From Biased Selective Labels to Pseudo-Labels: An Expectation-Maximization Framework for Learning from Biased Decisions

Trenton Chang [1] **Jenna Wiens** [1]

Abstract

Selective labels occur when label observations are subject to a decision-making process; *e.g.*, diagnoses that depend on the administration of laboratory tests. We study a clinically-inspired selective label problem called disparate censorship, where labeling biases vary across subgroups and unlabeled individuals are imputed as "negative" (*i.e.*, no diagnostic test = no illness). Machine learning models naïvely trained on such labels could amplify labeling bias. Inspired by causal models of selective labels, we propose Disparate Censorship Expectation-Maximization (DCEM), an algorithm for learning in the presence of disparate censorship. We theoretically analyze how DCEM mitigates the effects of disparate censorship on model performance. We validate DCEM on synthetic data, showing that it improves bias mitigation (area between ROC curves) without sacrificing discriminative performance (AUC) compared to baselines. We achieve similar results in a sepsis classification task using clinical data.

1. Introduction

Selective labels occur when a decision-making process determines access to ground truth (Lakkaraju et al., 2017). We study a practical case of selective labels: disparate censorship (Chang et al., 2022). Disparate censorship introduces two challenges: different labeling biases across subgroups and the assumption that unlabeled individuals have a *negative* label. For example, in healthcare, labels may depend on laboratory test results only available in some patients.

Past work has trained ML models to predict outcomes based on laboratory test results (e.g., sepsis (Seymour et al., 2016; Rhee & Klompas, 2020)). In this setting, patients with no test result are defined as negative (Hartvigsen et al., 2018; Teeple et al., 2020; Jehi et al., 2020; McDonald et al., 2021; Adams et al., 2022; Kamran et al., 2022). However, laboratory testing decisions may be biased. For example, women are undertested and underdiagnosed for cardiovascular disease (Beery, 1995; Schulman et al., 1999). ML models trained on such data may recommend women less often for diagnostic testing than men, reinforcing inequity.

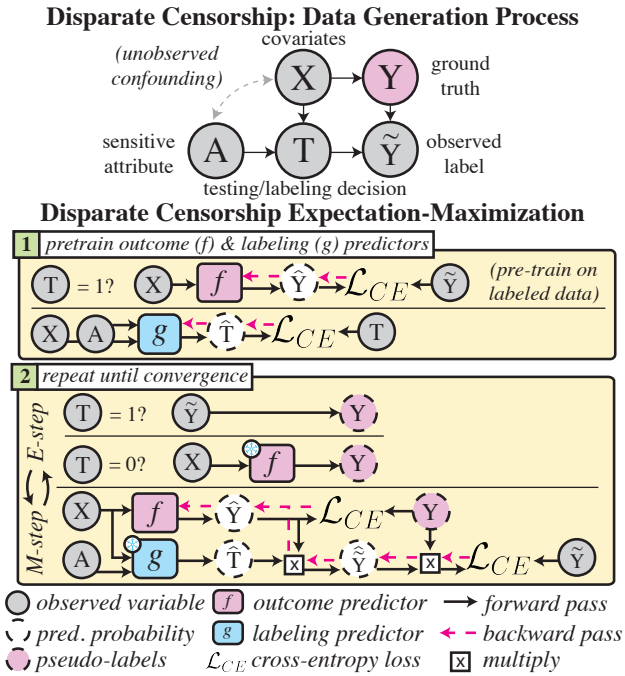

Figure 1. **Top:** Causal model of disparate censorship (**x**: covariates, y: ground truth, $\tilde{y}$: observed label, t: testing/labeling indicator, a: sensitive attribute). Shaded variables are fully observed. **Bottom:** Disparate Censorship Expectation-Maximization (DCEM). Dashed nodes are probabilistic estimates.

To address this bias, one option is to train only on tested individuals. Such an approach may discard a large subset of the data and may not generalize to untested patients. Another option is semi-supervised approaches that do not assume untested patients are negative, such as label propagation (Zhu & Ghahramani, 2002; Lee, 2013) or filtering (Li et al., 2020; Nguyen et al., 2020), or noisy-label learning methods (Blum & Stangl, 2020; Wang et al., 2021; Zhu

[1]Division of Computer Science & Engineering, University of Michigan, Ann Arbor, MI, USA. Correspondence to: Trenton Chang <ctrenton@umich.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

et al., 2021). However, such methods do not leverage causal models of label bias, a potential source of additional information. We aim to develop an approach that leverages all available signal while accounting for labeling biases.

Inspired by causal models of selective labeling (Laine et al., 2020; Chang et al., 2022; Guerdan et al., 2023a), we propose a simple method for mitigating bias when training models under disparate censorship: Disparate Censorship Expectation-Maximization (DCEM; Fig. 1). First, we show that DCEM regularizes model estimates to counterbalance disparate censorship. We validate DCEM in a simulation study and a sepsis classification task on clinical data. We find that our method mitigates bias (area between ROC curves) while maintaining competitive discriminative performance (AUC), and is generally more robust than baselines to changes in the data generation process.

2. Preliminaries: Disparate Censorship

We consider a dataset $\{\mathbf{x}^{(i)}, \tilde{y}^{(i)}, t^{(i)}, a^{(i)}\}_{i=1}^{N}$, with covariates $\mathbf{x}^{(i)} \in \mathbb{R}^d$, labeling/testing decision $t^{(i)} \in \{0, 1\}$, sensitive attribute $a^{(i)}$, and observed label $\tilde{y}^{(i)} \in \{0, 1\}$, a proxy for ground truth $y^{(i)} \in \{0, 1\}$. The proxy label $\tilde{y}^{(i)} = y^{(i)}$ when $t^{(i)} = 1$, and $\tilde{y}^{(i)} = 0$ otherwise (*i.e.*, $\tilde{y}^{(i)} = y^{(i)} t^{(i)}$).

What is disparate censorship? Disparate censorship models "double standards" in label collection decisions (Fig. 1, top). It is a variation of selective labeling or outcome measurement error (Lakkaraju et al., 2017; Guerdan et al., 2023b). Disparate censorship uniquely assumes that untested individuals are imputed as *negative.*

We consider disparate censorship in the context of binary classification (Chang et al., 2022) (Fig. 1, top). We justify the model by example. Consider a patient in an emergency room with characteristics $\mathbf{x}$ and sensitive attribute a. This patient may have some condition y (currently unobserved) caused by $\mathbf{x}$ but *not* a. A clinician may order a diagnostic test (set t to 1) to determine y. The decision is based on $\mathbf{x}$, but could be swayed by biases in a.

To simplify, suppose that tests are perfectly sensitive.[1] Then, we observe ground truth for tested individuals ($t = 1 \implies \tilde{y} = y$). Otherwise, the patient's label is imputed as negative ($t = 0 \implies \tilde{y} = 0$; *i.e.*, untested patients are presumed healthy). However, due to biases in testing decisions t, y may only be available in a biased subset of the data. The causal model of disparate censorship (Fig. 1, top) encodes this decision-making pipeline. Beyond healthcare, disparate censorship may arise whenever potentially biased decisions affect data labeling.

[1] If not, we can define T to indicate whether a label is *confirmed* correct. This definition captures differences in test sensitivity across groups (*i.e.*, spectrum bias (Mulherin & Miller, 2002)).

Learning under disparate censorship. We aim to learn a mapping $f_\theta : \mathbf{x} \to y$ parameterized by θ optimized for discriminative performance (*i.e.*, AUC), but only observe proxy labels $\tilde{y}$. The default approach for learning under disparate censorship is to assume $y = \tilde{y}$ and proceed using supervised learning. However, such an f_θ may encode labeling biases: estimates of $P(\tilde{Y} \mid X)$ may be inflated compared to $P(Y \mid X)$ for those more likely to be labeled. Thus, biased labeling could yield disproportionate impacts on performance across different subgroups of the data.

Note that we can interpret the estimand of interest as the causal effect of testing on the observed label, since, in the language of do-calculus (Pearl, 2009),

$$\begin{aligned}\mathbb{E}[Y \mid X] &= \mathbb{E}[Y \mid X, T = 1] = \mathbb{E}[\tilde{Y} \mid X, T = 1] \\ &= \mathbb{E}[\tilde{Y} \mid X, do(T = 1)], \end{aligned} \tag{1}$$

which follows from standard causal identifiability derivations given the causal graph of Fig. 1 (Imbens & Rubin, 2015). Intuitively, testing an individual ($do(T = 1)$) reveals their outcome. Thus, a model trained only on tested individuals could consistently estimate $P(Y|X)$, but may not correct for labeling bias. We discuss other approaches in semi-supervised learning in Section 6.

3. Methodology

We propose Disparate Censorship Expectation-Maximization (DCEM) as an approach for learning in the presence of disparate censorship. We first build intuition for how one could mitigate disparate censorship based on the causal model (Section 3.1). We then derive DCEM (Section 3.2) and show that it mitigates disparate censorship via a form of regularization (Section 3.3). We consider alternative designs and their limitations (Section 3.4). Detailed proofs and definitions are in Appendix B.

3.1. Towards mitigating disparate censorship

Recalling the causal model of disparate censorship, suppose that we are naively training a model f_θ to predict $\tilde{y}$. Define groups a and a' and $\mathbf{x} \sim \mathcal{X}$. Consider some $\mathcal{X}' \subseteq \mathcal{X}$ so that

$$\mathop{\mathbb{P}}_{\mathbf{x} \in \mathcal{X}'}[T \mid X, A = a] << \mathop{\mathbb{P}}_{\mathbf{x} \in \mathcal{X}'}[T \mid X, A = a'] \tag{2}$$

for all $\mathbf{x} \in \mathcal{X}'$. Define $\hat{t} \triangleq P(T \mid X = \mathbf{x}, A = a)$ (*e.g.*, probability of receiving a laboratory test) and $\hat{y} \triangleq P(Y \mid X = \mathbf{x})$. By assumption, $\mathbf{x}$ is sufficient for predicting y (*i.e.*, as in Fig. 1, top), such that the optimal $\hat{y}$ should be similar across a (within $\mathcal{X}'$). However, Eq. 2 states that group a is *undertested* relative to group a': they have a lower $\hat{t}$ within $\mathcal{X}'$. Equivalently, labeling bias *favors* group a'. Thus, our naive model would underestimate $\hat{y}$ in group a (lower $\hat{t}$ than group a' in $\mathcal{X}'$) relative to group a'.

Algorithm 1 Disparate Censorship Expectation-Maximization. $\mathcal{L}$: binary cross-entropy loss.

