Disentangling *Whether* from *When* in a Neural Mixture Cure Model for Failure Time Data

Matthew Engelhard **Ricardo Henao**
Department of Biostatistics and Bioinformatics, Duke University School of Medicine

Abstract

The mixture cure model allows failure probability to be estimated separately from failure timing in settings wherein failure never occurs in a subset of the population. In this paper, we draw on insights from representation learning and causal inference to develop a neural network based mixture cure model that is free of distributional assumptions, yielding improved prediction of failure timing, yet still effectively disentangles information about failure timing from information about failure probability. Our approach also mitigates effects of selection biases in the observation of failure and censoring times on estimation of the failure density and censoring density, respectively. Results suggest this approach could be applied to distinguish factors predicting failure occurrence versus timing and mitigate biases in real-world observational datasets.

1 INTRODUCTION

The mixture cure model was originally described by Farewell (1982) to model occurrence times of an event or failure of interest that does not occur in all individuals of the population. In contrast to standard approaches to the modeling of failure times, including the well-known Cox proportional hazards (Cox-PH) (Cox, 1972) and accelerated failure time (AFT) (Wei, 1992) models, the mixture cure model presumes that the population is divided into *susceptible* and *cured* individuals, and that failure can occur only in the former. Equivalently, susceptible individuals may be viewed as those who will fail in finite time, though not necessarily in a given observation window.

Proceedings of the 25[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

According to the mixture cure model, observations are generated by first drawing a binary random variable indicating susceptibility to failure. If the individual is susceptible, we then draw failure and censoring times, and the observed time is the minimum of the two, as in standard approaches for failure time. If the individual is not susceptible, on the other hand, we need only draw a censoring time, which will be observed. The probability of susceptibility *versus* cure typically follows a logistic model, though alternatives have been explored (Amico et al., 2019). The failure density is commonly modeled with Cox-PH, AFT (Xiang et al., 2011), or variants such as semi-parametric AFT (Zhang and Peng, 2007).

This model is appropriate in a wide variety of settings, from recommender systems that predict user interests, to industrial engineering models that predict tool defects. However, it is particularly useful in settings where learning is based on censored failure or event times, yet prediction of the times themselves is secondary to the effective prediction of failure susceptibility, or equivalently, lifetime event probability. In medical scenarios, for example, diagnosis times are often noisy and/or systematically biased (von Allmen et al., 2015; Dovidio and Fiske, 2012), therefore predicting the presence or absence of a condition of interest is more clinically meaningful than predicting the likely timing of diagnosis. Moreover, in this application and many others, we may wish to identify specific factors (predictors) that predict failure susceptibility versus timing, which requires a model that distinguishes between the two.

Because the mixture cure model has a failure time model at its core, it can take advantage of recent work to pair the failure time model with neural networks and stochastic gradient descent. Several studies have shown that replacing the usual linear formulation for the log-hazard ratio (*i.e.*, Cox-PH) or the parameters of a parametric failure time density (*e.g.*, AFT) with a neural network can improve performance, particularly on large datasets (Zheng et al., 2019; Katzman et al., 2018; Kvamme et al., 2019). More recently, however, this approach has been superseded by neural network based models that do not impose parametric assump-

tions on the form on the failure density (Ren et al., 2019; Lee et al., 2018; Chapfuwa et al., 2018; Tjandra et al., 2021).

Building on these developments, the first contribution of this work is to introduce a mixture cure model that is highly flexible in this same sense. Its flexibility allows our neural mixture cure (NMC) model to make more accurate failure time predictions, which in turn allows it to more accurately estimate failure susceptibility.

At the same time, however, this flexibility comes at a cost: the failure time prediction component of NMC can place arbitrary mass – including mass corresponding to the probability that a given individual is *cured* – at the end of the observation window, which in turn prevents us from distinguishing factors predicting failure timing from those affecting failure susceptibility.

Moreover, accurately modeling the failure density in this setting is challenging due to selection biases similar to those known to compromise prediction of individualized treatment effects from observational data. Specifically, selection biases in the observation of failure times results from the effects of features on censoring times and failure susceptibility, leading to poor prediction of failure times in affected regions of the domain.

To overcome these challenges, we draw on recent work in representation learning for causal inference shown to improve the accuracy of predicted treatment effects by balancing predictive features associated with treatment versus no treatment (Hassanpour and Greiner, 2019; Assaad et al., 2021).

In the same way, balancing features associated with susceptibility versus lack thereof is key to accurately predict both failure times and censoring times, which in turn improves prediction of event occurrence. So motivated, we propose the disentangled mixture cure model (DNMC) to improve the accuracy of failure time predictions via a learned representation in which effects of selection biases are reduced. By limiting information about failure susceptibility in the learned representation used to predict failure timing, this approach also allows factors predictive of the former to be easily distinguished from those predictive of the latter.

In summary, our contributions are as follows:

- Present a neural mixture cure (NMC) model free of strong parametric assumptions.
- Motivate, develop, and present a disentangled neural mixture cure (DNMC) model.
- Show that DNMC can identify features predicting failure timing, failure susceptibility, and both.
- Show that DNMC results in equal or better performance compared to alternative methods.
- Using risk of stroke as real-world scenario, show that DNMC identifies factors that differentially predict susceptibility versus timing.

All code needed to replicate our results is available at `https://github.com/mengelhard/dnmc`.

We begin by describing the existing mixture cure framework, then extend it to NMC and DNMC.

2 MIXTURE CURE MODEL

Consider triplets of random variables $\{\boldsymbol{X}, Y, S\}$, where $\boldsymbol{X} \in \mathcal{X} \subseteq \mathbb{R}^d$ is a d-dimensional feature vector, $Y \in (0, \infty)$ is an associated time, and $S \in \{0, 1\}$ denotes whether Y is a failure time or a right-censoring time. Unlike the standard failure time model, the cure model further supposes that individuals may or may not be *susceptible* to failure, and that a latent variable $E \in \{0, 1\}$ indicates their susceptibility or lack thereof. If $E = 0$, then the individual is *cured* and cannot fail.

