Health Care Breakout Session 4:
Adversarial Time-to-Event Modeling

Matt Engelhard

Today

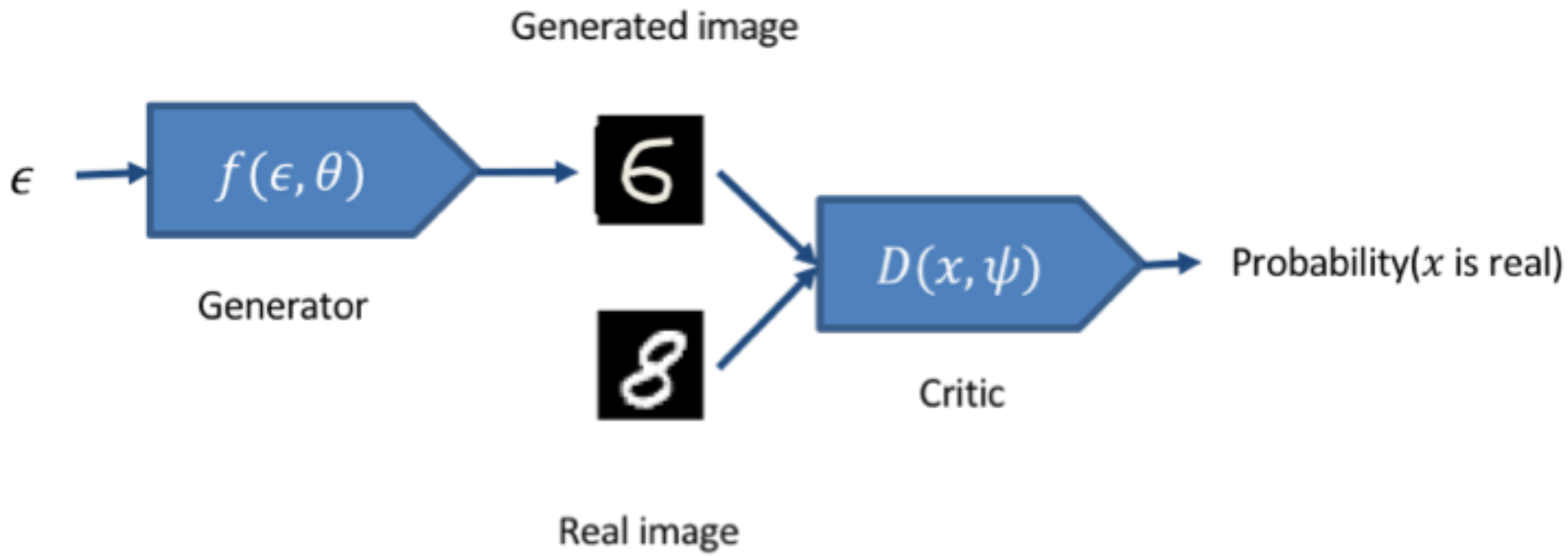
- How can we use GAN and adversarial learning in medicine?

- "Adversarial Time-to-Event Modeling" (Chapfuwa et al., 2018)
  - Critically evaluate despite complexity of methods
  - Highlight as a novel, medically-relevant application of adversarial learning

Medical Applications of

ADVERSARIAL LEARNING

Generating Realistic Samples



Critic: real or fake?

Generator: make the critic's job difficult

Beyond Celebrities' Faces?

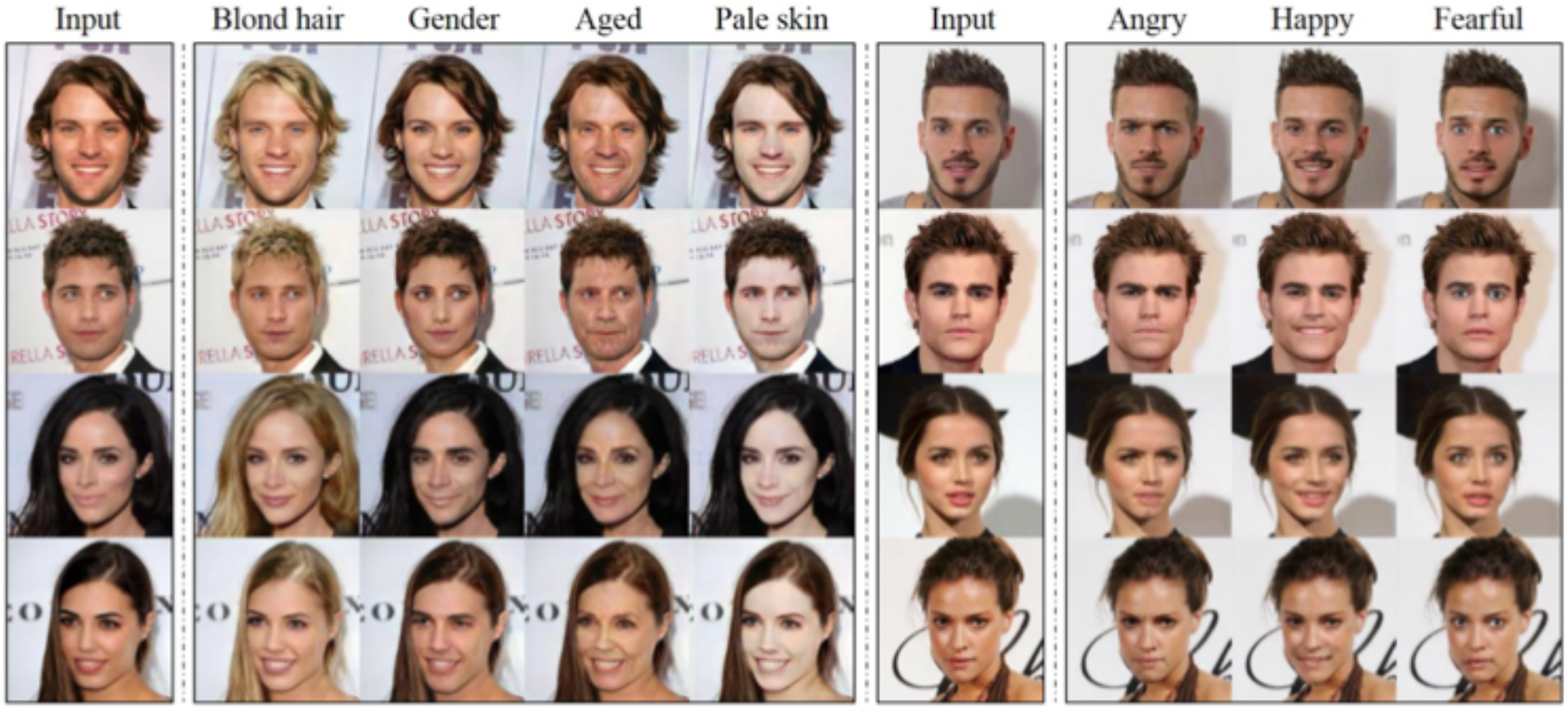


Figure 1. Multi-domain image-to-image translation results on the CelebA dataset via transferring knowledge learned from the RaFD dataset. The first and sixth columns show input images while the remaining columns are images generated by StarGAN. Note that the images are generated by a single generator network, and facial expression labels such as angry, happy, and fearful are from RaFD, not CelebA.

