Supplementary Materials

A Supplementary figures

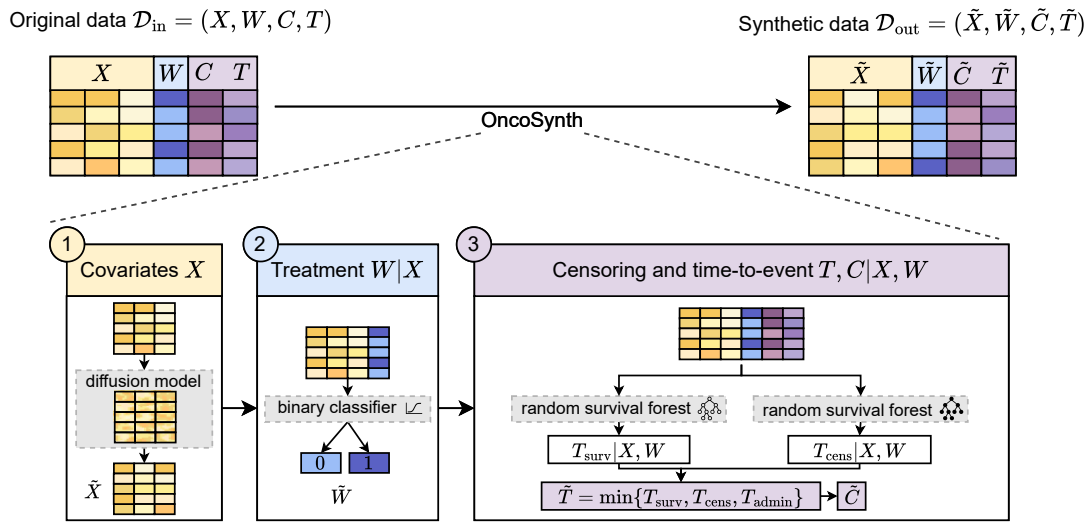

Figure S1: **Causally-aware diffusion-based architecture of OncoSynth for generation of synthetic oncology data.** OncoSynth takes original patient-level data $\mathcal{D}_{\text{in}}$ consisting of covariates X, treatment W, censoring indicator C, and time T, and generates a fully synthetic dataset $\mathcal{D}_{\text{out}} = (\tilde{X}, \tilde{W}, \tilde{C}, \tilde{T})$ that preserves both fidelity and utility for downstream tasks. It models the causal data-generating process through three steps: (1) patient covariates X are generated using a tabular diffusion model; (2) treatment assignment W is generated using a binary classifier conditioned on the covariates ($W \mid X$); (3) time T is derived as the minimum of latent survival time (T_{surv}), latent censoring time (T_{cens}) (generated using random survival forests conditioned on X and W), and administrative censoring time (T_{admin}). The censoring indicator C is given by $C = 1$, if $T = T_{surv}$, and $C = 0$ otherwise.