Variable	Description
$\boldsymbol{X}$	d-dimensional feature vector
T	failure time
C	right-censoring time
Y	observed time
S	indicates failure (1) vs censoring (0)
E	indicates susceptible (1) vs cured (0)

Table 1: Definitions of Random Variables

Let $T \in (0, \infty)$ and $C \in (0, \infty)$ be unobserved (*i.e.,* latent) random variables associated with failure and censoring, respectively. In susceptible individuals, S indicates whether failure occurs before censoring, and Y is the minimum of T and C. In cured individuals, failure cannot occur, and therefore we have $S = 0$ and $Y = C$. Thus, S and Y are determined by E, T, and C as follows:

$$S = \mathbf{1}((E = 1) \wedge (T < C)), \tag{1}$$

$$Y = \begin{cases} \min(T, C) & \text{if } E = 1 \\ C & \text{if } E = 0 \end{cases}, \tag{2}$$

If censoring times are independent of $\boldsymbol{X}$, T, and E, we have the conditional dependence structure shown in Figure 1.

Suppose T is drawn from the failure density $f_T(t|\boldsymbol{x})$, which has associated survival function $F_T(t|\boldsymbol{x}) = 1 - \int_0^t f_T(\tau|\boldsymbol{x})d\tau$. Similarly, let C be drawn from the censoring density $f_C(c|\boldsymbol{x})$, which has associated "survival" function $F_C(c|\boldsymbol{x}) = 1 - \int_0^c F_C(\tau|\boldsymbol{x})d\tau$.

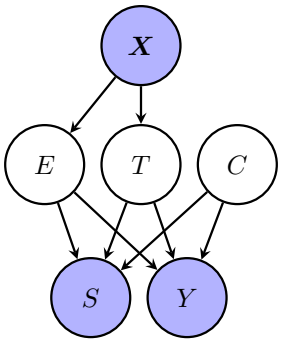


Figure 1: Conditional dependence structure among observed (shaded) and unobserved (not shaded) random variables under independent censoring.

In the mixture cure model, the density $p(y, s|\boldsymbol{x})$ is a mixture with components corresponding to susceptible versus cured individuals, respectively. In susceptible individuals, the model reduces to the standard failure time model. In cured individuals, on the other hand, failure cannot occur, so that $\forall y \forall \boldsymbol{x}$, we have $p(y, S = 1|\boldsymbol{x}, E = 1) = 0$, and the model reduces to the censoring density $f_C(y|\boldsymbol{x})$. This may be summarized as follows:

$$p(y,S = s|\boldsymbol{x}, E = e) = \begin{cases} f_T(y|\boldsymbol{x})F_C(y|\boldsymbol{x}) & \text{if } E = 1 \text{ and } S = 1 \\ F_T(y|\boldsymbol{x})f_C(y|\boldsymbol{x}) & \text{if } E = 1 \text{ and } S = 0 \\ f_C(y|\boldsymbol{x}) & \text{if } E = 0 \text{ and } S = 0 \\ 0 & \text{if } E = 0 \text{ and } S = 1 \end{cases}. \quad (3)$$

Having specified the density conditioned on possible values of E, we complete our model specification by supposing that $E \sim \text{Bern}(\sigma(h(\boldsymbol{x}))$, where $\sigma(\cdot)$ denotes the logistic (*i.e.,* sigmoid) function and $h(\cdot) : \mathcal{X} \to \mathbb{R}$. Marginalizing over E, we may write the full model as:

$$\begin{aligned} p(y, s|\boldsymbol{x}) = &\Big(\sigma(h(\boldsymbol{x}))f_T(y|\boldsymbol{x})F_C(y|\boldsymbol{x})\Big)^s \\ &\times\Big(\sigma(h(\boldsymbol{x}))F_T(y|\boldsymbol{x})f_C(y|\boldsymbol{x}) \\ &\quad + (1 - \sigma(h(\boldsymbol{x}))f_C(y|\boldsymbol{x})\Big)^{1-s}. \end{aligned} \quad (4)$$

Once forms for $f_T(y|\boldsymbol{x})$, $f_C(y|\boldsymbol{x})$, and $h(\boldsymbol{x})$ have been specified, their parameters may be optimized to maximize the likelihood $\sum_{\mathcal{D}} \log(p(y, s|\boldsymbol{x}))$ for a given dataset $\mathcal{D} = \{\boldsymbol{X}_i, Y_i, S_i\}_{i=1}^N$.

3 DISENTANGLING FAILURE PROBABILITY FROM FAILURE TIMING

Supposing independent censoring (see Figure 1), we may omit $f_C(y|\boldsymbol{x})$ and $F_C(y|\boldsymbol{x})$ from the likelihood in (4) when selecting parameters of $f_T(y|\boldsymbol{x})$, $f_C(y|\boldsymbol{x})$. Our goals are to (*a*) learn parameters θ_h of $h(\boldsymbol{x})$, *i.e.*, our model of failure susceptibility; (*b*) learn parameters θ_T of $f_T(y|\boldsymbol{x})$, *i.e.*, our model of the failure density; and (*c*) identify factors within $\boldsymbol{X}$ that are relevant to each.

However, our goal to disentangle failure probability from failure timing is complicated by several factors.

First, we wish to use a flexible model for f_T that does not place assumptions on the distribution of failure times, which has been shown to yield superior performance across a range of failure time benchmark tasks (Miscouridou et al., 2018; Lee et al., 2018; Tjandra et al., 2021). In contrast to a standard failure time model, however, this lack of distributional assumptions makes the mixture cure model non-identifiable: f_T can ignore h by shifting mass corresponding to $\Pr(E = 0)$ (i.e., not susceptible) to the end of the observation window and scaling the remainder of its mass to compensate. This has no effect on the likelihood, but results in a degenerate mixture cure model that does not distinguish failure susceptibility from failure timing.

Second, the learning of f_T is challenging due to two, closely related forms of selection bias:

1. In any failure time model, T is observed more often in individuals with earlier failure times, leading to poor estimation of f_T in regions of $\mathcal{X}$ associated with later failure times.

2. Unique to the cure model, T is observed more often in individuals who are susceptible to failure, leading to poor estimation of f_T in regions of $\mathcal{X}$ associated with low susceptibility.