Generating Multi-label Discrete Patient Records using Generative Adversarial Networks

Edward Choi[1] MP2893@GATECH.EDU

Siddharth Biswal[1] SBISWAL7@GATECH.EDU

Bradley Malin[2] BRADLEY.MALIN@VANDERBILT.EDU

Jon Duke[1] JON.DUKE@GATECH.EDU

Walter F. Stewart[3] STEWARWF@SUTTERHEALTH.ORG

Jimeng Sun[1] JSUN@CC.GATECH.EDU

[1] GEORGIA INSTITUTE OF TECHNOLOGY [2] VANDERBILT UNIVERSITY [3] SUTTER HEALTH

Abstract

Access to electronic health record (EHR) data has motivated computational advances in medical research. However, various concerns, particularly over privacy, can limit access to and collaborative use of EHR data. Sharing synthetic EHR data could mitigate risk.

In this paper, we propose a new approach, medical Generative Adversarial Network (medGAN), to generate realistic synthetic patient records. Based on input real patient records, medGAN can generate high-dimensional discrete variables (e.g., binary and count features) via a combination of an autoencoder and generative adversarial networks. We also propose minibatch averaging to efficiently avoid mode collapse, and increase the learning efficiency with batch normalization and shortcut connections. To demonstrate feasibility, we showed that medGAN generates synthetic patient records that achieve comparable performance to real data on many experiments including distribution statistics, predictive modeling tasks and a medical expert review. We also empirically observe a limited privacy risk in both identity and attribute disclosure using medGAN.

Synthetic Patient Data / Images

Low-Dose CT Denoising

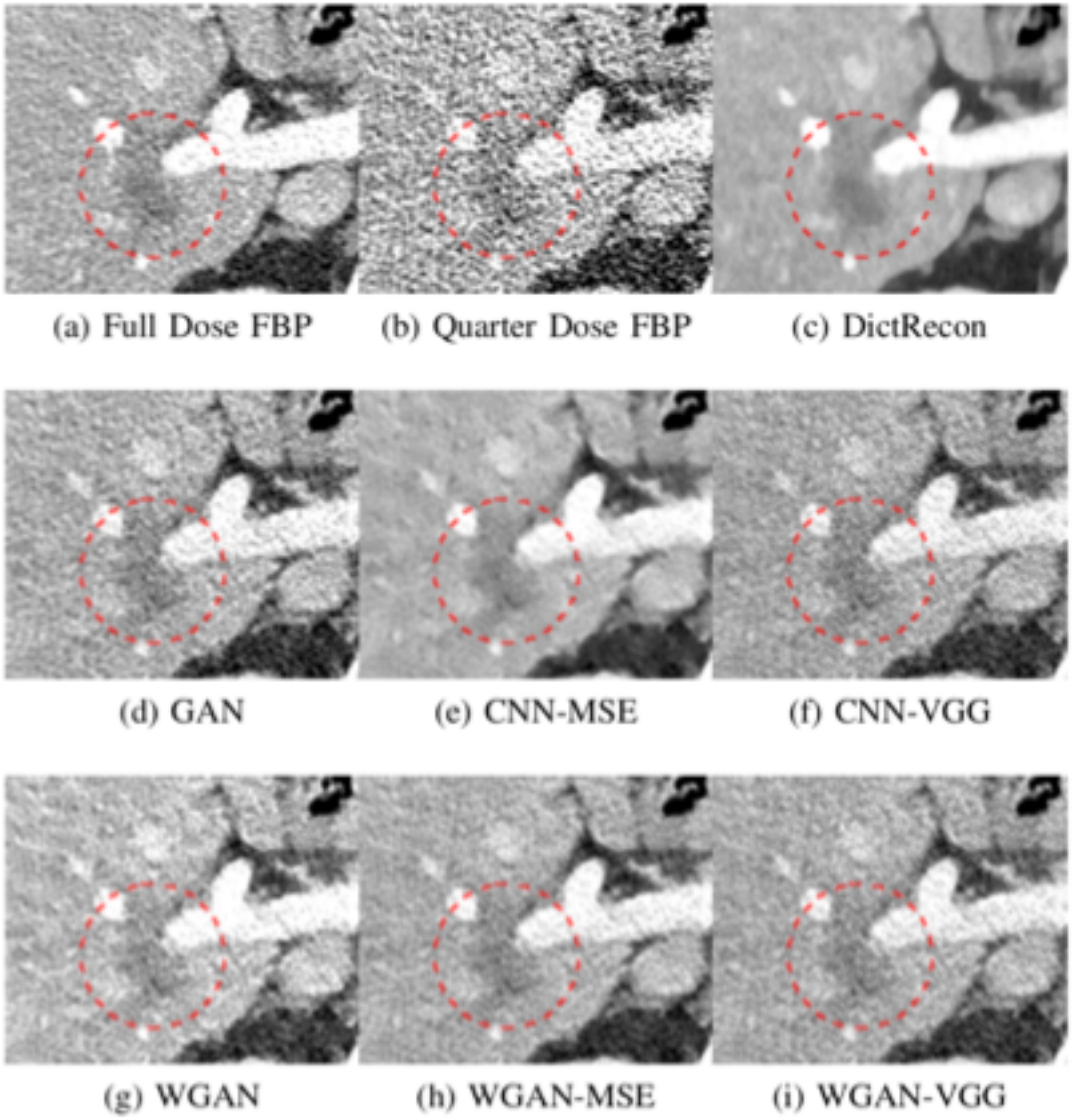


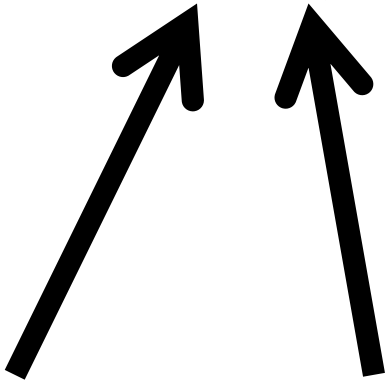
Fig. 6. Zoomed ROI of the red rectangle in Fig. 5. The low attenuation liver lesion with in the dashed circle represents metastasis. The lesion is difficult to assess on quarter dose FBP recon (b) due to high noise content. This display window is [-160, 240]HU.

Yang, Qingsong, et al. "Low dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss." *IEEE transactions on medical imaging*(2018).

