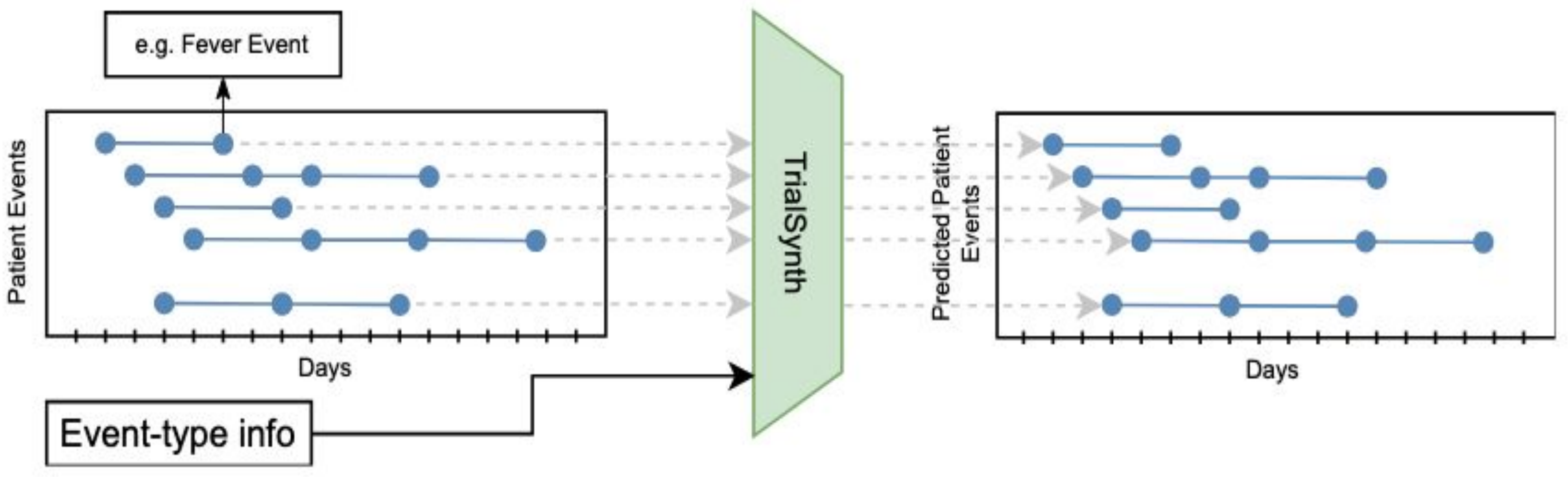
TrialSynth: Generation of Synthetic Sequential Clinical Trial Data

Chufan Gao[1], Mandis Beigi[2], Afrah Shafquat[2], Jacob Aptekar[2], Jimeng Sun[13]
[1]University of Illinois Urbana Champaign
[2]Medidata Solutions
[3]Carle Illinois College of Medicine
chufan2@illinois.edu

