Neural Conditional Event Time Models

Matthew Engelhard M.ENGELHARD@DUKE.EDU
Department of Psychiatry & Behavioral Sciences
Duke University
Durham, NC

Samuel Berchuck SAM.BERCHUCK@DUKE.EDU
Department of Statistical Science
Duke University
Durham, NC

Joshua D'Arcy JOSHUA.D.ARCY@DUKE.EDU
Department of Biostatistics & Bioinformatics
Duke University
Durham, NC

Ricardo Henao RICARDO.HENAO@DUKE.EDU
Department of Biostatistics & Bioinformatics
Duke University
Durham, NC

Abstract

Event time models predict occurrence times of an event of interest based on known features. Recent work has demonstrated that neural networks achieve state-of-the-art event time predictions in biomedical applications, where event time models are frequently used, as well as a variety of other settings. However, standard event time models suppose that the event occurs, eventually, in all cases. Consequently, no distinction is made between *a*) the probability of event occurrence, and *b*) the predicted time of occurrence. This distinction is critical when predicting medical diagnoses as well as social media posts, equipment defects, and other events that or may not occur; and for which the features affecting *a*) may be different from those affecting *b*). In this work, we develop a conditional event time model that distinguishes between these components, implement it as a neural network with a binary stochastic layer representing finite event occurrence, and show how it may be learned from right-censored event times via maximum likelihood estimation. Results demonstrate improved event occurrence and event time predictions on synthetic data, medical events (MIMIC-III), and social media posts (Reddit), including posts related to mental health, comprising 21 total prediction tasks.

1. Introduction

The modeling of event times, also known as failure or survival times, is ubiquitous in biostatistics and medicine, economics, operations research, and other fields. Common approaches include the Cox proportional hazards (Cox-PH) model (Cox, 1972), which assumes the effect of features is multiplicative on the hazard rate, and the accelerated failure time (AFT) model (Wei, 1992), in which features accelerate or decelerate the event time

density. A key characteristic of event time models, including Cox-PH and AFT, is that they are capable of learning from *censored* event times, particularly right-censored events, wherein the event time is known only to be above a given value. Right-censored events are common in real-world applications, in which events cannot be observed indefinitely.

A number of neural-network-based variations on established event time models have been shown to improve the resulting event time predictions, including several based on Cox-PH (Zheng et al., 2019; Katzman et al., 2018b; Kvamme et al., 2019b), and others designed for time-series features (Ren et al., 2019; Lee et al., 2018). Other neural-network-based models have used alternative loss functions; for example, Chapfuwa et al. (2018) used a nonparametric, adversarially trained model to obtain more accurate event time predictions. Recent interest in these models reflects the wide range of problems to which they can be applied, and the importance of learning from censored observations rather than discarding them.

However, the standard event time framework, which is shared by the examples above, makes the strong assumption that events of interest will occur, eventually, in all individuals (Kalbfleisch and Prentice, 2011). This assumption, while justified when predicting time of death, for example, limits effectiveness in settings in which censored event times are observed, but events occur only in a subset of the population. These settings include prediction of medical diagnoses, physical activities, social media activities, interest in specific media content, and many others. In each case, event time models cannot distinguish between a) the probability of event occurrence, and b) the time of occurrence, as well as the factors that impact the former versus the latter.

As a motivating example, we consider the problem of medical diagnosis, in which many patients are lost to follow-up, and consequently their subsequent diagnostic status is unknown. If a binary classifier is chosen to predict diagnosis, the population must be limited to individuals with adequate follow-up, resulting in substantial loss of training data. Moreover, these individuals may be systematically different from others, leading to biased predictions (von Allmen et al., 2015). If an event time model is applied, on the other hand, factors affecting the time to diagnosis, which include socioeconomic status, racial/ethnic status, and access to care (Dovidio and Fiske, 2012), are conflated with physiologic factors relevant to the underlying condition.

In this work, we address these limitations by formulating a novel conditional event time framework. Further, we develop a neural conditional event time model in which event occurrences are drawn from a multivariate Bernoulli distribution, *i.e.*, binary stochastic layer, and event times are predicted with a neural accelerated failure time model conditioned on event occurrence. This approach provides distinct event occurrence and event time predictions, leading to substantially improved prediction performance in both cases.

We evaluate our model on synthetic data, prediction of 10 clinically important events from MIMIC-III (Johnson et al., 2016), and prediction of user submissions to popular subreddits, *e.g.*, r/worldnews, from reddit.com, a leading news and web content aggregator. Model predictions are compared to a standard, *i.e.*, not conditional, neural event time model as well as binary classification of observed event occurrences, with emphasis on predicted event probabilities. This work, a novel generalization of the event time framework, leverages gradient estimation methods to predict medical conditions, user preferences, and other

characteristics not yet observed; and to distinguish the presence of these characteristics from the rate at which they manifest.

Generalizable Insights about Machine Learning in the Context of Healthcare

Event time models are advantageous in healthcare applications largely because they *a*) predict event risk over time, and *b*) account for *censoring*, in which information about event occurrence and timing are incomplete because patients are lost to follow-up, for example. However, standard event time models assume that the event will occur in all individuals. While valid when predicting death/survival, this key assumption causes event time models to conflate event timing with event occurrence when predicting diagnoses and other medical events that may never occur. For example, an event time model might make similar predictions for individuals who will be diagnosed promptly or never diagnosed, and individuals certain to be diagnosed eventually.

In this work, we argue that conditional event time models should be preferred when predicting diagnoses and (non-death) medical events. This approach provides improved event occurrence and event timing predictions compared to deep event time (*i.e.,* survival) models for events that may never occur, and allows factors affecting timing versus occurrence to be analyzed separately. Our results demonstrate benefits when predicting diagnostic testing in an ICU setting (MIMIC-III), and when predicting social media posts related to mental health.

2. Related Work

Neural-network-based (not conditional) event time models have been used to stratify patient risk (Ranganath et al., 2016) and recommend treatment based on electronic health record data and other clinical data (Katzman et al., 2018b), detect online fraud (Zheng et al., 2019), and predict survival based on blood serum biomarkers (Kvamme et al., 2019b). Many of these examples are based on Cox-PH (Cox, 1972), but the effect of features is modeled via neural networks (Zheng et al., 2019; Katzman et al., 2018b). In contrast, Ranganath et al. (2016) develop a generative model incorporating the Weibull distribution, whereas Chapfuwa et al. (2018) use an adversarial approach to generate nonparametric event time distributions. Additionally, Ren et al. (2019) use a recurrent neural network to predict event risk based on time-series data, and Lee et al. (2018) use a concordance-based loss function that accounts for competing risks, in which only one of several events of interest may occur.

Conditional event time distributions were explored by Elandt-Johnson (1976) and later used in biostatistics to predict long-term versus short-term survival (Farewell, 1982) and oncology outcomes (Gaynor et al., 1993). However, scaling the conditional event time framework to large datasets with many interrelated events of interest requires gradients of the event occurrence model to be backpropagated across a multivariate Bernoulli distribution, *i.e.*, binary stochastic layer. While high-variance gradient estimates can be obtained using the score function estimator (Williams, 1992), a number of lower-variance yet unbiased estimators have been developed more recently (Tucker et al., 2017; Grathwohl et al., 2017; Yin and Zhou, 2019). Alternatively, Jang et al. (2016) and Maddison et al. (2016) introduce a continuous relaxation of the categorical distribution that results in biased gradient estimates, but allows gradients to be backpropagated directly.