Input: $\{(\mathbf{x}^{(i)}, \tilde{y}^{(i)}, t^{(i)}, a^{(i)})\}_{i=1}^N$
Output: $f_\theta : \mathcal{X} \to [0,1]$
$f_\theta \leftarrow \arg\min_{f_\theta} \frac{1}{|\{i:t^{(i)}=1\}|} \sum_{i:t^{(i)}=1} \mathcal{L}(\tilde{y}^{(i)}, f_\theta(\mathbf{x}^{(i)}))$
$g_{\zeta^*} \leftarrow \arg\min_{g_\zeta} \frac{1}{N} \sum_{i=1}^N \mathcal{L}(t^{(i)}, g_\zeta(\mathbf{x}^{(i)}, a^{(i)}))$
$\hat{t}^{(i)} \leftarrow g_{\zeta^*}(\mathbf{x}^{(i)}, a^{(i)})$
while not converged do
  $Q(y^{(i)}) \leftarrow t^{(i)}\tilde{y}^{(i)} + (1-t^{(i)}) f_\theta(\mathbf{x}^{(i)})$ // E-step
  $f_\theta \leftarrow \arg\min_{f_\theta} \frac{1}{N} \sum_{i=1}^N \mathcal{L}(Q(y^{(i)}), f_\theta(\mathbf{x}^{(i)}))$
    $+ Q(y^{(i)}) \mathcal{L}(\tilde{y}^{(i)}, f_\theta(\mathbf{x}^{(i)}) \cdot \hat{t}^{(i)})$ // M-step
end while
return f_θ

To counterbalance this bias, one could increase $\hat{y}$ (within $\mathcal{X}'$) where group a is more prevalent than group a'; *i.e.*, lower-$\hat{t}$ regions. Since we are interested in discriminative performance, this is analogous to decreasing $\hat{y}$ where $\hat{t}$ is higher, from which the proposed method follows. More broadly, variables associated with labeling bias (A causally affects T) but not the outcome of interest (A does not causally affect Y) may be useful for mitigating labeling bias.

Given our causal model with latent variable Y (Fig. 1, top), we base our approach on expectation-maximization (EM) (Dempster et al., 1977). We can write:

$$\begin{aligned} &P(\tilde{Y}, Y, T, X, A, U) \\ &= P(\tilde{Y} \mid Y, T) P(Y \mid X) P(T \mid X, A) P(X, A, U). \end{aligned} \tag{3}$$

Since y is not fully observed, Eq. 3 cannot be optimized via standard supervised objectives. Dropping terms that do not involve Y, we can write the maximization of Eq. 3 as

$$\max \; P(\tilde{Y} \mid Y, T) P(Y \mid X). \tag{4}$$

Optimizing Eq. 4 proceeds via EM. We show that the resulting objectives align with reducing $\hat{y}$ in higher-$\hat{t}$ regions and maintain discriminative performance on tested individuals.

3.2. Disparate Censorship Expectation-Maximization

Informal overview. EM alternates an expectation step (E-step), which imputes guesses for the latent variable(s) (*i.e.*, Y in Eq. 4), and a maximization step that optimizes likelihood given the imputed estimates (M-step, *i.e.*, Eq. 4). Our E-step imputes preliminary estimates of $P(Y \mid X)$ for untested individuals. Our M-step updates the estimates to counteract labeling biases when $t^{(i)} = 0$, and is equivalent to full supervision when $t^{(i)} = 1$. The E- and M-steps alternate until convergence. Fig. 1 (bottom) shows a schematic of DCEM, with pseudocode in Algorithm 1.

E-step. The posterior of $y^{(i)}$ given the observed data is:

$$Q(y^{(i)}) \triangleq \mathbb{E}[y^{(i)} \mid \tilde{y}^{(i)}, t^{(i)}, \mathbf{x}^{(i)}, a^{(i)}]. \tag{5}$$

We can rewrite Eq. 5 as follows:

Theorem 3.1 (E-step)**.** *The posterior distribution of $y^{(i)}$ given the observed data is equivalent to*

$$Q(y^{(i)}) = \begin{cases} \tilde{y}^{(i)} & t^{(i)} = 1 \\ \mathbb{E}[y^{(i)} = 1 \mid \mathbf{x}^{(i)}] & \textit{otherwise} \end{cases} \tag{6}$$

Intuitively, the E-step uses $\tilde{y}$ as the label when we have complete label information (recall $t^{(i)} = 1 \implies \tilde{y}^{(i)} = y^{(i)}$); otherwise, we use the posterior estimate $f_\theta(\mathbf{x}^{(i)})$ as a smoothed label. Equivalently, the E-step imputes *soft pseudo-labels* for unlabeled data, *i.e.*, probabilistic estimates $\hat{y}^{(i)} \in [0,1]$. Motivated by approaches that train a pseudo-labeling model on labeled data (Arazo et al., 2020; Rizve et al., 2021), we pre-train f_θ on tested individuals.

M-step. The M-step maximizes the log-likelihood of Eq. 4 given E-step estimates $Q(y^{(i)})$ (Eq. 6). There are two terms to model, which is done via an estimator for $y^{(i)}$ trained using $Q(y^{(i)})$ and an estimator for $\tilde{y}^{(i)}$. The latter is obtained by combining an estimate of $t^{(i)}$ with $Q(y^{(i)})$. Concretely, let $\hat{y}^{(i)} \triangleq f_\theta(\mathbf{x}^{(i)})$, and let h_ϕ be a model of $P(\tilde{Y} \mid Y, T)$. Maximizing the log-likelihood of Eq. 4 reduces to

$$\begin{aligned} \max_\theta \; &\sum_{i=1}^N Q(y^{(i)}) \log \hat{y}^{(i)} + (1 - Q(y^{(i)})) \log(1 - \hat{y}^{(i)}) \\ &+ Q(y^{(i)}) \Big[\tilde{y}^{(i)} \log h_\phi(\hat{y}^{(i)}, \hat{t}^{(i)}) \\ &+ (1 - \tilde{y}^{(i)}) \log(1 - h_\phi(\hat{y}^{(i)}, \hat{t}^{(i)})) \Big]. \end{aligned} \tag{7}$$

This leads to the following result:

Theorem 3.2 (M-step, informal)**.** *Maximizing Eq. 7 also maximizes the evidence-based lower bound of Eq. 3.*

In practice, we set $h_\phi(\hat{y}^{(i)}, \hat{t}^{(i)}) \triangleq \hat{y}^{(i)}\hat{t}^{(i)}$, a smoothed version of the assumption $\tilde{y} = yt$. Defining $\mathcal{L}$ as binary cross-entropy loss, we can rewrite Eq. 7:

$$\min_\theta \sum_{i=1}^N \mathcal{L}(Q(y^{(i)}), \hat{y}^{(i)}) + Q(y^{(i)}) \mathcal{L}(\tilde{y}^{(i)}, \hat{y}^{(i)}\hat{t}^{(i)}). \tag{8}$$

Eq. 8 can be interpreted as a regularized cross-entropy loss with respect to pseudo-label $Q(y^{(i)})$. The first term pushes $\hat{y}^{(i)}$ towards $Q(y^{(i)})$, while the second "encourages" $\hat{y}^{(i)}$ to be consistent with the causal model. To obtain $\hat{t}$, we pre-train and freeze a binary classifier for t, and take the probabilistic estimates as $\hat{t}$.

3.3. DCEM counterbalances disparate censorship

We show that DCEM imposes a form of "causal regularization" that lowers $\hat{y}$ in untested individuals.

DCEM is a form of causal regularization. By analogy to regularized risk minimization, consider an objective

$$\mathcal{L}(\theta) + \lambda r(\theta), \tag{9}$$

for $\lambda \in \mathbb{R}_+$ (regularization strength) and a regularizer $r : \Theta \to \mathbb{R}$, where Θ is the parameter space of θ.

Without loss of generality, setting $\lambda = 1$ and matching terms between Eq. 9 and Eq. 8 yields $r(\theta) = Q(y^{(i)})\mathcal{L}(\tilde{y}^{(i)}, \hat{y}^{(i)}\hat{t}^{(i)})$. While $\hat{t}^{(i)}$ affects the optimization of Eq. 8, it is not a multiplier (*e.g.*, λ in Eq. 9). To interpret the effect of $\hat{t}^{(i)}$, we propose a definition of causal regularization strength based how the optimal $\hat{y}^{(i)}$ changes.[2]

Definition 3.3 (Causal regularization strength, informal)**.** Let $\hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)})$ be the minimizer of Eq. 8. For $\mathcal{L}$ finite & convex on $\hat{y}^{(i)}$ in [0, 1], the causal regularization strength is $R(\cdot) \triangleq |Q(y^{(i)}) - \hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)})|$.

Definition 3.3 quantifies the tradeoff between matching $\hat{y}^{(i)}$ to the E-step estimates and optimizing Eq. 8. While $\hat{y}^{(i)}$ is not an optimization parameter, analyzing the optimal $\hat{y}^{(i)}$ can clarify the inductive bias of the M-step. We proceed by considering how causal regularization impacts untested vs. tested individuals. When $t^{(i)} = 0$, the M-step is

$$\min_{\theta} \sum_{i=1}^{N} \mathcal{L}(Q(y^{(i)}), \hat{y}^{(i)}) - Q(y^{(i)}) \log(1 - \hat{y}^{(i)}\hat{t}^{(i)}). \tag{10}$$

Since $-\log(1 - \hat{y}^{(i)}\hat{t}^{(i)})$ increases in $\hat{y}^{(i)}$, the regularization term "encourages" $\hat{y}^{(i)}$ to decrease when $\hat{t}^{(i)} > 0$. The regularization term is *constant* if $\hat{t}^{(i)} = 0$, such that the M-step would not change the E-step estimate. This matches the intuition that one cannot learn about $y^{(i)}$ from individuals that are very different from labeled individuals (*i.e.*, when the *overlap* assumption in causal inference is violated). The regularization strength depends on $\hat{t}^{(i)}$ as follows:

Theorem 3.4 (informal)**.** *If $t^{(i)} = 0$, causal regularization strength increases in $\hat{t}^{(i)}$.*

The result implies that **causal regularization counterbalances disparate censorship.** Recall that lowering $\hat{y}^{(i)}$ in regions where $\hat{t}^{(i)}$ is higher can mitigate bias. Equivalently, causal regularization must strengthen as $\hat{t}^{(i)}$ increases, which follows from Theorem 3.4.

Causal regularization aligns with full supervision in tested individuals. When $t^{(i)} = 1$, the M-step is

$$\min_{\theta} \sum_{i=1}^{N} \mathcal{L}(y^{(i)}, \hat{y}^{(i)}) + y^{(i)}\mathcal{L}(y^{(i)}, \hat{y}^{(i)}\hat{t}^{(i)}), \tag{11}$$

substituting $y^{(i)}$ for $Q(y^{(i)})$ and $\tilde{y}^{(i)}$. Thus:

[2] "Causal regularization" has been defined in the context of causal discovery (Bahadori et al., 2017; Janzing, 2019). Our usage is unrelated: we use a causal model to regularize an estimator.