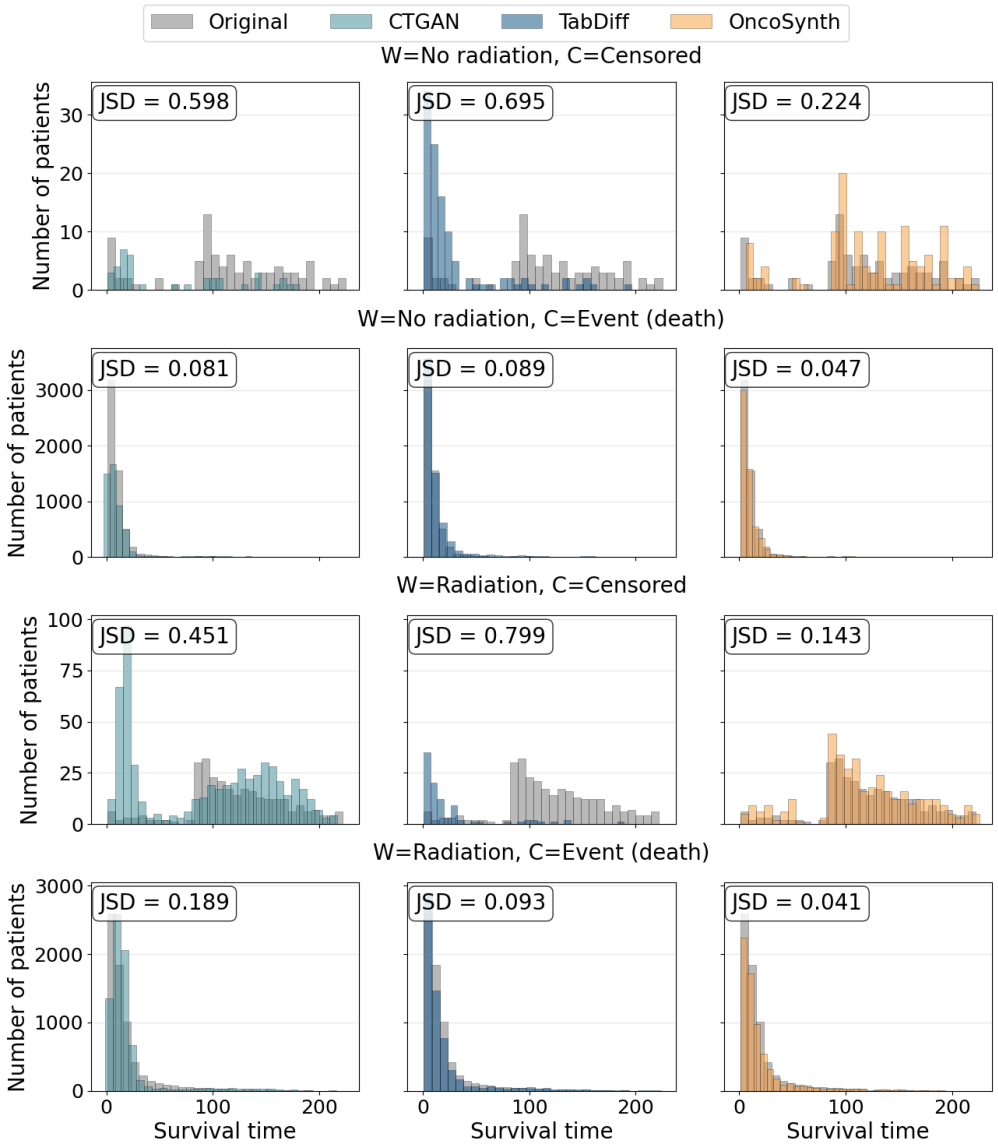

Figure S2: **Fidelity of survival time distributions in the lung cancer cohort.** Histograms show the distribution of survival times in the original data (gray) vs. the synthetic data generated by CTGAN, TabDiff, and OncoSynth. Results are stratified by treatment assignment (no radiotherapy vs. radiotherapy) and censoring status (censored vs. event/death). Lower Jensen-Shannon distance (JSD) indicates higher similarity to original. Across all strata, OncoSynth captures the original survival distributions and outperforms baseline methods.