Introduction: Lack of Data Availability in Clinical Trials

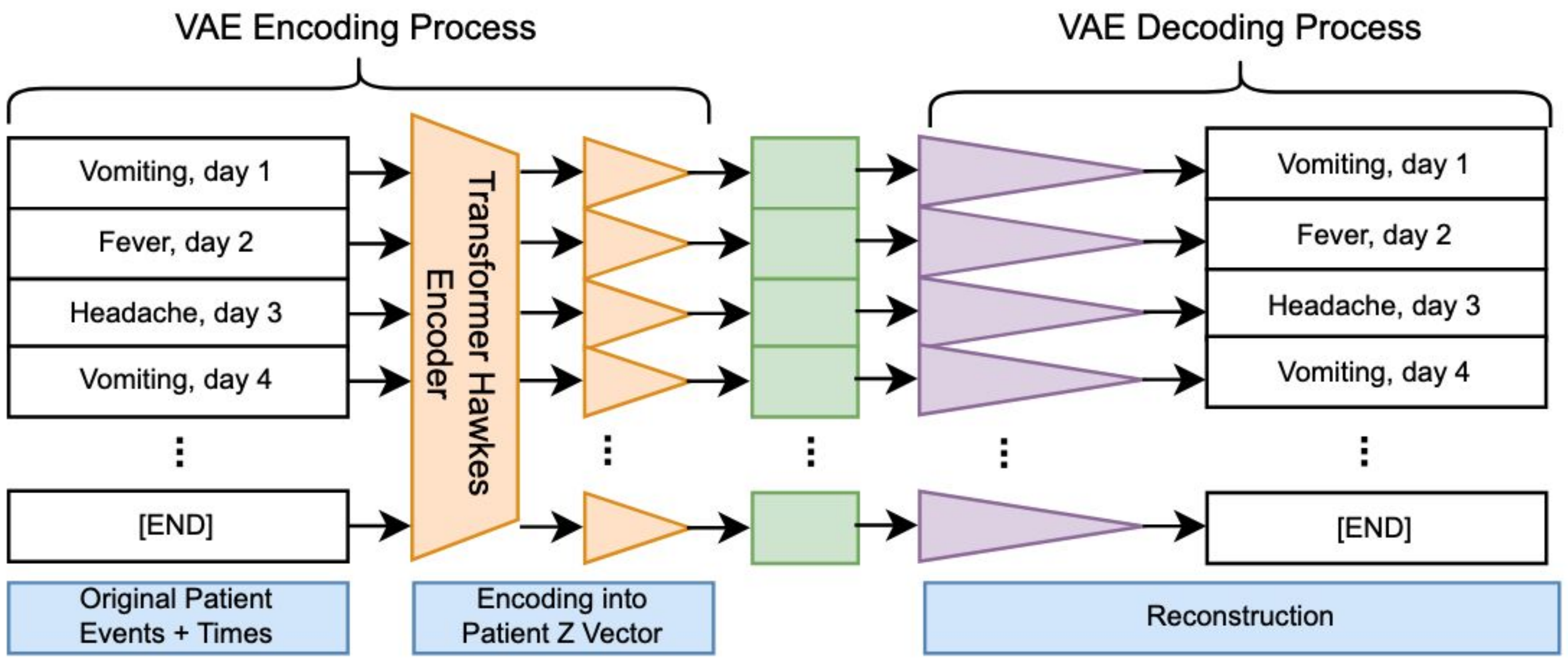
- Analyzing past clinical trial data is crucial for optimizing new trials and expediting drug development.
- However, challenges such as patient privacy, industry competition, and small dataset sizes hinder data availability.
- TrialSynth addresses these challenges by generating synthetic, high-fidelity sequential clinical trial data that mimics real-world patient trajectories.

Task: Generating Sequential Patient Events



- Input: A real patient event sequences with timestamps
- Output: A synthetic patient event sequences, where similarity to original input can be controlled for fidelity / privacy tradeoff

Methodology: Hawkes Process + VAE



- Transformer Hawkes Process [1] processes input sequential patient events
- Variational Autoencoder (VAE) allows for controlled randomness (by varying variance around latent vector (green))
- *VAE Latent Dimension is a function of maximum number of events per patient (padded to the max # events / subject)*

Methodology: Loss Functions

- Combined loss L = L_{hawkes} + L_{elbo} + L_{length}
- The L_{hawkes} is the log-likelihood of the sequence given the Hawkes process, given the predicted likelihood of each event at each time

$$\ln P_\theta(\{(t_1,k_1),\ldots,(t_L,k_L)\}|\boldsymbol{z}) = \sum_{j=1}^{L}\log(\lambda_\theta(t_j|\mathcal{H}_{t_j,z}))-\int_{t_1}^{t_L}\lambda_\theta(t|\mathcal{H}_{t,z})dt$$

λ_θ is the intensity function of any event occurring at time t, given previously predicted events $\mathcal{H}_{t,z}$

- L_{elbo} is the VAE loss consisting of 3 parts:
  - KL divergence from a standard Gaussian
  - Mean-squared error reconstruction loss of the event times
  - Cross-entropy loss of the event types
- L_{length} to ensure the model learns proper stopping criterion
  - Cross Entropy Loss of a [End] event (appended to the to the input sequence)

Experiments: Baseline Models

- LSTM VAE: is the same as our proposed model, except with an LSTM instead of a Transformer encoder
- PARSynthesizer: Specifically tailored for synthesizing sequential tabular data
- TabDDPM: Diffusion-based SOTA general tabular synthesizer
- CTGAN: Tabular GAN for general tabular synthesizer
- HALO: SOTA EHR generation using transformers

Experiments: Clinical Trial Datasets

Obtained from Project Data Sphere [2] (freely available for researchers after creating an account). Note the small # of data points and large label imbalance.