Proposition 3.5. *Eq. 11 is minimized when $\hat{y}^{(i)} = y^{(i)}$.*

Proposition 3.5 states that causal regularization does not change the M-step optimum from matching ground truth when $t^{(i)} = 1$ (*i.e.*, regularization strength = 0). Thus, the M-step objective aligns with fully-supervised loss.

Thus, the M-step (Eq. 8) counterbalances disparate censorship by regularizing $\hat{y}^{(i)}$ towards 0 as $\hat{t}^{(i)}$ increases. For $t^{(i)} = 1$, the M-step optimum stays constant, and DCEM should maintain discriminative performance.

3.4. Alternative designs and their limitations

We consider two alternative designs and their limitations: directly using $t^{(i)}$ in DCEM and propensity score adjustment.

Why not use $t^{(i)}$ directly? We substitute $\hat{t}^{(i)} = t^{(i)}$ into Eq. 8 and analyze one summand (without loss of generality):

$$\mathcal{L}(y^{(i)}, \hat{y}^{(i)}) + y^{(i)}\mathcal{L}(y^{(i)}, \hat{y}^{(i)}) \quad t^{(i)} = 1 \tag{12}$$

$$\mathcal{L}(Q(y^{(i)}), \hat{y}^{(i)}) \quad t^{(i)} = 0 \tag{13}$$

Both losses use the E-step estimate $Q(y^{(i)})$ as supervision. When $t^{(i)} = 1$ (Eq. 12), the M-step adds $y^{(i)}\mathcal{L}(y^{(i)}, \hat{y}^{(i)})$, penalizing false negatives 2x as heavily as false positives. This does not affect ranking metrics (*e.g.*, AUC). When $t^{(i)} = 0$ (Eq. 13), the M-step drops causal regularization, and thus cannot counterbalance disparate censorship. Directly using $t^{(i)}$ would only help if counterbalancing disparate censorship is unnecessary for good estimation, *i.e.*, when tested individuals are representative of the population.

Why not propensity score adjustment/related causal approaches? Recall that estimating the effect of T on the observed label yields a consistent estimate of $P(Y \mid X)$ (Eq. 1, Section 2). Indeed, $\hat{t}^{(i)}$ is an estimate of $P(T \mid X, A)$, *i.e.*, a *propensity score*, motivating the usage of causal effect estimators that leverage $\hat{t}^{(i)}$. However, propensity score adjustment (*e.g.*, IPW (Rosenbaum & Rubin, 1983) or doubly-robust variations (Robins et al., 1994; Van Der Laan & Rubin, 2006; Hu et al., 2022)) require an "overlap" assumption $\eta < \hat{t}^{(i)} < 1 - \eta$ for some $\eta = (0, \frac{1}{2})$ and have asymptotic variance of order $O(1/(\eta \cdot (1 - \eta)))$, which is sensitive to extreme $\hat{t}^{(i)}$ (*e.g.*, as in AIPW (Glynn & Quinn, 2010)).

However, in finite-sample settings, "sharp" testing decisions lead to weak overlap. Such extreme $\hat{t}^{(i)}$ may arise in threshold-based decisions (Djulbegovic et al., 2014; Pierson et al., 2018). For example, a patient either exhibits or does not exhibit the requisite symptoms to warrant testing. This is analogous to inducing covariate shift between tested and untested individuals. In other words, "holes" in the training data emerge when using only labeled examples. Thus, systematic testing bias could exacerbate model performance gaps across population subgroups. While low over-

lap still impacts DCEM (since DCEM cannot learn when $\hat{t}^{(i)} = 0$), our method instead leverages an evidence-based lower bound to model y under disparate censorship. We further discuss potential improvements in overlap-robustness of the proposed approach in Appendix B.

4. Experimental Setup

We validate DCEM with **synthetic data** across different data-generation processes on simulated binary classification tasks (Section 4.1) and in a pseudo-synthetic sepsis classification task using real clinical data **(MIMIC-III)** (Johnson et al., 2016), across potential laboratory testing policies (Section 4.2). We then discuss our chosen baselines (Section 4.3) and evaluation metrics (Section 4.4).

4.1. Synthetic Datasets

By definition, y is not fully observed under disparate censorship. Thus, we design a simulation study in order to evaluate various methods with respect to ground truth. The data generation process follows from the assumed causal model of disparate censorship (Fig. 1, top):

$$a^{(i)} \sim Ber(0.5), \mathbf{x}^{(i)} \sim \mathcal{N}(\mu_a \cdot \mathbf{1}_2, 0.03^2 \mathbf{I}_{2\times 2})$$
$$t^{(i)} \sim Ber(\sigma(30 \cdot s_T(\mathbf{x}^{(i)}, a^{(i)})))$$
$$y^{(i)} \sim Ber(\sigma(10 \cdot s_Y(\mathbf{x}^{(i)}) - c_y)),\ \tilde{y}^{(i)} = y^{(i)} t^{(i)}$$

where $\mathbf{I}_{2\times 2}$ is the identity matrix, and $s_T : \mathbf{x}, a \to \mathbb{R}$, $s_Y : \mathbf{x} \to \mathbb{R}$, and $\mu_a \in \mathbb{R}, c_y \in \mathbb{R}$ are simulation parameters. We set $P(A = 0) = 0.5$ and induce confounding between $\mathbf{x}^{(i)}$ and $a^{(i)}$ by setting $u^{(i)} = a^{(i)}$. We draw $\mathbf{x}^{(i)} \in \mathbb{R}^2$ from group-specific Gaussians, and assume Bernoulli-distributed $t^{(i)}$ and $y^{(i)}$ with parameters defined via $s_T : \mathbf{x}, a \to \mathbb{R}$ and $s_Y : \mathbf{x} \to \mathbb{R}$, respectively. Intuitively, s_T (s_Y) is a soft "decision boundary" for T (Y). Inspired by observations that clinician testing is can be represented by simpler functions than observed outcomes (Mullainathan & Obermeyer, 2022), we choose a non-linear s_Y and a linear s_T.

We simulate $N = 20,000$ individuals for training, validation, and testing each (*i.e.*, 60,000 total). We define settings in terms of testing disparity $q_t = \frac{P(T|A=0)}{P(T|A=1)}$, prevalence disparity $q_y = \frac{P(Y|A=0)}{P(Y|A=1)}$, and testing multiple $k = \frac{P(T=1)}{P(Y=1)}$. Intuitively, q_t controls labeling biases, q_y controls differences between groups, and k controls testing rate. We consider $q_t \in \{1/4, 1/3, 1/2, 1, 2, 3, 4\}$, $q_y \in \{1/4, 1/3, 1/2, 1\}$, and $k \in \{1/4, 1/3, 1/2, 1, 2, 3, 4\}$, and set simulation parameters to yield the desired q_t, q_y, k.[3]

Since s_Y is unknown in practice, we replicate the main experiments across various s_Y as a robustness check. The simulation makes simplifying assumptions (*e.g.*, low dimensionality and $P(A = 0) = 0.5$) but allows full control over y and t. Additional simulation details are in Appendix C.1.

[3]We skip settings where q_t, q_y, k yield infeasible testing rates.

4.2. Clinical data: MIMIC-III

Multiple sepsis definitions, such as Sepsis-3 (Singer et al., 2016), are based on laboratory tests (blood culture) such that patients without a test result are *by definition* negative. Thus, sepsis classification is a potential real-world case of disparate censorship. We curate a sepsis classification task using the MIMIC-III Sepsis-3 cohort (Johnson et al., 2016; 2018), an electronic health record dataset.

We aim to distinguish patients who never develop sepsis from those who develop sepsis within 8 hours of an initial 3-hour observation period. If a patient met the Sepsis-3 criteria between 3-11 hours of the first chart measurement, we set $y = 1$, and $y = 0$ if the patient never develops sepsis during their hospital stay. We exclude patients with onset times outside this range and include only White and Black patients to simplify the analysis of bias mitigation. We choose $\mathbf{x} \in \mathbb{R}^{13}$ following an existing sepsis prediction model (Delahanty et al., 2019), and exclude patients where all features are missing. This yields $N = 5,301$ patients, from which we create a 60-20-20 train-validation-test split. This is a simplified version of a real clinical task, since we exclude patients who develop sepsis later during their hospitalization. Nonetheless, it is helpful for probing the strengths and weaknesses of the proposed approach.

To evaluate model performance, we assume that the observed y reflects ground truth, since $\approx 90\%$ of patients were tested (*i.e.*, received a blood culture) in our cohort. To generate label proxies $\tilde{y}$, we simulate multiple potential labeling biases via a clinically-inspired testing function s_T. We specify a linear s_T based on qSOFA, a score for triaging patients at risk of sepsis (Seymour et al., 2016). Inspired by observations that clinicians over-weight representative symptoms in diagnostic test decisions (Mullainathan & Obermeyer, 2022), we create different versions of s_T via different weightings of qSOFA features. We examine $k \in \{1/4, 1/3, 1/2, 1, 2, 3, 4, 5\}$ and $q_t \in \{1/2, 2/3, 1, 3/2, 2\}$. Details of the sepsis cohort are in Appendix C.2.

4.3. Models

As naive baselines, we test a **y-obs model** (training on $\tilde{y}$) and training on **group a only**. We select similarly-motivated or applicable baselines from related settings:

- **Noisy-label learning**: Group peer loss (Wang et al., 2021) (Appendix: Peer loss (Liu & Guo, 2020), truncated ℓ_q loss (Zhang & Sabuncu, 2018) and generalized Jensen-Shannon loss (Englesson & Azizpour, 2021)),
- **Semi-supervised learning**: SELF (Nguyen et al., 2020) (Appendix: DivideMix (Li et al., 2020)),

- **Causal inference**: **tested-only** (training on examples where $t = 1$), and DragonNet (Shi et al., 2019), using the treatment effect of the sensitive attribute on testing to correct disparate censorship (*i.e.*, learn a correction for $P(Y \mid X) - P(\tilde{Y} \mid X, A)$),
- **Positive-unlabeled learning:** Selected-At-Random EM (SAREM) (Bekker et al., 2020).

As an oracle, we compare to training on y ("y-model"). We use neural networks for all approaches. Training details, such as hyperparameters, are in Appendix D.

4.4. Evaluation metrics

We consider bias mitigation and discriminative performance metrics with respect to y, and measure the robustness of both metrics to changes in the data-generation process.