3. Conditional Event Time Models

Here we introduce the conditional event time (CET) framework, distinguish it from other event time models, and show how conditional event time models may be implemented via neural network with a binary stochastic layer to predict the occurrence of multiple interrelated events on large datasets.

3.1. Event Time Framework

Suppose we have N data points in triplets of the form $\mathcal{D} = \{\boldsymbol{x}_i, t_i, s_i\}_{i=1}^N$, where the $\boldsymbol{x}_i \in \mathbb{R}^d$ are d features associated with individual i, the $t_i \in (0, \infty)$ are associated event times, and the $s_i \in \{0, 1\}$ denote whether the t_i are true event times or right-censoring times. We begin with a single event of interest to simplify our notation, then extend to the more general case in which there are M events of interest.

Let $\mathcal{F}_i \in (0, \infty)$ and $\mathcal{G}_i \in (0, \infty)$ denote random variables associated with events and censoring, respectively, for individual i. We suppose the $\mathcal{F}_i$ are drawn independently from event time distribution $f_\theta(t|\boldsymbol{x}_i)$, which has associated survivor function $F_\theta(t|\boldsymbol{x}_i) = 1 - \int_0^t f_\theta(\tau|\boldsymbol{x}_i)d\tau$. Similarly, the $\mathcal{G}_i$ are drawn independently from the unknown censoring density $g_i(t_i)$, which has associated survivor function $G_i(t) = 1 - \int_0^t g_i(\tau)d\tau$.

Our time observations correspond to random variables $\mathcal{T}_i \in (0, \infty)$ and $\mathcal{S}_i \in \{0, 1\}$, where $\mathcal{T}_i = \min(\mathcal{F}_i, \mathcal{G}_i)$ and $\mathcal{S}_i = \mathbf{1}(\mathcal{T}_i = \mathcal{F}_i)$ indicates whether $\mathcal{T}_i$ corresponds to an event time ($s_i = 1$) or a censoring time ($s_i = 0$). Following standard practice, we suppose the $\mathcal{F}_i$ and $\mathcal{G}_i$ are mutually independent given $\boldsymbol{x}_1, ..., \boldsymbol{x}_N$, implying that a) event times for individuals i and j are conditionally independent given $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, and b) censoring is *non-informative*, meaning that observing $s_i = 0$ implies only that the event occurred after t_i.

The likelihood of observing a particular $\{t_i, s_i\}$ conditioned on features $\boldsymbol{x}_i$ is then given by the following:

$$\begin{aligned} p_\theta(t_i, s_i = 1|\boldsymbol{x}_i) &= f_\theta(t_i|\boldsymbol{x}_i)G_i(t_i), \\ p_\theta(t_i, s_i = 0|\boldsymbol{x}_i) &= g_i(t_i)F_\theta(t_i|\boldsymbol{x}_i), \\ p_\theta(t_i, s_i|\boldsymbol{x}_i) &= p_\theta(t_i, s_i = 1|\boldsymbol{x}_i)^{s_i} p_\theta(t_i, s_i = 0|\boldsymbol{x}_i)^{1-s_i}. \end{aligned} \tag{1}$$

Note that the $g_i(\cdot)$ do not depend on θ, therefore parameters θ of the event time model may be chosen to maximize the likelihood as follows:

$$\theta_{\text{ML}} = \underset{\theta}{\text{argmax}} \sum_{i=1}^N \{s_i \log f_\theta(t_i|\boldsymbol{x}_i) + (1 - s_i) \log F_\theta(t_i|\boldsymbol{x}_i)\}.$$

When there are M events of interest, we suppose the event times $\boldsymbol{t}_i \in (0, \infty)^M$ are independent given $\boldsymbol{x}_1, ..., \boldsymbol{x}_N$, resulting in the following joint density $p_\theta(\boldsymbol{t}_i, \boldsymbol{s}_i|\boldsymbol{x}_i)$:

$$p_\theta(\boldsymbol{t}_i, \boldsymbol{s}_i|\boldsymbol{x}_i) = \prod_{j=1}^M p_\theta^j(t_i^j, s_i^j|\boldsymbol{x}_i). \tag{2}$$

The corresponding maximum likelihood estimate is then:

$$\theta_{\text{ML}} = \underset{\theta}{\text{argmax}} \sum_{i=1}^N \sum_{j=1}^M \log p_\theta^j(t_i^j, s_i^j|\boldsymbol{x}_i). \tag{3}$$

For further details, see Kalbfleisch and Prentice (2011). Importantly, events $1, \ldots, M$ are viewed as *independent* (given $\boldsymbol{x}_i$), rather than *competing* events.

3.2. Conditional Event Time

In the conditional event time framework, we are interested in the hidden variable $c_i \in \{0, 1\}$, which indicates whether an event of interest will *ever* occur in individual i. This variable may be viewed as an underlying medical condition, equipment defect, or other characteristic of interest that will eventually manifest given sufficient time. When $c_i = 1$, the associated event time $\mathcal{F}_i$ is finite, whereas when $c_i = 0$, it is not. As before, we begin with a single event of interest to simplify notation.

Since t_i may not be finite, we augment the domain of $\mathcal{F}_i$ such that $\mathcal{F}_i \in (0, \infty) \cup \{\infty\}$, whereas the censoring time $\mathcal{G}_i$ remains finite.

We would like to have a model $p_\phi(c_i|\boldsymbol{x}_i)$, parameterized by ϕ, for the probability $P(\mathcal{F}_i < \infty|\boldsymbol{x}_i)$ that the event will ever occur in individual i. We suppose the c_i depend on $\boldsymbol{x}_i$ and follow a Bernoulli distribution:

$$c_i \mid \boldsymbol{x}_i \sim \text{Bern}\left(\sigma(h_\phi(\boldsymbol{x}_i))\right), \tag{4}$$

where $\sigma(\cdot)$ denotes the logistic function and $h_\phi(\cdot) : \mathbb{R}^d \to \mathbb{R}$ is a function with parameters ϕ to be learned along with θ, *i.e.*, the parameters of the event time function.

When $c_i = 1$, the standard event time model (described previously) applies. Alternatively, when $c_i = 0$ the observed time t_i is guaranteed to be a censoring time, therefore $P(s_i = 1|c_i = 0, \boldsymbol{x}_i) = 0$ and $p(t_i, s_i = 1|c_i = 0, \boldsymbol{x}_i) = 0$ for all t_i. Moreover, since $c_i = 0$ implies that $P(\mathcal{G}_i < \mathcal{F}_i) = P(\mathcal{G}_i < \infty) = 1$, the density of $\mathcal{T}_i$ is simply $g_i(\cdot)$, the density of censoring times. Consequently, $p_\theta(t_i, s_i|c_i, \boldsymbol{x}_i)$ consists of the following four terms:

$$\begin{aligned}
p_\theta(t_i, s_i = 1|c_i = 1, \boldsymbol{x}_i) &= f_\theta(t_i|\boldsymbol{x}_i)G_i(t_i), \\
p_\theta(t_i, s_i = 0|c_i = 1, \boldsymbol{x}_i) &= g_i(t_i)F_\theta(t_i|\boldsymbol{x}_i), \\
p(t_i, s_i = 1|c_i = 0, \boldsymbol{x}_i) &= 0, \\
p(t_i, s_i = 0|c_i = 0, \boldsymbol{x}_i) &= g_i(t_i).
\end{aligned} \tag{5}$$

