Supplemental Material for “Adversarial Time-to-Event Modeling”

A. Missing data and DATE-AE

DATE-AE extends DATE by jointly learning the mapping $\boldsymbol{x} \rightarrow \boldsymbol{z} \rightarrow t$, where $\boldsymbol{z}$ is modeled as an adversarial autoencoder. For imputation, the covariates (entries of $\boldsymbol{x}$) in the encoder are set to zero if the entry is missing. When evaluating the reconstruction loss γ_3 in (1), we only do so for observed covariates; in this way the autoencoder can learn the correlation structure of the observed data despite missingness and without the need for imputation, while letting the decoder, $\boldsymbol{x} = \mathrm{decoder}(\boldsymbol{z})$, handle the imputation if needed. Note that for time-to-event prediction, at test time, we do not have to impute missing values as we can directly evaluate $\boldsymbol{x} \rightarrow \boldsymbol{z} \rightarrow t$. DATE-AE, extends DATE formulation with additional autoencoder discriminator and generator losses shown below:

$$\begin{aligned}
\gamma_1(\boldsymbol{\theta_x}, \boldsymbol{\theta_z}, \boldsymbol{\psi}; \mathcal{D}) &= \mathbb{E}_{(\boldsymbol{x}, \tilde{\boldsymbol{z}})}[D_{\boldsymbol{\psi}}(\boldsymbol{x}, \tilde{\boldsymbol{z}})] \\
&\quad + \mathbb{E}_{(\tilde{\boldsymbol{x}}, \boldsymbol{z})}[1 - D_{\boldsymbol{\psi}}(\tilde{\boldsymbol{x}}, \boldsymbol{z})], \\
\gamma_2(\boldsymbol{\theta_x}, \boldsymbol{\theta_z}; \mathcal{D}) &= \mathbb{E}_{(\boldsymbol{z} \sim p(\boldsymbol{z}), \hat{\boldsymbol{z}})}[d(\boldsymbol{z}, \hat{\boldsymbol{z}})], \\
\gamma_3(\boldsymbol{\theta_x}, \boldsymbol{\theta_z}; \mathcal{D}) &= \mathbb{E}_{(\boldsymbol{x} \sim p(\boldsymbol{x}), \hat{\boldsymbol{x}})}[d(\boldsymbol{x}, \hat{\boldsymbol{x}})], \\
\min_{\boldsymbol{\theta_x}, \boldsymbol{\theta_z}} \max_{\boldsymbol{\psi}} \gamma(\boldsymbol{\theta_x}, \boldsymbol{\theta_z}, \boldsymbol{\psi}; \mathcal{D}) &= \gamma_1(\boldsymbol{\theta_x}, \boldsymbol{\theta_z}, \boldsymbol{\psi}; \mathcal{D}) \\
&\quad + \zeta_2 \gamma_2(\boldsymbol{\theta_x}, \boldsymbol{\theta_z}; \mathcal{D}) \\
&\quad + \zeta_3 \gamma_3(\boldsymbol{\theta_x}, \boldsymbol{\theta_z}; \mathcal{D}) \,, \qquad (1)
\end{aligned}$$

where $\boldsymbol{x} \sim p(\boldsymbol{x})$, $\tilde{\boldsymbol{z}} = G_{\boldsymbol{\theta_x}}(\boldsymbol{x}, \boldsymbol{\epsilon_x})$, $\boldsymbol{z} \sim p(\boldsymbol{z})$, $\tilde{\boldsymbol{x}} = G_{\boldsymbol{\theta_z}}(\boldsymbol{z}, \boldsymbol{\epsilon_z})$, $\boldsymbol{\epsilon}$ is the noise source, d is the distortion measure and $\{\zeta_2, \zeta_3\}$ are reconstruction tuning parameters.

Tables 1 and 2 compares the effects of randomly introducing missing values on the Flchain relative absolute error and concordance-index respectively.

B. Concordance index and relative absolute error

Tables 3 and 4 show comparisons on concordance-index and relative absolute error across all datasets.