Dataset	Description	# Rows	# Subjects	# Events	Events / Subject	Positive Label Proportion
NCT00003299 (LC1)	Small Cell Lung Cancer	20210	548	34	36.880	0.951
NCT00041119 (BC1)	Breast Cancer	2983	425	150	7.019	0.134
NCT00079274 (CC)	Colon Cancer	316	70	18	4.514	0.184
NCT00174655 (BC2)	Breast Cancer	7002	953	21	7.347	0.019
NCT00312208 (BC3)	Breast Cancer	2193	378	182	5.802	0.184
NCT00694382 (VTE)	Venous Thromboembolism in Cancer Patients	7853	803	746	9.780	0.456
NCT03041311 (LC2)	Small Cell Lung Cancer	1043	47	207	22.192	0.622

Utility Evaluation: Downstream Binary Mortality Prediction

Models were trained on Original Data vs Synthetic Data from each method, TrialSynth performs the best

Dataset	Original Data	LSTM VAE	PAR	CTGAN	TabDDPM	HALO	TrialSynth
LC1	$0.689_{\pm 0.105}$	$0.563_{\pm 0.053}$	$0.504_{\pm 0.066}$	$0.508_{\pm 0.122}$	$0.557_{\pm 0.055}$	$0.457_{\pm 0.079}$	$\mathbf{0.672}_{\pm 0.061}$
BC1	$0.678_{\pm 0.078}$	$0.617_{\pm 0.036}$	$0.573_{\pm 0.043}$	$0.550_{\pm 0.046}$	$\mathbf{0.630}_{\pm 0.045}$	$0.461_{\pm 0.184}$	$\mathbf{0.651}_{\pm 0.046}$
CC	$0.657_{\pm 0.140}$	$0.481_{\pm 0.092}$	$0.567_{\pm 0.096}$	$0.448_{\pm 0.023}$	$\mathbf{0.583}_{\pm 0.098}$	$0.446_{\pm 0.02}$	$\mathbf{0.652}_{\pm 0.015}$
BC2	$0.660_{\pm 0.128}$	$0.535_{\pm 0.073}$	$0.523_{\pm 0.074}$	$0.523_{\pm 0.11}$	$0.513_{\pm 0.078}$	$0.503_{\pm 0.075}$	$\mathbf{0.599}_{\pm 0.042}$
BC3	$0.632_{\pm 0.072}$	$0.454_{\pm 0.039}$	$0.463_{\pm 0.039}$	$0.493_{\pm 0.013}$	$0.503_{\pm 0.043}$	$0.535_{\pm 0.183}$	$\mathbf{0.620}_{\pm 0.038}$
VTE	$0.640_{\pm 0.038}$	$0.490_{\pm 0.019}$	$0.549_{\pm 0.022}$	$0.508_{\pm 0.113}$	$0.531_{\pm 0.021}$	$0.485_{\pm 0.066}$	$\mathbf{0.618}_{\pm 0.024}$
LC2	$0.738_{\pm 0.149}$	$0.563_{\pm 0.097}$	$0.507_{\pm 0.087}$	$0.573_{\pm 0.118}$	$0.574_{\pm 0.096}$	$0.534_{\pm 0.078}$	$\mathbf{0.729}_{\pm 0.044}$

Privacy Evaluation: ML Inference Score

Models were trained to predict whether data was real vs synthetic. TrialSynth performs well here as well