Discriminative performance. We use the area under the receiver operating characteristic curve (AUC), a standard discriminative performance metric.

Mitigating bias. We use the ROC gap (also called ROC fairness (Vogel et al., 2021) or ABROCA (Gardner et al., 2019)), the absolute area between the ROC curves for each group a. The ROC gap is in [0, 1]. Lower values indicate better bias mitigation. Intuitively, the ROC gap is zero when a classifier with some fixed false positive rate in each group obtains equal true positive rates across groups. Under disparate censorship, a zero ROC gap is achievable if a model *perfectly* predicts y from $\mathbf{x}$.

Robustness. We consider the median AUC and ROC gap over all s_Y (synthetic data setting) or s_T (sepsis classification) and the empirical **worst-case** (AUC: min.; ROC gap: max.) and **range**.

5. Experiments & Discussion

Our experiments aim to substantiate our main claims:

- In synthetic data, DCEM mitigates bias, maintains competitive discriminative performance and improves robustness, while achieving better tradeoffs between performance and bias mitigation compared to baselines (Section 5.1).
- On a sepsis classification task, DCEM improves discriminative performance while maintaining good tradeoffs with bias mitigation, and is more robust compared to baselines (Section 5.2).

We also report full results (Appendix E.1) and an ablation study of DCEM (Appendix E.2). We also benchmark causal effect estimators (*i.e.*, as alternatives to the tested-only model) and their overlap robustness compared to DCEM (Appendix E.3). Further sensitivity analyses can be found in Appendix E.4 (smoothed $\hat{t}^{(i)}$) and E.5 (E-step initialization).

5.1. Results on simulated disparate censorship

Fig. 2 shows ROC gaps (left) and AUCs (right) of the baselines most competitive with our approach (DCEM, magenta) at $q_y = 0.5, k = 1, q_t = 2$. In this setting, 25% (*i.e.*, $k \cdot P(Y = 1)$) of individuals are tested, and the base rate of Y in group $a = 0$ is $1/2$ that of group $a = 1$, but group $a = 0$ is twice as likely to be tested. Each point is an ROC gap/AUC value achieved under one decision boundary s_Y. Results for the remaining baselines are in Appendix E.1. The takeaways align with the main results.

DCEM mitigates bias more effectively than baselines. DCEM achieves a median ROC gap of 0.030 (2nd-best, SELF: 0.034), suggesting that it mitigates bias more effectively than baselines (Fig. 2, left). We show similar trends for $q_t \geq 1$, q_y, and $k \in [1/3, 2]$ (Appendix E.1). At low testing rates, all models mitigate bias poorly. At high testing rates, the tested-only model is sufficient.

For $q_t < 1$ (Appendix E.1), DCEM mitigates bias compared to the default approach (y-obs model) but no longer dominates the baselines. We hypothesize that DCEM has similar bias mitigation capabilities as baselines, since there is less bias to mitigate. Recalling that $q_y < 1$, since $q_t < 1$, testing probability, $P(Y \mid X)$ *and* $P(\tilde{Y} \mid X)$ are correlated. Learning to predict $\tilde{Y}$ would preserve ordering in $P(Y \mid X)$, reducing impacts on ranking metrics (*e.g.*, ROC gap).[4]

DCEM is more robust than baselines to changes in the data-generating process. Fig. 2 (left) shows that the maximum ROC gap is lower for DCEM compared to baselines (ours: 0.060 vs. 2nd-best, tested-only: 0.083). We report similar results for the minimum AUC (Fig. 2, right; ours: 0.768 vs. 2nd-best, tested-only: 0.623). DCEM also achieves a tighter ROC gap and AUC range. Fig. 2 also shows that our method has the tightest ROC gap range (left, DCEM: 0.048 vs. tested-only, 2nd-tightest: 0.063) and AUC range (right, DCEM: 0.055 vs. DragonNet: 0.199).

The results suggest that DCEM maintains robust bias mitigation and discriminative performance across different data-generation processes (s_Y). This is expected, as DCEM optimizes likelihood under the disparate censorship data-generation process by design. In contrast, the baselines may experience selection bias or misspecification, since they discard data or assume certain noise structures/variable dependencies that disparate censorship violates.

[4] Such settings are related to boundary-consistent noise; see Proposition 1 of (Menon et al., 2018).

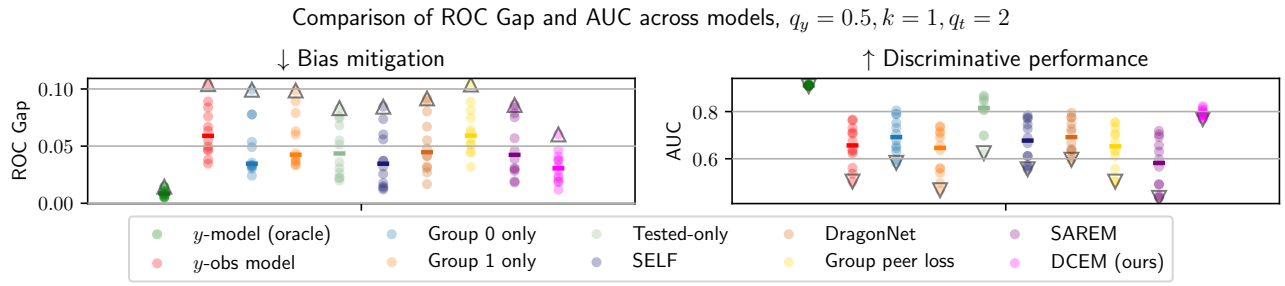

Figure 2. Comparison of ROC gap (left) and AUC (right) of selected models at $q_y = 0.5, k = 1, q_t = 2$. Each point represents a different s_Y. Our method (DCEM, magenta) mitigates bias while maintaining competitive AUC compared to baselines, with a tighter range and improved empirical worst-case for both metrics. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

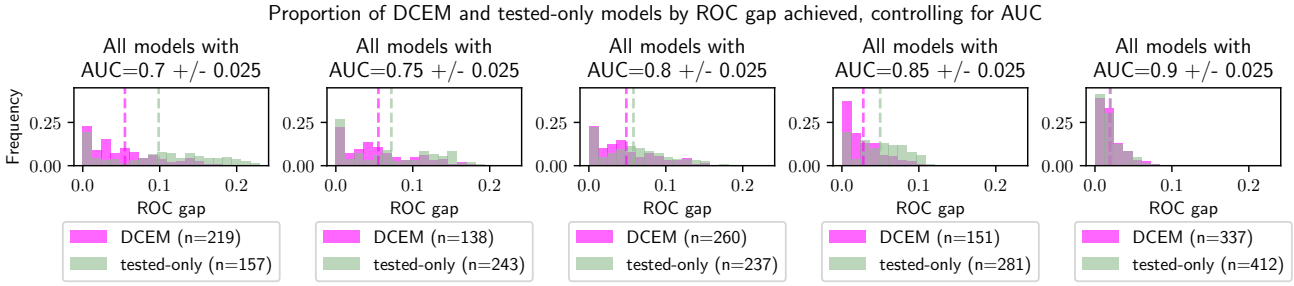

Figure 3. Relative frequencies of ROC gaps for DCEM vs. tested-only models at similar AUC (increasing to the right), pooling models across all k, q_y, q_t tested. Dashed lines = mean ROC gap by model. DCEM improves bias mitigation among models with similar AUC.

DCEM maintains competitive discriminative performance. Fig. 2 (right) shows that DCEM outperforms all baselines except for the tested-only model, which our approach lags by 0.028 AUC (DCEM: 0.787 vs. tested-only: 0.815). Other causal estimators achieve similar discriminative performance to the tested-only approach (Appendix E.3). However, our method improves on the "default" y-obs model, increasing the median AUC by 0.130 (y-obs: 0.657). SELF, which has a similar median ROC gap to DCEM, underperforms DCEM by 0.110 AUC (SELF: 0.677 vs. DCEM: 0.787). Other baselines also underperform. This is expected, since some methods ignore label bias: training on $\tilde{Y}$ alone is misspecified for $E(Y \mid X)$, since it incorrectly assumes that if $T = 0$, then $Y = 0$. The same argument applies to Group 0/1 only approaches.

Some baselines account for label noise/bias, but are misspecified under disparate censorship since they make different independence assumptions. Group peer loss assumes $T \perp\!\!\!\perp X \mid (Y, A)$, and SELF assumes $T \perp\!\!\!\perp X \mid Y$, ignoring the dependence of biased selective labeling on X. DragonNet accounts for X by adding $P(T \mid X, A = 1) - P(T \mid X, A = 0)$ to the default model's estimates (i.e., $P(\tilde{Y} \mid X)$) as a "correction factor." However, the correction factor may be biased for true negatives: the oracle is zero, because $P(\tilde{Y} \mid X) = P(Y \mid X)$, but $P(T \mid X, A = 1) - P(T \mid X, A = 0) \neq 0$ in general under systematic labeling bias.

SAR-EM, which is most similar to the proposed approach, models missingness at random (i.e., $T \not\perp\!\!\!\perp X, Y$), but discards reliable negatives. In contrast, the proposed approach incorporates reliable negatives in alignment with our assumptions about labeling bias, allowing it to counterbalance biased selective labeling. Trends are similar for other q_t, q_y and $k \in [1/2, 2]$. Since the tested-only model is a strong baseline, we now compare it directly to DCEM.

DCEM achieves better tradeoffs between discriminative performance and bias mitigation. Among models with similar AUC where AUC < 0.875, DCEM reduces ROC gaps compared to the tested-only model (Fig. 3). For example, for AUC $\in (0.825, 0.875)$ (Fig. 3, 2nd from right), DCEM improves the average ROC gap by 0.022 (0.028 vs. 0.050), with similar trends at lower AUCs. Among the best-performing models (AUC > 0.875; Fig. 3, 1st from right), both methods have similar ROC gaps.

The results suggest that DCEM is not trading discriminative performance for bias mitigation. At a given AUC, DCEM more often yields models with a lower ROC gap than the tested-only model. Since the tested-only approach does not account for label bias, it can achieve relatively high AUC without mitigating bias. In contrast, DCEM explicitly counteracts disparate censorship. A similar comparison to SELF shows that, at low ROC gaps, DCEM likewise finds higher-AUC solutions than SELF (Appendix E.6).