To mitigate effects of selection bias and ensure information about failure susceptibility is captured by h, not f_T, we aim to learn an intermediate representation $\Omega(\cdot) : \mathcal{X} \to \boldsymbol{\Omega}$ that will serve as the input to f_T, and that has three important, related properties: (a) $\boldsymbol{\Omega}$ does not contain information about failure susceptibility; (b) susceptible versus cured individuals have similar distributions over $\boldsymbol{\Omega}$; and (c) individuals with earlier versus later failure times have similar distributions over $\boldsymbol{\Omega}$. More precisely, we would like $p_{\boldsymbol{\Omega}}(\boldsymbol{\omega}|S = 0)$ to be similar to $p_{\boldsymbol{\Omega}}(\boldsymbol{\omega}|S = 1)$; and $p_{\boldsymbol{\Omega}}(\boldsymbol{\omega}|E = 0)$ to be similar to $p_{\boldsymbol{\Omega}}(\boldsymbol{\omega}|E = 1)$, where $\boldsymbol{\omega} = \Omega(\boldsymbol{x})$.

Along with $\boldsymbol{\Omega}$, we design additional representations $\Phi(\cdot) : \mathcal{X} \to \boldsymbol{\Phi}$ and $\Psi : \mathcal{X} \to \boldsymbol{\Psi}$ to capture information predictive of failure susceptibility and information predictive of *both* susceptibility and timing, respectively, while applying regularization to limit the presence of redundant information across representations. The dependence of E and T on the representations $\boldsymbol{\Phi}$, $\boldsymbol{\Psi}$, and $\boldsymbol{\Omega}$ is summarized in Figure 2(a).

Returning to $\mathbf{\Omega}$, we enforce properties (a)-(c) by minimizing $d\big(p_{\mathbf{\Omega}}(\boldsymbol{\omega}|S=1), p_{\mathbf{\Omega}}(\boldsymbol{\omega}|S=0)\big)$, where $d(p,q)$ denotes the maximum mean discrepancy (MMD) (Gretton et al., 2012) between the densities p and q. This ensures $p_{\mathbf{\Omega}}(\boldsymbol{\omega}|S=0)$ is similar to $p_{\mathbf{\Omega}}(\boldsymbol{\omega}|S=1)$. We would also like $p_{\mathbf{\Omega}}(\boldsymbol{\omega}|E=0)$ to be similar to $p_{\mathbf{\Omega}}(\boldsymbol{\omega}|E=1)$, but in contrast to S, which is observed, the failure susceptibility E is *not* observed, therefore we cannot minimize $d\big(p_{\mathbf{\Omega}}(\boldsymbol{\omega}|E=1), p_{\mathbf{\Omega}}(\boldsymbol{\omega}|E=0)\big)$ directly. However, we will show in Section 3.1 that under typical conditions, $d\big(p_{\mathbf{\Omega}}(\boldsymbol{\omega}|S=1), p_{\mathbf{\Omega}}(\boldsymbol{\omega}|S=0)\big)$ is an upper bound on $d\big(p_{\mathbf{\Omega}}(\boldsymbol{\omega}|E=1), p_{\mathbf{\Omega}}(\boldsymbol{\omega}|E=0)\big)$.

Writing the complete collection of model parameters as $\boldsymbol{\theta} = \{\theta_T, \theta_h, \theta_\Phi, \theta_\Psi, \theta_\Omega\}$, we solve for:

$$\begin{aligned}\underset{\boldsymbol{\theta}}{\operatorname{argmin}} \Bigg(& \frac{1}{N}\sum_{i=1}^{N} -w_i \log p(y_i, s_i|\boldsymbol{x}_i;\boldsymbol{\theta}) \\ & +\lambda_d d\big(p_{\mathbf{\Omega}}(\boldsymbol{\omega}|S=1;\theta_\Omega), p_{\mathbf{\Omega}}(\boldsymbol{\omega}|S=0;\theta_\Omega)\big) \\ & +\lambda_R R(\boldsymbol{\theta})\Bigg),\end{aligned} \tag{5}$$

where λ_d and λ_R are hyperparameters weighting components of the loss corresponding to the distance $d(\cdot,\cdot)$ and a regularizer $R(\cdot)$, respectively. In our experiments, $R(\cdot)$ is an L2 penalty applied to all neural network parameters.

3.1 Bounding the Distance between Conditional Densities

We wish to minimize a distance between $p_{\mathbf{\Omega}}(\boldsymbol{\omega}|S=0)$ and $p_{\mathbf{\Omega}}(\boldsymbol{\omega}|S=1)$, as well as between $p_{\mathbf{\Omega}}(\boldsymbol{\omega}|E=0)$ and $p_{\mathbf{\Omega}}(\boldsymbol{\omega}|E=1)$, where $\boldsymbol{\omega} = \Omega(\boldsymbol{x})$. Here we use the maximum mean discrepancy (MMD), a distance metric defined as follows:

$$\begin{aligned}&\mathrm{MMD}[\mathcal{F}, p, q] \\ &\quad := \sup_{f\in\mathcal{F}} \big(\mathbb{E}_{z\sim p(z)}(f(z)) - \mathbb{E}_{z\sim q(z)}(f(z))\big).\end{aligned} \tag{6}$$

Claim 1. *Suppose $p(x)$ is a mixture of a finite number of densities $p_1, \ldots, p_K$, with corresponding weights $w_1, \ldots, w_K$. Then $d(p,q) \le \sum_{k=1}^K w_k d(p_k, q)$.*

Proof. Writing $\mathbb{E}_{z\sim p(z)}(f(z))$ as $\mathbb{E}_p(f(z))$, we have:

$$\begin{aligned} d(p,q) &= \sup_{f\in\mathcal{F}} \Big(\mathbb{E}_p(f(z)) - \mathbb{E}_q(f(z))\Big) \\ &= \sup_{f\in\mathcal{F}} \Big(\big[\textstyle\sum_{k=1}^K w_k\, \mathbb{E}_{p_k}(f(z))\big] - \mathbb{E}_q(f(z))\Big) \\ &= \sup_{f\in\mathcal{F}} \Big(\textstyle\sum_{k=1}^K w_k \big[\mathbb{E}_{p_k}(f(z)) - \mathbb{E}_q(f(z))\big]\Big) \\ &\le \textstyle\sum_{k=1}^K w_k \sup_{f\in\mathcal{F}} \Big(\mathbb{E}_{p_k}(f(z)) - \mathbb{E}_q(f(z))\Big),\end{aligned}$$

where the equality between the second and third lines follows from the fact that $\sum_{i=1}^K w_k = 1$. □

To make our notation more concise, let $p_e = p_{\mathbf{\Omega}}(\boldsymbol{\omega}|E=1)$, let $p_{\neg s} = p_{\mathbf{\Omega}}(\boldsymbol{\omega}|S=0)$, let $p_{\neg s,e} = p_{\mathbf{\Omega}}(\boldsymbol{\omega}|S=0, E=1)$, and so on. Note that $p_s = p_{s,e}$, since $\Pr(S=1, E=0) = 0$, and similarly $p_{\neg e} = p_{\neg s,\neg e}$, since $\Pr(S=0, E=1) = 0$.