C. Normalized Relative Error (NRE)

Figures 2, 3, 4 and 5, show comparison on NRE distributions for both censored and non-censored events.

D. Test set time-to-event distributions

We randomly draw best and worst observation samples based on the NRE metric. Figures 6 , 7, 8 and 9, show the corresponding distributions comparisons relative to the ground truth or censored time $t^\star$.

E. Effects of noise source and stochastic layers

Figure 1 shows the contribution effects of stochastic layers for noise $\mathrm{Uniform}(0,1)$ on both censored and non-censored time-to-event distributions. Tables 5 and 6 compares noise sources on relative absolute error and CI.

Table 1. Introduced proportion of missing values comparison on Flchain relative absolute error. Ranges in parentheses are 50% empirical ranges over (median) test-set predictions.

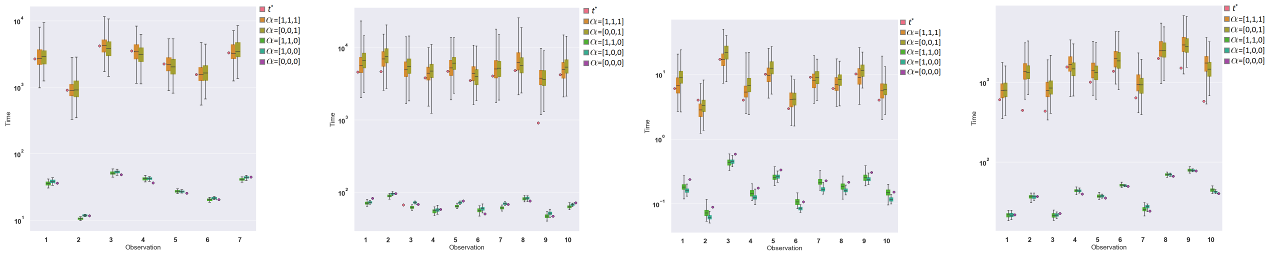
	0.10	0.20	0.30	0.50
Non-Censored				
DATE	$19.9_{(9.6,32.7)}$	$\mathbf{19.8}_{(9.1,33.7)}$	$\mathbf{19.7}_{(10.8,33.2)}$	$19.7_{(10.3,33.5)}$
DATE-AE	$\mathbf{19.2}_{(9.6,34.9)}$	$21.9_{((9.5,33.4)}$	$20.6_{(9.7,32.8)}$	$\mathbf{18.3}_{(9.5,32.9)}$
DRAFT	$32.9_{(10.0,92.3)}$	$34.1_{(11.5,119.8)}$	$\mathbf{19.7}_{(10.3,33.5)}$	$19.7_{(10.3,33.5)}$
Censored				
DATE	$\mathbf{0}_{(0,20.4)}$	$1.9_{(0,19.4)}$	$2.7_{(0,20.1)}$	$7.3_{(0,21.8)}$
DATE-AE	$\mathbf{0}_{(0,12.9)}$	$3_{(0,19)}$	$\mathbf{2.1}_{(0,16.5)}$	$\mathbf{6}_{(0,21.3)}$
DRAFT	$\mathbf{0}_{(0,0)}$	$\mathbf{0}_{(0,0)}$	$7.3_{(0,21.8)}$	$7.3_{(0,21.8)}$

Table 2. Introduced proportion of missing values comparison on FLCHAIN Concordance-Index.

	0.10	0.20	0.30	0.50
DATE	0.815	0.803	**0.803**	0.784
DATE-AE	0.814	0.804	0.799	**0.785**
DRAFT	**0.822**	**0.807**	0.801	0.783

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



(a) Non-Censored (b) Censored (c) Non-Censored (d) Censored

Figure 1. Effects of stochastic layers on uncertainty estimation on 10 randomly selected test-set subjects from the FLCHAIN ((a) and (b)) and SUPPORT ((c) and (d)) datasets. Ground truth times are denoted as t^* and box plots represent time-to-event distributions from a 2-layer model, where $\boldsymbol{\alpha} = [\alpha_0, \alpha_1, \alpha_2]$ indicates whether the corresponding noise source, $\{\boldsymbol{\epsilon}_0, \boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2\}$, is active. For example $\boldsymbol{\alpha} = [1, 0, 0]$ indicates noise on the input layer only.

