scientific reports

Check for updates

OPEN

Challenges of analyzing synthetic tabular data generated from 115 phase 3 oncology trials

Alexander Decruyenaere[1,2]✉, Christiaan Polet[1], Johan Decruyenaere[1], Heidelinde Dehaene[1], Paloma Rabaey[1,3], Thomas Demeester[1,3], Stijn Vansteelandt[1,4] & Sylvie Rottey[1,2]

Synthetic data have been proposed to facilitate data sharing in privacy-sensitive contexts, including clinical trials. It remains unclear, however, how well original treatment effect estimates can be replicated in synthetic data analyses. Therefore, we synthesized and reanalyzed 128 treatment comparisons from 115 phase 3 randomized oncology trials using sixteen different generative models. Our findings demonstrate that careful methodological choices are essential for drawing valid statistical conclusions from synthetic data analyses. Naive analyses frequently yield falsely significant treatment effects, occurring in up to half of the trials created by deep generative models. Correcting standard errors to reflect the uncertainty inherent in synthetic data generation reduces these false positives, but primarily suffices for trials generated by parametric models. Although this correction entails some power loss, it can be mitigated by increasing the synthetic sample size. Thus, at present, large synthetic trials generated by parametric models and analyzed with corrected standard errors are more likely to preserve inferential utility. Advancing valid statistical inference from synthetic data created by deep generative models remains an important direction for future research.

Keywords Clinical trial, Generative model, Oncology, Statistical inference, Synthetic data, Utility

The call from peers, funders, journals, and regulators for data sharing in clinical research has never been louder[1–6]. Greater access to raw data enables researchers to validate the original findings, perform secondary analyses, test alternative hypotheses, and conduct individual patient data meta-analyses, ultimately contributing to open science, avoiding duplication of efforts and accelerating research. Nevertheless, legal obstacles remain to be overcome before allowing public access to privacy-sensitive health data. Many data privacy regulations, including the General Data Protection Regulation and the Health Insurance Portability and Accountability Act, rightfully impose restrictions on accessing and sharing such data to protect patient identities[1,7], especially in an era where reidentification is increasingly feasible through data linkage[8,9]. Unsurprisingly, the prevalence of data sharing in medical research therefore remains extremely low[10].

The use of synthetic data has been proposed as a compromise[1,11–13]. These are artificial data created by a generative model trained on the original data, attempting to replicate their statistical properties without disclosing original records. In the most ambitious scenario, synthetic data could replace the original data in statistical analyses, yielding similar results while simultaneously addressing privacy concerns. This may create various opportunities across multiple areas of medical research, including clinical cancer research, some of which have already been explored. For instance, synthetic versions of the Netherlands Cancer Registry[14] and England's National Cancer Registration and Analysis Service[15] have been made available for research purposes. The German Centre for Cancer Registry Data is currently exploring this approach as well[16]. Much attention is also being given to the development of synthetic control arms in cancer trials, potentially reducing costs and ensuring that all participants receive the experimental treatment[17–20]. Furthermore, synthetic samples may augment datasets to improve predictive performance of models[21,22] and to enhance generalizability to underrepresented subgroups[23]. Additional use cases in clinical cancer research have been outlined elsewhere[24]. The growing importance of synthetic data is underscored by the establishment of the HealthData4EU cluster, a Horizon Europe initiative comprising seven funded projects dedicated to operationalizing synthetic data generation for secure use of health data in research across Europe, including applications in oncology[25].

[1]Synthetic Data for Research Acceleration (SYNDARA), Ghent University Hospital, Ghent, Belgium. [2]Department of Medical Oncology, Ghent University Hospital, Ghent, Belgium. [3]Internet and Data Science Lab (IDLab), Ghent University - imec, Ghent, Belgium. [4]Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium. ✉email: alexander.decruyenaere@ugent.be

Synthetic data can be created by different classes of generative models[13,26], the choice of which entails a paradigmatic bias-variance trade-off. Parametric models fit a prespecified and well-characterized distribution to the original data, whereas more flexible machine learning approaches like tree-based and deep generative models aim to capture the underlying data structure directly from the original data, without imposing a fixed distributional form. Therefore, the latter may potentially avoid so-called generative model misspecification and safeguard against bias in synthetic data analyses, especially in high-dimensional settings. Regardless of the choice of generative model class made, the process of learning and generating synthetic data introduces non-negligible uncertainty, which complicates fundamental calculations like p-values and confidence intervals, undermining the inferential utility through an increased risk of false-positive findings[27,28]. For synthetic data generated by parametric models, this can be mitigated via application of a previously proposed (yet often ignored) correction factor to the standard error (SE)[29]. Such correction does not exist for synthetic data created by tree-based or deep generative models[27], as these data also contain non-negligible bias resulting from the bias-variance trade-off made by these approaches.

This study reanalyzed data from phase 3 oncology randomized controlled trials and investigated whether their results could be replicated using a synthetic version of the original data. We thereby developed parametric generative models and extended a correction factor to the SE to handle time-to-event data, a modality commonly used in clinical cancer research. In addition, we aimed to provide guidance to clinicians and applied researchers on how to properly perform statistical inference from synthetic data.

Results

This retrospective, cross-sectional study compared 128 experimental vs. control treatments for solid tumors from 115 (reconstructed) original phase 3 trials published in seven high-impact journals in 2023. Sixteen different generative models were trained on each original trial dataset to create a single synthetic trial dataset per model. The two parametric models (that fit a predefined distribution to each variable in a prespecified order) were developed specifically for this study to process time-to-event data, whereas the four tree-based models (that either sequentially fit classification or regression trees to each variable, or jointly model the variables via adversarial random forests) and the ten deep neural network-based models (that rely on jointly modeling variables via generative adversarial networks, variational autoencoders, or diffusion models) were adopted from prior work[30–35] and customized as needed for this study. The (unadjusted) hazard ratio (HR) for overall survival (OS) was estimated in the original and synthetic trials. Inferential utility of the synthetic data was assessed through the standardized difference between the original and synthetic point estimates, probability overlap between the original and synthetic confidence intervals (CIs), and the true (TPR) and false positive rate (FPR) in terms of decision agreement between the original and synthetic trials. Corrected SEs, which inflate the default SEs as a function of the synthetic-to-original sample size, have been proposed for synthetic data analyses to account for the uncertainty in the fitted generative models. Accordingly, the impact of default vs. corrected SEs (to construct the CIs) and synthetic sample size was investigated per generative model class. Finally, the relationship between the statistical signal in the original trial relative to the noise (as reflected by the absolute z-value of the original HR) and the likelihood of decision agreement was modeled. Methodological details are described in a subsequent section.

We present here the results for the best-performing model per generative model class: *Weibull-XET* (parametric), *CART-ETX* (tree-based), and *SurvivalVAE+* (deep). These had the lowest average standardized difference between the original and synthetic estimates across all trials (of same original and synthetic sample size) and thus seemed unbiased. The performance of the other models is given in the Supplementary Information (Table C1 and Figures C1-5), leaving the interpretation below unchanged. We first provide the key insights from our study on decision agreement and argue that caution is warranted when drawing statistical conclusions from synthetic data analyses. We then examine the empirical evidence supporting these insights in different subsections that cover the original trial characteristics, the baseline setting with synthetic trials of same sample size as the original ones analyzed with default SEs, the impact of corrected SEs, the impact of synthetic sample size, and the relationship between the z-value of the original trial and decision agreement. Sensitivity analyses excluding original trials with smaller sample sizes, fewer observed events, and higher levels of censoring are reported in the Supplementary Information (Figures C7-9). Although our results focus on the marginal HR, the implication of potential treatment effect heterogeneity is explored in a case study in the Supplementary Information (Tables C2-3 and Figures C10-11).