Sampling Time-to-Event

draw from:

$$p(t|\boldsymbol{x})$$



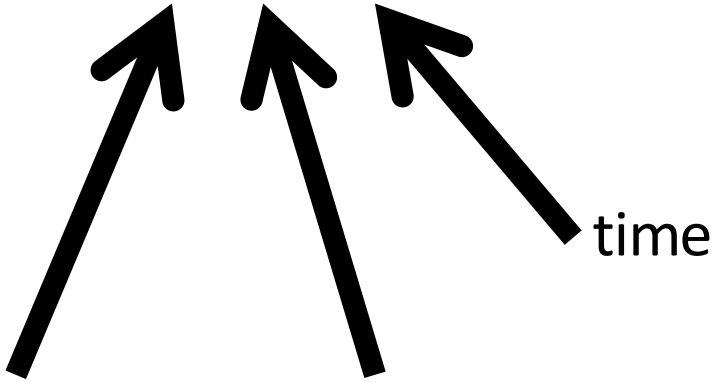
time to a specific event of interest

covariates (i.e. patient characteristics)

Prognosis by Sampling Outcomes?

draw from:

$$p(o|\boldsymbol{x}, t)$$



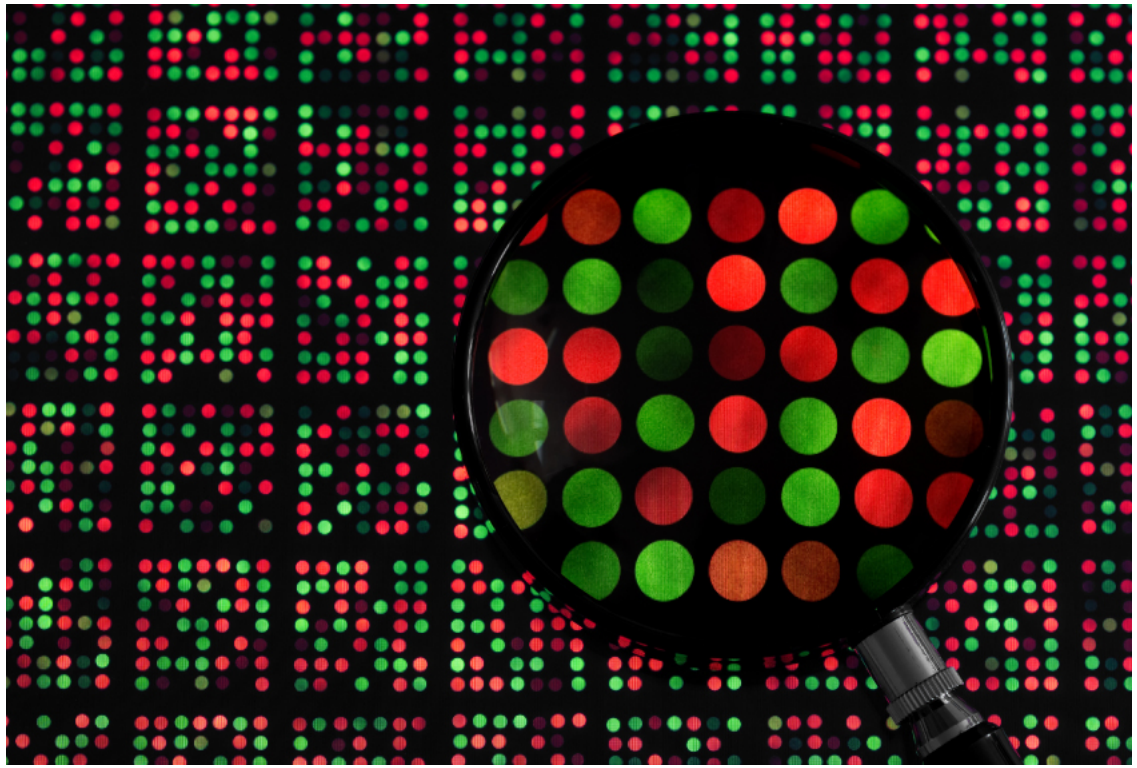
time

outcome(s) of interest
(e.g. ADHD symptoms)

covariates
(i.e. patient characteristics)

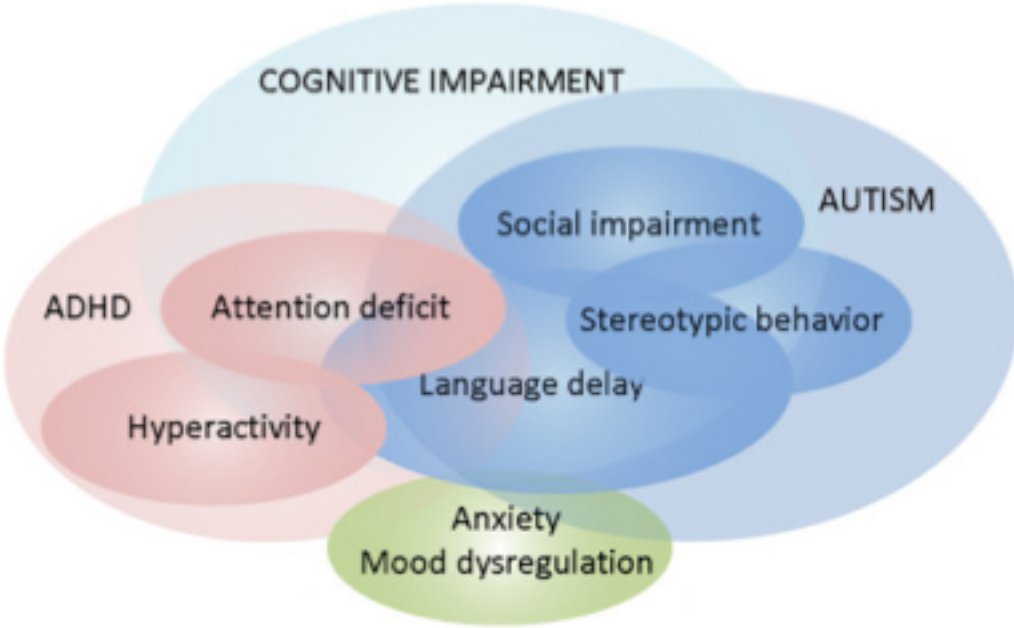
Broader Applications…

Microarray Data



https://www.spectrumnews.org/news/toolbox/online-tool-can-mix-match-gene-expression-data/

Psychiatric Symptoms



https://www.researchgate.net/publication/46392638_Fragile_X_and_autism_Intertwined_at_the_molecular_level_leading_to_targeted_treatments