In practice, we penalize incorrect prediction of $s_i = 1$ when $c_i = 0$ by assigning a small probability $0 < \epsilon \ll 1$ to $P(\mathcal{F}_i < \infty|c_i = 0)$, where ϵ is a hyperparameter of our model tuned on the validation set. Combining the four terms in (5) yields the following expression for $p_\theta(t_i, s_i|c_i, \boldsymbol{x}_i)$:

$$\begin{aligned}
p_\theta(t_i, s_i|c_i, \boldsymbol{x}_i) = &\; p_\theta(t_i, s_i = 1|c_i = 1, \boldsymbol{x}_i)^{s_i c_i} \\
&\times p_\theta(t_i, s_i = 0|c_i = 1, \boldsymbol{x}_i)^{(1-s_i)c_i} \\
&\times p(t_i, s_i = 1|c_i = 0, \boldsymbol{x}_i)^{s_i(1-c_i)} \\
&\times p(t_i, s_i = 0|c_i = 1, \boldsymbol{x}_i)^{(1-s_i)(1-c_i)},
\end{aligned} \tag{6}$$

which may be simplified (see Appendix) as follows after removing terms that do not depend on θ or c_i, including $g_i(\cdot)$ and $G_i(\cdot)$:

$$p_\theta(t_i, s_i|c_i, \boldsymbol{x}_i) \propto \epsilon^{s_i(1-c_i)} f_\theta(t_i|\boldsymbol{x}_i)^{s_i} F_\theta(t_i|\boldsymbol{x}_i)^{(1-s_i)c_i}. \tag{7}$$

We then use Jensen's inequality to maximize a lower bound on the expected log-likelihood over the latent variables c_i:

$$\log p_{\theta,\phi}(\mathcal{D}) = \sum_{i=1}^{N} \log \mathbb{E}_{c_i \sim p_\phi(c_i|\boldsymbol{x}_i)}[p_\theta(t_i, s_i \mid c_i, \boldsymbol{x}_i)]$$
$$\geq \sum_{i=1}^{N} \mathbb{E}_{c_i \sim p_\phi(c_i|\boldsymbol{x}_i)}[\log p_\theta(t_i, s_i \mid c_i, \boldsymbol{x}_i)]. \tag{8}$$

When there are M events of interest, rather than a single one, we suppose the c_i^j are independent given x_i and drawn from a multivariate Bernoulli distribution:

$$\boldsymbol{c}_i \mid \boldsymbol{x}_i \sim \prod_{j=1}^{M} \text{Bern}\left(\sigma(h_\phi^j(\boldsymbol{x}_i))\right), \tag{9}$$

where $h_\phi(\cdot) : \mathbb{R}^d \to \mathbb{R}^M$ describes the log-odds of all M events. In parallel with equations (2) and (3), we expand $p_\theta(\boldsymbol{t}_i, \boldsymbol{s}_i|\boldsymbol{c}_i, \boldsymbol{x}_i)$ to obtain the following lower bound on the log-likelihood:

$$\log p_{\theta,\phi}(\mathcal{D}) \geq \sum_{i=1}^{N} \mathbb{E}_{c_i \sim p_\phi(\boldsymbol{c}_i|\boldsymbol{x}_i)} \left[\sum_{j=1}^{M} \log p_\theta^j(t_i^j, s_i^j|\boldsymbol{c}_i, \boldsymbol{x}_i) \right].$$

Importantly, when M events are present, the fact that one event will (eventually) occur, *i.e.*, $c_i^j = 1$ for some j, affects the timing of other events. Thus we have $f_\theta(t_i|\boldsymbol{x}_i, \boldsymbol{c}_i)$ rather than $f_\theta(t_i|\boldsymbol{x}_i)$. This is critical when predicting medical diagnoses, for example, wherein the presence of a given condition may affect health services use or providers' ability to recognize other conditions. Importantly, however, this dependence requires that $\nabla_\phi \log(p_{\theta,\phi}(\mathcal{D}))$ be backpropagated through samples from a multivariate Bernoulli distribution.

3.3. Event Occurrence as a Binary Stochastic Layer

We instantiate $h_\phi(\boldsymbol{x}_i)$ and the parameters of the event time distribution $f_\theta(t_i|\boldsymbol{x}_i, \boldsymbol{c}_i)$ as neural networks, allowing our conditional event time model to be learned via backpropagation. The form of $f_\theta(\cdot)$ chosen for our experiments is described in the next section, however, the conditional event time framework permits a range of parametric distributions to be used. Learning the parameters θ and ϕ therefore requires us to calculate both $\nabla_\theta \log p_{\theta,\phi}(\mathcal{D})$ and $\nabla_\phi \log p_{\theta,\phi}(\mathcal{D})$ from equation (10). The former may be estimated directly based on samples of $\boldsymbol{c}$, but the latter must be backpropagated across these samples, drawn from a multivariate Bernoulli distribution, which is not differentiable.

To estimate $\nabla_\phi \log p_{\theta,\phi}(\mathcal{D})$, we take advantage of recently developed gradient estimators for categorical and Bernoulli random variables. Specifically, we explore both the Gumbel-Softmax estimator developed concurrently by Jang et al. (2016) and Maddison et al. (2016), which is a continuous (and differentiable) relaxation of the categorical distribution; as well as the Augment-Reinforce-Merge (ARM) estimator (Yin and Zhou, 2019), which provides an unbiased, low-variance gradient estimate for the multivariate Bernoulli distribution specifically. Although conditional event time models have been proposed in the past,

as previously described, these developments allow them to be applied to large datasets containing a large number of features and interrelated event occurrences. This is critical to their application to the problems we have described, including diagnosis of multiple medical conditions from the electronic health record, and prediction of user interests from social media activity or in recommender systems.

3.4. Accelerated Failure Time

We model the event time distribution $f_\theta(\cdot)$ using the accelerated failure time (AFT) model originally proposed by Wei (1992). This model supposes that a baseline survival function $F_0(t)$ is scaled uniformly by the effect of features $\boldsymbol{x}$ such that $F_\theta(t_i) = F_0(\mu(\boldsymbol{x}_i) \cdot t_i)$. Consequently, the density $f_\theta(t_i|\boldsymbol{x}_i)$ may be written as $\mu(\boldsymbol{x}_i) f_0(\mu(\boldsymbol{x}_i) \cdot t_i)$, and the log-transformed event time random variable $\mathcal{F}_i$ satisfies:

$$\log(\mathcal{F}_i) = \mu(\boldsymbol{x}_i) + \nu_i \varepsilon. \tag{10}$$

When ε is chosen to be normally distributed, *i.e.*, $\varepsilon \sim \mathcal{N}(0, 1)$, f_θ is log-normal with mean and standard deviation given by $\mu(\boldsymbol{x}_i)$ and ν_i, respectively.

To account for the dependency of both the scale and uncertainty of event time predictions on $\boldsymbol{x}$, we instantiate $\mu(\boldsymbol{x}_i)$ and $\nu(\boldsymbol{x}_i)$ in (10) using neural networks with parameters θ_μ and θ_ν, respectively, where $\theta = \{\theta_\mu, \theta_\nu\}$, $\mu(\boldsymbol{x}_i) = \mathrm{NN}(\boldsymbol{x}_i; \theta_\mu)$, and $\nu(\boldsymbol{x}_i) = \exp\left(\mathrm{NN}(\boldsymbol{x}_i; \theta_\nu)\right)$.