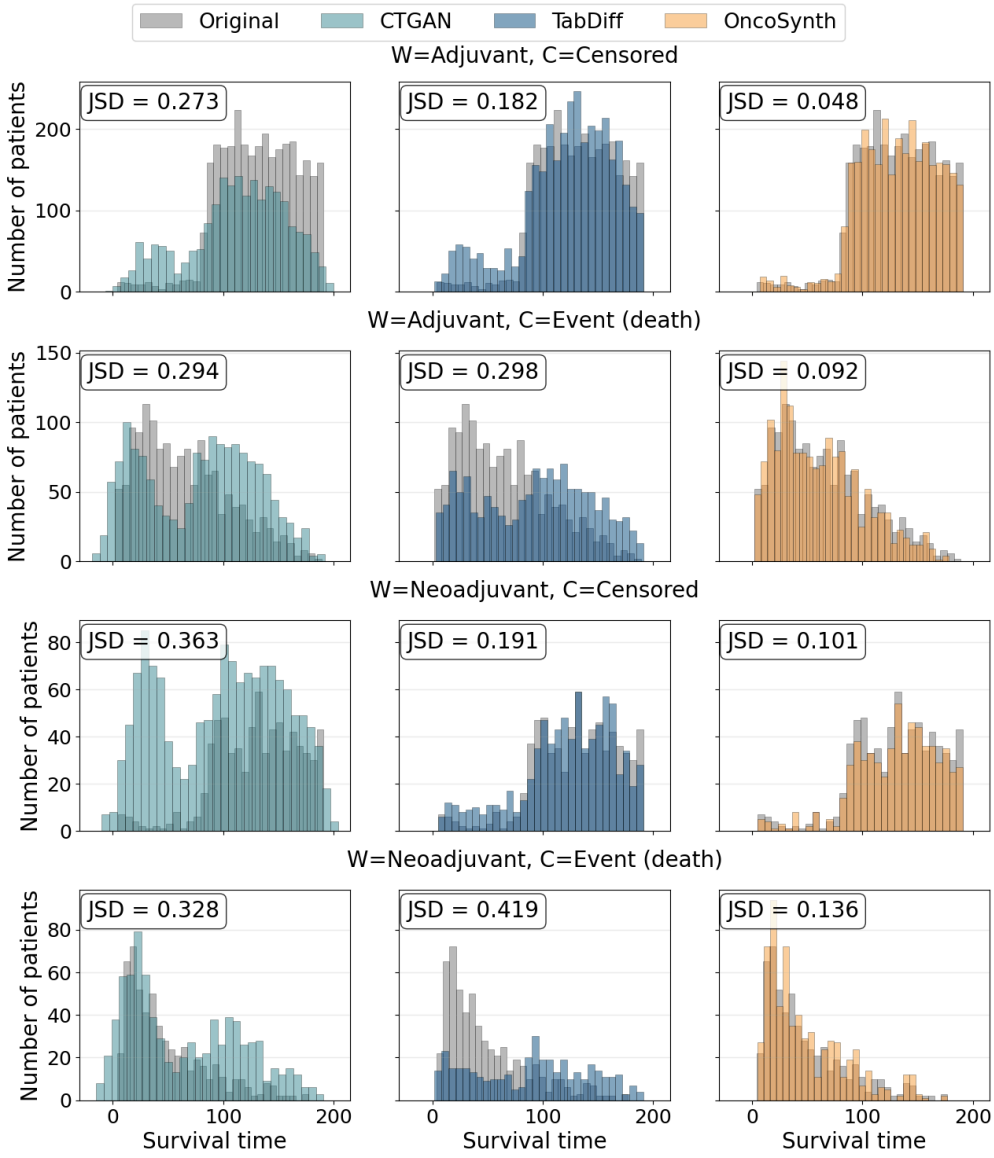

Figure S3: **Fidelity of survival time distributions in the breast cancer cohort.** Histograms show the distribution of survival times in the original data (gray) vs. the synthetic data generated by CTGAN, TabDiff, and OncoSynth. Results are stratified by treatment assignment (adjuvant vs. neoadjuvant) and censoring status (censored vs. event/death). Lower Jensen–Shannon distance (JSD) indicates higher similarity to original. Across all strata, OncoSynth captures the original survival distributions and outperforms baseline methods.