Key insights on decision agreement

Ideally, to draw the exact same statistical conclusions from synthetic data analyses as from original data analyses, the 95% CIs of the original and synthetic HR should overlap perfectly. This requires the original and synthetic point estimates to coincide and the estimated SEs (which reflect the uncertainty associated with a particular point estimate) used to construct the CIs to be identical. Our results revealed a sobering insight: currently, this may only be achieved when parametric models are deployed to generate synthetic data of sufficiently large size (relative to the original sample size) and corrected SEs are applied in the synthetic data analysis, given the (rather ambitious) assumption that the parametric model is able to adequately capture the underlying original data distribution (i.e., no generative model misspecification).

The main findings from our study are graphically summarized in Fig. 1, which depicts the correlation (represented by the amount of scatter around the diagonal) and decision agreement (indicated by the color of the point estimates) between the original and synthetic HR of each trial per generative model. Since synthetic data generation introduced additional variability, the correlation between the original and synthetic point estimates was less than perfect, even in the absence of generative model misspecification. This additional variability depended on the generative model itself and the synthetic sample size (which influenced the sampling variability

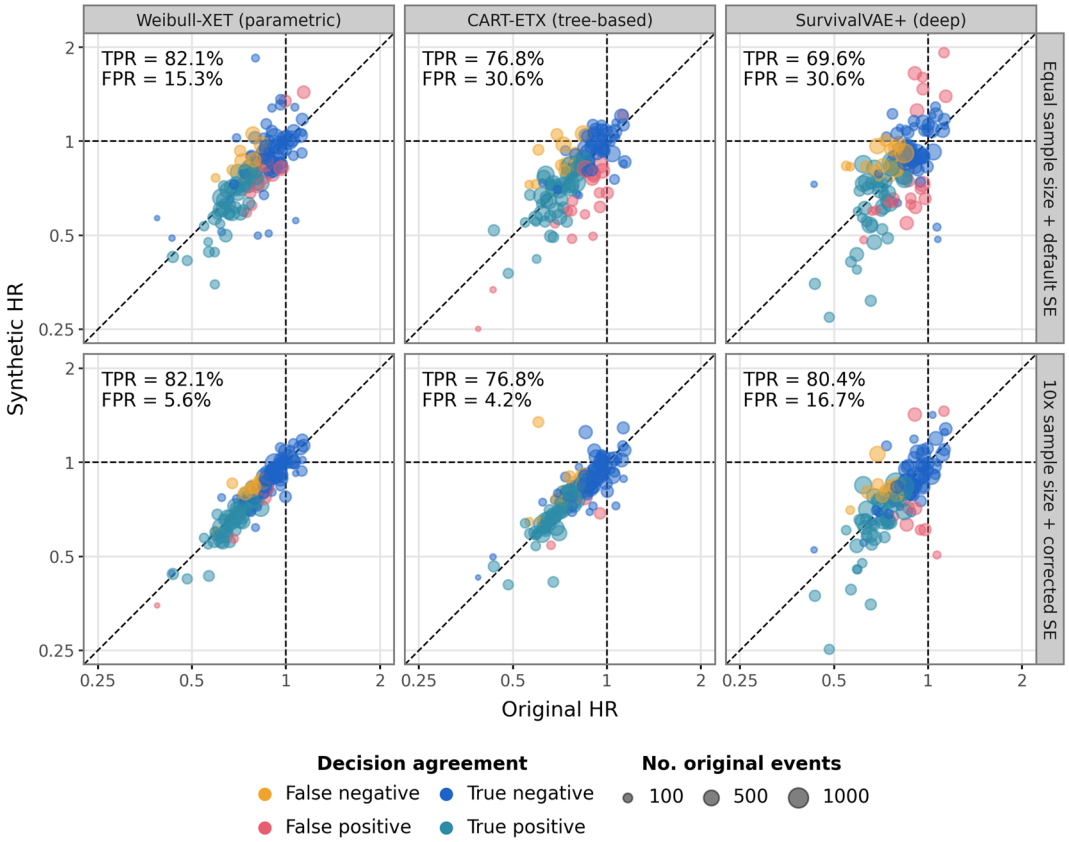

Fig. 1. Correlation and decision agreement between original and synthetic estimates of the hazard ratio (HR). The baseline setting (upper row) included synthetic trials of same sample size as the original ones analyzed with default standard errors (SEs), whereas the alternative setting (lower row) considered a synthetic-to-original sample size ratio of ten and the application of corrected SEs. The color coding represents the type of decision agreement. Each dot is the point estimate of the HR in a single original vs. corresponding synthetic trial. The size of the dots is proportional to the number of original events. Extreme values (HRs of <0.25 or HRs of >2.00) are not plotted (0 values for *Weibull-XET*, 0 values for *CART-ETX*, and 2 values for *SurvivalVAE+* in each row). FPR, false positive rate; TPR, true positive rate.

from the fitted generative model). Indeed, the more flexible the generative model (parametric < tree-based < deep) and the smaller the synthetic sample size, the more the additional variability from learning the synthetic data model. If the corresponding SE failed to acknowledge this additional variability, decision disagreement would occur.

In the baseline setting (upper row in Fig. 1), the synthetic sample sizes were set equal to the original ones and default SEs were applied in synthetic data analyses. Matching the sample size retained the dimensionality of the original data, which might be perceived as more convenient by supporting one-to-one replacement between original and synthetic data in the analysis and reporting. Default SEs (erroneously) treated the synthetic data as if they were actually observed, thereby ignoring the additional uncertainty inherent in synthetic data generation, as noted in prior work[27,29]. This yielded both false positives and false negatives, obviously compromising decision agreement. A false positive finding might have been generated, even for similar original and synthetic point estimates of the HR, as the CI around the synthetic point estimate had been overly optimistic (i.e., too narrow) and no longer contained the null effect (i.e., HR of 1). Conversely, the synthetic point estimate might also have become much less extreme than the (for instance, statistically significant) original one, just by chance due to the additional variability of synthetic data generation, potentially shifting the synthetic CI such that it covered the null effect (as opposed to the original CI) and produced a false negative result.

The alternative setting (lower row in Fig. 1) considered a remedy against decision disagreement by both increasing the synthetic sample size (to ten times the original one) and using corrected SEs. Applying either measure alone is not sufficient, regardless of the generative model class, as demonstrated in the ablation study presented in the Supplementary Information (Figure C6). A larger synthetic sample size reduced the sampling variability from the fitted generative model, thereby decreasing one source of the additional variability (reflected by less scatter around the diagonal). The additional variability remained nonetheless substantial for synthetic data created by deep generative models and, to a much lesser extent, by tree-based models, as the process of learning such models introduces more uncertainty. Corrected SEs acknowledged the additional variability of synthetic data generation when the generative model was parametric. Consequently, the excess variability observed with tree-based and deep generative models was left unaccounted for. Decision agreement could therefore only be

Characteristic	Summary	Characteristic	Summary
Journal		**Tumor type**	
Ann Oncol	11 (8.6%)	Breast	14 (10.9%)
J Clin Oncol	48 (37.5%)	Gastrointestinal	23 (18.0%)
JAMA	3 (2.3%)	Genitourinary	27 (21.1%)
JAMA Oncol	4 (3.1%)	Gynaecological	17 (13.3%)
Lancet	21 (16.4%)	Head & neck	11 (8.6%)
Lancet Oncol	20 (15.6%)	Melanoma	7 (5.5%)
N Engl J Med	21 (16.4%)	Other	1 (0.8%)
		Thoracic	28 (21.9%)
Sample size	571 (82–5,637)	**OS primary endpoint**	65 (50.8%)
OS event size	224.5 (18–1,230)	**HR for OS**	0.80 (0.39–1.14)
OS % censoring	52.1 (3.7–97.6)	**Significant OS**	56 (43.8%)

Table 1. Original trial characteristics (128 treatment comparisons). Categorical variables are presented as count (proportion) and continuous variables as median (range). HR, hazard ratio; OS, overall survival.