When predicting M events of interest, we have $\mu(\cdot) : \mathbb{R}^{d+M} \to \mathbb{R}^M$ and $\nu(\cdot) : \mathbb{R}^{d+M} \to \mathbb{R}^M$, where $\mu^j(\boldsymbol{x}_i, \boldsymbol{c}_i)$ and $\nu^j(\boldsymbol{x}_i, \boldsymbol{c}_i)$ specify the parameters of the distribution $f^j_\theta(t^j_i|\boldsymbol{x}_i, \boldsymbol{c}_i)$.

This approach provides a simple, flexible event time distribution capable of making accurate event time predictions, as we will show. Having described the conditional event time model, we now present experimental results.

4. Experiments

We describe our experimental methods, including performance metrics, baseline models, datasets, and training and evaluation procedures. We perform experiments on one synthetic and two real-world datasets, comprising a total of 21 distinct prediction tasks.

4.1. Performance Metrics

AUC The area under the receiver operating characteristic (AUC) assesses binary classification performance of the learned $p_\phi(c|x)$ in predicting whether events of interest will ever occur. It is calculated using standard methods based on the predicted $p_\phi(c|x)$ and true c, on the test set. For baseline models predicting event times only, AUC was calculated by sweeping a threshold across the full range of predicted times.

Mean Relative Absolute Error (MRAE) The accuracy of event time predictions was assessed on the test set by normalizing the absolute error of predictions by the event range, *i.e.*, $|t - \hat{t}|/t_{\max}$, where $\hat{t}$ is the predicted event time. For censored events, predictions are penalized only if the predicted time is before the censoring time, therefore the relative absolute error is defined as $\max(0, t - \hat{t})/t_{\max}$.

Concordance Index (CI) Correct ordering of event time predictions was assessed using the concordance index (CI) developed by Harrell Jr et al. (1984), which quantifies the degree to which the order of predicted event times is consistent with the true event times. Pairs of event times contribute to the CI only if *a*) both event times are known, or *b*) one event time is known, the other is censored, and the known event time occurs before the censoring time.

4.2. Baseline Models

We compare the performance of our neural conditional event time model (CET) to *a*) a neural event time model (ET), and *b*) a binary classifier (BC) trained to predict whether events are observed, *i.e.*, s. These represent the available alternatives to CET. All three performance metrics are evaluated on the ET models, but only the AUC can be evaluated on the binary classifier, which does not predict event times. The ET model matches the baseline model used in Chapfuwa et al. (2018) and is similar to the deep survival models used by Katzman et al. (2018a) and Kvamme et al. (2019a), but we use the accelerated failure time model from CET rather than a Cox proportional hazards framework (Cox, 1972).

Our aim is to evaluate differences between CET, ET, and BC rather than the impact of specific neural network architectures or hyperparameters, therefore, all neural network layers and model hyperparameters are identical between the CET model and the two baselines. Thus, the ET model $p_{\theta_{\text{ET}}}(\boldsymbol{t}, \boldsymbol{s}|\boldsymbol{x})$ matches the event time component $p_{\theta_{\text{CET}}}(\boldsymbol{t}, \boldsymbol{s}|\boldsymbol{c}, \boldsymbol{x})$ of CET with the exception of the additional input $\boldsymbol{c}$, and the BC model matches $p_\phi(\boldsymbol{c}|\boldsymbol{x})$ from CET.

4.3. Datasets

Here we describe the three datasets used in our experiments. Experimental results are presented in the next section.

4.3.1. Synthetic

To illustrate the advantage of the CET model over alternative approaches when learning from censored data, we construct a simple, synthetic dataset with five features and two events of interest. The eventual occurrence of the first event depends only on the first two features, as shown in the top left panel of Figure 2, whereas the eventual occurrence of the second event depends only on the second two features, as shown in the bottom left panel of Figure 2. The timing of both events (expected log-time), however, depends linearly on a fifth feature drawn from a standard normal distribution. Training, validation, and test sets contain 24k, 8k, and 8k samples, respectively. Censoring times are uniformly distributed over the full range of event times.

4.3.2. MIMIC-III

MIMIC-III (Medical Information Mart for Intensive Care), is a de-identified, accessible dataset of intensive care unit stays at the Beth Israel Deaconess Medical Center between 2001 and 2012 (Johnson et al., 2016). With this dataset, we aim to predict whether and when each of 10 important but non-routine laboratory measurements will be collected for the first time based on physiologic and other measurements from the first 24 hours. Laboratory measurements were selected among those rarely observed in the first 24 hours based on

Table 1: Relevance and occurrence rates for MIMIC-III events.

Lab Measure	Relevance	Rate
WBC, CSF	Lumbar Puncture	4.4%
Troponin T	Heart Damage	35.2%
Intubated	Intubate Patient	38.5%
WBC, Pleural	Pleural Fluid	2.8%
TSH	Thyroid Function	20.5%
D-Dimer	Thromboses	5.1%
Urobilinogen	Urinalysis	54.8%
ANA	Autoimmune	1.8%
Ammonia	Liver Function	3.6%
Lipase	Pancreatic Func.	34.8%

our assessment of their diagnostic and clinical relevance. For example, observing a "WBC, CSF" measurement suggests that a lumbar puncture has been performed. All 10 laboratory measurements and their rates of occurrence among MIMIC-III stays are presented in Table 1.

The most common chart events (80 total), lab measurements (30 total), and output events (10 total) occurring within the first 24 hours of admission among all stays in the training set were used as features for the prediction tasks. We ensured that lab measurements selected as events were excluded, but these measurements were not among the 30 most common and were typically observed beyond 24 hours. All measurements were aggregated by patient by taking the sum and count of all output events; the mean, minimum, maximum, and count of other numeric measurements; and the count of all categorical measurements, resulting in 346 total features.

Event times were censored uniformly over the interval $(0, 2 \cdot t^j_{\text{median}})$, where t^j_{median} is the median event time at which measurement j was first collected. Note that artificial censoring is critical to our performance evaluation, which requires ground truth event occurrence labels that are distinct from observed events in the training data. MIMIC-III was chosen for its completeness, which allows this ground truth to be determined. In contrast, CET is necessary only when substantial censoring is present – which is common in many health datasets – otherwise separate classification and event time models can be trained to predict event occurrence and timing, respectively.

4.3.3. Reddit

Reddit is a web content aggregator and discussion forum with approximately 330 million users as of April 2018 (Pardes, 2018). With this dataset, we aim to predict whether and when users will post to each of 9 different subreddits for the first time based on their prior comment history. Subreddits were hand-selected among those with at least 100k subscribed Reddit users, and all data were collected using the pushshift.io API. Submission histories prior to Jan 2020 were collected and grouped by user, and individual comment histories from June 2005 to Nov 2017 were collected for all users that posted to at least one of the 9 subreddits. Users with 20 or more comments prior to their first submission to any of the 9 subreddits were included in the final dataset, which included 492,059 total Reddit users.

The number of total subscribers to each subreddit and the proportion of our sample who posted to it are presented in Table 2.

Table 2: Popularity and submission rate for each Subreddit.