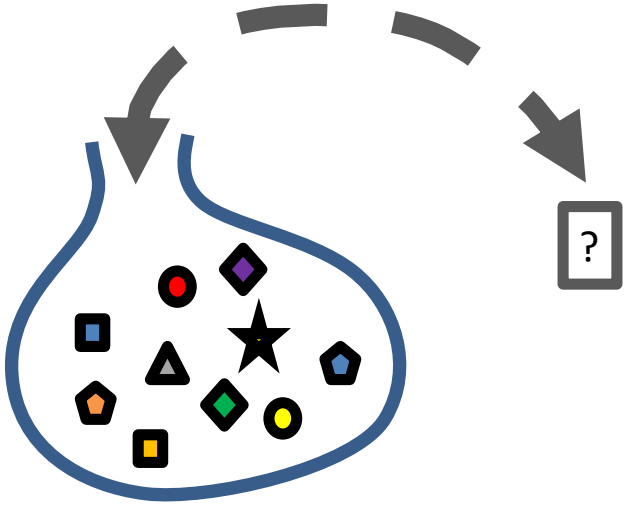
Adversarial Time-to-Event Modeling

Chapfuwa P, Tao C, Li C, Page C, Goldstein B, Carin L, Henao R

arXiv:1804.03184. 2018 Apr 9

Q: In this work, the authors learn to sample from a specific distribution using a conditional generative adversarial network (GAN). What is the distribution in question, what is it conditioned on, and what do samples from it represent?

Drawing Event Times



draw from a bag:
which object?
(discrete)

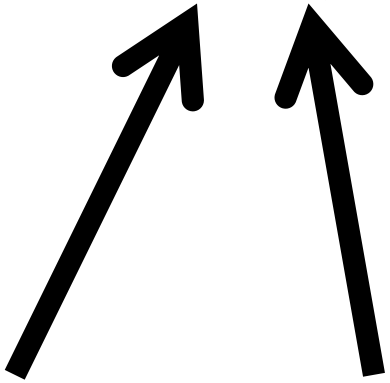
spin a top:
how long will it stay up?
(continuous)

drive to work:
how long will it take?
(continuous and
depends on covariates)

Sampling Time-to-Event

draw from:

$$p(t|\boldsymbol{x})$$



time to a specific event of interest

covariates (i.e. patient characteristics)

Parametric Distribution: Less Flexible

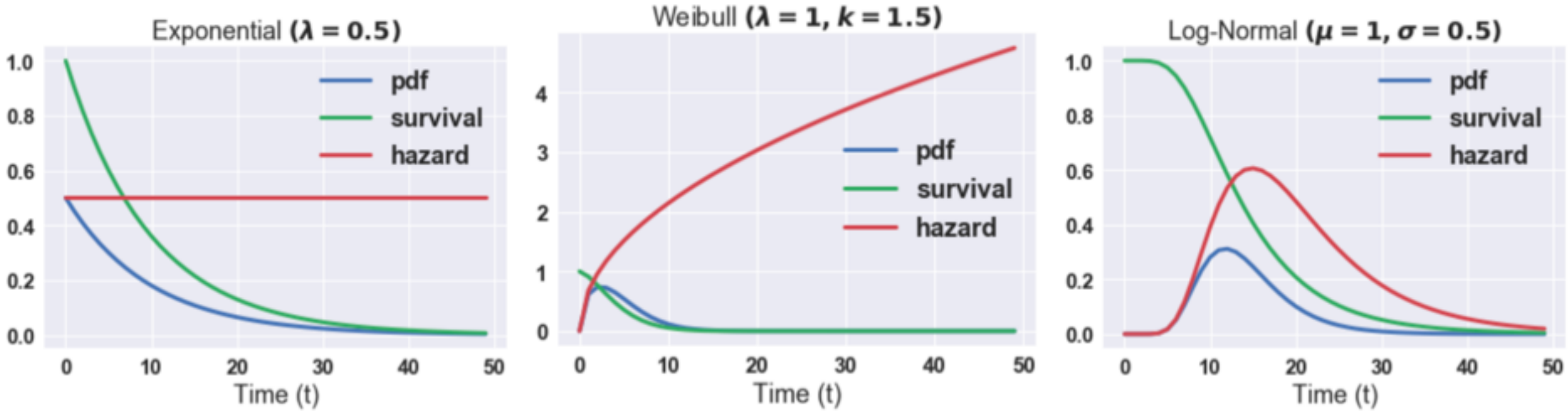
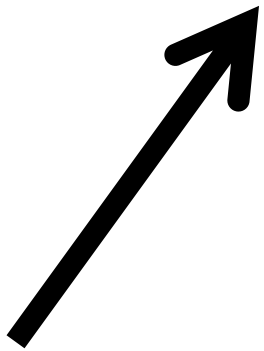


Figure 10. Popular parametric characterizations: exponential (left), Weibull (middle) and log-normal (right).

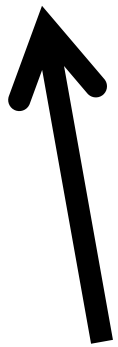
Cox Proportional Hazards

hazard rate:

$$h(t|\boldsymbol{x}) = h_0(t)\exp(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})$$



time-varying
hazard rate

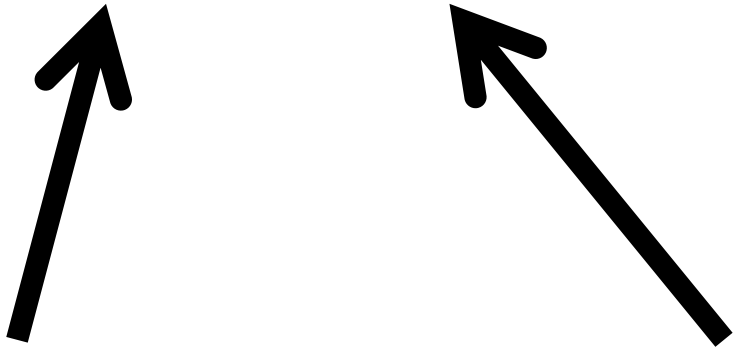


time-independent
covariate effect

Accelerated Failure Time (AFT)

hazard rate:

$$h(t|\boldsymbol{x}) = h_0\big(\exp(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})t\big)\exp(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})$$



parametric distribution

acceleration by covariates

Deep Adversarial Time-to-Event

$$p(t|\boldsymbol{x})$$

time to a specific event of interest

covariates (i.e. patient characteristics)

approximate via conditional GAN

Deep Adversarial Time-to-Event

$p(t|\boldsymbol{x})$

approximate via conditional GAN

$t = G_\theta(\boldsymbol{x}, \boldsymbol{\epsilon}; l = 1)$

$\boldsymbol{\epsilon} \sim p_\epsilon(\boldsymbol{\epsilon})$

Generator (DNN)