Consider $d(p_e, p_{\neg e})$, the distance between densities conditioned on the value of e; it is this distance we wish to minimize. Since e is not observed, we cannot minimize this quantity directly. However, there are two conditions under which we may justify minimizing $d(p_s, p_{\neg s})$ instead.

Claim 2. *Writing p_e as $\alpha p_s + (1-\alpha)p_{\neg s,e}$, where $\alpha = Pr(S=1|E=1)$; and $p_{\neg s}$ as $\beta p_{\neg e} + (1-\beta)p_{\neg s,e}$, where $\beta = Pr(E=0|S=0)$, the distance $d(p_e, p_{\neg e})$ is bounded as follows:*

$$\begin{aligned} d(p_e, p_{\neg e}) \le & \alpha d(p_s, p_{\neg s}) \\ & +[\alpha(1-\beta) + (1-\alpha)] d(p_{\neg s,e}, p_{\neg e}).\end{aligned} \tag{7}$$

Proof. See Appendix; the proof is straightforward from Claim 1 and properties of the metric $d(\cdot,\cdot)$. □

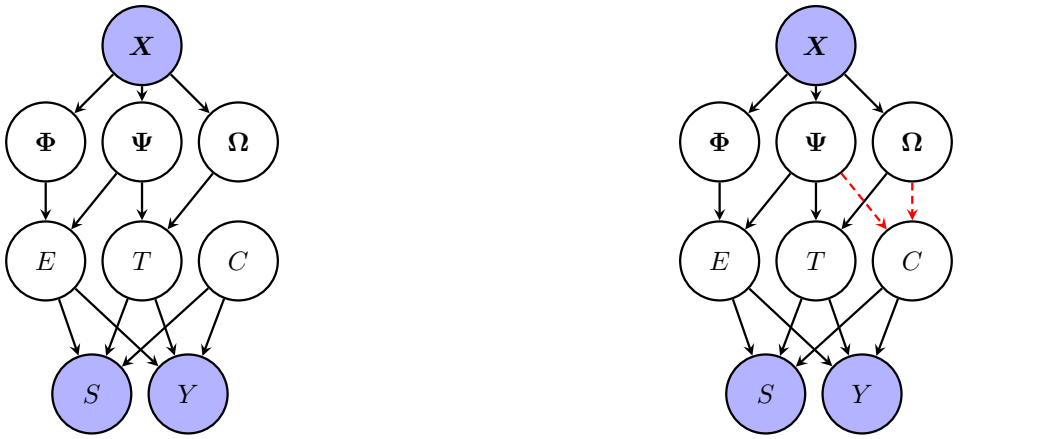
Note that $\beta = \Pr(E=0)/\big(\Pr(E=0) + (1-\alpha)\Pr(E=1)\big)$ (see Appendix), so that $\alpha \approx 1$ implies $\beta \approx 1$, and $\alpha \approx 0$ implies $\beta \approx \Pr(E=0)$. From this, we see that $d(p_s, p\neg s)$ becomes an upper bound on $d(p_e, p_{\neg e})$ as $\alpha \to 1$. However, this condition on α implies that almost all failures are observed, in which case a standard failure time model would more suitable.

Importantly, there is a second, more applicable condition under which $d(p_s, p_{\neg s})$ provides an upper bound on $d(p_e, p_{\neg e})$, namely when $d(p_{\neg s,e}, p_{\neg e}) \le d(p_s, p_{\neg s})$. Recalling that $p_{\neg s} = \beta p_{\neg e} + (1-\beta)p_{\neg s,e}$, this is satisfied when the representations $\Omega(\boldsymbol{x})$ are more similar between the two *subpopulations* of individuals without observed events – in other words, those who are versus are not susceptible – than between individuals with observed events versus those without. Although this condition is not guaranteed to hold, we anticipate it will hold under typical conditions, which allows us to approximately match $p_{\mathbf{\Omega}}(\boldsymbol{\omega}|E=0)$ to $p_{\mathbf{\Omega}}(\boldsymbol{\omega}|E=1)$ by using S, which is observed, as a proxy for E, which is not. Our empirical results support this approach.

4 DEPENDENT CENSORING

In settings with covariate-dependent censoring, which is common in medical applications, accurately modeling the censoring density f_C is challenging due to selection bias similar to that described in the previous section:

1. C is observed more often for regions of $\mathcal{X}$ associated with (a) later failure times, and (b) earlier



(a) $\mathbf{\Phi}$, $\mathbf{\Psi}$, and $\mathbf{\Omega}$ under independent censoring

(b) $\mathbf{\Phi}$, $\mathbf{\Psi}$, and $\mathbf{\Omega}$ under dependent censoring

Figure 2: Conditional dependence structure among observed (shaded) and unobserved (not shaded) random variables, including the learned representations $\mathbf{\Phi}$, $\mathbf{\Psi}$, and $\mathbf{\Omega}$.

censoring times, leading to poor estimation of f_C in regions of $\mathcal{X}$ associated with earlier failure or later censoring.

2. Again, unique to the cure model, C is observed more often in individuals who are *not* susceptible to failure, leading to poor estimation of f_C in regions of $\mathcal{X}$ with high susceptibility in settings where early failure is common.