Subreddit	Total Subscribed	Rate
ADHD	613k	6.3%
Anxiety	325k	8.4%
Books	17.5m	13.4%
Depression	597k	20.0%
Fitness	7.7m	34.6%
LifeProTips	17.2m	25.9%
Mental Health	144k	2.8%
Suicide Watch	180k	7.3%
World News	23.1m	0.3%

For the prediction tasks, the first 20 comments from each Reddit user were encoded using Google's Universal Sentence Encoder (Cer et al., 2018). Embedded comments were refined via a single fully-connected layer with tanh activation, then aggregated via max and average pooling (Shen et al., 2018). The average time between comments and average comment length (batch-normalized) were used as additional features. Submission times were censored uniformly over the interval $(0, 2 \cdot t^j_{\text{median}})$, where t^j_{median} is the median submission time to subreddit j. Similar to MIMIC-III (4.3.2), this provides ground truth event occurrence labels that are distinct from observed events in the training data.

4.4. Training and Evaluation

For all tasks, data were partitioned into training (60%), validation (20%), and test (20%) sets. Our aim is to illustrate differences between CET and alternative approaches, therefore we utilize simple multilayer perceptron architectures with a single hidden layer (ReLU activations) for the functions $h_\phi(\cdot)$, $\mu(\cdot)$, and $\nu(\cdot)$. All hyperparameters including hidden layer width, Gumbel-Softmax temperature, number of c_i samples, Gumbel-Softmax versus ARM estimator, and the penalty ϵ (see 3.2) were tuned to maximize AUC on the validation set. Hyperparameters were then fixed, and all models (CET and baselines) were evaluated on the test set. Due to stochasticity in the training procedure (from sampling in the binary stochastic layer), we trained and evaluated 10 models (with identical hyperparameters) per condition. Reported performance measures are the mean and standard deviation of each measure over all 10 runs. All models were implemented in Tensorflow 1.10 (Abadi et al., 2016) and trained via backpropagation with the Adam optimizer (Kingma and Ba, 2014) and a batch size of 400, learning rate of 3×10^{-4}, and dropout rate of 0.5.

5. Experimental Results

Prediction performance (AUC, MRAE) aggregated across all tasks in each dataset is shown in Figure 1. Results show that CET effectively predicts event occurrence despite learning from censored events, with superior performance (AUC, MRAE) compared to ET and BC. All code used to generate results is available at https://github.com/mengelhard/cft.

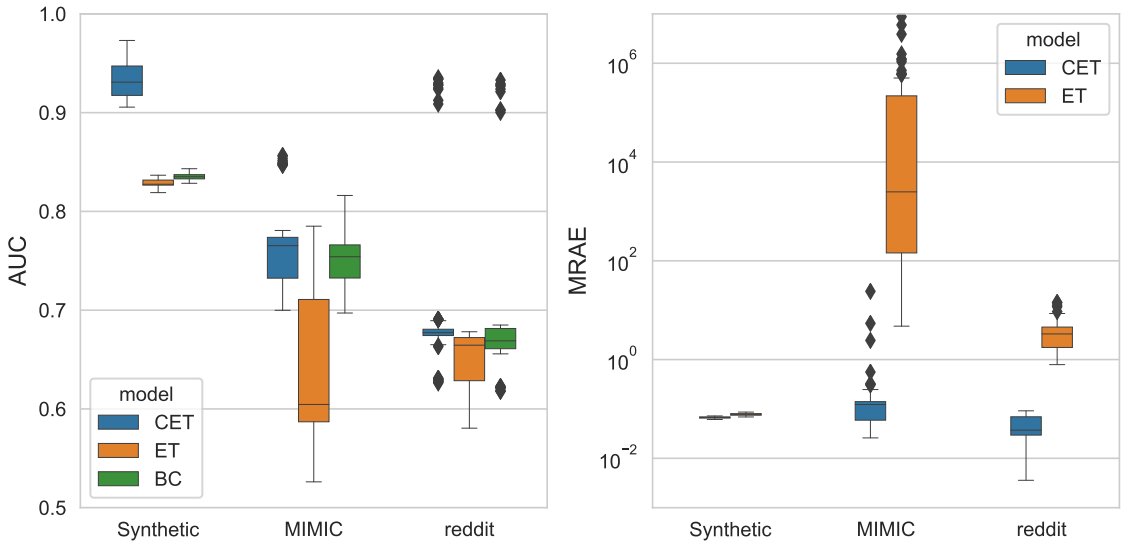

Figure 1: AUC (left) and MRAE (right) for all prediction tasks.

5.1. Synthetic

Results on our synthetic dataset (see Table 3), illustrate superior performance of the CET framework compared to existing baselines in a) predicting the probability of event occurrence, and b) making accurate event time predictions despite using a simple, parametric event time model. The left panels of Figure 2 show that a simple multilayer perceptron classifier trained directly on known event occurrences (*i.e.,* $\boldsymbol{c}$) effectively separates (AUC≈1) individuals in whom the event does versus does not occur in both tasks. Importantly, this information is not available to the CET and baseline models, which are trained on censored event times. The middle left panels show that CET also separates these groups effectively despite learning from censored event times only. In contrast, BC (right panels) cannot distinguish between cases that have been censored and cases in which the event never occurs. Similarly, although the ET model is able to learn from censored events, it conflates low event probabilities with high event times, leading to poor classification performance (middle right panels).

Figure 2 was generated with a lower noise setting compared to the quantitative results, providing clearer separation between groups that allows classification performance to be visualized more effectively.

Compared to the ET model, CET also makes substantially more accurate event time predictions, as shown in Table 3. This results from the fact that ET must predict a high event time, rather than a low event probability, for individuals in whom the event is not likely to occur. Consequently, when events do occur in these individuals, the event times predicted by ET are highly inaccurate. In contrast, CET distinguishes between event probabilities and event times, allowing it to maintain accurate predictions in these cases.

The CI is similar between the CET and ET models, but consistently higher for ET. This suggests that the ET model is more effective in correctly ordering observed, *i.e.*, non-censored, events. These results are consistent with the fact that the ET objective is designed solely

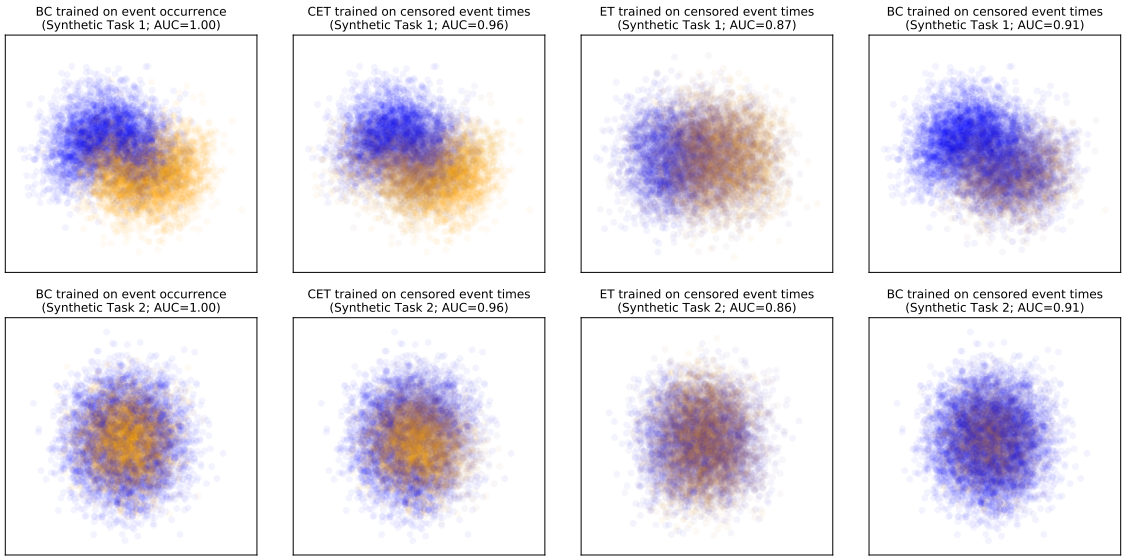

Figure 2: Prediction of event occurrence on synthetic data.

