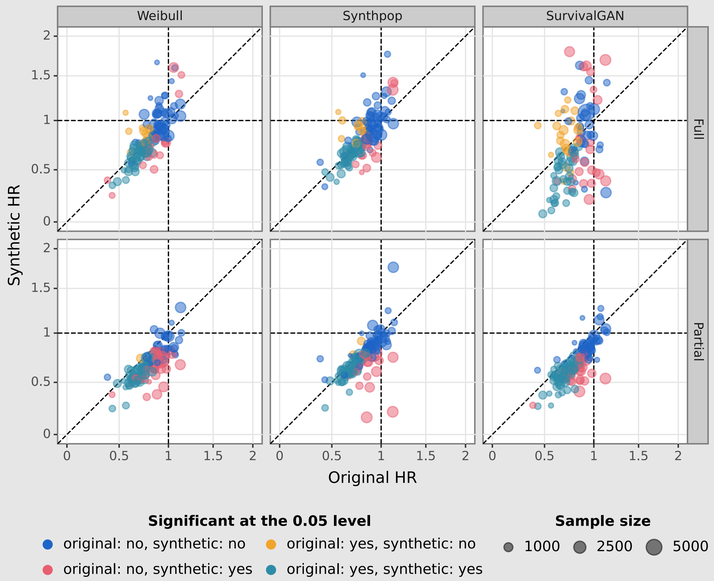
Original trial characteristics (n = 128).

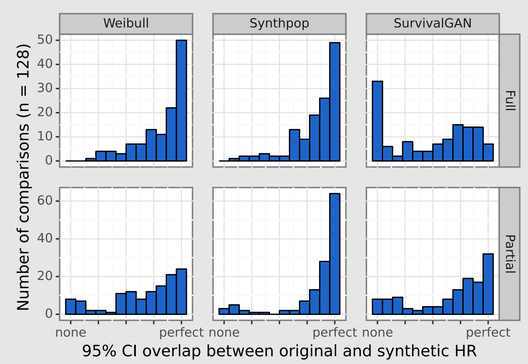
Sample size	571 (82-5637)	**Tumor type**	
Journal		breast	14 (10.9%)
Ann Oncol	11 (8.6%)	gastrointestinal	23 (18.0%)
J Clin Oncol	48 (37.5%)	genitourinary	27 (21.1%)
JAMA	3 (2.3%)	gynaecological	17 (13.3%)
JAMA Oncol	4 (3.1%)	head & neck	11 (8.6%)
Lancet	21 (16.4%)	melanoma	7 (5.5%)
Lancet Oncol	20 (15.6%)	other	1 (0.8%)
N Engl J Med	21 (16.4%)	thoracic	28 (21.9%)
HR for OS	0.80 (0.39-1.14)	**Significant OS**	56 (43.8%)

Categorical variables are presented as count (proportion) and continuous variables as median (range).

Utility metrics

The (inferential) utility is assessed here by the degree of confidence interval (CI) probability overlap (0: none, 1: perfect) for the hazard ratio (HR) obtained in the original and synthetic trial. The HR of a synthetic trial is considered truly positive when the HR is also significant (with same direction) in the original trial, and falsely positive when it was only significant in the synthetic trial or also significant in the original trial but with different direction.





Inferential utility of synthetic trials.

	True positive rate (%)			False positive rate (%)		
	Weibull	**Synthpop**	**SurvivalGAN**	**Weibull**	**Synthpop**	**SurvivalGAN**
Full	79.6	80.4	61.8	26.5	25.0	50.0
Partial	98.2	96.4	100.0	55.9	27.8	41.7

CONCLUSIONS

While synthetic data hold great promise for privacy protection, their statistical analysis poses significant challenges that necessitate innovative solutions. Synthetic data generation is known to induce considerable bias and imprecision into synthetic data analyses, compromising their (inferential) utility as opposed to original data analyses. This bias and uncertainty can be substantial enough to complicate fundamental calculations like p-values and confidence intervals, with no straightforward remedy currently available. In particular, we show that naive inference from phase 3 oncology RCTs with synthetic arms may lead to overly optimistic or even wrong conclusions. Before publishing synthetic trial data, it is therefore essential to develop statistical inference tools for such data.

GHENT UNIVERSITY

UZ GENT

SYNDARA

Alexander.Decruyenaere@UGent.be
https://syndara.ai

DOI none