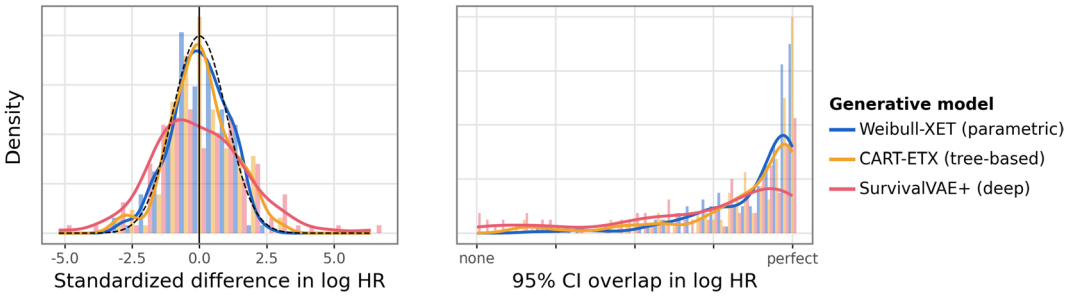

Fig. 2. **(Left) Distribution of the standardized difference between original and synthetic estimates of the log hazard ratio (HR).** If the original and synthetic sample size are equal, the standardized metric follows a normal distribution $\mathcal{N}(0,1)$ (represented by the dashed black curve) under standard assumptions, so that extreme values suggest misspecification and/or excess variability of the generative model. **(Right) Distribution of the 95% confidence interval (CI) overlap between original and synthetic estimates of the log hazard ratio (HR).** Perfect overlap indicates that the 95% CI was exactly the same in the original and synthetic trials.

sufficiently improved for parametric models: using corrected SE reduced false positives, albeit at the expense of power (i.e., more false negatives), which could be restored by choosing larger synthetic sample sizes. However, and somewhat surprisingly, some tree-based methods performed nearly on par with parametric models in our study.

In the following subsections, we describe in greater detail the impact of the generative model class, type of SE, and synthetic sample size on the statistical conclusions drawn from the synthetic data analyses of 128 treatment comparisons from 115 phase 3 oncology trials. The relationship between the z-value of the original trial and decision agreement is also explored.

Original trial characteristics

The original trial characteristics are presented in Table 1. The median sample size was 571 (range 82–5,637) with a median number of observed OS events of 224.5 (range 18–1,230) and a median proportion of censoring of 52.1% (range 3.7–97.6). The median HR was 0.80 (range 0.39–1.14), with 56 (43.8%) reaching statistical significance at the 0.05 level, all showing superiority of the experimental treatment.

Default standard errors and same sample size

These original trials were synthesized using sixteen different generative models. Each model generated synthetic trials in which event times might be censored, as in the original trials. The HR for OS was estimated in the original and synthetic trials using a univariate Cox proportional hazards model with treatment arm indicator as predictor. The baseline setting considered default SEs and same synthetic sample sizes as the original ones. The distributions of the standardized difference and 95% CI overlap between the original and synthetic log HR are shown in Fig. 2. The standardized difference expresses the deviation of the synthetic log HR from the original one in terms of original SE units, while the 95% CI overlap quantifies the proportion of overlap (standardized with range from 0, indicating no overlap, to 1, indicating perfect overlap) between the original and synthetic CI for the log HR. All utility metrics, including decision agreement, are summarized in Table 2.

As stated previously, only the results for *Weibull-XET* (parametric), *CART-ETX* (tree-based), and *SurvivalVAE+* (deep) are reported here. Results for the other models can be found in the Supplementary

Model	Std. difference in log HR: mean (sd)	Significant std. difference: %	95% CI overlap: median (Q1-Q3)	TPR: %	FPR: %
Weibull-XET	−0.01 (1.03)	3.9	0.94 (0.79–0.99)	82.1	15.3
CART-ETX	0.05 (1.16)	10.9	0.93 (0.79–0.99)	76.8	30.6
SurvivalVAE+	−0.02 (1.76)	23.4	0.84 (0.58–0.96)	69.6	30.6

Table 2. Utility metrics of the synthetic trials (of same sample size as the original trials and analyzed with default standard errors). CI, confidence interval; FPR, false positive rate; HR, hazard ratio; Q1, first quartile; Q3, third quartile; sd, standard deviation; std., standardized; TPR, true positive rate.

Information (Table C1 and Figures C1-5). On average, the HR of the original trial could be replicated in the synthetic version generated by all three models, with a mean standardized difference in log HR of -0.01 (standard deviation [sd] 1.03) for *Weibull-XET*, 0.05 (sd 1.16) for *CART-ETX*, and -0.02 (sd 1.76) for *SurvivalVAE+*. Even though *CART-ETX* and *SurvivalVAE+* seemed thus unbiased (across trials), they produced more variable HR estimates (within trials) than the parametric *Weibull-XET* model, as reflected by the larger spread (and thus larger sd) in standardized differences (see Fig. 2 left). This is due to their data-adaptive training process, leading to excess variability. Indeed, standardized differences deviated significantly from zero (i.e., the synthetic HR differed at least two original SEs from the original one on the log scale; see Appendix B in the Supplementary Information) in 3.9% of synthetic trials generated by *Weibull-XET*, 10.9% by *CART-ETX*, and 23.4% by *SurvivalVAE+*.

The misalignment of the synthetic HRs with the original ones was further complicated by the use of default SEs to construct the CIs around the synthetic HRs. In fact, ignoring the additional variability of synthetic data generation by applying default SEs (which naively assumed that the synthetic data were originally observed data) yielded too optimistically narrow synthetic CIs, frequently leading to insufficient overlap with the original CIs (see Fig. 2 right) and thereby impacting decision agreement (see Table 2). The median (Q1-Q3) 95% CI overlap in log HR was 0.94 (0.79–0.99) for *Weibull-XET*, 0.93 (0.79–0.99) for *CART-ETX*, and 0.84 (0.58–0.96) for *SurvivalVAE+*. The highest TPR was achieved for *Weibull-XET* (82.1%), followed by *CART-ETX* (76.8%) and *SurvivalVAE+* (69.6%), while the reverse order was observed in terms of FPR, ranging from 15.3% for *Weibull-XET* to 30.6% for both *CART-ETX* and *SurvivalVAE+*. Thus, when analyzing synthetic trials of same size as the original ones with default SEs, our results indicated that the decision agreement between the original and synthetic trials was unsatisfactory in general, and even unacceptable in the case of tree-based and deep generative models.

Impact of corrected standard errors

Corrected SEs have previously been proposed to analyze synthetic data as these account for (some of) the uncertainty in the fitted generative models[29]. Here, we adapted them to time-to-event data. The corrected SE of the HR was obtained by multiplying the default SE with the correction factor $\sqrt{1 + m_{events}/n_{events}}$, where m_{events} and n_{events} denote the number of observed events in the synthetic and original trial, respectively. As a result, the corresponding corrected CIs were wider, impacting decision agreement. Indeed, use of corrected SEs resulted in a considerably lower FPR of 4.2% for *Weibull-XET*, 6.9% for *CART-ETX*, and 12.5% for *SurvivalVAE+* (compared to the baseline setting shown in Table 2). However, it came at the expense of power, with also a lower TPR of 60.7% for *Weibull-XET*, 46.4% for *CART-ETX*, and 50.0% for *SurvivalVAE+*.