Table 3: Performance metrics on synthetic data.

	Task	AUC	MRAE	CI
CET	T1	**0.93±0.02**	**0.07±0.00**	0.88±0.01
	T2	**0.94±0.02**	**0.07±0.00**	0.89±0.00
	Avg	**0.93±0.02**	**0.07±0.00**	0.88±0.00
ET	T1	0.83±0.00	0.08±0.00	**0.90±0.00**
	T2	0.83±0.00	0.08±0.00	**0.90±0.00**
	Avg	0.83±0.00	0.08±0.00	**0.90±0.00**
BC	T1	0.84±0.00		
	T2	0.83±0.00		
	Avg	0.84±0.00		

to optimize this ordering, whereas the CET objective also seeks to optimize the predicted probability of event occurrence.

Figure 3 shows that the event probabilities predicted by CET and ET are effectively calibrated, whereas those predicted by BC are not.

Results on all datasets use the Gumbel-Softmax estimator with temperature fixed to 0.3, which was found to optimize AUC on the validation sets.

5.2. MIMIC-III

MIMIC-III results are consistent with the synthetic dataset: the CET model predicts event occurrence more effectively than ET or BC, and also predicts event times more accurately than ET, but with lower concordance index compared to ET (see Table 4). However, the degree of these differences is larger than found on the synthetic data. In particular, event

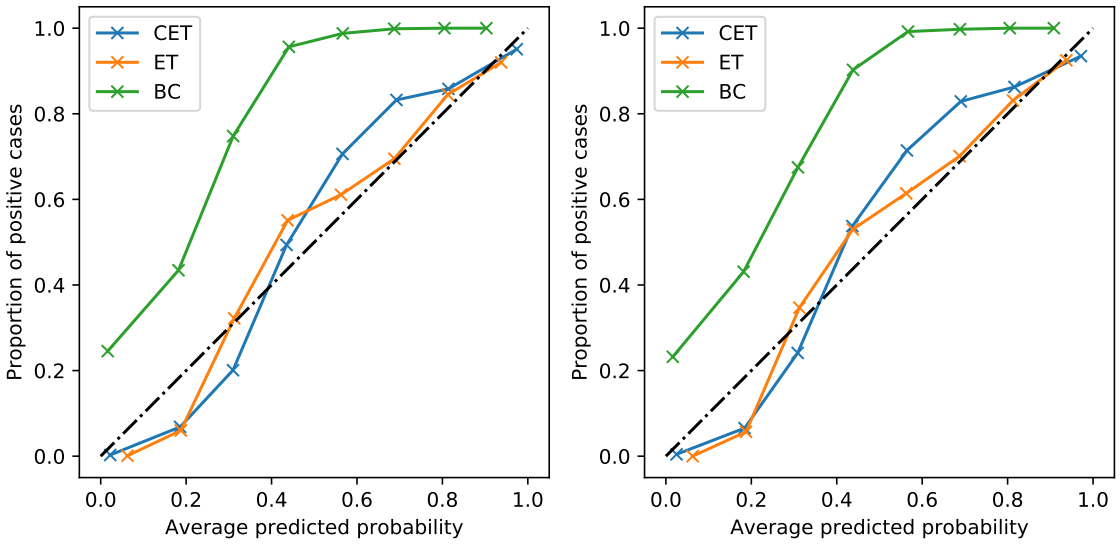

Figure 3: Calibration curves on synthetic data.

time predictions made by ET are highly inaccurate, which may be due to the high variance and long tail of event times in most of the tasks.

Direct prediction of known event occurrences yields AUCs ranging from 0.74 (TSH) to 0.90 (Intubation) with an average of 0.80. It is notable that performance on most tasks is high, demonstrating that important diagnostic tests ordered by care providers can be effectively predicted based on patient profiles over the first 24 hours, even when many events are censored. Although censoring is artificial on MIMIC-III, it is natural in most medical prediction settings, wherein many patients are lost to follow-up before events of interest can be observed. Moreover, follow-up rates are often correlated with events of interest, leading to biased results when these patients are removed from the dataset. The CET framework allows event occurrence to be predicted effectively in all patients, not just those who have been followed for a sufficiently long period.

5.3. Reddit

Reddit results remain consistent with previous experiments. Event occurrence probabilities predicted by CET are superior to those predicted by ET and BC, as measured via AUC, and event time predictions are more accurate than those predicted by ET. On the other hand, ET orders events more effectively than CET, as measured via CI (see Table 5).

Direct prediction of known event occurrences, *i.e.*, subreddit posts, yields AUCs ranging from 0.64 (r/ADHD) to 0.94 (r/worldnews), with an average of 0.71. Good prediction performance, although not as high compared to MIMIC-III, suggests that Reddit users' tendency to post to specific subreddits – including several related to mental health, (*e.g.*, r/ADHD, r/depression, r/mentalhealth, r/SuicideWatch), can be predicted effectively from a small number of early comments. Prediction performance may be substantially higher

Table 4: Performance metrics on MIMIC-III dataset.

	Lab	AUC	MRAE	CI
CET	CSF	**0.77±0.01**	**3.09±7.55**	0.53±0.04
	Trop.	**0.78±0.00**	**0.11±0.01**	0.59±0.03
	Intub.	**0.85±0.00**	**0.14±0.01**	0.67±0.01
	Pleur.	**0.76±0.00**	**0.05±0.01**	0.47±0.03
	TSH	**0.70±0.00**	**0.22±0.14**	0.52±0.03
	D-Dim.	**0.77±0.00**	**0.07±0.03**	0.50±0.05
	Urob.	**0.76±0.00**	**0.13±0.01**	0.62±0.03
	ANA	**0.74±0.01**	**0.04±0.01**	0.50±0.04
	Amm.	**0.77±0.01**	**0.33±0.74**	0.50±0.05
	Lipase	**0.73±0.00**	**0.14±0.01**	0.60±0.02
	Avg	**0.76±0.00**	**0.43±0.78**	0.55±0.02
ET	CSF	0.58±0.02	3.1e9±5.0e9	**0.59±0.02**
	Trop.	0.72±0.01	9.1e5±1.8e6	**0.71±0.01**
	Intub.	0.74±0.02	1.7e1±5.6e0	**0.73±0.02**
	Pleur.	0.59±0.02	5.6e2±1.3e3	**0.59±0.02**
	TSH	0.60±0.01	3.1e5±3.4e5	**0.60±0.01**
	D-Dim.	0.60±0.02	3.9e3±8.7e3	**0.56±0.01**
	Urob.	0.72±0.01	8.0e3±1.0e4	**0.68±0.01**
	ANA	0.56±0.02	4.1e3±1.2e4	**0.56±0.03**
	Amm.	0.60±0.02	4.3e3±9.5e3	**0.61±0.02**
	Lipase	0.66±0.01	1.3e6±2.7e6	**0.68±0.02**
	Avg	0.64±0.01	3.1e8±5.0e8	**0.63±0.01**
BC	CSF	0.75±0.01		
	Trop.	0.78±0.00		
	Intub.	0.81±0.00		
	Pleur.	0.75±0.00		
	TSH	0.70±0.00		
	D-Dim.	0.76±0.00		
	Urob.	0.75±0.00		
	ANA	0.73±0.01		
	Amm.	0.77±0.00		
	Lipase	0.71±0.00		
	Avg	0.75±0.00		

when using a more sophisticated natural language model, whereas our current aim was to demonstrate the advantages of CET compared to alternative learning frameworks.