$D_\phi(\boldsymbol{x}, t)$

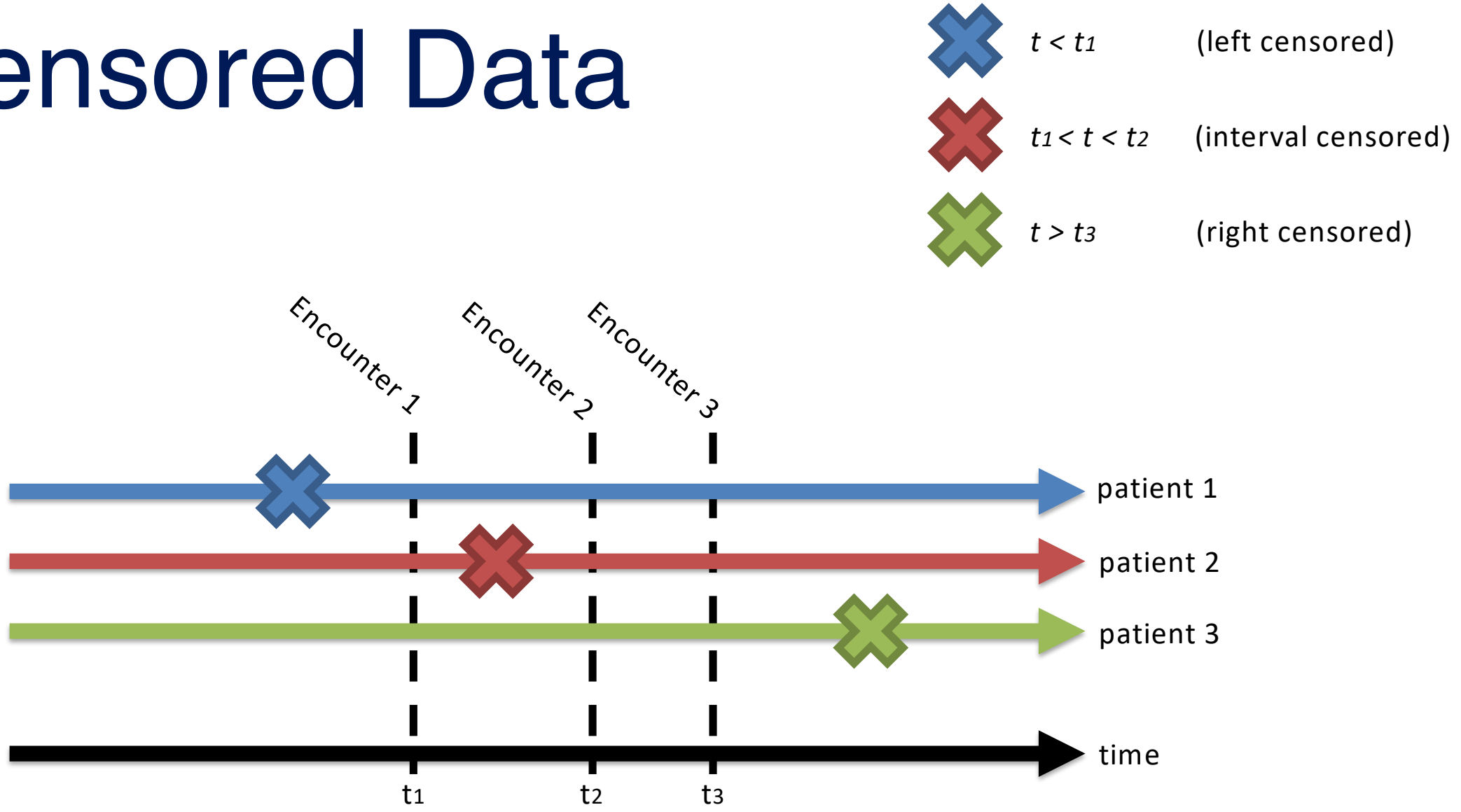
Discriminator (DNN)

censoring indicator

deep neural network parameters

Q: What is censored data, and why is it common in medical applications? How does censoring complicate this analysis?

Censored Data



EVALUATION STRATEGY

Comparison (Baseline) Methods

- Non deep learning:
  - Cox-Efron (Cox-PH modified for censored data)
  - Random Survival Forests

- Deep Learning:
  - Deep Regularized Accelerated Failure Time

Q: What metric is used to evaluate the performance of DATE, and how is it calculated (for both censored and non-censored events)?

Performance Measures

Relative Absolute Error	$\lvert t - \hat{t} \rvert / t_{max}$ (uncensored)	$\max(0, t - \hat{t}) / t_{max}$ (censored)
Concordance Index	quantifies the degree to which the order of the predicted times is consistent with the ground truth	
Qualitative Inspection	common with generative models to subjectively evaluate the quality of generated samples	

Evaluation on Four Datasets

- FLCHAIN: predict survival time from non-clonal serum immunoglobin free light chains
- SUPPORT: predict survival time of seriously-ill hospitalized adults
- SEER: 10-year breast cancer follow-up

publicly-available

- EHR: predict inpatient visits among patients with type-2 diabetes

from Duke

Q: Competition to achieve best results on publicly-available datasets has helped to push ML forward as a field. How can we encourage this? What is the "grand challenge" of ML in medicine?

Evaluation on Four Datasets

Table 1. Summary of datasets used in experiments.

	EHR	FLCHAIN	SUPPORT	SEER
Events (%)	23.9	27.5	68.1	51.0
N	394,823	7,894	9,105	68,082
p (cat)	729 (106)	26 (21)	59 (31)	789 (771)
NaN (%)	1.9	2.1	12.6	23.4
$t_{\max}$	365 days	5,215 days	2,029 days	120 months

Train, Validation, Test

All Datasets:

80% train

10% validation (early stopping, hyperparameters)

10% test

- stratified by non-censored / censored event proportion

"To avoid bias due to multiple encounters per patient, we split the training, validation and test sets so that a given patient can only be in one of the sets."