Notably, the FPR remained high for synthetic trials generated by deep generative models like *SurvivalVAE+*, since the corrected SE just acknowledged the additional variability of synthetic data generation that resulted from the fitting of and sampling from parametric generative models. As such, it did not suffice in the analysis of synthetic trials created by deep generative models, whose predicted functionals converged much more slowly to their counterparts calculated on the original trials. This poor agreement between synthetic and original data functionals produced excess variability in the downstream estimator (e.g., the HR) derived from synthetic data. In these trials, the corrected CI was therefore still too narrow. Interestingly, the excess variability induced by the tree-based *CART-ETX* model turned out to be limited in our analysis, such that the corrected SE only fell slightly short compared to the parametric *Weibull-XET* model.

Impact of synthetic sample size

Increasing the synthetic sample size reduces the sampling variability from the fitted generative model. This may lead to better alignment between the original and synthetic point estimates. In addition, the corrected SE obtained from synthetic data of large size (relative to the original sample size) will then converge to the original SE. Together, these might contribute to better overlap between the original and synthetic CIs and hence, to improved decision agreement. Figure 3 illustrates how the 95% CI overlap in log HR between the original and synthetic trials was distributed across varying synthetic sample size (0.5 to 10 times the original sample size with steps of 0.5), using default vs. corrected SEs. Each dot represents the CI overlap for a trial at a specific synthetic-to-original sample size ratio.

In line with statistical theory, our results showed that larger synthetic samples led to worse overlap in the case of default SEs, regardless of the type of generative model. By contrast, better overlap could be obtained with corrected SEs, albeit mainly for the parametric *Weibull-XET* model and, to a lesser extent, for the tree-based *CART-ETX* model. Although the overlap may also improve with corrected SEs for deep generative models like *SurvivalVAE+*, the additional gain from larger synthetic samples appeared more limited and bounded. Moreover, certain deep generative models, whose predicted functionals underlying the HR converged more slowly, still

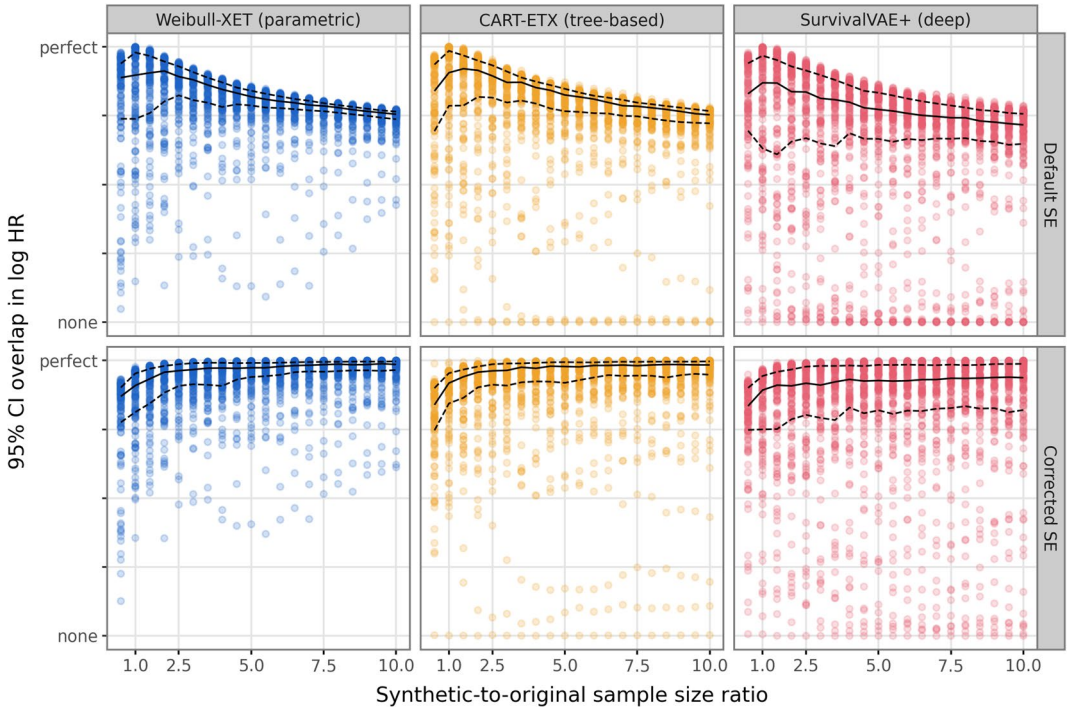

Fig. 3. 95% confidence interval (CI) overlap between original and synthetic estimates of the log hazard ratio (HR) by synthetic-to-original sample size ratio and by default vs. corrected standard errors (SEs). Each dot represents a score for the CI overlap between the original and synthetic trial at a specific synthetic-to-original sample size ratio. To reduce computational burden, only synthetic trials with a synthetic-to-original sample size ratio of ten were generated and then subsampled to obtain a synthetic trial with a smaller ratio. The solid line marks the median overlap at each ratio, while the dashed lines indicate the first and third quartile.

exhibited worse overlap with increasing synthetic sample size, despite the use of corrected SEs (see Figure C5 in the Supplementary Information). At a synthetic-to-original sample size ratio of ten and with corrected SEs, the median (Q1-Q3) 95% CI overlap became 0.99 (0.97–1.00) for *Weibull-XET*, 0.98 (0.95–1.00) for *CART-ETX*, and 0.94 (0.82–0.97) for *SurvivalVAE+*. This enabled more trustworthy statistical inference from synthetic trials, especially when generated by a parametric or tree-based model, by increasing again the TPR while keeping control of the FPR. TPR and FPR became 82.1% and 5.6% for *Weibull-XET*, 76.8% and 4.2% for *CART-ETX*, and 80.4% and 16.7% for *SurvivalVAE+*, respectively.

Relationship between original z-value and decision agreement

The relationship between the statistical signal in the original trial relative to the noise (as reflected by the absolute z-value of the original HR) and the likelihood of decision agreement is plotted in Fig. 4. The upper row represents the baseline setting with synthetic trials of same sample size as the original ones analyzed with default SEs, whereas the lower row considered the alternative setting with a ten times larger synthetic sample size and the application of corrected SEs. Ideally, in the case of perfect agreement, non-significant original trials (plotted left of the vertical dashed line which marks the statistical significance threshold) would always deliver true negative synthetic trials (which would have been indicated by the blue line remaining constant at 100% up to the vertical dashed line) and significant original trials (plotted right of the vertical dashed lined) would always deliver true positive synthetic trials (which would have been indicated by the green line remaining constant at 100% from the vertical dashed line onward). False negative and false positive synthetic trials would not occur (which would have been indicated by the yellow and red lines staying flat at 0%).

Such a scenario was not observed here as decision disagreement occurred, peaking around the statistical significance threshold: false positive synthetic trials were most likely to arise from borderline non-significant original trials, while false negative synthetic trials tended to result from borderline significant original trials. The further the absolute z-value was from the threshold, the lower the likelihood of decision disagreement, with a decay rate that again depended on the generative model class (fastest in the parametric *Weibull-XET* model, intermediate in the tree-based *CART-ETX* model, and slowest in deep generative models like *SurvivalVAE+*). When using corrected SE and larger synthetic sample sizes, the decay rate accelerated further, though it remained slow for deep generative models.