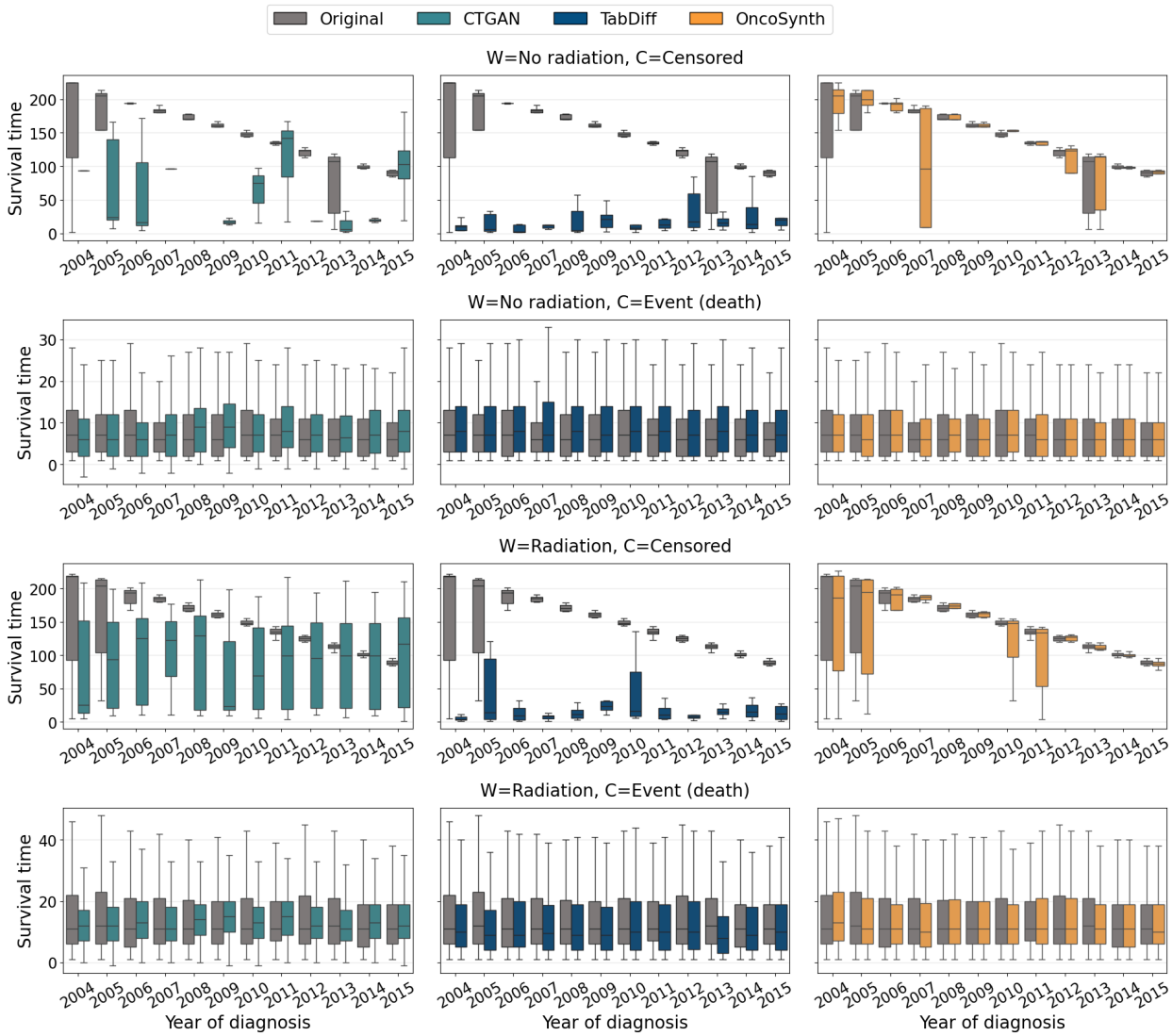

Figure S4: **Fidelity in capturing the impact of administrative censoring on survival times in the lung cancer cohort.** Boxplots compare survival time distributions per diagnosis year between the original data (gray) and synthetic cohorts generated by CTGAN, TabDiff, and OncoSynth, stratified by treatment (no radiotherapy vs. radiotherapy) and censoring status (censored vs. event/death). Administrative censoring influences the distribution of survival times, since patients diagnosed more recently have shorter follow-up periods. OncoSynth faithfully captures the impact of administrative censoring, while baseline methods fail to reproduce this temporal pattern.