5.2. Results on sepsis classification in MIMIC-III

DCEM has better discriminative performance than baselines. Fig. 4 compares the ROC gap and AUC of DCEM

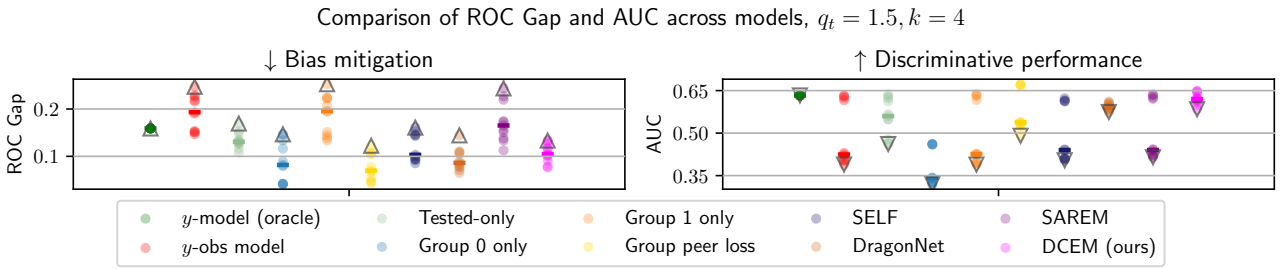

Figure 4. ROC gaps (left) and AUC (right) of baselines and DCEM on sepsis classification task at $q_t = 1.5, k = 4$. Each dot represents a different s_T. Our method (DCEM, magenta) maintains competitive or better bias mitigation and discriminative performance compared to baselines. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

to selected baselines at testing disparity $q_t = 1.5$, and testing rate multiplier $k = 4$. Each dot corresponds to one variation of s_T (laboratory testing policy). Our method has the highest median AUC among baselines (ours: 0.620 vs. DragonNet: 0.593), nearing the oracle (y-model, 0.633). Note that DCEM has better discriminative performance than the tested-only approach, suggesting that extrapolation from tested to untested individuals is more difficult on the sepsis classification task than the fully synthetic tasks.

DCEM achieves good tradeoffs with bias mitigation. DCEM achieves a smaller median ROC gap compared to five of eight baselines tested. Group peer loss, DragonNet and the Group 0 approach achieve lower median ROC gaps of 0.070, 0.088 and 0.082, respectively (DCEM: 0.105). However, the Group 0 approach catastrophically fails (median AUC: 0.342). Models may perform arbitrarily poorly under disparate censorship if labeling biases sufficiently "conceal" the true decision boundary. Group peer loss (among many other baselines) also exhibits a much wider AUC range than the proposed approach (Group peer loss: 0.182 vs. DCEM: 0.065), suggesting that its discriminative utility may be limited. DragonNet appears competitive (0.027 AUC lower than DCEM), but would only perform well when the effect of race on testing is close to $|P(Y \mid X) - P(\tilde{Y} \mid X, A)|$, which is violated if labeling biases (large effect of race on testing) are present in negative patients ($P(Y \mid X) \approx P(\tilde{Y} \mid X, A)$).

Many approaches, including DCEM, obtain a lower ROC gap than training on y. Although the oracle obtains the highest median AUC, optimizing discriminative performance on y is not always guaranteed to mitigate bias. DCEM uses labeling probabilities to mitigate bias via causal regularization, while DragonNet directly uses an estimate of the labeling bias as a correction factor. Thus, the results validate that the labeling bias can provide signal for bias mitigation.

DCEM is more robust than most baselines to changes in s_T. DCEM maintains robust bias mitigation capabilities across s_T; *i.e.*, differences in how labelers weigh features in their decisions. Fig. 4 shows that DCEM attains a maximum ROC gap of 0.133 (left; DragonNet: 0.144), and a minimum AUC of 0.584 (right; DragonNet: 0.574). Fig. 4 also shows that DCEM achieves the tightest ROC gap range (left; DCEM: 0.094 vs. 2nd-best: 0.102) and 2nd-tightest AUC range (right; DCEM: 0.065 vs. DragonNet: 0.018). Many baselines also exhibit a *bimodal* empirical AUC distribution and only perform well under specific labeling behaviors. We examine the sensitivity of baselines to s_T by plotting AUC and ROC gaps against coefficients of s_T (Appendix E.7).

While DragonNet is competitive on this dataset, its robustness and performance capabilities may not generalize (*e.g.*, simulation results, Fig. 2). DCEM is the only approach tested that achieved competitive discriminative performance and bias mitigation on both datasets. Trends in performance and robustness are similar for other k, q_t (Appendix E.1).

Overall takeaways. In a simulation study of disparate censorship, DCEM mitigates bias while achieving similar or better discriminative performance compared to baselines. The proposed approach is empirically more robust than baselines to changes in the data-generating process. On a sepsis classification task, DCEM mitigates bias while improving or maintaining discriminative performance compared to baselines across different labeling behaviors. Thus, DCEM can potentially mitigate bias with less impact on discriminative performance than existing methods.

6. Related Work

Selective labeling/disparate censorship. Disparate censorship is a variation of selective labeling (Lakkaraju et al., 2017; Kleinberg et al., 2018) and outcome measurement error (Guerdan et al., 2023b). Selective labeling problems have been studied in clinical settings (Farahani et al., 2020; Shanmugam et al., 2024; Chang et al., 2022; Mullainathan & Obermeyer, 2022; Balachandar et al., 2024), social/public policy (Saxena et al., 2020; Kontokosta & Hong, 2021; Laufer et al., 2022; Liu & Garg, 2022; Kiani et al., 2023), and finance (Björkegren & Grissen, 2020; Henderson et al., 2023), among other domains. For an extended literature review of selective labeling problems, see Appendix A.

Past work has trained ML models under disparate censorship, directly encoding untested individuals as negative (Henry et al., 2015; Jehi et al., 2020; McDonald et al., 2021; Adams et al., 2022; Kamran et al., 2022). Previous approaches for learning under selective labels leverage heterogeneity in human decisions to recover outcome estimates (Lakkaraju et al., 2017; Kleinberg et al., 2018; Chen et al., 2023), or use domain-specific adjustments (Gholami et al., 2018; Balachandar et al., 2024). We propose DCEM, a complementary approach for mitigating bias under disparate censorship without such restrictions.

Semi-supervised learning. Semi-supervised approaches do not assume labels for untested individuals. However, many causally-motivated methods diverge from the causal model of disparate censorship (Madras et al., 2019; Yao et al., 2021; Garg et al., 2023; Guerdan et al., 2023a; Gong et al., 2021; Kato et al., 2023; Sportisse et al., 2023) via different independence/causal relationships between variables. Filtering methods (Han et al., 2018; Li et al., 2020; Nguyen et al., 2020; Chen et al., 2020; Zhang et al., 2021; Zhao et al., 2022) assume specific model behavior on noisy examples (*e.g.,*, noise is learned late in training (Arpit et al., 2017)) or labeling bias (randomness/class-dependence), which disparate censorship violates. We also highlight historical expectation-maximization approaches for learning with missing data (Ghahramani & Jordan, 1993; Ghahramani et al., 1996; Ambroise & Govaert, 2000), which place parametric assumptions on the data-generation process. We use neural networks to target the estimands of interest to circumvent parametric assumptions.

Other alternatives include positive-unlabeled learning approaches that assume labeling depends on covariates (*e.g.*, missing not at random) (Bekker et al., 2020; Furmańczyk et al., 2022; Gerych et al., 2022; Wang et al., 2024). However, these methods do not leverage correctly-labeled negatives, and naively-incorporating negative examples without causal assumptions may potentially harm model performance or bias mitigation. Other methods for noisy-label learning make assumptions incompatible with our setting, *e.g.* uniform noise within subgroups (Wang et al., 2021), almost-surely clean & noisy examples (Liu & Tao, 2015; Patrini et al., 2017; Tjandra & Wiens, 2023), different variable independence/directionality relationships (Wu et al., 2022), that noisy (*i.e.*, out of distribution) examples are rare (Wald & Saria, 2023), or other noise constraints (Li et al., 2021; Zhu et al., 2021). Our approach complements existing work by jointly modeling selective and biased labeling via causal assumptions tailored to a biased decision-making pipeline.

7. Conclusion

When biased human decisions affect observations of ground truth, applying standard supervised learning techniques to data exhibiting disparate censorship can amplify the harm of ML models to marginalized groups. We propose Disparate Censorship Expectation-Maximization (DCEM), a novel approach to classification, to mitigate such harm. In a simulation study and a sepsis classification task, DCEM mitigates bias and maintains competitive discriminative performance compared to baselines. Limitations of DCEM include potential slow convergence, since EM is iterative. Model evaluation under disparate censorship is also inherently difficult due to the difficulty of obtaining ground truth, motivating future work in dataset curation. Furthermore, DCEM does not learn a full generative model with all variables. While such a model could target a wider range of estimands, it would also increase the number of terms that need to be modeled. Ultimately, DCEM is a step towards mitigating the disproportionate impacts of disparate censorship. Our work aims to raise awareness of disparate censorship and motivate the study of bias mitigation methods.

Impact Statement

This paper addresses disparate censorship, a realistic source of label bias in ML, and proposes a method that mitigates its harms. Since the goal of the paper is aligned with reducing inequity in decision-making, practical use cases of DCEM are inherently high-stakes settings. Thus, we believe that the ethical usage of DCEM (or any bias mitigation approach) in the real-world requires prospective model evaluation in the context of use (*e.g.*, shadowing human decision-makers) to assess unforeseen negative impacts. Our work provides a general choice of bias mitigation (area between ROC curves) and discriminative performance metrics (AUC), which are motivated by clinical tasks where equitably ranking individuals in terms of resource needs is important. Practitioners should ensure their evaluation metrics align with domain-specific criteria for bias mitigation and performance.

Acknowledgements

We thank (in alphabetical order) Donna Tjandra, Divya Shanmugan, Fahad Kamran, Jung Min Lee, Maggie Makar, Meera Krishnamoorthy, Michael Ito, Sarah Jabbour, Shengpu Tang, Stephanie Shepard, and Winston Chen for helpful conversations and proofreading, and the anonymous reviewers for their constructive feedback. T.C. and J.W. are supported by the U.S. National Heart, Lung, and Blood Insitute of the National Institutes of Health (Grant No. 5R01HL158626-03). The views and conclusions in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied of the U.S. National Institutes of Health.

References

Adams, R., Henry, K. E., Sridharan, A., Soleimani, H., Zhan, A., Rawat, N., Johnson, L., Hager, D. N., Cosgrove, S. E., Markowski, A., et al. Prospective, multi-site study of patient outcomes after implementation of the trews machine learning-based early warning system for sepsis. *Nature Medicine*, pp. 1–6, 2022.

Ambroise, C. and Govaert, G. EM algorithm for partially known labels. In *Data Analysis, Classification, and Related Methods*, pp. 161–166. Springer, 2000.

Arazo, E., Ortego, D., Albert, P., O'Connor, N. E., and McGuinness, K. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2020.