Discussion

Synthetic data may facilitate data sharing in privacy-sensitive contexts, including healthcare, yet their statistical analysis remains challenging. In particular, their inferential utility, which concerns whether similar conclusions

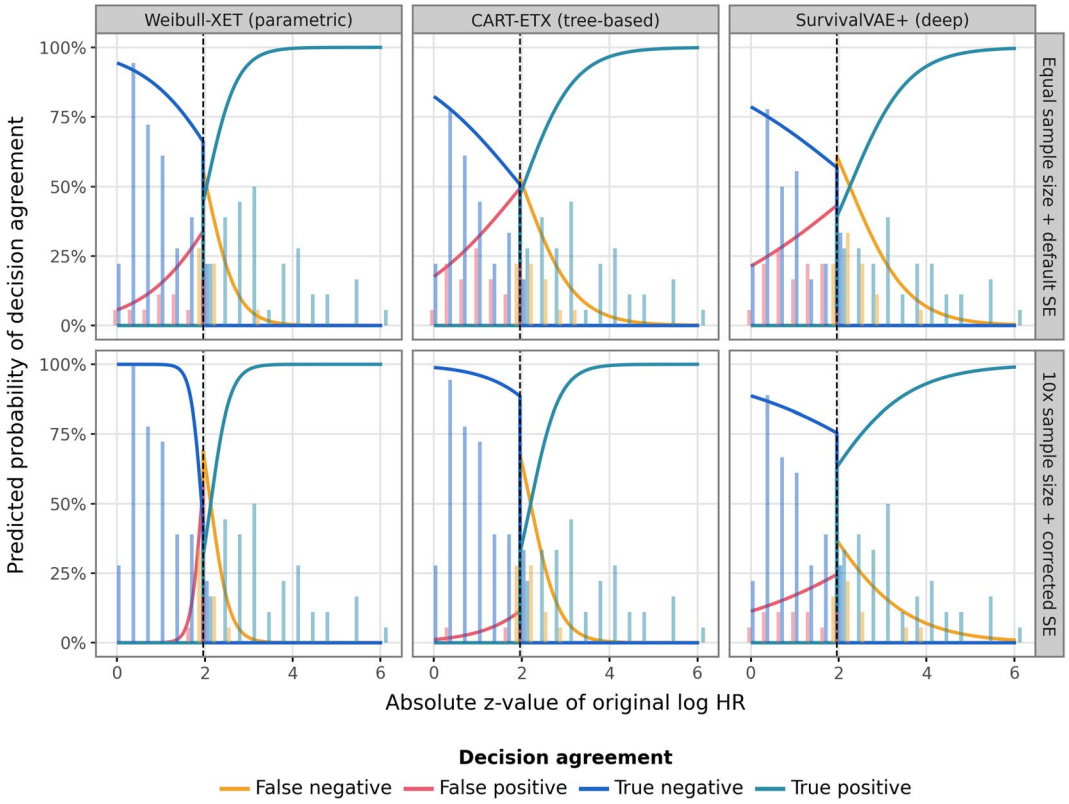

Fig. 4. Relationship between the absolute z-value of the original log hazard ratio (HR) and decision agreement. The baseline setting (upper row) included synthetic trials of same sample size as the original ones analyzed with default standard errors (SEs), whereas the alternative setting (lower row) considered a synthetic-to-original sample size ratio of ten and the application of corrected SEs. The curves represent the predicted conditional probability of decision agreement given the original absolute z-value, while the histogram shows the observed marginal frequency of decision agreement. The color coding represents the type of decision agreement. The vertical dashed line marks the threshold for statistical significance at the 0.05 level.

can be drawn from synthetic vs. original data, highly depends on the choice of generative model, synthetic sample size, and type of SE[27–29]. Prior research on synthetic data directed at clinicians and applied researchers, including the oncology readership, mainly focused on the assessment of summary statistics and predictive performance. Moreover, some misconceptions still persist in the literature, where benefits beyond privacy protection are wrongly attributed to synthetic data. Claims are made, for example, that generative models could filter out the statistical noise in the original data such that synthetic data contain more signal[36], or that synthetic data could provide a larger sample size for the analysis of rare diseases[18,37], in both cases leading to narrower CIs and hence, artificially improving power.

In this reanalysis of synthetic trials generated from 128 treatment comparisons from 115 phase 3 oncology trials, we showed that the number of false positive and false negative findings was unacceptably high, even when the generative models seemed unbiased. This resulted from ignoring the variability of synthetic data generation, which increased as the generative model became more flexible (parametric < tree-based < deep)[27]. Indeed, excess variability of deep generative models can also be observed in other studies investigating the application of synthetic data in oncology[19,20,38–40].

Interestingly, despite their flexibility, the inferential utility of the tree-based generative models examined in this study was more closely aligned with the parametric models than with the deep generative models. Their strong performance has already been noted by other authors[19,36,38,39]. This suggests that, in our relatively simple study setting, some tree-based models only introduced a small amount of excess variability that was left unaccounted for by corrected SEs, while still being very flexible in approximating the underlying data structure, leading to a quite favorable bias-variance trade-off. This likely resulted from their underlying algorithm, which involves conditional resampling of original values (structured by classification or regression trees), followed by smoothing to prevent direct reidentification and avoid exact duplicates. Nevertheless, their excess variability is not guaranteed to remain small in each use case, as it will depend on the dimension of the data and the complexity of the original data distribution[27]. Novel strategies are therefore being developed that entirely remove the excess variability induced by tree-based and deep generative models to enable valid inference from the resulting synthetic data[28].

A correction factor to the SE, which we extended in this study to time-to-event data regimes, only suffices in capturing the additional variability of parametric models. We confirmed that it may fall short for tree-based and

deep generative models, particularly for the latter[27]. Surprisingly, recent oncology studies deploying synthetic data[17,19,20,36,41–43] continue to ignore the use of corrected SEs, although these have initially been proposed many years ago[29,44]. Not using corrected SEs gives a false impression of high power at the expense of a high risk of false positives. Instead, by using synthetic data generated by parametric models along with corrected SEs, it becomes clear—as theoretically expected—that an analysis of synthetic data is of less quality, and in particular less powerful, than an analysis of the original data. However, we demonstrated that this power loss could be recovered by increasing the synthetic sample size. An infinitely large synthetic sample should then deliver the same analysis results on average as the original data (provided that the generative model was correctly specified)[28]. Conversely, in the case of default SEs, larger synthetic sample sizes will further artificially inflate the power while also producing more false positives. Although we did not assess the effect of larger synthetic sample sizes on privacy, our results already highlight the challenges in generating synthetic data of sufficient inferential utility, even without imposing any hard privacy constraints.

Finally, while we appreciate previous efforts directed at the clinical readership that address the inferential utility of synthetic data[38,39], we believe their guidance (that "analysis of synthetic data should be based on at least ten generated datasets of the same size as the original whose analysis results are combined" and that further "amplification, i.e., increasing the synthetic sample size, had only a marginal benefit") prioritizes the number of synthetic datasets unnecessarily over the synthetic sample size. It follows from the combining rule for the pooled variance derived by Raab et al.[29] that changing either the number or the size of synthetic datasets (sampled from the same generative model) with the same factor has the same impact: ten synthetic trials of original sample size provide the same statistical information as a single synthetic trial of ten times the original sample size. From a practical viewpoint, we prefer the latter since a single dataset is arguably more convenient for analysis. In addition, the different implications for inferential utility from applying parametric vs. flexible machine learning approaches to synthetic data generation have been largely overlooked in the current literature. We hope that this study has clarified some of the ongoing controversies surrounding synthetic data and that it will promote good practice. Otherwise, we fear that the release of synthetic data would ironically defeat the intent of accelerating research.