Table 4. Median relative absolute errors (as percentages of $t_{\max}$), on non-censored and censored data. Ranges in parentheses are 50% empirical ranges over (median) test-set predictions.

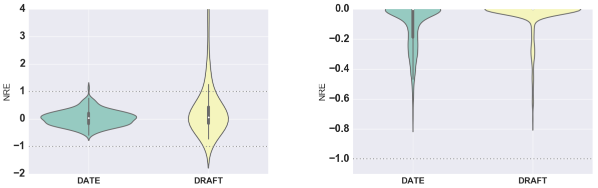
	DATE	DATE-AE	DRAFT
Non-censored			
EHR	**23.6**$_{(11.1,43.0)}$	24.5$_{(12.4,44.0)}$	36.7$_{(16.1,81.3)}$
FLCHAIN	19.5$_{(9.5,31.1)}$	**19.3**$_{(8.9,32.4)}$	26.2$_{(9.0,53.5)}$
SUPPORT	2.7$_{(0.4,16.1)}$	**1.5**$_{(0.4,19.2)}$	2.0$_{(0.2,35.3)}$
SEER	**18.6**$_{(8.3,34.1)}$	20.2$_{(10.3,35.8)}$	23.7$_{(9.9,51.2)}$
Censored			
EHR	12.4 $_{(0,38.7)}$	1.6$_{(0,34.)}$	**0** $_{(0,0)}$
FLCHAIN	0$_{(0,18.8)}$	0$_{(0,15.6)}$	**0** $_{(0,0)}$
SUPPORT	0$_{(0,13.0)}$	0$_{(0,8.8)}$	**0** $_{(0,0)}$
SEER	**0** $_{(0,0)}$	**0** $_{(0,0)}$	**0** $_{(0,0)}$

Table 5. Effects of noise source and stochastic layers on SUPPORT Median relative absolute error. Ranges in parentheses are 50% empirical ranges over (median) test-set predictions.

	Uniform(-1,1)	Uniform(0,1)	Gaussian(0,1)
Non-censored			
All	2.4$_{(0.4,19.9)}$	2.2$_{(0.5,19.2)}$	1.9$_{(0.4,17.)}$
Input	2.2$_{(0.4,18.)}$	**1.8**$_{(0.4,16.1)}$	1.9$_{(0.4,14.9)}$
Output		2.6$_{(0.4,21.1)}$	
Censored			
All	**0**$_{(0,14.6)}$	**0** $_{(0,13.7)}$	**0** $_{(0,16.4)}$
Input	**0**$_{(0,15.3)}$	1.2$_{(0,22.4)}$	0.8 $_{(0,21.2)}$
Output		**0**$_{(0,8.2)}$	

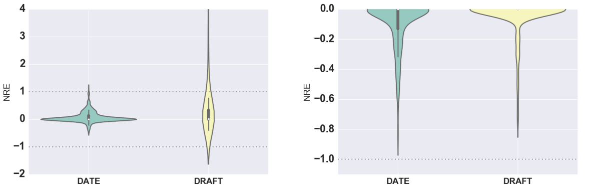
Table 6. Effects of noise source and stochastic layers on SUPPORT concordance-index.

	Uniform(-1,1)	Uniform(0,1)	Gaussian(0,1)
All	0.825	**0.835**	0.826
Input	0.841	0.829	0.825
Output		**0.836**	



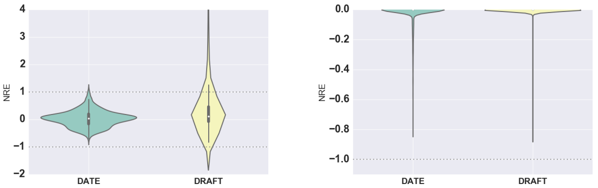
(a) Non-Censored (b) Censored

Figure 2. Normalized relative error on FLCHAIN test data.