RESULTS

Qualitative Evaluation: 200 Samples

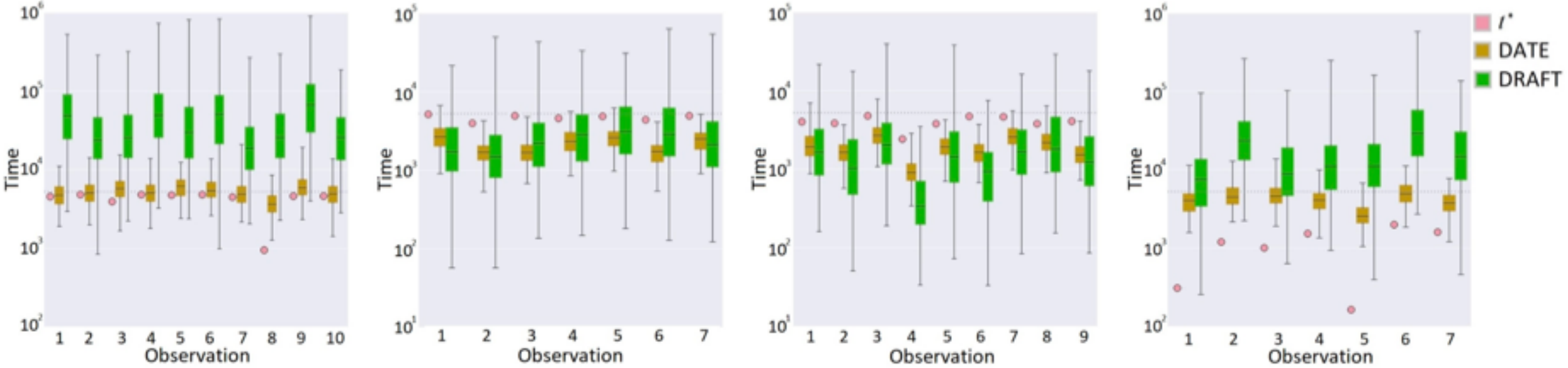


Figure 2. Example test-set predictions on FLCHAIN data. Top best (left) and worst (middle-left) predictions on censored events, and top best (middle-right) and worst (right) predictions on non-censored events. Circles denote ground-truth events or censoring points, while box-plots represent distributions over 200 samples for both DATE and DRAFT. The horizontal dashed line represents the range ($t_{\max} = 5,215$ days) of the events.

Relative Absolute Error

Table 2. Median relative absolute errors (as percentages of $t_{\max}$), on non-censored data. Ranges in parentheses are 50% empirical ranges over (median) test-set predictions.

	DATE	DATE-AE	DRAFT
EHR	**23.6**$_{(11.1,43.0)}$	24.5$_{(12.4,44.0)}$	36.7$_{(16.1,81.3)}$
FLCHAIN	19.5$_{(9.5,31.1)}$	**19.3**$_{(8.9,32.4)}$	26.2$_{(9.0,53.5)}$
SUPPORT	2.7$_{(0.4,16.1)}$	**1.5**$_{(0.4,19.2)}$	2.0$_{(0.2,35.3)}$
SEER	**18.6**$_{(8.3,34.1)}$	20.2$_{(10.3,35.8)}$	23.7$_{(9.9,51.2)}$

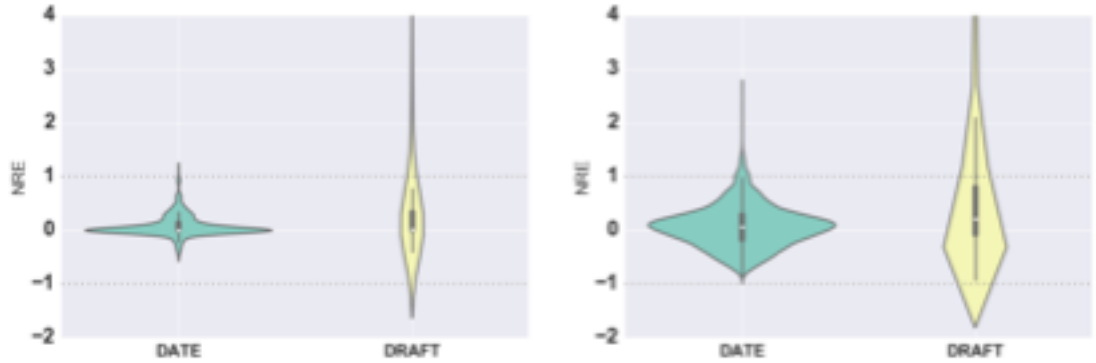


Figure 3. Normalized Relative Error (NRE) distribution for SUPPORT (top) and EHR (bottom), test-set non-censored events. The horizontal dashed lines represent the range of the events, $t_{\max} = 120$ months and $t_{\max} = 365$ days, respectively.

Concordance Index

Table 3. Concordance-Index results on test data.

	DATE	DATE-AE	DRAFT	Cox-Efron	RSF
EHR	**0.78**	**0.78**	0.76	0.75	–
FLCHAIN	**0.83**	**0.83**	**0.83**	**0.83**	0.82
SUPPORT	0.84	0.83	**0.86**	0.84	0.80
SEER	**0.83**	**0.83**	**0.83**	0.82	0.82

Q: How could the DATE model be used in clinical practice? More specifically, what steps would be required to sample times to a specific health event of interest for a given patient, and how could the model's output be summarized?

THANK YOU!