The CET model learns from censored event times to predict the probability that users will post to a given subreddit. This is particularly advantageous when predicting mental health status, as many users with mental health problems may discontinue social media activity before they might otherwise decide to post. Good prediction performance also suggests that CET might be effective for predicting other social media activity, or in recommender systems that predict user interest in specific media content.

Table 5: Performance metrics on Reddit dataset.

	SubR	AUC	MRAE	CI
CET	ADHD	**0.63±0.00**	**0.04±0.00**	0.58±0.01
	Anx.	**0.69±0.00**	**0.04±0.00**	0.59±0.01
	Books	**0.67±0.00**	**0.06±0.00**	0.59±0.00
	Dep.	**0.68±0.00**	**0.07±0.00**	0.62±0.01
	Fit.	**0.68±0.00**	**0.09±0.00**	0.64±0.00
	LPT	**0.68±0.00**	**0.08±0.00**	0.60±0.00
	MH	**0.68±0.00**	**0.02±0.00**	0.55±0.01
	SW	**0.68±0.00**	**0.03±0.00**	0.57±0.01
	WN	**0.93±0.01**	**0.01±0.00**	**0.73±0.02**
	Avg	**0.70±0.00**	**0.05±0.00**	0.61±0.01
ET	ADHD	0.59±0.01	3.74±0.41	**0.64±0.01**
	Anx.	0.68±0.00	3.37±0.32	**0.72±0.00**
	Books	0.66±0.00	2.50±0.36	**0.67±0.00**
	Dep.	0.67±0.00	1.86±0.19	**0.70±0.00**
	Fit.	0.67±0.00	0.95±0.10	**0.66±0.00**
	LPT	0.68±0.00	1.36±0.26	**0.65±0.00**
	MH	0.63±0.00	6.45±0.96	**0.65±0.01**
	SW	0.66±0.00	4.18±0.50	**0.70±0.00**
	WN	0.37±0.05	10.10±2.91	0.39±0.05
	Avg	0.62±0.01	3.83±0.52	**0.65±0.01**
BC	ADHD	0.62±0.00		
	Anx.	0.68±0.00		
	Books	0.66±0.00		
	Dep.	0.67±0.00		
	Fit.	0.66±0.00		
	LPT	0.67±0.00		
	MH	0.67±0.00		
	SW	0.68±0.00		
	WN	0.92±0.01		
	Avg	0.69±0.00		

6. Conclusion

In this work we have presented conditional event time models, argued that they are advantageous when modeling event occurrence and event times in a variety of real-world settings, and described how they can be implemented as a neural network with a binary stochastic layer representing the unknown, eventual occurrence of each event of interest. Results demonstrate that CET yields superior event occurrence probabilities and event time predictions compared to alternative approaches across one synthetic and two real-world datasets comprising a total of 21 distinct prediction tasks. Learning of CET models on large-scale datasets is facilitated by recent, improved methods for estimating gradients across categorical variables in neural networks. We believe CET, rather than alternative event time models, should be preferred when learning from multiple censored events, particularly when accurate prediction of eventual event occurrence is a primary goal. Future work will

focus on evaluating CET in additional real-world settings, including prediction of medical diagnoses, wherein learning event occurrence probabilities from censored events is critical to avoid selection biases that may otherwise confound results.

Acknowledgments

This research was supported in part by NIH/NIBIB R01-EB025020 and by Duke Forge, Duke University's center for health data science.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283, 2016.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018.

Paidamoyo Chapfuwa, Chenyang Tao, Chunyuan Li, Courtney Page, Benjamin Goldstein, Lawrence Carin, and Ricardo Henao. Adversarial time-to-event modeling. In *International Conference on Machine Learning*, pages 735–744, 2018.

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

John F Dovidio and Susan T Fiske. Under the radar: how unexamined biases in decision-making processes in clinical interactions can contribute to health care disparities. *American journal of public health*, 102(5):945–952, 2012.

Regina C. Elandt-Johnson. Conditional failure time distributions under competing risk theory with dependent failure times and proportional hazard rates. *Scandinavian Actuarial Journal*, 1976(1):37–51, January 1976. ISSN 0346-1238. doi: 10.1080/03461238.1976.10405934.

V. T. Farewell. The Use of Mixture Models for the Analysis of Survival Data with Long-Term Survivors. *Biometrics*, 38(4):1041–1046, 1982. ISSN 0006-341X. doi: 10.2307/2529885.

Jeffrey J. Gaynor, Erick J. Feuer, Claire C. Tan, Danny H. Wu, Claudia R. Little, David J. Straus, Bayard D. Clarkson, and Murray F. Brennan. On the Use of Cause-Specific Failure and Conditional Failure Probabilities: Examples From Clinical Oncology Data. *Journal of the American Statistical Association*, 88(422):400–409, 1993. ISSN 0162-1459. doi: 10.2307/2290318.

Will Grathwohl, Dami Choi, Yuhuai Wu, Geoffrey Roeder, and David Duvenaud. Backpropagation through the void: Optimizing control variates for black-box gradient estimation, 2017.

Frank E Harrell Jr, Kerry L Lee, Robert M Califf, David B Pryor, and Robert A Rosati. Regression modelling strategies for improved prognostic prediction. *Statistics in medicine*, 3(2):143–152, 1984.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, 2011.

Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1):24, 2018a.

Jared L. Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1):24, December 2018b. ISSN 1471-2288. doi: 10.1186/s12874-018-0482-1.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-event prediction with neural networks and cox regression. *Journal of Machine Learning Research*, 20(129):1–30, 2019a.

Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-Event Prediction with Neural Networks and Cox Regression. *arXiv:1907.00825 [cs, stat]*, September 2019b. arXiv: 1907.00825.

Changhee Lee, William R Zame, Jinsung Yoon, and Mihaela van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

Arielle Pardes. The Inside Story of Reddit's Redesign. *Wired*, 2018. ISSN 1059-1028.

Rajesh Ranganath, Adler Perotte, Noémie Elhadad, and David Blei. Deep survival analysis. *arXiv preprint arXiv:1608.02158*, 2016.

Kan Ren, Jiarui Qin, Lei Zheng, Zhengyu Yang, Weinan Zhang, Lin Qiu, and Yong Yu. Deep recurrent survival analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4798–4805, 2019.

Dinghan Shen, Guoyin Wang, Wenlin Wang, Martin Renqiang Min, Qinliang Su, Yizhe Zhang, Chunyuan Li, Ricardo Henao, and Lawrence Carin. Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms. *arXiv preprint arXiv:1805.09843*, 2018.

George Tucker, Andriy Mnih, Chris J Maddison, John Lawson, and Jascha Sohl-Dickstein. Rebar: Low-variance, unbiased gradient estimates for discrete latent variable models. In *Advances in Neural Information Processing Systems*, pages 2627–2636, 2017.

Regula S von Allmen, Salome Weiss, Hendrik T Tevaearai, Christoph Kuemmerli, Christian Tinner, Thierry P Carrel, Juerg Schmidli, and Florian Dick. Completeness of follow-up determines validity of study findings: results of a prospective repeated measures cohort study. *PLoS One*, 10(10), 2015.