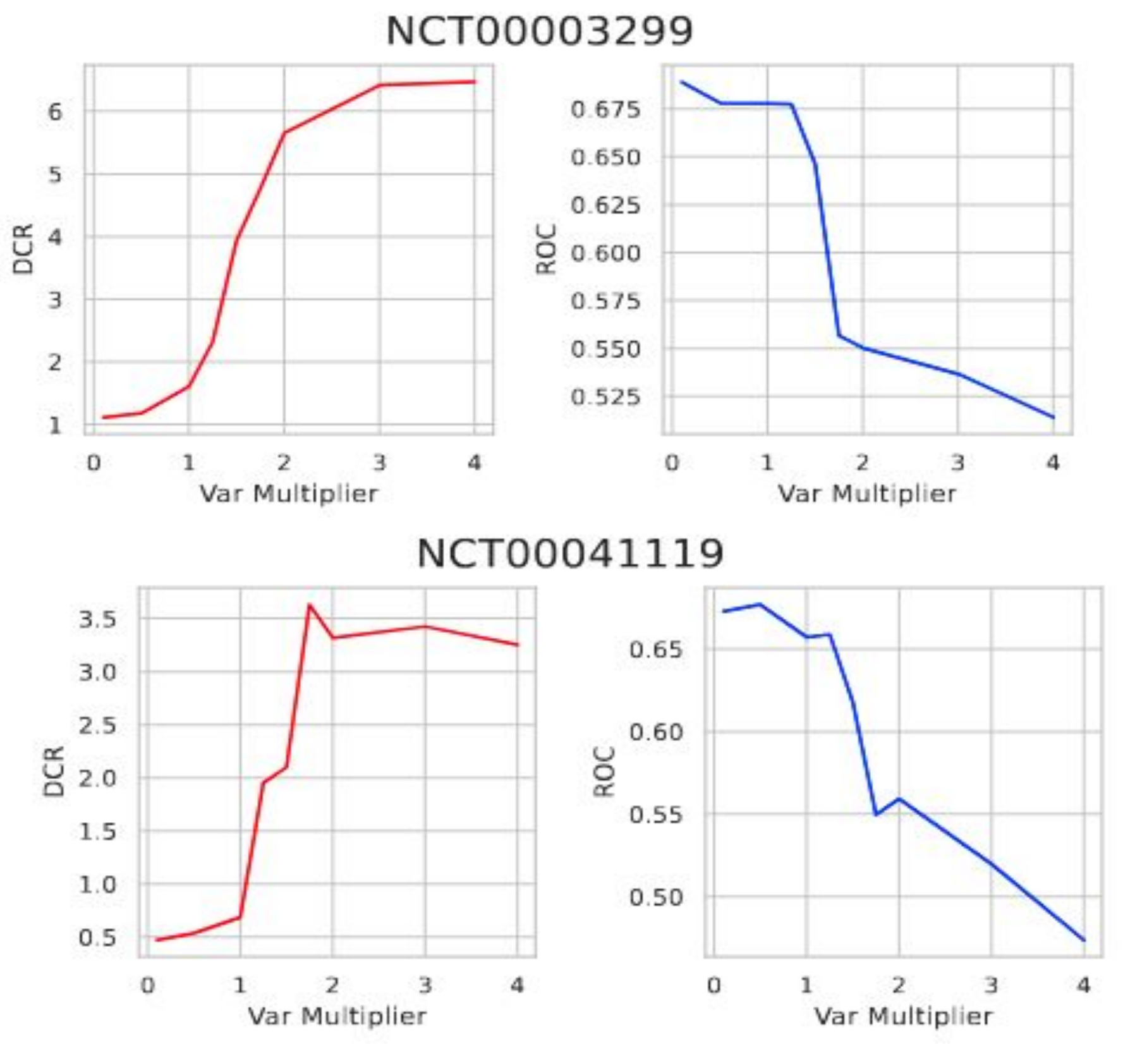
Dataset	LSTM VAE	PAR	CTGAN	TabDDPM	HALO	TrialSynth
LC1	$1.000_{\pm 0.000}$	$0.968_{\pm 0.010}$	$0.952_{\pm 0.056}$	$0.762_{\pm 0.024}$	$1.000_{\pm 0.004}$	$\mathbf{0.613}_{\pm 0.024}$
BC1	$0.932_{\pm 0.017}$	$0.998_{\pm 0.002}$	$0.973_{\pm 0.082}$	$0.926_{\pm 0.017}$	$1.000_{\pm 0.001}$	$\mathbf{0.616}_{\pm 0.025}$
CC	$1.000_{\pm 0.000}$	$0.807_{\pm 0.082}$	$0.935_{\pm 0.056}$	$0.894_{\pm 0.050}$	$0.998_{\pm 0.005}$	$0.711_{\pm 0.051}$
BC2	$1.000_{\pm 0.000}$	$0.999_{\pm 0.001}$	$0.998_{\pm 0.075}$	$0.998_{\pm 0.001}$	$0.999_{\pm 0.001}$	$\mathbf{0.605}_{\pm 0.048}$
BC3	$0.994_{\pm 0.007}$	$0.874_{\pm 0.026}$	$0.895_{\pm 0.098}$	$0.729_{\pm 0.035}$	$0.992_{\pm 0.008}$	$\mathbf{0.689}_{\pm 0.023}$
VTE	$1.000_{\pm 0.000}$	$0.923_{\pm 0.012}$	$0.879_{\pm 0.119}$	$0.992_{\pm 0.005}$	$0.000_{\pm 0.004}$	$\mathbf{0.871}_{\pm 0.014}$
LC2	$1.000_{\pm 0.000}$	$0.651_{\pm 0.112}$	$0.982_{\pm 0.038}$	$0.374_{\pm 0.021}$	$0.000_{\pm 0.003}$	$\mathbf{0.573}_{\pm 0.111}$

Privacy / Fidelity Tradeoff Curves

Distance to Closest Record (DCR) compares distance of synthetic data to real data. Higher is more private. Variance denotes the variance parameter in VAE sampling.

- DCR is computed on event features (e.g. mean time per event)

Performance trends inversely to higher privacy, as expected



Discussion

- TrialSynth balances optimal performance on small datasets while offering control over privacy and utility.
- Shows promise for future applications that demand high-quality, secure synthetic datasets
- Future Work: Evaluation in general sequential tabular data domain, extension to very long sequences

References:
[1] Zuo, S., Jiang, H., Li, Z., Zhao, T., & Zha, H. (2020, November). Transformer hawkes process. In International conference on machine learning (pp. 11692-11702). PMLR.
[2] https://data.projectdatasphere.org/