Limitations of this study included the low-dimensional setting, for which parametric models were well suited. Although these models can be readily extended to the high-dimensional setting, they might fail to preserve complex multivariate relationships, leading to bias and thus lower inferential utility, for instance, in subgroup analyses. A too restrictive parametric model that does not include an interaction between treatment and subgroup, will then never generate synthetic data that contain evidence for treatment effect heterogeneity. This was illustrated by a case study presented in the Supplementary Information (Tables C2-3 and Figures C10-11). In addition, we did not consider privacy since the datasets evaluated in our study only had a limited number of (non-sensitive) attributes. Although the trade-off between privacy and utility of synthetic data is well-established[45], where improvements in one may negatively impact the other, we demonstrated that the generation of high-quality synthetic data in a real-world application considering 115 trials was already difficult to attain without additional privacy constraints. Finally, we cannot rule out some dependence between the synthetic estimates due to the inclusion of eight three-arm studies with a shared control arm. We believe, however, that the impact of this dependence on our results is limited, since (1) only 8 of the 128 (6.3%) treatment comparisons may be partially correlated with another one, (2) the training of a generative model and the sampling of synthetic data for one experimental vs. control comparison were performed independently of the corresponding comparison involving the other experimental arm, and (3) the estimates obtained from synthetic data created by the parametric generative model in Fig. 2 follow the theoretical reference distribution, thereby providing no empirical indication that the assumption of independent and identically distributed observations is violated.

In conclusion, while synthetic data are promising for data sharing, their analysis is challenging as they may introduce false positive and false negative findings. After carefully analyzing synthetic versions of a large number of real-world trials, generated by parametric, tree-based, and deep generative models, we found that mainly synthetic data of sufficiently large size generated by parametric models and analyzed with corrected standard errors may preserve inferential utility.

Methods

Original trial data

This retrospective, cross-sectional study included 128 experimental vs. control treatment comparisons from 115 phase 3 oncology randomized controlled trials published in seven high-impact journals (Ann Oncol, J Clin Oncol, JAMA, JAMA Oncol, Lancet, Lancet Oncol, N Engl J Med) in 2023. A PubMed literature search was performed on April 11th, 2024 with "neoplasm", "phase 3", and "randomized" as keywords. The exact search query and selection criteria are given in the Supplementary Information (Figure A1), while a detailed description of the selected trials is presented in the Supplementary Data. When trials included multiple experimental arms or reported results only for (mutually exclusive) subpopulations, multiple treatment comparisons were allowed per trial. These were treated as separate trials in our study.

Ideally, access to the original individual patient data for each trial would have been obtained. As this was not feasible, we reconstructed them from the published Kaplan–Meier curves of OS by first extracting the survival coordinates (either automatically via the PostScript vector[46] if attainable or manually via DigitizeIt software[47] otherwise), and subsequently, applying the algorithm described by Guyot et al.[48] to derive the individual patient time-to-event data from the extracted survival coordinates given the numbers at risk. Note that this algorithm also recovers the censoring pattern. This resulted in reconstructed datasets with treatment arm indicator, survival time, and event indicator as variables. The reconstruction error was deemed minimal (see Figure A2 in the Supplementary Information).

Model name	Model type	Data type	Synthesis type	Synthesis order	References
Weibull-XET	Parametric	Time-to-event	Sequential	X→E→T	Custom
Weibull-ETX	Parametric	Time-to-event	Sequential	E→T→X	Custom
survctree-XET	Tree-based	Time-to-event	Sequential	X→ (E,T)	[30,49]
CART-XET	Tree-based	Tabular	Sequential	X→E→T	[30,49]
CART-ETX	Tree-based	Tabular	Sequential	E→T→X	[30,49]
ARF	Tree-based	Tabular	Joint	(X,T,E)	[31]
SurvivalGAN	Deep	Time-to-event	Mixed*	E→X→T	[32]
SurvivalGAN+	Deep	Time-to-event	Mixed*	E→X→T	[32,50], custom
SurvivalVAE	Deep	Time-to-event	Mixed*	E→X→T	[32]
SurvivalVAE+	Deep	Time-to-event	Mixed*	E→X→T	[32,50], custom
CTGAN-ETX	Deep	Time-to-event	Mixed*	E→T→X	[33,35]
TVAE-ETX	Deep	Time-to-event	Mixed*	E→T→X	[33,35]
DDPM-ETX	Deep	Time-to-event	Mixed*	E→T→X	[34,35]
CTGAN	Deep	Tabular	Joint	(X,T,E)	[33]
TVAE	Deep	Tabular	Joint	(X,T,E)	[33]
DDPM	Deep	Tabular	Joint	(X,T,E)	[34]

Table 3. Overview of generative models. E, event indicator; T, survival time; X, baseline characteristics. *Multivariate X are synthesized jointly by a deep generative model conditional on the preceding variable(s) E and/or T.

Synthetic trial data generation

These original data were then used to train sixteen generative models, from which a single synthetic dataset of varying size (equal to or ten times the original sample size) was sampled. An overview of all models is presented in Table 3, while a detailed explanation can be found in the Supplementary Information (Appendix A). A custom name was assigned to each model: its prefix referred to the base model and its suffix to the synthesis order (only specified in the case of sequential generation), where the treatment arm indicator, survival time, and event indicator were denoted by *X*, *T*, and *E*, respectively.

We developed two parametric models based on the Weibull distribution (*Weibull-XET* and *Weibull-ETX*) to handle time-to-event data. The four tree-based models (*survctree-XET, CART-XET*, *CART-ETX*, and *ARF*) and the ten deep neural network-based models (*SurvivalGAN*, *SurvivalGAN+*, *SurvivalVAE*, *SurvivalVAE+*, *CTGAN-ETX*, *TVAE-ETX*, *DDPM-ETX*, *CTGAN*, *TVAE*, and *DDPM*) were adopted from earlier work by other investigators. The parametric and three tree-based models sequentially fit either a parametric model (*Weibull-XET* and *Weibull-ETX*) or a classification or regression tree (*survctree-XET, CART-XET*, and *CART-ETX*[30]), respectively, to each variable in a prespecified order, whereas one tree-based model and the deep generative models jointly fit variables via adversarial random forests (*ARF*[31]), generative adversarial networks (*SurvivalGAN*[32] and *CTGAN*[33]), variational autoencoders (*SurvivalVAE*[32] and *TVAE*[33]), or diffusion models (*DDPM*[34]). We also considered a variant of these deep generative models, named *CTGAN-ETX*, *TVAE-ETX*, and *DDPM-ETX*, respectively, following a recent proposal to adapt them to time-to-event data[35]. The *SurvivalGAN+* and *SurvivalVAE+* models are our custom extensions to the *SurvivalGAN* and *SurvivalVAE* models, replacing the XGBoost algorithm with the NGboost algorithm[50] to compute the parameters of a Weibull distribution. All models produce synthetic datasets with potentially censored event times, matching the original data structure.

Statistical analysis

The HR for OS between the experimental vs. control arm was calculated using a univariate Cox proportional hazards model with treatment arm indicator as predictor, both for the original and synthetic trials. Default (model-based) SEs were extracted to construct the default 95% CIs. Statistical significance was tested at the 0.05 level (as such, it may have deviated from the original trial report).

The defaults SEs were inflated by $\sqrt{1 + m_{events}/n_{events}}$ to obtain corrected SEs and CIs that account for (some of) the uncertainty in the fitted synthetic data models, where m_{events} and n_{events} denote the number of observed events in the synthetic and original trial, respectively. This correction factor extends the factor $\sqrt{1 + m/n}$ previously proposed by Raab et al.[29], where m and n are the synthetic and original sample size, respectively, to time-to-event data (see Appendix B in the Supplementary Information).