Arpit, D., Jastrzebski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., and Lacoste-Julien, S. A closer look at memorization in deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 233–242, 2017.

Bahadori, M. T., Chalupka, K., Choi, E., Chen, R., Stewart, W. F., and Sun, J. Causal regularization. *arXiv preprint arXiv:1702.02604*, 2017.

Balachandar, S., Garg, N., and Pierson, E. Domain constraints improve risk prediction when outcome data is missing. In *12th International Conference on Learning Representations*, 2024.

Beery, T. A. Gender bias in the diagnosis and treatment of coronary artery disease. *Heart & Lung*, 24(6):427–435, 1995.

Bekker, J., Robberechts, P., and Davis, J. Beyond the selected completely at random assumption for learning from positive and unlabeled data. In *Machine Learning and Knowledge Discovery in Databases*, pp. 71–85, Cham, 2020.

Bergman, P., Kopko, E., and Rodriguez, J. E. A seven-college experiment using algorithms to track students: Impacts and implications for equity and fairness. Technical report, National Bureau of Economic Research, 2021.

Binns, R., Veale, M., Van Kleek, M., and Shadbolt, N. Like trainer, like bot? Inheritance of bias in algorithmic content moderation. In *Social Informatics: 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings, Part II 9*, pp. 405–415, 2017.

Björkegren, D. and Grissen, D. Behavior revealed in mobile phone usage predicts credit repayment. *The World Bank Economic Review*, 34(3):618–634, 2020.

Blum, A. and Stangl, K. Recovering from biased data: Can fairness constraints improve accuracy? In *1st Symposium on Foundations of Responsible Computing*, 2020.

Chang, T., Sjoding, M. W., and Wiens, J. Disparate censorship & undertesting: A source of label bias in clinical machine learning. In *Proceedings of the 7th Machine Learning for Healthcare Conference*, volume 182 of *Proceedings of Machine Learning Research*, pp. 343–390, Aug 2022.

Chen, J., Li, Z., and Mao, X. Learning under selective labels with heterogeneous decision-makers: An instrumental variable approach. *arXiv preprint arXiv:2306.07566*, 2023.

Chen, P., Ye, J., Chen, G., Zhao, J., and Heng, P.-A. Beyond class-conditional assumption: A primary attempt to combat instance-dependent label noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 2020.

Delahanty, R. J., Alvarez, J., Flynn, L. M., Sherwin, R. L., and Jones, S. S. Development and evaluation of a machine learning model for the early identification of patients at risk for sepsis. *Annals of Emergency Medicine*, 73(4): 334–344, 2019.

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (methodological)*, 39(1):1–22, 1977.

Djulbegovic, B., Elqayam, S., Reljic, T., Hozo, I., Miladinovic, B., Tsalatsanis, A., Kumar, A., Beckstead, J., Taylor, S., and Cannon-Bowers, J. How do physicians decide to treat: an empirical evaluation of the threshold model. *BMC Medical Informatics and Decision Making*, 14:1–10, 2014.

Englesson, E. and Azizpour, H. Generalized jensen-shannon divergence loss for learning with noisy labels. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 30284–30297, 2021.

Farahani, N. Z., Sundaram, D. S. B., Enayati, M., Arunachalam, S. P., Pasupathy, K., and Arruda-Olson, A. M. Explanatory analysis of a machine learning model to identify hypertrophic cardiomyopathy patients from EHR using diagnostic codes. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1932–1937, 2020.

Furmańczyk, K., Mielniczuk, J., Rejchel, W., and Teisseyre, P. Joint estimation of posterior probability and propensity score function for positive and unlabelled data. *arXiv preprint arXiv:2209.07787*, 2022.

Gardner, J., Brooks, C., and Baker, R. Evaluating the fairness of predictive student models through slicing analysis. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, pp. 225–234, 2019.

Garg, A., Nguyen, C., Felix, R., Do, T.-T., and Carneiro, G. Instance-dependent noisy label learning via graphical modelling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2288–2298, 2023.

Gerych, W., Hartvigsen, T., Buquicchio, L., Agu, E., and Rundensteiner, E. Recovering the propensity score from biased positive unlabeled data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 6694–6702, 2022.

Ghahramani, Z. and Jordan, M. Supervised learning from incomplete data via an EM approach. *Advances in Neural Information Processing Systems*, 6, 1993.

Ghahramani, Z., Hinton, G. E., et al. The EM algorithm for mixtures of factor analyzers. Technical report, Technical Report CRG-TR-96-1, University of Toronto, 1996.

Gholami, S., Mc Carthy, S., Dilkina, B., Plumptre, A., Tambe, M., Driciru, M., Wanyama, F., Rwetsiba, A., Nsubaga, M., Mabonga, J., et al. Adversary models account for imperfect crime data: Forecasting and planning against real-world poachers. In *International Conference on Autonomous Agents and Multiagent Systems*, 2018.

Glynn, A. N. and Quinn, K. M. An introduction to the augmented inverse propensity weighted estimator. *Political Analysis*, 18(1):36–56, 2010.

Gong, C., Wang, Q., Liu, T., Han, B., You, J. J., Yang, J., and Tao, D. Instance-dependent positive and unlabeled learning with labeling bias estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:4163–4177, 2021.

Guerdan, L., Coston, A., Holstein, K., and Wu, Z. S. Counterfactual prediction under outcome measurement error. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1584–1598, 2023a.

Guerdan, L., Coston, A., Wu, Z. S., and Holstein, K. Ground(less) truth: A causal framework for proxy labels in human-algorithm decision-making. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 688–704, 2023b.

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I. W., and Sugiyama, M. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 8536–8546, 2018.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.

Hartvigsen, T., Sen, C., Brownell, S., Teeple, E., Kong, X., and Rundensteiner, E. A. Early Prediction of MRSA Infections using Electronic Health Records. In *HEALTHINF*, pp. 156–167, 2018.

Henderson, P., Chugg, B., Anderson, B., Altenburger, K., Turk, A., Guyton, J., Goldin, J., and Ho, D. E. Integrating reward maximization and population estimation: Sequential decision-making for internal revenue service audit selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 5087–5095, 2023.

Henry, K. E., Hager, D. N., Pronovost, P. J., and Saria, S. A targeted real-time early warning score (trewscore) for septic shock. *Science Translational Medicine*, 7(299): 299ra122–299ra122, 2015.

Hu, X., Niu, Y., Miao, C., Hua, X.-S., and Zhang, H. On non-random missing labels in semi-supervised learning. In *10th International Conference on Learning Representations*, 2022.

Imbens, G. W. and Rubin, D. B. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Janzing, D. Causal regularization. *Advances in Neural Information Processing Systems*, 32, 2019.

Jehi, L., Ji, X., Milinovich, A., Erzurum, S., Rubin, B. P., Gordon, S., Young, J. B., and Kattan, M. W. Individualizing risk prediction for positive coronavirus disease 2019 testing: results from 11,672 patients. *Chest*, 158(4): 1364–1375, 2020.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1): 1–9, 2016.

Johnson, A. E., Aboab, J., Raffa, J. D., Pollard, T. J., Deliberato, R. O., Celi, L. A., and Stone, D. J. A comparative analysis of sepsis identification methods in an electronic database. *Critical Care Medicine*, 46(4):494–499, 2018.

Kamran, F., Tang, S., Ötleş, E., McEvoy, D. S., Saleh, S. N., Gong, J., Li, B. Y., Dutta, S., Liu, X., Medford, R. J., Valley, T. S., West, L. R., Singh, K., Blumberg, S., Donnelly, J. P., Shenoy, E. S., Ayanian, J. Z., Nallamothu, B. K., Sjoding, M. W., and Wiens, J. Early identification of patients admitted to hospital for covid-19 at risk of clinical deterioration: model development and multisite external validation study. *The BMJ*, 376, 2022.

Kato, M., Wu, S., Kureishi, K., and Yasui, S. Automatic debiased learning from positive, unlabeled, and exposure data. *arXiv preprint arXiv:2303.04797*, 2023.

Kennedy, E. H. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049, 2023.

Kiani, S., Barton, J., Sushinsky, J., Heimbach, L., and Luo, B. Counterfactual prediction under selective confounding. *arXiv preprint arXiv:2310.14064*, 2023.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*, 2015.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1): 237–293, 2018.

Kontokosta, C. E. and Hong, B. Bias in smart city governance: How socio-spatial disparities in 311 complaint behavior impact the fairness of data-driven decisions. *Sustainable Cities and Society*, 64:102503, 2021.

Laine, R., Hyttinen, A., and Mathioudakis, M. Evaluating decision makers over selectively labelled data: A causal modelling approach. In *Discovery Science: 23rd International Conference, DS 2020, Thessaloniki, Greece, October 19–21, 2020, Proceedings 23*, pp. 3–18, 2020.

Lakkaraju, H., Kleinberg, J., Leskovec, J., Ludwig, J., and Mullainathan, S. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 275–284, 2017.

Laufer, B., Pierson, E., and Garg, N. End-to-end auditing of decision pipelines. In *ICML Workshop on Responsible Decision-Making in Dynamic Environments.*, pp. 1–7, 2022.

Lee, D.-H. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3(2), pp. 896, 2013.

Li, J., Socher, R., and Hoi, S. C. H. Dividemix: Learning with noisy labels as semi-supervised learning. In *8th International Conference on Learning Representations, ICLR*, 2020.

Li, X., Liu, T., Han, B., Niu, G., and Sugiyama, M. Provably end-to-end label-noise learning without anchor points. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 6403–6413, 2021.

Liu, T. and Tao, D. Classification with noisy labels by importance reweighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):447–461, 2015.

Liu, Y. and Guo, H. Peer loss functions: Learning from noisy labels without knowing noise rates. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Liu, Z. and Garg, N. Equity in resident crowdsourcing: Measuring under-reporting without ground truth data. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pp. 1016–1017, 2022.

Madras, D., Creager, E., Pitassi, T., and Zemel, R. Fairness through causal awareness. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019.

McDonald, S. A., Medford, R. J., Basit, M. A., Diercks, D. B., and Courtney, D. M. Derivation with internal validation of a multivariable predictive model to predict covid-19 test results in emergency department patients. *Academic Emergency Medicine*, 28(2):206–214, 2021.

Menon, A. K., Van Rooyen, B., and Natarajan, N. Learning from binary labels with instance-dependent noise. *Machine Learning*, 107(8):1561–1595, 2018.

Mulherin, S. A. and Miller, W. C. Spectrum bias or spectrum effect? subgroup variation in diagnostic test evaluation. *Annals of Internal Medicine*, 137(7):598–602, 2002.