Questions or ideas? Please contact me at m.engelhard@duke.edu

200 Samples: FLCHAIN

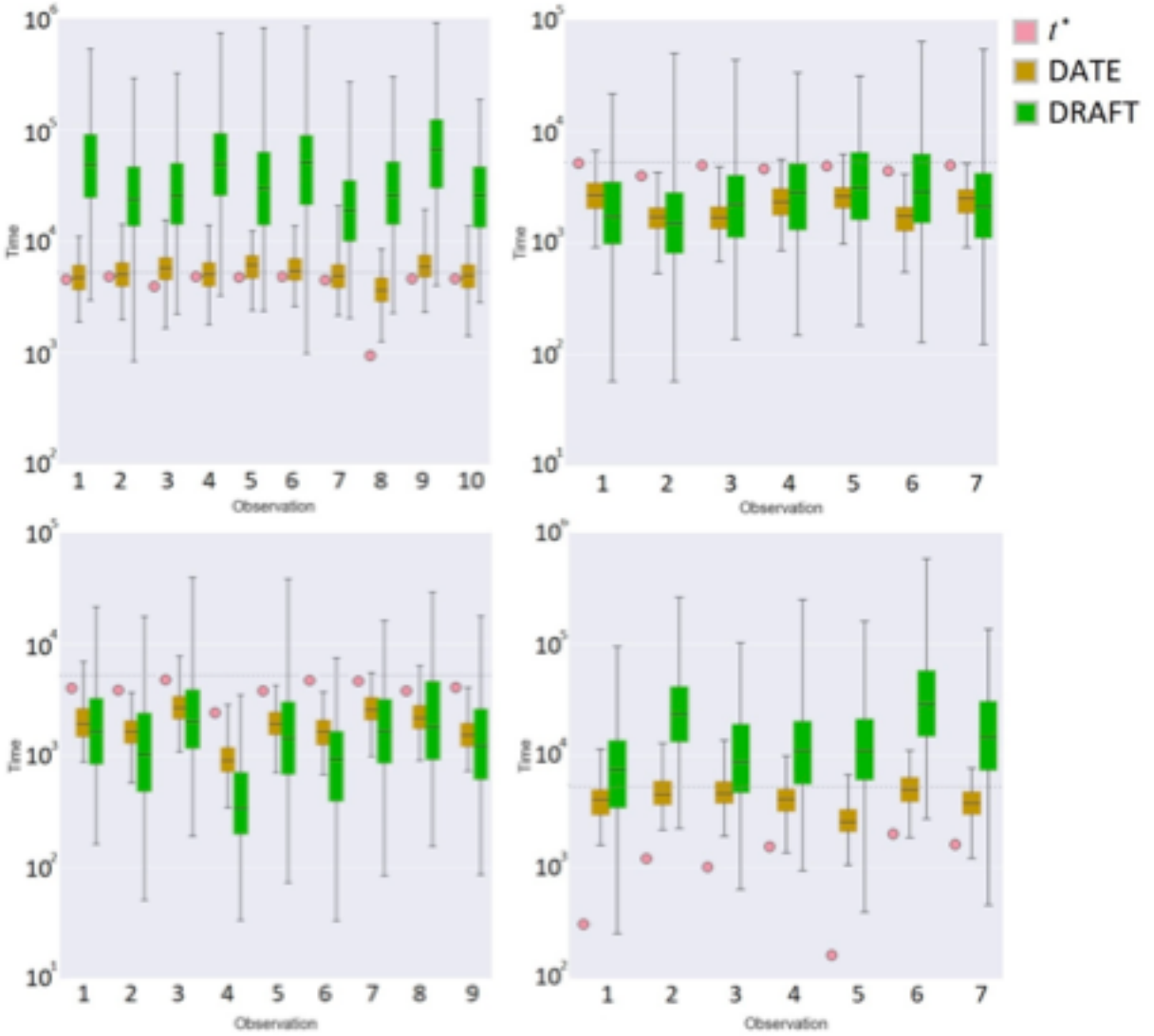


Figure 6. Comparison on FLCHAIN Censored best (top-left), worst (top-right) and Non-Censored best (bottom-left), worst (bottom-right).

200 Samples: SUPPORT

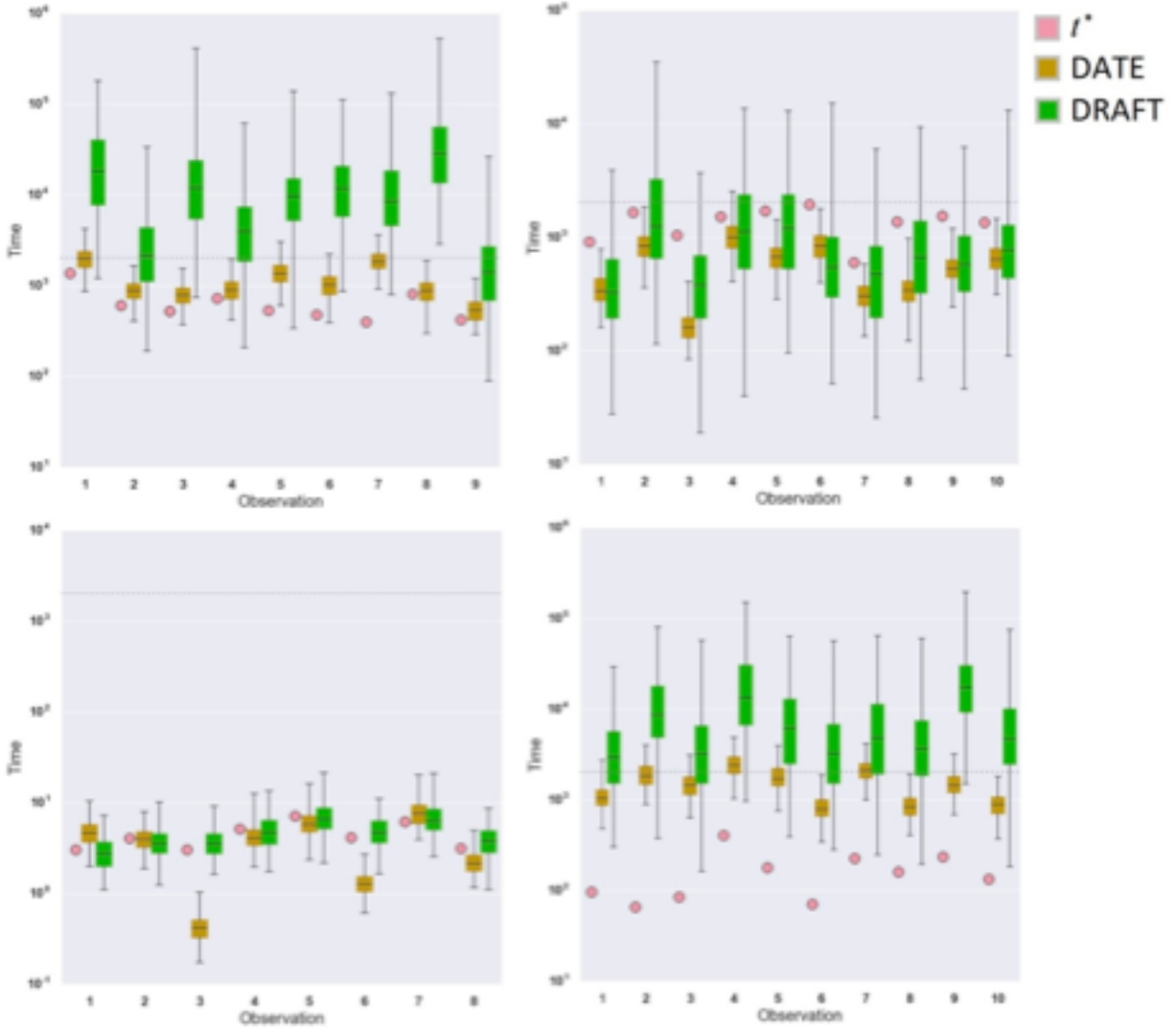


Figure 7. Comparison on SUPPORT Censored best (top-left), worst (top-right) and Non-Censored best (bottom-left), worst (bottom-right).