Utility metrics

The inferential utility of the synthetic data was assessed through the following metrics:

- Standardized difference between the original and synthetic estimates of the log HR, defined as $(\widehat{\beta}_{orig} - \widehat{\beta}_{syn})/SE(\widehat{\beta}_{orig})$, where $\widehat{\beta}_{orig}$ and $\widehat{\beta}_{syn}$ are the estimated log HRs in the original and synthetic trials, respectively, and $SE(\widehat{\beta}_{orig})$ is the estimated SE for the log HR in the original trial. It expresses the deviation of the synthetic log HR estimate from the original one in terms of original SE units (positive and negative values indicate overestimation in favor of the experimental and control treatment, respectively).

Moreover, we show in the Supplementary Information (Appendix B) that this metric follows a normal distribution $\mathcal{N}(0, n_{events}/m_{events})$ under standard assumptions, so that extreme values suggest misspecification and/or excess variability of the generative model.

- CI probability overlap between the original and synthetic estimates of the log HR, defined as $\frac{1}{2}\left[\int_{LL_{syn}}^{UL_{syn}} f_{orig}(x)dx + \int_{LL_{orig}}^{UL_{orig}} f_{syn}(x)dx\right]/(1-\alpha)$, following Karr et al.[51], where (LL_{orig}, UL_{orig}) and (LL_{syn}, UL_{syn}) are the CI bounds for the log HR in the original and synthetic trials, respectively, $f_{orig}(x)$ is a normal density function with the original log HR estimate as mean and its corresponding estimated SE as standard deviation, $f_{syn}(x)$ is a normal density function with the synthetic log HR estimate as mean and its corresponding estimated SE as standard deviation, and $1-\alpha$ represents the confidence level. The first term expresses how much of the probability mass of the original HR distribution lies within the synthetic CI, and vice versa for the second term. Averaging these two terms ensures that the metric is symmetric and does not depend on which is considered the reference. For convenience, we standardized this metric (by dividing it by the confidence level) to obtain a value between 0 (no overlap) and 1 (perfect overlap).
- Decision agreement, that is, whether the same inferential conclusion would be drawn from the synthetic trial as from the original one based on the direction of the HR and its statistical significance at the 0.05 level. A synthetic trial is considered truly positive (TP) when the HR was significant (with same direction) in both the original and synthetic trial, falsely negative (FN) when it was only significant in the original trial, falsely positive (FP) when it was only significant in the synthetic trial or also significant in the original trial but with different direction, or truly negative (TN) when it was not significant in both the original and synthetic trial. The TPR is then defined as $TP/(TP+FN)$ and the FPR as $FP/(FP+TN)$.

The baseline setting considered default SEs and same synthetic sample sizes as the original ones. Subsequently, the impact of corrected SEs and increasing synthetic sample size was investigated. In addition, the relationship between the absolute z-value of the original log HR and the likelihood of decision agreement between the original and synthetic trial was modeled. The absolute z-value is defined as $|\widehat{\beta}_{orig}|/SE(\widehat{\beta}_{orig})$, where $\widehat{\beta}_{orig}$ and $SE(\widehat{\beta}_{orig})$ are the estimated log HR and corresponding SE in the original trial, respectively. It reflects the statistical signal in the original trial relative to the noise, and increases with larger observed effect size and larger sample size. A multinomial logistic regression model, with decision agreement as the outcome and the absolute z-value as the predictor, was used to calculate the predicted probabilities of decision agreement for each z-value. Because discontinuity in decision agreement was expected around the statistical significance threshold (for instance, false negatives and true positives could not arise from synthetic versions of non-significant original trials, nor could true negatives result from significant original trials), this model was fitted separately for non-significant and significant original trials. Finally, privacy assessment was beyond the scope of this work.

Original trials were reconstructed in R version 4.2.2, while synthetic data generation and evaluation was performed in Python version 3.11, relying on the Synthcity[52] and Synthpop[49] packages.

Data availability

An overview of the selected trials is presented in the Supplementary Data. The reconstructed data of the original trials are available from the corresponding author on reasonable request. The code to reproduce all results is available on GitHub:https://github.com/syndara-lab/synthetic-trials-challenges.

Received: 31 October 2025; Accepted: 9 July 2026
Published online: 23 July 2026

References

1. Bentzen, H. B. et al. Remove obstacles to sharing health data with researchers outside of the European Union. *Nat. Med.* **27**, 1329–1333 (2021).
2. Naudet, F. et al. Medical journal requirements for clinical trial data sharing: Ripe for improvement. *PLoS Med.* **18**, e1003844 (2021).
3. Locher, C., Goff, G. L., Louarn, A. L., Mansmann, U. & Naudet, F. Making data sharing the norm in medical research. *BMJ* **382**, p1434 (2023).
4. Loder, E., Macdonald, H., Bloom, T. & Abbasi, K. Mandatory data and code sharing for research published by The BMJ. *BMJ* **384**, q324 (2024).
5. Taichman, D. B. et al. Data sharing statements for clinical trials—A requirement of the International Committee of Medical Journal Editors. *N. Engl. J. Med.* **376**, 2277–2279 (2017).
6. Johnson, A. L. et al. Clinical trial data-sharing policies among journals, funding agencies, foundations, and other professional organizations: A scoping review. *J. Clin. Epidemiol.* **154**, 42–55 (2023).
7. Hoffman, S. Privacy and security—Protecting patients' health information. *N. Engl. J. Med.* **387**, 1913–1916 (2022).
8. Ohm, P. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Rev.* **5**, 1701 (2010).
9. Rocher, L., Hendrickx, J. M. & de Montjoye, Y.-A. Estimating the success of re-identifications in incomplete datasets using generative models. *Nat. Commun.* **10**, 3069 (2019).
10. Hamilton, D. G. 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* **382**, e075767 (2023).
11. Kokosi, T. & Harron, K. Synthetic data in medical research. *BMJ Med.* **1**, e000167 (2022).
12. Giuffrè, M. & Shung, D. L. Harnessing the power of synthetic data in healthcare: Innovation, application, and privacy. *Npj Digit. Med.* **6**, 1–8 (2023).
13. van Breugel, B., Liu, T., Oglic, D. & van der Schaar, M. Synthetic data in biomedicine via generative artificial intelligence. *Nat. Rev. Bioeng.* **2**, 991–1004 (2024).
14. Netherlands Comprehensive Cancer Organisation (IKNL). Synthetic dataset Netherlands Cancer Registry (NCR). Available from: https://iknl.nl/en/ncr/synthetic-dataset