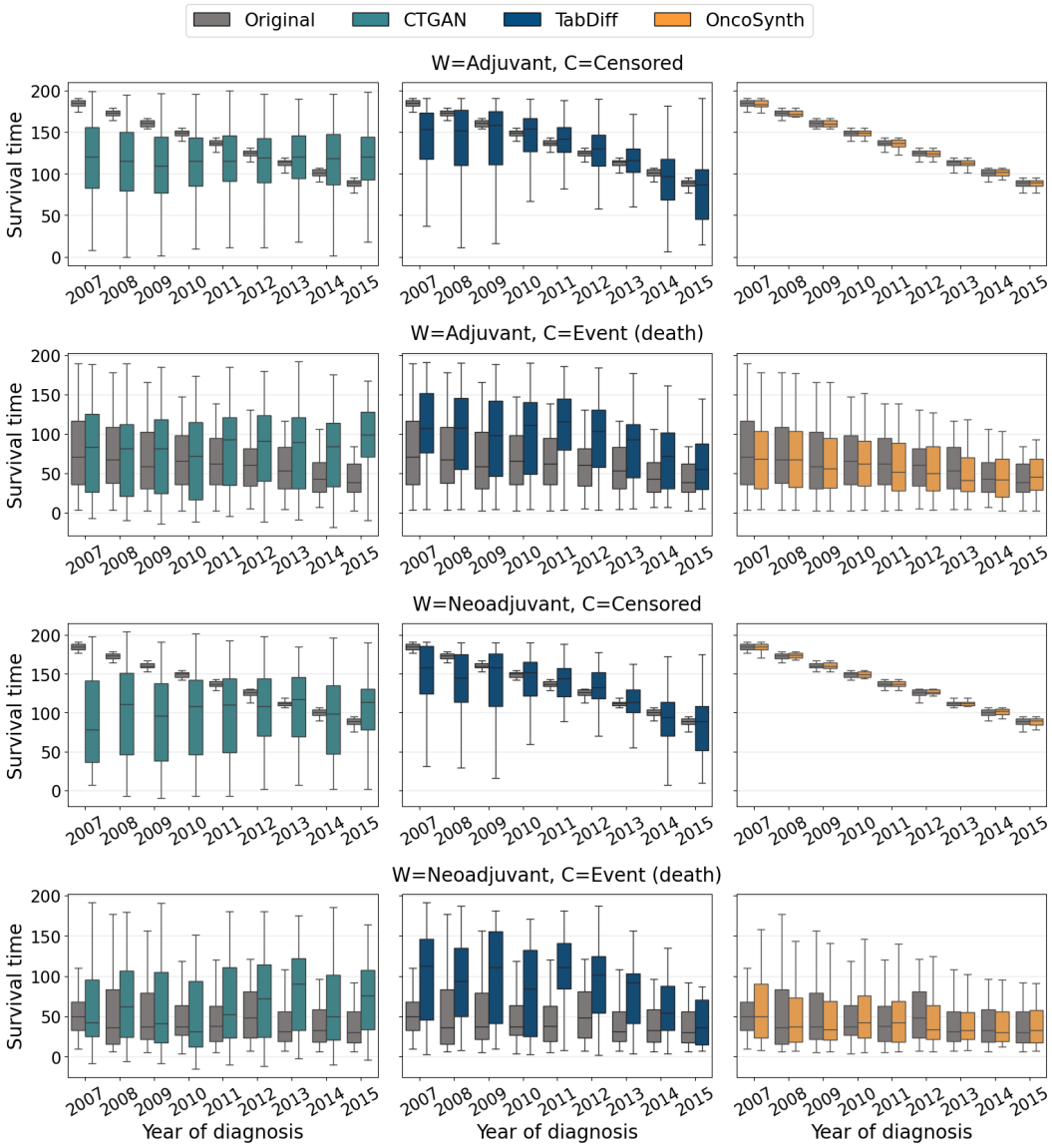

Figure S5: **Fidelity in capturing the impact of administrative censoring on survival times in the breast cancer cohort.** Boxplots compare survival time distributions per diagnosis year between the original data (gray) and synthetic cohorts generated by CTGAN, TabDiff, and OncoSynth, stratified by treatment (adjuvant vs. neoadjuvant) and censoring status (censored vs. event/death). Administrative censoring influences the distribution of survival times, since patients diagnosed more recently have shorter follow-up periods. OncoSynth faithfully captures the impact of administrative censoring, while baseline methods fail to reproduce this temporal pattern.