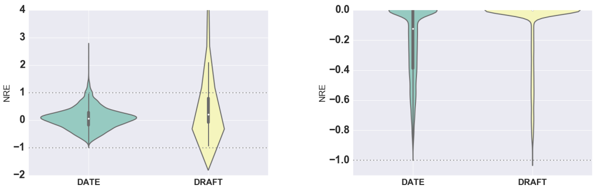
(a) Non-Censored (b) Censored

Figure 3. Normalized relative error on SUPPORT test data.



(a) Non-Censored (b) Censored

Figure 4. Normalized relative error on SEER test data.



(a) Non-Censored (b) Censored

Figure 5. Normalized relative error on EHR test data.

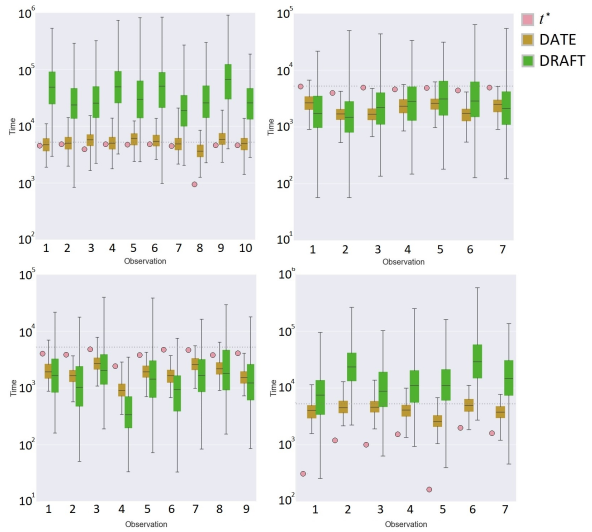


Figure 6. Comparison on FLCHAIN Censored best (top-left), worst (top-right) and Non-Censored best (bottom-left), worst (bottom-right).

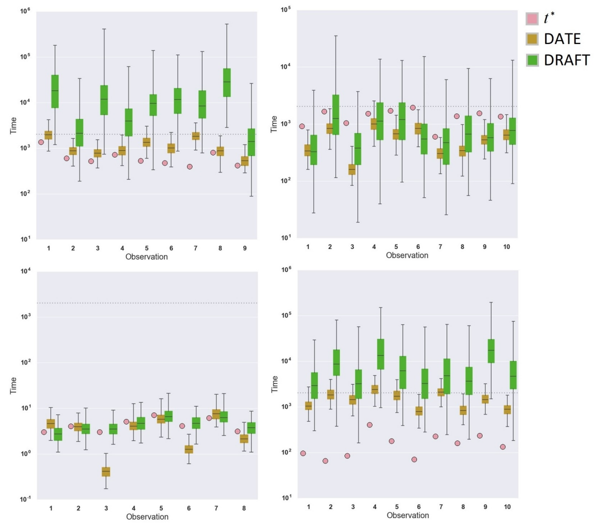


Figure 7. Comparison on SUPPORT Censored best (top-left), worst (top-right) and Non-Censored best (bottom-left), worst (bottom-right).

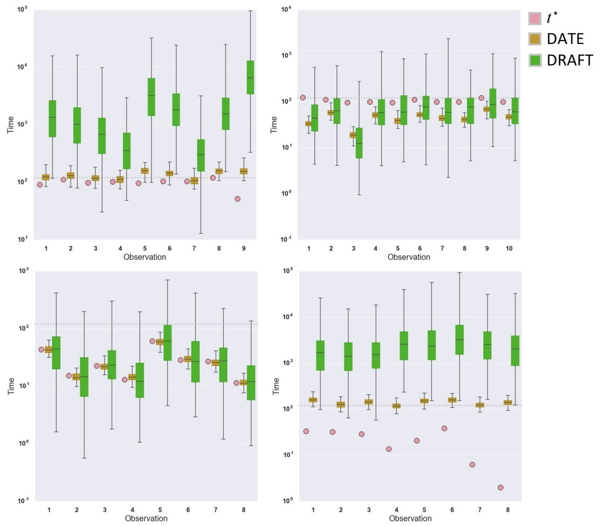


Figure 8. Comparison on SEER Censored best (top-left), worst (top-right) and Non-Censored best (bottom-left), worst (bottom-right).