15. National Disease Registration Service, NHS England. Simulacrum, a synthetic dataset developed by Health Data Insight CiC and derived from anonymous cancer data. Available from: https://simulacrum.healthdatainsight.org.uk
16. Escudié, J.-B. et al. Generation of synthetic data for the German cancer registries. *Stud. Health Technol. Inform.* **329**, 1846–1847 (2025).
17. D'Amico, S. et al. Synthetic data generation by artificial intelligence to accelerate research and precision medicine in hematology. *JCO Clin. Cancer Inform.* **5**, e2300021 (2023).
18. Serrano, C. et al. Rethinking placebos: Embracing synthetic control arms in clinical trials for rare tumors. *Nat. Med.* **29**, 2689–2692 (2023).
19. Akiya, I., Ishihara, T. & Yamamoto, K. Comparison of synthetic data generation techniques for control group survival data in oncology clinical trials: Simulation study. *JMIR Med. Inform.* **12**, e55118 (2024).
20. Elvatun, S., Knoors, D., Brant, S., Jonasson, C. & Nygård, J. F. Synthetic data as external control arms in scarce single-arm clinical trials. *PLoS Digit. Health* **4**, e0000581 (2025).
21. Chen, R. J., Lu, M. Y., Chen, T. Y., Williamson, D. F. K. & Mahmood, F. Synthetic data in machine learning for medicine and healthcare. *Nat. Biomed. Eng.* **5**, 493–497 (2021).
22. Liu, D. *et al.* Synthetic data generation for augmenting small samples. Preprint at https://doi.org/10.48550/arXiv.2501.18741 (2025).
23. Koul, A., Duran, D. & Hernandez-Boussard, T. Synthetic data, synthetic trust: Navigating data challenges in the digital revolution. *Lancet Digit. Health* https://doi.org/10.1016/j.landig.2025.100924 (2025).
24. Jacobs, F. et al. Opportunities and challenges of synthetic data generation in oncology. *JCO Clin. Cancer Inform.* **8**, e2300045 (2023).
25. Declerck, J. et al. Rethinking trust in synthetic health data: Lessons from 7 European Research Initiatives. *J. Med. Internet Res.* **28**, e83369 (2026).
26. Loni, M., Poursalim, F., Asadi, M. & Gharehbaghi, A. A review on generative AI models for synthetic medical text, time series, and longitudinal data. *Npj Digit. Med.* **8**, 1–10 (2025).
27. Decruyenaere, A. *et al.* The real deal behind the artificial appeal: Inferential utility of tabular synthetic data. in *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence* 966–996 (PMLR, 2024).
28. Decruyenaere, A. et al. Debiasing synthetic data generated by deep generative models. *Adv. Neural Inf. Process. Syst.* **37**, 41539–41576 (2024).
29. Raab, G. M., Nowok, B. & Dibben, C. Practical data synthesis for large samples. *J. Priv. Confid.* **7**, 67–97 (2016).
30. Drechsler, J. & Reiter, J. P. An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Comput. Stat. Data Anal.* **55**, 3232–3243 (2011).
31. Watson, D. S., Blesch, K., Kapar, J. & Wright, M. N. Adversarial random forests for density estimation and generative modeling. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics* 5357–5375 (PMLR, 2023).
32. Norcliffe, A., Cebere, B., Imrie, F., Lió, P. & Schaar, M. van der. SurvivalGAN: Generating time-to-event data for survival analysis. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics* 10279–10304 (PMLR, 2023).
33. Xu, L., Skoularidou, M., Cuesta-Infante, A. & Veeramachaneni, K. Modeling tabular data using conditional GAN. In *Advances in Neural Information Processing Systems* vol. 32 (Curran Associates, Inc., 2019).
34. Kotelnikov, A., Baranchuk, D., Rubachev, I. & Babenko, A. TabDDPM: Modelling tabular data with diffusion models. In *Proceedings of the 40th International Conference on Machine Learning* 17564–17579 (PMLR, 2023).
35. Ashhad, M. & Henao, R. Conditioning on time is all you need for synthetic survival data generation. Preprint at https://doi.org/10.48550/arXiv.2405.17333 (2024).
36. Azizi, Z., Zheng, C., Mosquera, L., Pilote, L. & Emam, K. E. Can synthetic data be a proxy for real clinical trial data? A validation study. *BMJ Open* **11**, e043497 (2021).
37. Trotsyuk, A. A. et al. Toward a framework for risk mitigation of potential misuse of artificial intelligence in biomedical research. *Nat. Mach. Intell.* **6**, 1435–1442 (2024).
38. El Kababji, S. et al. Evaluating the utility and privacy of synthetic breast cancer clinical trial data sets. *JCO Clin. Cancer Inform.* https://doi.org/10.1200/CCI.23.00116 (2023).
39. El Emam, K., Mosquera, L., Fang, X. & El-Hussuna, A. An evaluation of the replicability of analyses using synthetic health data. *Sci. Rep.* **14**, 6978 (2024).
40. Qian, Z. et al. Synthetic data for privacy-preserving clinical risk prediction. *Sci. Rep.* **14**, 25676 (2024).
41. Hogenboom, J. et al. Actionability of synthetic data in a heterogeneous and rare health care demographic: Adolescents and young adults with cancer. *JCO Clin. Cancer Inform.* **8**, e2400056 (2024).
42. Perez, I. M., Movahedi, P., Nieminen, V., Airola, A. & Pahikkala, T. Does differentially private synthetic data lead to synthetic discoveries?. *Methods Inf. Med.* **63**, 35–51 (2024).
43. Eckardt, J.-N. et al. Mimicking clinical trials with synthetic acute myeloid leukemia patients using generative artificial intelligence. *Npj Digit. Med.* **7**, 1–11 (2024).
44. Raghunathan, T. E., Reiter, J. P. & Rubin, D. B. Multiple imputation for statistical disclosure limitation. *J. Off. Stat.* **19**, 1–16 (2003).
45. Kaabachi, B. et al. A scoping review of privacy and utility metrics in medical synthetic data. *Npj Digit. Med.* **8**, 1–9 (2025).
46. Liu, Z., Rich, B. & Hanley, J. A. Recovering the raw data behind a non-parametric survival curve. *Syst. Rev.* **3**, 151 (2014).
47. Bormann, I. DigitizeIt. Available from: https://www.digitizeit.xyz/
48. Guyot, P., Ades, A., Ouwens, M. J. & Welton, N. J. Enhanced secondary analysis of survival data: Reconstructing the data from published Kaplan-Meier survival curves. *BMC Med. Res. Methodol.* **12**, 9 (2012).
49. Nowok, B., Raab, G. M. & Dibben, C. Synthpop: Bespoke creation of synthetic data in R. *J. Stat. Softw.* **74**, 1–26 (2016).
50. Duan, T. *et al.* NGBoost: Natural gradient boosting for probabilistic prediction. In *Proceedings of the 37th International Conference on Machine Learning* 2690–2700 (PMLR, 2020).
51. Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P. & Sanil, A. P. A framework for evaluating the utility of data altered to protect confidentiality. *Am. Stat.* **60**, 224–232 (2006).
52. Qian, Z. et al. Synthcity: A benchmark framework for diverse use cases of tabular synthetic data. In *Advances in Neural Information Processing Systems* Vol. 36 (ed. Oh, A.) 3173–3188 (Curran Associates, Inc., 2023).

Acknowledgements

The authors declare no acknowledgements beyond the funding.

Author contributions

AD conceived the study, performed the analysis, and drafted the manuscript. CP, JD, HD, PR, TD, SV, and SR were involved in the data interpretation and critically revised the draft versions of the manuscript. All authors read and approved the final manuscript.

Funding
This work was supported by Ghent University Hospital [Fund for Innovation and Clinical Research]. Paloma Rabaey's research is funded by the Research Foundation Flanders (FWO-Vlaanderen) [grant number 1170124N]. Thomas Demeester's research is partly funded by the Flemish government ["Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen"].

Declarations

Competing interests
The authors declare no competing interests.

Human ethics and consent to participate
This study was conducted in accordance with the Declaration of Helsinki. Because this study utilized publicly available, aggregated data (i.e., published Kaplan–Meier curves) that do not contain personally identifiable information, ethics committee submission and informed consent requirements were waived. These curves were converted into reconstructed datasets using a validated algorithm and subsequently synthesized using generative models, as described in the Methods section.

Clinical trial number
The clinical trial numbers of the selected trials are shown in the Supplementary Data.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-62212-z.

Correspondence and requests for materials should be addressed to A.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2026