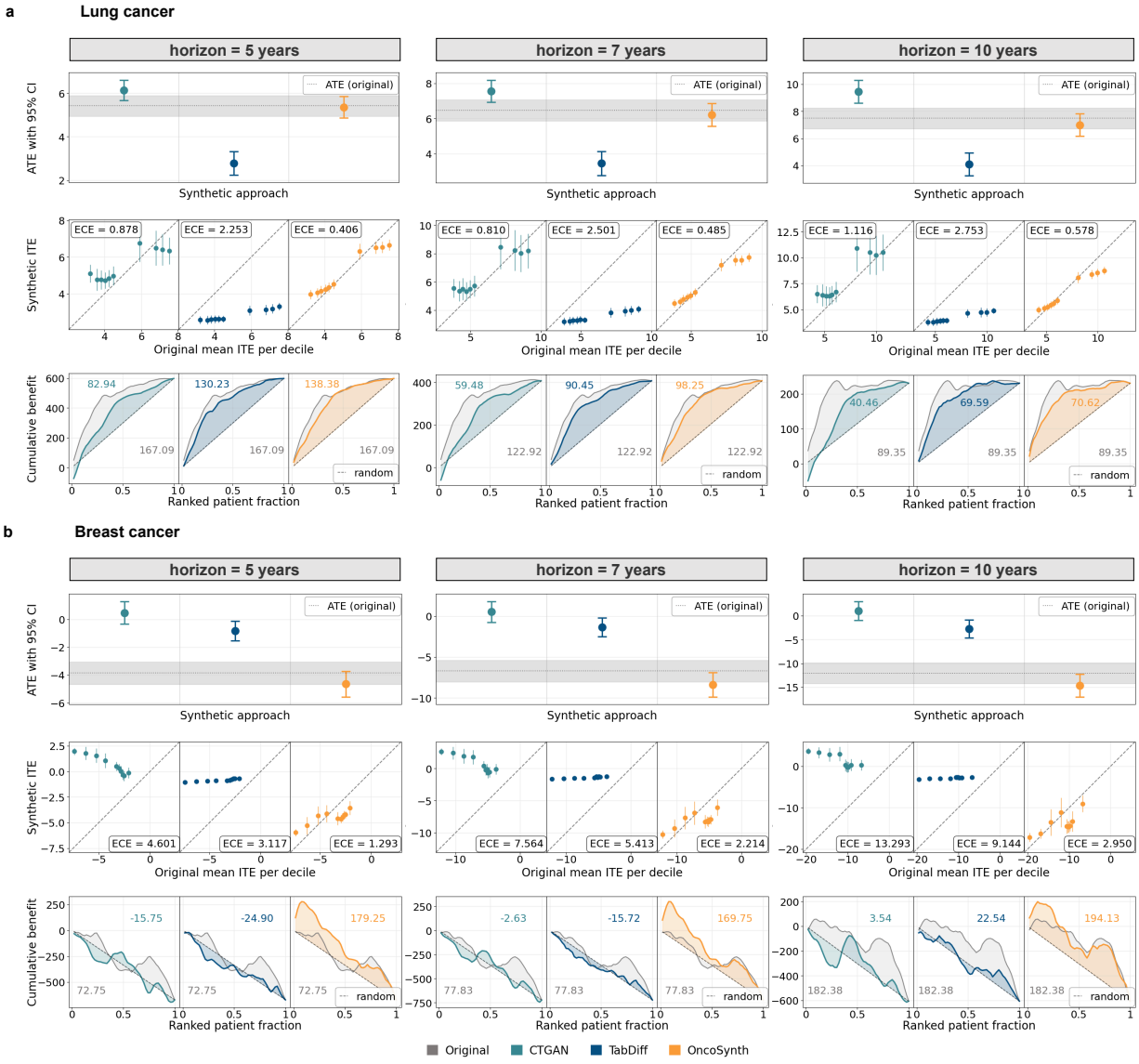

Figure S6: **Sensitivity of treatment effect estimates across follow-up horizons.** Results are shown for prediction horizons of 5, 7, and 10 years for **a,** lung cancer, and **b,** breast cancer. Consistency of population-level treatment effect comparing predicted ATEs with 95% confidence intervals (top row); agreement of patient-level treatment effects comparing mean ITEs per deciles (middle row); clinical utility for treatment allocation policy evaluated via Qini curves (bottom row). Across horizons, OncoSynth consistently produces treatment effect estimates that are closest to the original data, which indicates limited sensitivity to the choice of follow-up horizon.

B Supplementary tables

Table S1: **List of variables used for analysis.**

(a) Lung cancer cohort

Variable	Datatype
Covariates	
Age at diagnosis	numerical
Sex	binary
Race	categorical
Marital status	categorical
Year of diagnosis	categorical (ordinal)
Household income	categorical (ordinal)
Primary site	categorical
Laterality	categorical
AJCC T stage	categorical
AJCC N stage	categorical
AJCC M stage	categorical
Surgery	binary
Chemotherapy	binary
Treatment	
Radiotherapy vs. No radiotherapy	binary
Outcome	
Observed time	numerical
Event/censoring indicator	binary

Abbreviations: AJCC T/N/M stage = American Joint Committee on Cancer tumor/node/metastasis stage.

(b) Breast cancer cohort

Variable	Datatype
Covariates	
Age at diagnosis	numerical
Race	categorical
Marital status	categorical
Year of diagnosis	categorical (ordinal)
Household income	categorical (ordinal)
Tumor size	numerical
Tumor grade	categorical (ordinal)
Positive nodes	numerical
ER status	categorical
PR status	categorical
Treatment	
Adjuvant vs. neoadjuvant	binary
Outcome	
Observed time	numerical
Event/censoring indicator	binary

Abbreviations: ER = estrogen receptor; PR = progesterone receptor.

Table S2: **Patient characteristics for lung cancer cohort.** Continuous variables are reported as median (interquartile range), and categorical variables as counts (percentage). Age is reported in years, time-to-event in months.

Characteristics	No radiotherapy N=16,600	Radiotherapy N=20,528	Total N=37,128
Demographics			
Age	68 [61, 75]	65 [58, 71]	66 [59, 73]
Sex			
male	8,339 (50.2)	10,087 (49.1)	18,426 (49.6)
female	8,261 (49.8)	10,441 (50.9)	18,702 (50.4)
Race			
White	13,663 (82.3)	16,809 (81.9)	30,472 (82.1)
Black	1,393 (8.4)	1,899 (9.3)	3,292 (8.9)
Hispanic	837 (5.0)	898 (4.4)	1,735 (4.7)
other/unknown	707 (4.3)	922 (4.5)	1,629 (4.4)
Marital status			
married	7,987 (48.1)	10,790 (52.6)	18,777 (50.6)
single	2,209 (13.3)	2,792 (13.6)	5,001 (13.5)
divorced	2,558 (15.4)	3,402 (16.6)	5,960 (16.1)
widowed	3,303 (19.9)	2,954 (14.4)	6,257 (16.9)
unknown	543 (3.3)	590 (2.9)	1,133 (3.1)
Tumor characteristics			
Primary site			
main bronchus	1,953 (11.8)	2,502 (12.2)	4,455 (12.0)
upper lobe	7,545 (45.5)	10,545 (51.4)	18,090 (48.7)
middle lobe	691 (4.2)	841 (4.1)	1,532 (4.1)
lower lobe	3,556 (21.4)	3,968 (19.3)	7,524 (20.3)
overlap	297 (1.8)	283 (1.4)	580 (1.6)