These effects are mitigated by the same strategy described to improve learning f_T. As before, our goal is to learn a representation $\Omega(\cdot) : \mathcal{X} \to \mathbf{\Omega}$ such that the density is similar for individuals with earlier versus later failure times as well as for susceptible versus cured individuals. The former can be achieved directly by minimizing $d\big(p_{\mathbf{\Omega}}(\boldsymbol{\omega}|S=1), p_{\mathbf{\Omega}}(\boldsymbol{\omega}|S=0)\big)$. Under typical conditions this also achieves the latter objective, because $d\big(p_{\mathbf{\Omega}}(\boldsymbol{\omega}|S=1), p_{\mathbf{\Omega}}(\boldsymbol{\omega}|S=0)\big)$ is an upper bound on $d\big(p_{\mathbf{\Omega}}(\boldsymbol{\omega}|E=1), p_{\mathbf{\Omega}}(\boldsymbol{\omega}|E=0)\big)$ (Section 3.1).

Note that for cases in which censoring times are observed in *all* individuals, including those with observed failure times, we may be interested in estimating the causal effect of failure on the censoring time. In this case, there is a counterfactual outcome that is never observed, namely, the censoring time had failure *not* occurred in individuals for whom failure did occur, and the censoring time had failure occurred in individuals for whom it did not. Although we defer this direction for future work, this direct connection to causal methods makes the relationship between this work and previous work to learn disentangled representations to more accurately estimate causal effects (Hassanpour and Greiner, 2019) more clear.

5 MODEL DESCRIPTIONS

All models compared in this work follow the approach described by Miscouridou et al. (2018), Lee et al. (2018), Tjandra et al. (2021), and others, which does not assume a specific form for the failure density. Instead, failure times are discretized by partitioning the time horizon $(0, T_{\max})$ into K intervals such that all intervals contain approximately the same number of events. The probability of failure in each interval is then predicted by neural network with softmax activation.

Disentangled Neural Mixture Cure (DNMC) The DNMC model was instantiated via encoder networks for the representations Φ, Ψ, and Ω, which were parameterized by θ_Φ, θ_Ψ, and θ_Ω, respectively; as well as decoder networks f_T, f_C, and h, which were parameterized by θ_{f_T}, θ_{f_C}, and θ_h, respectively. In our experiments, each encoder was comprised of a single fully-connected layer with 256 units and ReLU activation. For all DMNC models, the maximum mean discrepancy used a Gaussian kernel with bandwidth set based on the median heuristic (Garreau et al., 2017), and loss as defined in (5). To evaluate the effect of the inclusion of Ψ on performance and identification of features, we also explored DNMC-Ψ, which omits the encoder for Ψ such that f_T and f_C are functions of Ω only, and h is a function of Φ only. DNMC-Ψ was otherwise identical to DNMC.

Neural Mixture Cure (NMC) The NMC model was instantiated via neural networks predicting f_T, f_C, and h, respectively, directly from the input $\boldsymbol{x}$. With the exception of the MMD term in (5), which is omitted, the loss is identical to the DMNC loss. In our experiments, each network was comprised of two fully-connected lay-

ers of 256 units each, with ReLU activation, such that the total number of layers and hidden units in NMC matched that of DNMC. However, DNMC contains approximately 3×256^2 additional parameters due to the fact that f_T, f_C, and h are each based on two of the three latent representations Φ, Ψ, and Ω.

Neural Survival (NSurv) The neural survival (NSurv) model is identical to NMC with the exception of the network predicting h, which is omitted. The NSurv loss corresponds to the standard failure time model. Equivalently, E is assumed to be 1 for all individuals. This model follows the approach to prediction of failure times used in Miscouridou et al. (2018), Lee et al. (2018), Tjandra et al. (2021), which does not assume a specific form for the failure density.

Multilayer Perceptron (MLP) Finally, we compared all models to a simple MLP trained to predict the failure indicator S. This model is identical to the NMC sub-network predicting h, and is trained to minimize the cross-entropy loss between $\sigma(h(\boldsymbol{x}))$ and S.

6 EXPERIMENTS

6.1 Datasets

Northern Alberta Cancer Dataset (NACD) This dataset was analyzed by Yu et al. (2011) when developing the multi-task logistic regression approach to the prediction of failure times. It is publicly available, and was originally derived from the Alberta Cancer Registry by the University of Alberta Cross Cancer Institute. The dataset contains survival times for 2,402 cancer patients along with demographics, laboratory measurements, reported symptoms, and other features collected before the patient's first chemotherapy treatment. The survival time is right-censored in 879 patients (36.6%).

Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT) This publicly available dataset, developed by Knaus et al. (1995), contains survival times for 9,105 seriously ill, hospitalized adult patients along with demographic information, medical history, physiologic and neurologic measures, and other features. The survival time is right-censored in 2904 patients (31.9%).

Pooled Stroke Risk Cohorts This dataset contains stroke occurrence and timing information pooled across three cohorts: the Framingham Offspring study (N=8,348) (Feinleib et al., 1975), the Atherosclerosis Risk in Communities (ARIC) Study (N=23,158) (investigators, 1989), and the Multi-Ethnic Study of Atherosclerosis (MESA) (N=6,390) (Bild et al., 2002). Features include demographic information and cardiovascular and other medical history. Stroke was observed in 1,543 of the 37,896 total patients (4.1%).

6.2 Performance Measures

AUC The area under the receiver operating characteristic (AUC) assesses binary classification performance of the learned $p_{\phi_h}(e|\boldsymbol{x})$ in predicting the probability of failure susceptibility. It is calculated using standard methods on the test set based on the predicted $p_{\phi_h}(e|\boldsymbol{x})$ and true E, when it is known. Since E is not directly observed, the performance of $p_{\phi_h}(e|\boldsymbol{x})$ can only be assessed for experiments with synthetic data, semi-synthetic data, or synthetic censoring. For the NSurv model, which predicts only the failure density, AUC was calculated based on predicted expected failure times by sweeping a threshold across the full range of these predictions.