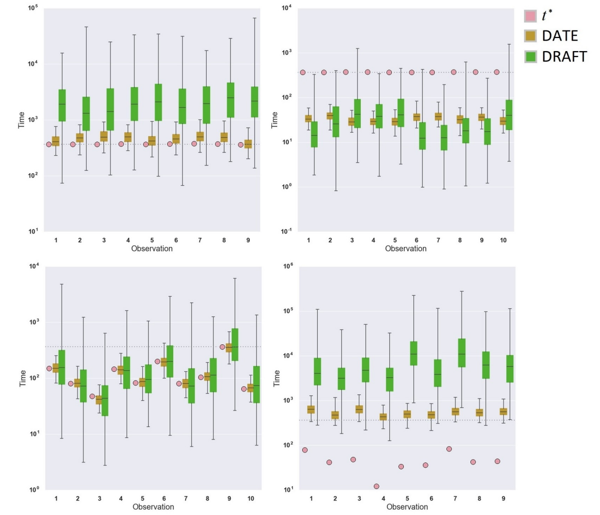


Figure 9. Comparison on EHR Censored best (top-left), worst (top-right) and Non-Censored best (bottom-left), worst (bottom-right).

F. Parametric examples of f, h and S relationships

Figure 10 shows examples of exponential, Weibull and log-normal time-to-event pdf $f_T(t|\boldsymbol{\theta})$ with corresponding survival function $S_T(t|\boldsymbol{\theta})$ and $h(t|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ are the pdf parameters and T is the time-to-event random variable.

G. Architecture of the neural network

In all experiments, DATE and DRAFT are specified in terms of two-layer MLPs of 50 hidden units with Rectified Linear Unit (ReLU) activation functions and batch normalization (Ioffe & Szegedy, 2015). The discriminator for DATE is a similarly defined MLP. As an optimizer, we use Adam (Kinga & Adam, 2015) with the following hyperparameters: learning rate 3×10^{-4}, first moment 0.9, second moment 0.99, and epsilon 1×10^{-8}. Further, we set the minibatch size to $M = 350$ and use dropout with $p = 0.8$ on all layers. All the network weights are initialized using *Xavier* (Glorot & Bengio, 2010). Datasets are split into training, validation and test sets as 80%, 10% and 10% partitions, respectively, stratified by non-censored event proportion. We use the validation set for early stopping and learning model hyper-parameters. DATE is executed using one NVIDIA P100 GPU with 16GB memory.

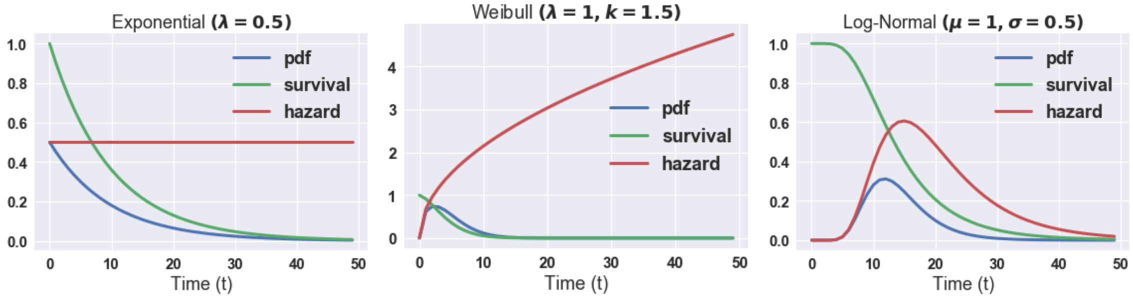


Figure 10. Popular parametric characterizations: exponential (left), Weibull (middle) and log-normal (right).

References

Glorot, Xavier and Bengio, Yoshua. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, 2010.

Ioffe, Sergey and Szegedy, Christian. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

Kinga, D and Adam, J Ba. A method for stochastic optimization. In *ICLR*, 2015.