Continued on next page

Table S2 continued from previous page

Characteristics	**No radiotherapy**	**Radiotherapy**	**Total**
	N=16,600	N=20,528	N=37,128
unspecified	2,558 (15.4)	2,389 (11.6)	4,947 (13.3)
Laterality			
left	6,868 (41.4)	8,270 (40.3)	15,138 (40.8)
right	9,066 (54.6)	11,729 (57.1)	20,795 (56.0)
other	666 (4.0)	529 (2.6)	1,195 (3.2)
Other therapies			
Surgery	720 (4.3)	549 (2.7)	1,269 (3.4)
Chemotherapy	11,619 (70.0)	18,440 (89.8)	30,059 (81.0)
Outcome			
Censoring			
Censored	256 (1.5)	823 (4.0)	1,079 (2.9)
Event (death)	16,344 (98.5)	19,705 (96.0)	36,049 (97.1)
Time-to-event	7 [2, 12]	12 [6, 23]	9 [4, 17]

Table S3: **Patient characteristics for breast cancer cohort.** Continuous variables are reported as median (interquartile range), and categorical variables as counts (percentage). Age is reported in years, time-to-event in months, and tumor size in mm. Positive nodes indicate the number of histologically confirmed lymph-node involvement. Abbreviations: ER = estrogen receptor, PR = progesterone receptor.

Characteristics	**Adjuvant** N=13,305	**Neoadjuvant** N=3,741	**Total** N=17,046
Demographics			
Age	54 [46, 63]	52 [44, 61]	54 [45, 63]
Race			
White	8,299 (62.4)	2,119 (56.6)	10,418 (61.1)
Black	1,632 (12.3)	667 (17.8)	2,299 (13.5)
Hispanic	2,055 (15.4)	598 (16.0)	2,653 (15.6)
other/unknown	1,319 (9.9)	357 (9.5)	1,676 (9.8)
Marital status			
married	7,809 (58.7)	1,994 (53.3)	9,803 (57.5)
single	2,261 (17.0)	826 (22.1)	3,087 (18.1)
divorced	1,749 (13.1)	478 (12.8)	2,227 (13.1)
widowed	1,022 (7.7)	287 (7.7)	1,309 (7.7)
unknown	464 (3.5)	156 (4.2)	620 (3.6)
Tumor characteristics			
Tumor size	35 [22, 55]	55 [34, 74]	38 [23, 60]
Tumor grade			
I	1,083 (8.1)	192 (5.1)	1,275 (7.5)
II	5,400 (40.6)	1,166 (31.2)	6,566 (38.5)
III	6,074 (45.7)	1,971 (52.7)	8,045 (47.2)
IV	58 (0.4)	35 (0.9)	93 (0.5)
unknown	690 (5.2)	377 (10.1)	1,067 (6.3)

Continued on next page

Table S3 continued from previous page

Characteristics	**Adjuvant** N=13,305	**Neoadjuvant** N=3,741	**Total** N=17,046
Positive nodes	5 [4, 6]	2 [1, 5]	4 [2, 6]
ER status			
negative	2,761 (20.8)	1,548 (41.4)	4,309 (25.3)
positive	10,371 (77.9)	2,126 (56.8)	12,497 (73.3)
other	58 (0.4)	23 (0.6)	81 (0.5)
unknown	115 (0.9)	44 (1.2)	159 (0.9)
PR status			
negative	4,176 (31.4)	1,966 (52.6)	6,142 (36.0)
positive	8,899 (66.9)	1,692 (45.2)	10,591 (62.1)
other	84 (0.6)	34 (0.9)	118 (0.7)
unknown	146 (1.1)	49 (1.3)	195 (1.1)
Outcome			
Censoring			
Censored	9,005 (67.7)	2,083 (55.7)	11,088 (65.0)
Event (death)	4,300 (32.3)	1,658 (44.3)	5,958 (35.0)
Time-to-event	114 [83, 150]	97 [40, 139]	111 [71, 147]

C Implementation details

C.1 Preprocessing

For each raw patient cohort extracted from SEER, variables were harmonized, which included converting continuous variables to numerical valid ranges, and recoding categorical values into clinically meaningful groups. Since CTGAN and TabDiff do not natively support missing values in the input, handling of missing values was conducted during preprocessing. Missing categorical variables were encoded as an "other/unknown" category. Patients with missing or invalid values in key continuous variables (e.g., tumor size) were excluded according to predefined inclusion criteria (Fig. 2), resulting in no remaining missingness in continuous variables. For downstream utility evaluation, we fitted a single preprocessing pipeline on the original subset used for downstream training and applied it to all synthetic datasets as well as the held-out test set, to ensure consistency during evaluation. Continuous variables were standardized and categorical ones were one-hot encoded.