Time-Dependent Concordance Index (CI) Correct ordering of failure time predictions is typically assessed using the concordance index (CI) developed by Harrell Jr et al. (1984), which quantifies the degree to which the order of predicted failure risks agrees with the observed failure times. However, our model allows relative risk between individuals to vary over time. Thus, we instead use the time-dependent concordance index developed by Antolini et al. (2005), which compares individuals' predicted risk at observed failure times to the predicted risk at that time for other individuals with later failure times. Pairs of failure times contribute to the CI only if (a) both failure times are known, or (b) one failure time is known, the other is censored, and the known failure time occurs before the censoring time.

6.3 Training and Evaluation

For all tasks, data were partitioned into training ($\sim$60%), validation ($\sim$20%), and test ($\sim$20%) sets. In all experiments, the time horizon $(0, T_{\max})$ was partitioned into $K = 10$ intervals. L2-regularization was used in all models with $\lambda_R = 0.03$, which was chosen because it maximized the performance of our primary baseline model, NSurv, on the NACD semi-synthetic validation set. This value was expected to perform well across all models due to the similarities in architectures and number of parameters. The strength of MMD regularization was also chosen based on performance on the NACD semi-synthetic validation set. Values above $\lambda_d = 10$ tended to make training less stable and consistent. All models were evaluated on 12 different versions of the semi-synthetic and syn-

Dataset	Model	AUC	CI
NACD Semi-Synthetic	**DNMC**	**.88±.02**	**.89±.02**
	DNMC−Ψ	**.88±.02**	**.89±.02**
	NMC	**.88±.02**	**.89±.02**
	NSurv	.80±.08	.88±.02
	MLP	.75±.08	NA
SUPPORT Semi-Synthetic	**DNMC**	.79±.06	**.88±.03**
	DNMC−Ψ	.76±.11	**.88±.02**
	NMC	**.80±.05**	.87±.04
	NSurv	.69±.09	**.88±.03**
	MLP	.71±.11	NA

Table 2: Performance with known failure susceptibility (semi-synthetic data).

Dataset	Model	AUC	CI
NACD Synthetic Censoring	**DNMC**	.87±.03	**.70±.02**
	DNMC−Ψ	.87±.03	.69±.02
	NMC	.87±.03	.69±.02
	NSurv	.87±.03	.69±.02
	MLP	**.88±.03**	NA
SUPPORT Synthetic Censoring	**DNMC**	**.77±.08**	**.64±.04**
	DNMC−Ψ	.76±.08	.63±.04
	NMC	.75±.10	.63±.05
	NSurv	**.77±.08**	**.64±.04**
	MLP	.76±.10	NA

Table 3: Performance with known failure susceptibility (synthetic censoring data).

thetic censoring datasets generated with three different random seeds, and with four different degrees of overlap between factors affecting failure susceptibility and other factors. Hyperparameters were identical across all runs. Reported performance measures are the mean and standard deviation of each measure over all runs. All models were implemented in Tensorflow 2.4 (Abadi et al., 2016), trained via backpropagation with the Adam optimizer (Kingma and Ba, 2014), a batch size of 100, learning rate of 1×10^{-3}, and no dropout.

6.4 Performance on Semi-Synthetic Data

Evaluation of NMC and DNMC on semi-synthetic data is critical because failure susceptibility is unknown in real datasets, but must be known in order to evaluate our models' performance in predicting it. Moreover, to evaluate performance in identifying factors predicting failure susceptibility, failure timing, and both, these factors must also be known. However, we use real features from NACD and SUPPORT, providing a more realistic data distribution.

As described in the previous section, a total of 24 distinct semi-synthetic datasets were generated based on (a) two datasets (*i.e.,* NACD and SUPPORT), (a) three random seeds, and (b) four degrees of overlap (4, 8, 12, 16) in the factors affecting failure susceptibility, failure and censoring timing, and both. Specific features affecting susceptibility, timing, and both were selected at random, and 20 features in total were used in each run. Coefficients corresponding to each feature were drawn from a standard normal distribution and applied to determine the probability of failure susceptibility, which was then drawn from a Bernoulli distribution, and the failure and censoring times.

The mean and standard deviation of the AUC and CI across all semi-synthetic datasets is shown in Table 2. Note that the MLP model predicts only failure susceptibility and not the failure density, so only the AUC and not the CI is applicable.

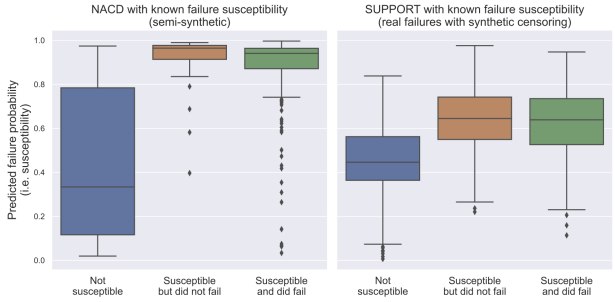


Figure 3: Predicted susceptibility to failure in (a) the cured population, (b) the portion of the susceptible population with no failure observed, and (c) the portion of the susceptible population with failure observed.

6.5 Performance on Real Data with Synthetic Censoring

While semi-synthetic data allows NMC and DNMC performance to be fully evaluated, the relationship between predictors and failure times may not be an adequate surrogate for real-world complexities. Using real failure times is therefore preferable, but failure susceptibility still must be known in order to evaluate the AUC of our models in predicting it. Thus, to further evaluate under more realistic conditions, we repeated our experiments with synthetic failure susceptibility, but real failure times. For individuals not susceptible to failure (synthetic) but with observed failure times (real), the failure time was changed to a censoring time.

Similar to the semi-synthetic datasets, a total of 24 synthetic censoring scenarios were considered. The mean and standard deviation of the AUC and CI across all synthetic censoring datasets is shown in Table 3.

Results across both the semi-synthetic datasets and real datasets with synthetic censoring show that NMC is consistently superior to NSurv and MLP. Furthermore, DNMC performance is equal to or better than NMC performance, while also permitting factors predicting failure susceptibility versus timing to be identified and

differentiated.

6.6 Identification of Factors

The extent to which DNMC correctly identified factors affecting failure susceptibility, failure timing, and both was evaluated on the semi-synthetic datasets, and is summarized in Figure 4.

The presented values are based on coefficients in the first layer of the encoders for Φ, Ψ, and Ω, respectively. Φ is intended to capture factors predicting only E, failure susceptibility, which are known in the semi-synthetic datasets. Similarly, Ω is intended to capture factors predicting only T, the failure time, and C, the censoring time, which are also known. Finally, Ψ is designed to capture factors predicting E as well as T and C.