Mullainathan, S. and Obermeyer, Z. Diagnosing physician error: A machine learning approach to low-value health care. *The Quarterly Journal of Economics*, 137(2):679–727, 2022.

Nguyen, D. T., Mummadi, C. K., Ngo, T., Nguyen, T. H. P., Beggel, L., and Brox, T. SELF: Learning to Filter Noisy Labels with Self-Ensembling. In *8th International Conference on Learning Representations*, 2020.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library, 2019.

Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., and Qu, L. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1944–1952, 2017.

Pearl, J. *Causality*. Cambridge university press, 2009.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Peng, A., Nushi, B., Kıcıman, E., Inkpen, K., Suri, S., and Kamar, E. What you see is what you get? the impact of representation criteria on human bias in hiring. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pp. 125–134, 2019.

Pierson, E., Corbett-Davies, S., and Goel, S. Fast threshold tests for detecting discrimination. In *International Conference on Artificial Intelligence and Statistics*, pp. 96–105, 2018.

Pierson, E., Simoiu, C., Overgoor, J., Corbett-Davies, S., Jenson, D., Shoemaker, A., Ramachandran, V., Barghouty, P., Phillips, C., Shroff, R., et al. A large-scale analysis of racial disparities in police stops across the united states. *Nature Human Behaviour*, 4(7):736–745, 2020.

Rambachan, A. and Roth, J. Bias in, bias out? Evaluating the folk wisdom. In *1st Symposium on Foundations of Responsible Computing*, 2020.

Rhee, C. and Klompas, M. Sepsis trends: increasing incidence and decreasing mortality, or changing denominator? *Journal of Thoracic Disease*, 12(Suppl 1):S89, 2020.

Rizve, M. N., Duarte, K., Rawat, Y. S., and Shah, M. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *9th International Conference on Learning Representations*, 2021.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, pp. 846–866, 1994.

Rockafellar, R. T. *Convex analysis*. Princeton University Press, Princeton, N. J., 1970.

Rosenbaum, P. R. and Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Saxena, D., Badillo-Urquiola, K., Wisniewski, P. J., and Guha, S. A human-centered review of algorithms used within the u.s. child welfare system. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–15, 2020.

Schulman, K. A., Berlin, J. A., Harless, W., Kerner, J. F., Sistrunk, S., Gersh, B. J., Dube, R., Taleghani, C. K., Burke, J. E., Williams, S., et al. The effect of race and sex on physicians' recommendations for cardiac catheterization. *New England Journal of Medicine*, 340(8):618–626, 1999.

Seymour, C. W., Liu, V. X., Iwashyna, T. J., Brunkhorst, F. M., Rea, T. D., Scherag, A., Rubenfeld, G., Kahn, J. M., Shankar-Hari, M., Singer, M., et al. Assessment of clinical criteria for sepsis: for the third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA*, 315(8):762–774, 2016.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

Shanmugam, D., Hou, K., and Pierson, E. Quantifying disparities in intimate partner violence: a machine learning method to correct for underreporting. *npj Women's Health*, 2(1), 2024.

Shi, C., Blei, D., and Veitch, V. Adapting neural networks for the estimation of treatment effects. *Advances in Neural Information Processing Systems*, 32, 2019.

Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, R., Bernard, G. R., Chiche, J.-D., Coopersmith, C. M., et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). *Jama*, 315(8):801–810, 2016.

Sportisse, A., Schmutz, H., Humbert, O., Bouveyron, C., and Mattei, P.-A. Are labels informative in semi-supervised learning? estimating and leveraging the missing-data mechanism. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Sühr, T., Hilgard, S., and Lakkaraju, H. Does fair ranking improve minority outcomes? Understanding the interplay of human and algorithmic biases in online hiring. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 989–999, 2021.

Teeple, E., Hartvigsen, T., Sen, C., Claypool, K. T., and Rundensteiner, E. A. Clinical performance evaluation of a machine learning system for predicting hospital-acquired clostridium difficile infection. In *HEALTHINF*, pp. 656–663, 2020.

The pandas development team. pandas-dev/pandas: Pandas, 2020.

Tjandra, D. and Wiens, J. Leveraging an alignment set in tackling instance-dependent label noise. In *Proceedings of the Conference on Health, Inference, and Learning*, 2023.

Van Der Laan, M. J. and Rubin, D. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

Vogel, R., Bellet, A., and Clémençon, S. Learning fair scoring functions: Bipartite ranking under ROC-based fairness constraints. In *International Conference on Artificial Intelligence and Statistics*, pp. 784–792, 2021.

Wager, S. Stats 361: Causal inference. Technical report, Stanford University, 2020.

Wald, Y. and Saria, S. Birds of an odd feather: Guaranteed out-of-distribution (OOD) novel category detection. In *Uncertainty in Artificial Intelligence*, pp. 2179–2191, 2023.

Wang, J., Liu, Y., and Levy, C. Fair classification with group-dependent label noise. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 526–536, 2021.

Wang, X., Chen, H., Guo, T., and Wang, Y. Pue: Biased positive-unlabeled learning enhancement by causal inference. *Advances in Neural Information Processing Systems*, 36, 2024.

Wu, S., Gong, M., Han, B., Liu, Y., and Liu, T. Fair classification with instance-dependent label noise. In *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pp. 927–943, Apr 2022.

Yao, Y., Liu, T., Gong, M., Han, B., Niu, G., and Zhang, K. Instance-dependent label-noise learning under a structural causal model. In *Advances in Neural Information Processing Systems*, volume 34, pp. 4409–4420, 2021.

Zhang, Y., Zheng, S., Wu, P., Goswami, M., and Chen, C. Learning with feature-dependent label noise: A progressive approach. In *9th International Conference on Learning Representations*, 2021.

Zhang, Z. and Sabuncu, M. R. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 8792–8802, 2018.

Zhao, G., Li, G., Qin, Y., Liu, F., and Yu, Y. Centrality and consistency: Two-stage clean samples identification for learning with instance-dependent noisy labels. In Avidan, S., Brostow, G., Cissé, M., Farinella, G. M., and Hassner, T. (eds.), *Computer Vision – ECCV 2022*, pp. 21–37, 2022.

Zhu, X. and Ghahramani, Z. Learning from labeled and unlabeled data with label propagation. Technical report, Center for Automated Learning and Discovery, Carnegie Mellon University, 2002.

Zhu, Z., Song, Y., and Liu, Y. Clusterability as an alternative to anchor points when learning with noisy labels. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 12912–12923, 2021.

A. Selective labeling in the literature

We enumerate domains in which our literature review found instances of selective label problems in the ML methods and applications literature:

- **Healthcare**: Laboratory/diagnostic testing (Chang et al., 2022; Mullainathan & Obermeyer, 2022) and diagnosis (Farahani et al., 2020; Shanmugam et al., 2024; Balachandar et al., 2024)
- **Social & public policy:** Child welfare assessment (Saxena et al., 2020; Kiani et al., 2023), urban planning/policy (Kontokosta & Hong, 2021; Laufer et al., 2022; Liu & Tao, 2015), hiring pipelines (Peng et al., 2019; Sühr et al., 2021), student placement (Bergman et al., 2021), and bias in policing (Rambachan & Roth, 2020; Pierson et al., 2020)
- **Finance:** Credit repayment (Björkegren & Grissen, 2020) and financial auditing (Henderson et al., 2023)
- **Others/miscellaneous:** Wildlife conservation (Gholami et al., 2018), social media content moderation (Binns et al., 2017)

We note that this is not an exhaustive list of all papers in the selective labeling literature or related problem settings. However, this list illustrates the broad applicability and relevance of our problem setting.

B. Omitted Proofs

For convenience, we restate all theorems and propositions here.

B.1. Theorem 3.1

Theorem (E-step derivation)**.** *The posterior conditional mean of $y^{(i)} = 1$ given the observed data, $Q(y^{(i)}) \triangleq \mathbb{E}[y^{(i)} = 1 \mid t^{(i)}, \tilde{y}^{(i)}, a^{(i)}, \mathbf{x}^{(i)}]$, is equal to*

$$Q(y^{(i)}) = \begin{cases} \tilde{y}^{(i)} & t^{(i)} = 1 \\ \mathbb{E}[y^{(i)} = 1 \mid \mathbf{x}^{(i)}] & \textit{otherwise} \end{cases}. \tag{14}$$

Proof. We drop superscripts $(\cdot)^{(i)}$ in the proof for clarity. Denote $Q(y)$ as posterior distribution of y given the observed data, $\mathbb{E}[y = 1 \mid t, \tilde{y}, a, \mathbf{x}]$ (*i.e.*, the E-step estimate). First, we can write

$$Q(y) \triangleq \mathbb{E}[y = 1 \mid t, \tilde{y}, a, \mathbf{x}] = \mathbb{E}[y = 1 \mid \tilde{y}, x, t] = P(y = 1 \mid \tilde{y}, x, t) \tag{15}$$

for simplicity, where we use the fact that $Y \perp\!\!\!\perp A \mid (T, X)$ and the fact that $\mathbb{E}[y = 1 \mid \tilde{y}, x, t]$ is binary. Proceeding, we can write:

$$\begin{aligned}
&= t \cdot P(y = 1 \mid \tilde{y}, x, t = 1) + (1 - t) \cdot P(y = 1 \mid \tilde{y}, x, t = 0) && (16)\\
&= t\tilde{y} + (1 - t)[\tilde{y}P(y = 1 \mid \tilde{y} = 1, x, t = 0) + (1 - \tilde{y})P(y = 1 \mid \tilde{y} = 0, x, t = 0)] && (17)\\
&= t\tilde{y} + (1 - t)(1 - \tilde{y})P(y = 1 \mid \tilde{y} = 0, x, t = 0) && (18)\\
&= t\tilde{y} + (1 - t)(1 - \tilde{y}) \cdot \frac{P(\tilde{y} = 0 \mid y = 1, x, t = 0)P(y = 1 \mid x, t = 0)}{P(\tilde{y} = 0 \mid x, t = 0)} && (19)\\
&= t\tilde{y} + (1 - t)(1 - \tilde{y})P(y = 1 \mid x). && (20)
\end{aligned}$$

The second equality follows since $t = 1 \implies \tilde{y} = y$. The third equality holds since $P(y = 1 \mid \tilde{y} = 1, x, t = 0) = 0$ by construction, since $\tilde{y} = yt$ under disparate censorship. The final step follows from three facts: (1) $P(\tilde{y} = 0 \mid y = 1, x, t = 0) = 1$ for all x, (2) $P(\tilde{y} = 0 \mid x, t = 0) = 1$ for all x, and (3) $Y \perp\!\!\!\perp T \mid X$. This is more succinctly rewritten as E-step is:

$$Q(y) = \begin{cases} \tilde{y} & t = 1 \\ \mathbb{E}[y = 1 \mid x] & \text{otherwise (i.e., } \tilde{y} = 0 \wedge t = 0) \end{cases}, \tag{21}$$

which is what we wanted to show. □

Remark B.1. Since $y^{(i)}$ is binary by assumption, this result fully determines the posterior distribution since $\mathbb{E}[y = 1 \mid \tilde{y}, x, t] = 1 - \mathbb{E}[y = 0 \mid \tilde{y}, x, t]$.