C.2 Configuration of baseline data generators

For TabDiff, we used a compact architecture (`num_layers=2`, `d_token=4`, `n_head=1`, `factor=4`, `dim_t=128`) and a reduced training budget (`batch_size=128`, `learning_rate=0.001`, `weight_decay=`10^{-4}, `steps=500`) to match the scale of the data and reduce computational costs. All remaining hyperparameters were kept at their default values.

Importantly, OncoSynth used the exact same TabDiff architecture and training configuration for its diffusion component. This ensures that any performance differences between OncoSynth and joint TabDiff are not the result of differences in model capacity or optimization, but arise solely from the causally structured, sequential factorization used by OncoSynth. For CTGAN, we used

its default hyperparameter configuration and trained the model for 300 epochs.

C.3 Configuration of OncoSynth

Covariate generation. For the first step, which models the marginal distribution of patient covariates, we used the same configuration used for the TabDiff baseline (Supplement C.2). This ensures that differences between OncoSynth and joint TabDiff were attributable to the causal, sequential factorization rather than differences in the underlying diffusion architecture.

Treatment generation. For the second step, which models the treatment assignment mechanism, we used a logistic regression model with the following hyperparameters: `solver=lbfgs`, `penalty = L2`, `C=1`, `max_iter=5000`. Predicted propensities were calibrated using isotonic regression.

Outcome generation. For the third step, we used four separate RSFs. For each treatment arm, we trained one model to predict survival time (where the 'event' was death and the 'censoring time' was the time to last follow-up in patients without an event), and one model to predict censoring time (where the 'event' was considered the censoring and the 'censoring time' was the survival time for uncensored patients). Each of the four RSFs used the following hyperparameters: `n_estimators=600`, `max_depth=None`, `min_samples_split=6`, `min_samples_leaf=5`, `max_features=0.7`, `bootstrap=True`, `max_leaf_nodes=None`, `max_samples=None`.

C.4 Hyperparameter tuning for downstream utility evaluation

To evaluate whether synthetic data preserved treatment-effect mechanisms, we tuned the hyperparameters of downstream models used for utility assessment using the Optuna [50] framework.

Propensity model. For the treatment-assignment component of the evaluation pipeline, we trained a logistic regression model to estimate propensities. Hyperparameters were optimized based on the binary cross-entropy in a 3-fold cross-validation on the downstream training subset

of the original data. The parameter search space as well as the best parameter for each cohort are listed below. Selected hyperparameters were used for evaluation of all synthetic models.

Causal survival forest for ITE estimation. For the downstream task of predicting treatment effects using causal survival forests [30, 47], we tuned hyperparameters on a subset (80%) of the original downstream training data, which was randomly sampled with stratification by treatment assignment and event status. Since for treatment effect estimation tasks, the ground truth cannot be observed (due to the central problem of causal inference), tuning was based on out-of-bag diagnostics from the forest. The optimization was performed for a fixed prediction horizon of 36 months. We optimized the hyperparameters using a composite objective that favored low orthogonal loss (i.e., mean squared out-of-bag orthogonal scores), good overlap (i.e., penalized extreme propensity estimates), and non-collapsed individualized treatment effect estimates. Hyperparameters were tuned over 50 trials and the resulting hyperparameter were then used consistently across all datasets and horizons for downstream treatment effect estimation. The hyperparameter search space and the selected hyperparameters are as follows:

Hyperparameter	**Search space**	**Selected hyperparameters**	
		Lung cancer cohort	**Breast cancer cohort**
Logistic regression			
`solver`	liblinear/lbfgs/saga	saga	liblinear
`penalty`	L_1/L_2	L_1	L_2
`C`	0.0001–100	0.075	0.021
`class_weight`	None/balanced	None	None
Causal survival forest			
`num_trees`	500–3,000 (step 500)	2,500	3,000
`min_node_size`	10–100 (step 10)	90	50
`mtry`	1–nr. of features	6	5
`honesty`	True/False	True	True
`honesty_fraction`	0.5–0.7 (step 0.1)	0.7	0.7
`alpha`	Fixed	0.05	0.05
`imbalance_penalty`	Fixed	0	0
`sample_fraction`	Fixed	0.5	0.5