200 Samples: SEER

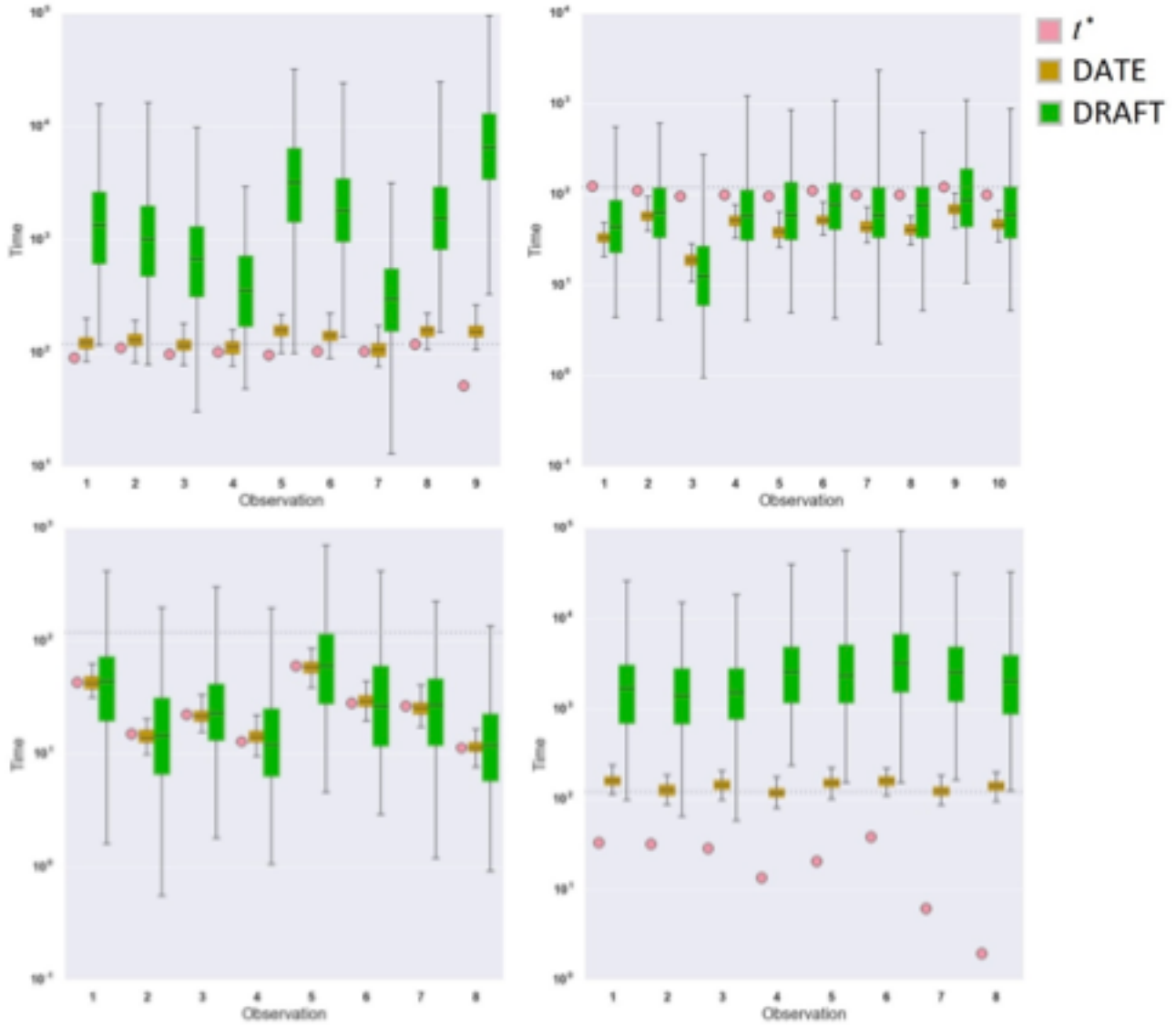


Figure 8. Comparison on SEER Censored best (top-left), worst (top-right) and Non-Censored best (bottom-left), worst (bottom-right).

200 Samples: EHR

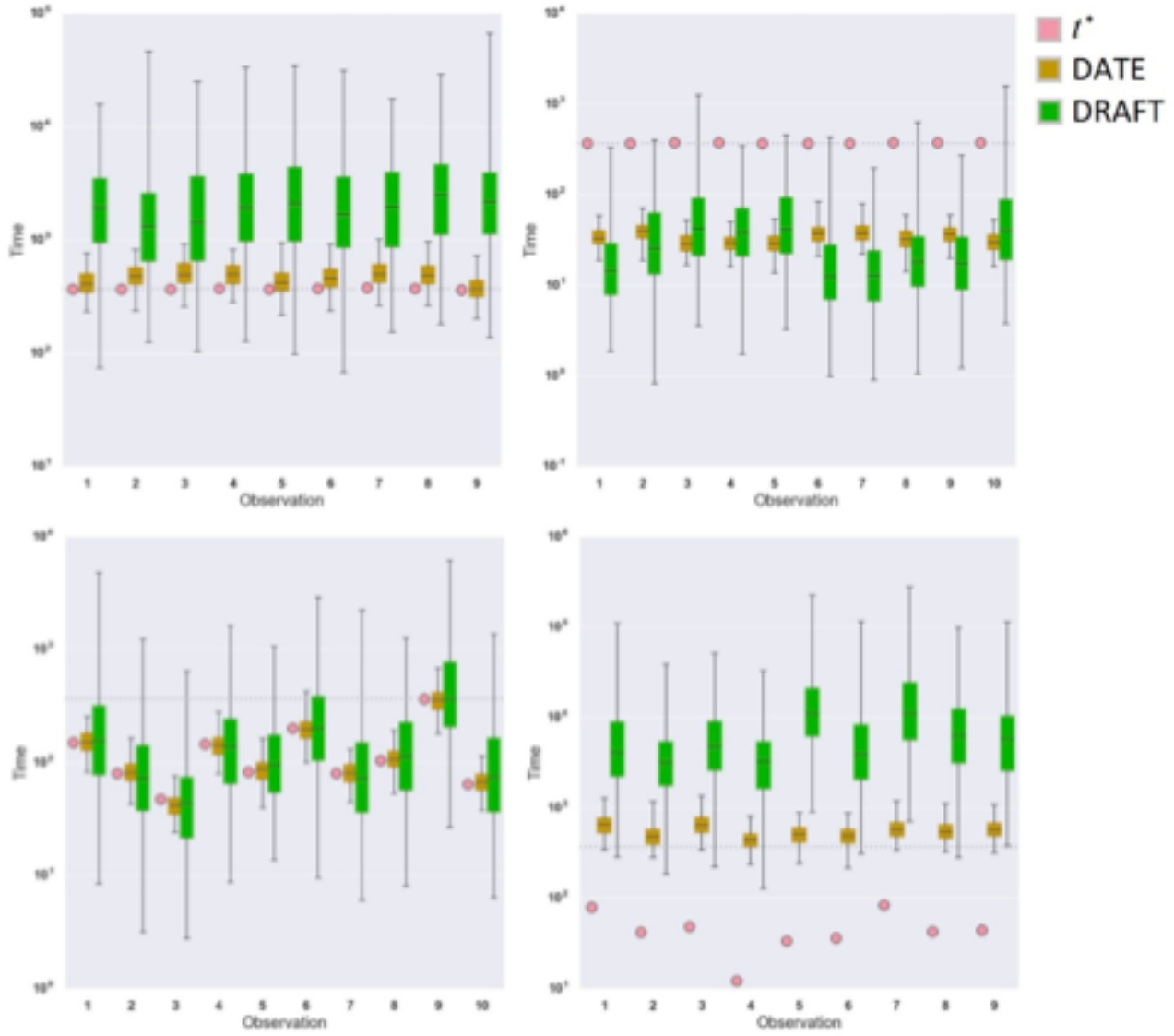


Figure 9. Comparison on EHR Censored best (top-left), worst (top-right) and Non-Censored best (bottom-left), worst (bottom-right).