More precisely, these values are the average of the coefficients in the slice of a given layer corresponding to the features it is intended to capture, which should be large. This is contrasted with the average value of coefficients in the complementary slice, which should be small. As Figure 4 shows, identification of these factors is more effective with stronger MMD penalty λ_d, as expected and intended. For λ_d values below 0.1, identification of factors was poor. For values of λ_d between 1 and 10, identification gradually improved without loss of performance. However, when values were increased beyond 10, training became less stable and consistent.

6.7 Analysis of Stroke Risk

Prediction of stroke is a suitable application for the mixture cure framework, because not all individuals are susceptible to stroke. Furthermore, understanding which risk factors predict stroke susceptibility, stroke timing, and both is of considerable clinical interest. Applying DNMC to the previously described pooled stroke risk cohorts therefore provided an opportunity to demonstrate its potential clinical value.

All models were applied and had similar prediction performance (CI = 0.72) on a held-out test set with approximately equal numbers of participants from each of the three cohorts. Additionally, average DNMC coefficients corresponding to each feature in the first layer of Ω and Φ, respectively, were used to quantify the importance of that feature on stroke timing and susceptibility, respectively. Results are shown in Table 4.

Not surprisingly, current age has the greatest effect on the predicted time until stroke occurrence, which is consistent with well-known increases in stroke rates with age. In contrast, age has much less effect on predicted susceptibility, which may be viewed as the individual's lifetime stroke probability. Other results are more difficult to interpret, but it is notable that aside from current age, total cholesterol has the most profound effect on the predicted time until stroke occurrence. Moreover, susceptibility is predominantly associated with other cardiovascular conditions such as left ventricular hypertrophy (L Vent Hyp), atrial fibrillation (Atrial Fib) and history of cardiovascular disease (Hx of CVD).

Predictor	Effect on ***Timing***	**Predictor**	Effect on ***Suscept.***
Age	1.0	L Vent Hyp	1.0
Total Chol	0.81	Systolic BP	0.97
Systolic BP	0.79	Hx of CVD	0.91
HDL Chol	0.70	Atrial Fib	0.87
Curr Smoker	0.56	On BP Med	0.85
Sex	0.56	Diabetes	0.80
Hx of CVD	0.37	Sex	0.79
Diabetes	0.36	Curr Smoker	0.73
On BP Med	0.35	Total Chol	0.69
L Vent Hyp	0.33	HDL Chol	0.64
Atrial Fib	0.15	Age	0.57

Table 4: Predictors ordered by their effects on stroke timing (left) and stroke susceptibility (right), as quantified by the average coefficient magnitude in the first layer of the encoder for Ω and Φ, respectively. Values are relative to the largest value to conserve space.

While AUC cannot be evaluated in this setting, since stroke susceptibility is unknown in patients without an observed stroke, comparing the predicted probability of stroke susceptibility between individuals who had a stroke versus others shows a large effect size between groups (Cohen's $d = 0.61$). This effect was statistically significant (Mann-Whitney $U = 907093$; $p < 10^{-26}$).

7 CONCLUSIONS

The mixture cure model is designed to predict failure or event times that occur in some individuals and not in others. In this paper, we introduced a neural mixture cure model that extends the recent development of flexible, neural network based failure time models to the mixture cure setting. Drawing on parallels to causal inference and representation learning, we then developed the disentangled neural mixture cure model, which ensures factors predicting failure susceptibility are distinguished from those predicting failure timing while also mitigating effects of selection bias to improve prediction of failure and censoring times. Results on semi-synthetic data showed that this approach allows factors predicting failure susceptibility versus timing to be identified. We then applied this approach to identify factors predicting stroke susceptibility versus timing and observed that such factors are distinct and consis-

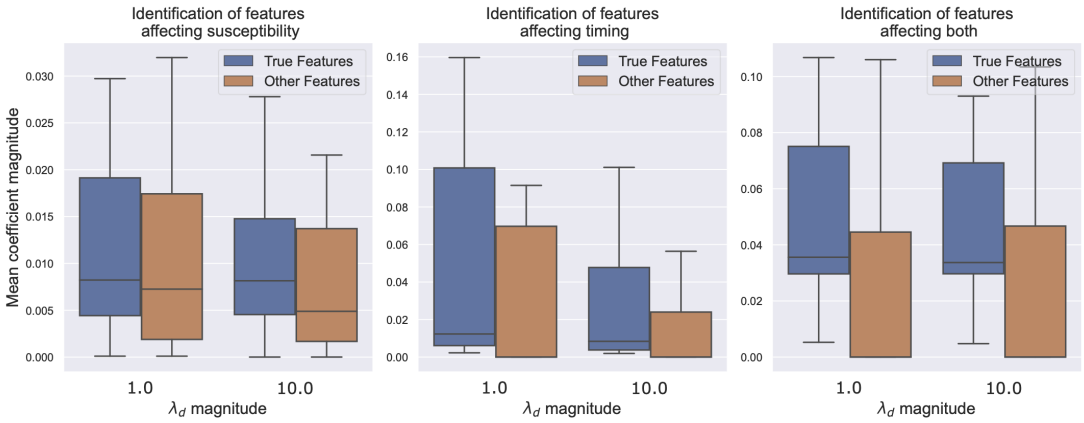


Figure 4: Correct identification of features affecting failure timing, failure susceptibility, and both in DNMC models with weak (λ_d=1.0) versus strong (λ_d=10.0) MMD regularization on Ω. Model coefficients corresponding to the true features are larger than other coefficients, and this effect is more pronounced with stronger regularization ($\lambda_d = 10.0$). Weak regularization ($\lambda_d \ll 1.0$) resulted in poor identification, and excessive regularization ($\lambda_d \gg 10.0$) led to inconsistent performance.

tent with clinical reasoning. As future work we will seek to extend the proposed approach for counterfactual prediction for comparative effectiveness applications.

Acknowledgements

This research was supported in part by NIH/NIDDK R01-DK123062, NIH/NINDS R61-NS120246, and ONR.