Lee-Jen Wei. The accelerated failure time model: a useful alternative to the cox regression model in survival analysis. *Statistics in medicine*, 11(14-15):1871–1879, 1992.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

Mingzhang Yin and Mingyuan Zhou. ARM: Augment-REINFORCE-merge gradient for stochastic binary networks. In *International Conference on Learning Representations*, 2019.

Panpan Zheng, Shuhan Yuan, and Xintao Wu. SAFE: A Neural Survival Analysis Model for Fraud Early Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33: 1278–1285, July 2019. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v33i01.33011278.

Appendix A: Derivation of Equation (8) (Section 3.2)

From equation (7) of section 3.2, we have the following expression for $p_\theta(t_i, s_i|c_i, \boldsymbol{x}_i)$:

$$\begin{aligned} p_\theta(t_i, s_i|c_i, \boldsymbol{x}_i) = & \, p_\theta(t_i, s_i = 1|c_i = 1, \boldsymbol{x}_i)^{s_i c_i} \\ & \times p_\theta(t_i, s_i = 0|c_i = 1, \boldsymbol{x}_i)^{(1-s_i)c_i} \\ & \times p(t_i, s_i = 1|c_i = 0, \boldsymbol{x}_i)^{s_i(1-c_i)} \\ & \times p(t_i, s_i = 0|c_i = 1, \boldsymbol{x}_i)^{(1-s_i)(1-c_i)}. \end{aligned} \tag{11}$$

We assign a small probability $0 < \epsilon \ll 1$ to $P(\mathcal{F}_i < \infty|c_i = 0)$, so that:

$$p(t_i, s_i = 1|c_i = 0, \boldsymbol{x}_i) = \epsilon f_\theta(t_i|\boldsymbol{x}_i) G_i(t_i) \tag{12}$$

$$\begin{aligned} p(t_i, s_i = 0|c_i = 0, \boldsymbol{x}_i) = & (1-\epsilon) g_i(t_i) + \epsilon g_i(t_i) F_\theta(t_i|\boldsymbol{x}_i) \\ \approx & \, g_i(t_i). \end{aligned} \tag{13}$$

This allows us to expand (11):

$$\begin{aligned} p_\theta(t_i, s_i|c_i, \boldsymbol{x}_i) = &\big(f_\theta(t_i|\boldsymbol{x}_i)G_i(t_i)\big)^{s_i c_i} \\ &\times \big(g_i(t_i)F_\theta(t_i|\boldsymbol{x}_i)\big)^{(1-s_i)c_i} \\ &\times \big(\epsilon f_\theta(t_i|\boldsymbol{x}_i)G_i(t_i)\big)^{s_i(1-c_i)} \\ &\times g_i(t_i)^{(1-s_i)(1-c_i)}. \end{aligned} \tag{14}$$

Simplifying, we obtain:

$$\begin{aligned} p_\theta(t_i, s_i|c_i, \boldsymbol{x}_i) = &\big(f_\theta(t_i|\boldsymbol{x}_i)G_i(t_i)\big)^{s_i} \\ &\times F_\theta(t_i|\boldsymbol{x}_i)^{(1-s_i)c_i} \\ &\times g_i(t_i)^{(1-s_i)}, \end{aligned} \tag{15}$$

$$\times \epsilon^{s_i(1-c_i)} \tag{16}$$

We then remove terms that do not depend on θ or c_i, including $g_i(\cdot)$ and $G_i(\cdot)$, to obtain equation (8) from section 3.2:

$$p_\theta(t_i, s_i|c_i, \boldsymbol{x}_i) \propto \epsilon^{s_i(1-c_i)} f_\theta(t_i|\boldsymbol{x}_i)^{s_i} F_\theta(t_i|\boldsymbol{x}_i)^{(1-s_i)c_i}. \tag{17}$$

Appendix B: Descriptive Statistics, MIMIC-III

MIMIC-III may be accessed, following approval, at https://mimic.physionet.org. A complete description of this dataset, including descriptive statistics for all tables used in this work, may be found in (Johnson et al., 2016).

Appendix C: Descriptive Statistics, Reddit

Reddit data was accessed via the pushshift.io API.

Our final dataset included the earliest 20 comments and first subreddit submissions to each of the nine chosen subreddits from 492,059 unique Reddit users active between 2005 and 2020. Supplementary table (6) shows the breakdown of comments and submissions by year:

Supplementary figure (4) shows the number of users who posted to each subreddit. Supplementary table (7) shows that the majority of users posted to only one of the nine subreddits, and none posted to eight or all nine.

Appendix D: Additional Experiment Details

All models were trained in Tensorflow 1.10 (Abadi et al., 2016) using a single NVIDIA Titan XP GPU.

Hyperparameters were explored via random search, selected uniformly in the ranges listed in supplementary table (8), and tuned to optimize AUC of the CET model on the validation set.

The Gumbel-Softmax estimator with a temperature of approximately .3 and was found to be optimal on all three datasets. 100 samples were adequate on all datasets; further

Year	Submissions	Comments
2005	0	69
2006	0	5228
2007	0	18118
2008	136	42624
2009	985	123788
2010	5285	301978
2011	14294	748479
2012	30456	1254574
2013	43994	1339624
2014	57023	1471171
2015	81687	1559287
2016	93793	1613979
2017	105302	1362261
2018	89941	0
2019	58704	0
2020	3717	0

Table 6: Comments and submissions by year in Reddit dataset

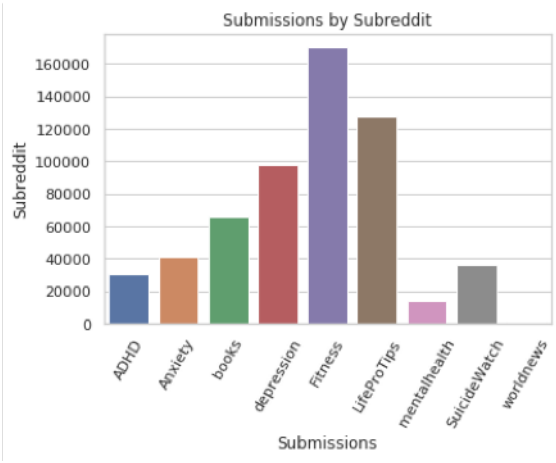

Figure 4: Total submissions to each subreddit

increasing the number of samples did not improve performance. Optimal values of $\log \epsilon$ were approximately -2 on all datasets. Layer widths of 750 (for $h_\phi(\boldsymbol{x}_i)$, $\mu_\theta(\boldsymbol{x}_i, \boldsymbol{c}_i)$, and $\nu_\theta(\boldsymbol{x}_i, \boldsymbol{c}_i)$) were used in the final MIMIC-III and Reddit models, whereas widths of 100 were used in the final Synthetic model.

Num. subreddits	Num. users
1	416965
2	60795
3	11236
4	2408
5	525
6	113
7	17
8	0
9	0

Table 7: Reddit users by the number of distinct subreddits to which they posted

Hyperparameter	Range
Estimator	{Gumbel-Softmax, ARM}
Num. c_i samples	$\{30, ..., 200\}$
$\log \epsilon$	$(-4, 0)$
Hidden units	$\{100, ..., 1000\}$
Gumbel-SM Temp.	$(0, 1)$

Table 8: Hyperparameter ranges for random search