References

Vern T Farewell. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, pages 1041–1046, 1982.

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

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.

Maïlis Amico, Ingrid Van Keilegom, and Catherine Legrand. The single-index/cox mixture cure model. *Biometrics*, 75(2):452–462, 2019.

Liming Xiang, Xiangmei Ma, and Kelvin KW Yau. Mixture cure model with random effects for clustered interval-censored survival data. *Statistics in Medicine*, 30(9):995–1006, 2011.

Jiajia Zhang and Yingwei Peng. A new estimation method for the semiparametric accelerated failure time mixture cure model. *Statistics in medicine*, 26 (16):3157–3171, 2007.

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.

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.

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.

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 2018. ISSN 1471-2288. doi: 10.1186/s12874-018-0482-1.

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

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.

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.

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.

Donna Tjandra, Yifei He, and Jenna Wiens. A hierarchical approach to multi-event survival analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 591–599, 2021.

Negar Hassanpour and Russell Greiner. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*, 2019.

Serge Assaad, Shuxi Zeng, Chenyang Tao, Shounak Datta, Nikhil Mehta, Ricardo Henao, Fan Li, and Lawrence Carin Duke. Counterfactual representation learning with balancing weights. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1972–1980. PMLR, 13–15 Apr 2021. URL `https://proceedings.mlr.press/v130/assaad21a.html`.

Xenia Miscouridou, Adler Perotte, Noemie Elhadad, and Rajesh Ranganath. Deep survival analysis: Nonparametrics and missingness. In *Machine Learning for Healthcare Conference*, page 244–256. PMLR, 2018.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL `http://jmlr.org/papers/v13/gretton12a.html`.

Damien Garreau, Wittawat Jitkrittum, and Motonobu Kanagawa. Large sample analysis of the median heuristic. *arXiv preprint arXiv:1707.07269*, 2017.

Chun-Nam Yu, Russell Greiner, Hsiu-Chin Lin, and Vickie Baracos. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. *Advances in neural information processing systems*, 24:1845–1853, 2011.

William A Knaus, Frank E Harrell, Joanne Lynn, Lee Goldman, Russell S Phillips, Alfred F Connors, Neal V Dawson, William J Fulkerson, Robert M Califf, Norman Desbiens, et al. The support prognostic model: Objective estimates of survival for seriously ill hospitalized adults. *Annals of internal medicine*, 122(3):191–203, 1995.

Manning Feinleib, William B Kannel, Robert J Garrison, Patricia M McNamara, and William P Castelli. The framingham offspring study. design and preliminary data. *Preventive medicine*, 4(4):518–525, 1975.

The ARIC investigators. The atherosclerosis risk in communities (aric) study: Design and objectives. *American Journal of Epidemiology*, 129(4):687–702, 1989.

Diane E Bild, David A Bluemke, Gregory L Burke, Robert Detrano, Ana V Diez Roux, Aaron R Folsom, Philip Greenland, David R JacobsJr, Richard Kronmal, Kiang Liu, et al. Multi-ethnic study of atherosclerosis: objectives and design. *American journal of epidemiology*, 156(9):871–881, 2002.

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.

Laura Antolini, Patrizia Boracchi, and Elia Biganzoli. A time-dependent discrimination index for survival data. *Statistics in medicine*, 24(24):3927–3944, 2005.

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.

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

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*, 2019.

Supplementary Material: Disentangling *Whether* from *When* in a Neural Mixture Cure Model for Failure Time Data

A SOCIETAL IMPACT

This work was motivated by the need to mitigate effects of bias when developing predictive models from observational health datasets, including electronic health records. In healthcare, disparities exist in both in *rates* of diagnosis and in the *timing* of diagnosis. This is true for autism, ADHD, and many other mental health and physical health conditions with profound effects on long-term health and quality of life. This work does not address disparities in rates of diagnosis, but it does address disparities in the timing of diagnosis by disentangling information about diagnosis probability from information about diagnosis timing. Specifically, our approach limits the degree to which disparities in diagnosis timing present in the training data compromise the model's diagnosis risk predictions, which allows individuals at risk for these conditions to be identified more equitably, potentially contributing to more equitable treatment and outcomes.

Furthermore, by improving methods to accurately estimate the censoring density in the mixture cure setting, this work improves our ability to identify individuals who are most likely to be lost to follow-up before their health conditions are recognized and treated, which disproportionately affects disadvantaged groups. Identifying these individuals permits interventions that aim to provide them with access to care and other resources that may improve their long-term health outcomes.

B CO2 FOOTPRINT

All experiments were conducted using a single Tesla V100-PCIE-16GB GPU, which has an estimated carbon efficiency of 0.432 $\text{kgCO}_2\text{eq/kWh}$. A total of 96 hours of compute time were required, resulting in estimated total emissions of 12.44 kgCO_2eq, which is comparable to driving 50.3 kilometers in a typical car.

These estimates were generated using the MachineLearning Impact calculator presented by Lacoste et al. (2019).

C DETAILS OF EQUATION (7)

Additional details related to equation (7) are found below.

$$
\begin{aligned}
d(p_e,p_{\neg e}) &\leq \alpha d(p_s, p_{\neg e}) + (1-\alpha) d(p_{\neg s,e}, p_{\neg e}) \\
&\leq \alpha\big(d(p_s, p_{\neg s}) + d(p_{\neg s}, p_{\neg e})\big) + (1-\alpha) d(p_{\neg s,e}, p_{\neg e}) \\
&= \alpha\big(d(p_s, p_{\neg s}) + d(\beta p_{\neg e} + (1-\beta) p_{\neg s,e}, p_{\neg e})\big) + (1-\alpha) d(p_{\neg s,e}, p_{\neg e}) \\
&\leq \alpha d(p_s, p_{\neg s}) + \alpha\beta d(p_{\neg e}, p_{\neg e}) + \alpha(1-\beta) d(p_{\neg e}, p_{\neg s,e}) + (1-\alpha) d(p_{\neg s,e}, p_{\neg e}) \\
&= \alpha d(p_s, p_{\neg s}) + [\alpha(1-\beta) + (1-\alpha)] d(p_{\neg s,e}, p_{\neg e})
\end{aligned}
\tag{8}
$$