B.2. Theorem 3.2

Theorem (M-step derivation)**.** *Let $P(U, A, X, T, \tilde{Y}; \theta)$ be a model for the joint data distribution parameterized by some arbitrary $\theta \in \Theta$ in some parameter space Θ, which factorizes according to the disparate censorship DAG =(Fig. 1). Let $Q(y) \triangleq \mathbb{E}[y = 1 \mid t, \tilde{y}, a, \mathbf{x}]$ be the posterior expectation that $y = 1$ given the observed data. Then (replacing random variables with their realized counterparts), we have*

$$\begin{aligned}\max_\theta \frac{1}{N}\sum_{i=1}^{N} \log P(u^{(i)}, a^{(i)}, \mathbf{x}^{(i)}, t^{(i)}, \tilde{y}^{(i)}; \theta) \geq \max_\theta \frac{1}{N}\sum_{i=1}^{N} & Q(y^{(i)}) \log P(y^{(i)} \mid \mathbf{x}^{(i)}; \theta_{Y|X}) \\ & + (1 - Q(y^{(i)})) \log(1 - P(y^{(i)} \mid \mathbf{x}^{(i)}; \theta_{Y|X})) \\ & + Q(y^{(i)}) \log P(\tilde{y}^{(i)} \mid y^{(i)}, t^{(i)}; \theta_{\tilde{Y}|Y,T})\end{aligned} \tag{22}$$

where $\theta = [\theta_{Y|X}\ \theta_{\tilde{Y}|Y,T}]^\top$.

Proof. We first construct the evidence-based lower bound (ELBO) of the LHS in the theorem statement. First, for a single example indexed by i, we can write:

$$\log P(u^{(i)}, a^{(i)}, \mathbf{x}^{(i)}, t^{(i)}, \tilde{y}^{(i)}; \theta) = \log \sum_{y^{(i)} \in \{0,1\}} Q(y^{(i)}) \frac{P(u^{(i)}, a^{(i)}, \mathbf{x}^{(i)}, t^{(i)}, \tilde{y}^{(i)}, y^{(i)}; \theta)}{Q(y^{(i)})} \tag{23}$$

$$\geq \sum_{y^{(i)} \in \{0,1\}} Q(y^{(i)}) \log \frac{P(u^{(i)}, a^{(i)}, \mathbf{x}^{(i)}, t^{(i)}, \tilde{y}^{(i)}, y^{(i)}; \theta)}{Q(y^{(i)})} \tag{24}$$

via Jensen's inequality. Then, we note that

$$\begin{aligned}&\max_\theta \frac{1}{N}\sum_{i=1}^{N} \sum_{y^{(i)} \in \{0,1\}} Q(y^{(i)}) \log \frac{P(u^{(i)}, a^{(i)}, \mathbf{x}^{(i)}, t^{(i)}, \tilde{y}^{(i)}, y^{(i)}; \theta)}{Q(y^{(i)})} \\ &= \max_\theta \frac{1}{N}\sum_{i=1}^{N} \sum_{y^{(i)} \in \{0,1\}} Q(y^{(i)}) \log P(u^{(i)}, a^{(i)}, \mathbf{x}^{(i)}, t^{(i)}, \tilde{y}^{(i)}, y^{(i)}; \theta),\end{aligned} \tag{25}$$

dropping $Q(y^{(i)}) \log Q(y^{(i)})$, which is constant with respect to θ, after expanding the log term. We can then use the DAG to factorize the joint distribution of all variables (including latent variable Y), which is given by

$$P(\tilde{Y}, Y, T, X, A, U) = P(\tilde{Y} \mid Y, T) P(Y \mid X) P(T \mid X, A) P(X, A, U). \tag{26}$$

Note that we need only model the first two terms for estimation of $y^{(i)}$. The first two terms do not involve $y^{(i)}$, are not parameterized, and can be dropped from the maximization problem. Hence, we proceed to write

$$= \max_\theta \frac{1}{N}\sum_{i=1}^{N} \sum_{y^{(i)} \in \{0,1\}} Q(y^{(i)}) \log P(y^{(i)} \mid \mathbf{x}^{(i)}; \theta_{\tilde{Y}|X}) P(\tilde{y}^{(i)} \mid t^{(i)}, y^{(i)}; \theta_{\tilde{Y}|Y,T}) \tag{27}$$

where $\theta = [\theta_{Y|X}\ \theta_{\tilde{Y}|Y,T}]^\top$. This can be rewritten as

$$= \max_\theta \frac{1}{N}\sum_{i=1}^{N} \sum_{y^{(i)} \in \{0,1\}} Q(y^{(i)}) \log P(y^{(i)} \mid \mathbf{x}^{(i)}; \theta_{\tilde{Y}|X}) + Q(y^{(i)}) P(\tilde{y}^{(i)} \mid t^{(i)}, y^{(i)}; \theta_{\tilde{Y}|Y,T}), \tag{28}$$

at which point we note that it is sufficient to show that

$$(1 - Q(y^{(i)})) \log(1 - P(\tilde{y}^{(i)} \mid y^{(i)}, t^{(i)}; \theta_{\tilde{Y}|Y,T})) \tag{29}$$

is constant in θ. We can rewrite the above as

$$(1 - Q(y^{(i)}))[\tilde{y}^{(i)} \log(P(\tilde{y}^{(i)} = 1 \mid y^{(i)} = 0, t^{(i)}; \theta_{\tilde{Y}|Y,T})) + (1 - \tilde{y}^{(i)}) \log(P(\tilde{y}^{(i)} = 0 \mid y^{(i)} = 0, t^{(i)}; \theta_{\tilde{Y}|Y,T}))]. \tag{30}$$

First, note that the event $\{\tilde{y}^{(i)} = 1 \mid y^{(i)} = 0\}$ occurs with probability zero by definition (recall $\tilde{y}^{(i)} = y^{(i)}t^{(i)}$). Thus, $P(\tilde{y}^{(i)} = 1 \mid y^{(i)} = 0, t^{(i)}; \theta_{\tilde{Y}|Y,T})$) cannot change with respect to θ; we drop it from the maximization problem. Similarly, $P(\tilde{y}^{(i)} = 0 \mid y^{(i)} = 0, t^{(i)}) = 0$ by definition, so $(1 - Q(y^{(i)})) \log(1 - P(\tilde{y}^{(i)} \mid y^{(i)}, t^{(i)})) = 0$ which is constant as needed, from which the theorem follows. □

Remark B.2. In the theorem statement, replacing $P(y^{(i)} \mid \mathbf{x}^{(i)}; \theta_{Y|X})$ with $\hat{y}^{(i)}$ and $P(\tilde{y}^{(i)} \mid y^{(i)}, t^{(i)}; \theta_{\tilde{Y}|Y,T})$ with $h_\phi(\cdot)$, assuming $y^{(i)}$ and $\tilde{y}^{(i)}$ are binary, and writing the explicit form of negative binary cross-entropy (*e.g.*, $y \log y + (1-y) \log \hat{y}$) recovers the form of the M-step objective seen in Eq. 7. Note that the optimization problem flips from a maximization to a minimization due to the relationship between maximizing log-likelihood of binary variable(s) and minimizing cross-entropy loss.

B.3. Theorem 3.4

Theorem (Strength of the causal regularizer in $\hat{t}^{(i)}$)**.** *For an example indexed by i, $Q(y^{(i)}) \in [0, 1)$, and $J^{(i)}$ defined as in Eq. 88, $R(J^{(i)})$ is monotonically increasing in $\hat{t}^{(i)}$ on $(0, 1]$.*

Proof. As a proof outline, we first show the closed-form of $\hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)})$ by solving the first-order optimality condition of Eq. 88. Then, we show that $\hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)})$ is decreasing in $\hat{t}^{(i)}$, and attains a maximum of $Q(y^{(i)})$ as $\hat{t}^{(i)} \to 0^+$. We conclude by showing that the preceding implies that $R(J^{(i)})$ is monotonically increasing in $\hat{t}^{(i)}$ on $(0, 1]$, as desired.

The first-order optimality condition of Eq. 88 is

$$-\left(\frac{Q(y^{(i)})}{\hat{y}^{(i)}} - \frac{1 - Q(y^{(i)})}{1 - \hat{y}^{(i)}}\right) + \frac{\hat{t}^{(i)} Q(y^{(i)})}{1 - \hat{y}^{(i)}\hat{t}^{(i)}} = 0. \tag{31}$$

By assumption (convexity of $\mathcal{L}$), the minimizer is unique. Some algebra yields

$$-Q(y^{(i)})(1 - \hat{y}^{(i)})(1 - \hat{y}^{(i)}\hat{t}^{(i)}) + (1 - Q(y^{(i)}))\hat{y}^{(i)}(1 - \hat{y}^{(i)}\hat{t}^{(i)}) + \hat{t}^{(i)}Q(y^{(i)})\hat{y}^{(i)}(1 - \hat{y}^{(i)}) = 0 \tag{32}$$

$$\iff (\hat{t}^{(i)} + Q(y^{(i)})\hat{t}^{(i)})\hat{y}^{(i)2} - (2Q(y^{(i)})\hat{t}^{(i)} + 1)\hat{y}^{(i)} + Q(y^{(i)}) = 0, \tag{33}$$

from which we can apply the quadratic formula. Define $B(Q(y^{(i)}), \hat{t}^{(i)}) \triangleq 2Q(y^{(i)})\hat{t}^{(i)} + 1$ and $D(Q(y^{(i)}), \hat{t}^{(i)}) \triangleq (-B(Q(y^{(i)}), \hat{t}^{(i)}))^2 - 4Q(y^{(i)})(Q(y^{(i)})\hat{t}^{(i)} + \hat{t}^{(i)})$.[5] The quadratic formula yields solutions

$$\hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)}) = \frac{B(Q(y^{(i)}), \hat{t}^{(i)}) \pm \sqrt{D(Q(y^{(i)}, \hat{t}^{(i)})}}{2(\hat{t}^{(i)} + Q(y^{(i)})\hat{t}^{(i)})}. \tag{34}$$

We use the fact that $\hat{y}^{(i)}$ must be in [0, 1] and the constraints that $\hat{t}^{(i)} \in (0, 1]$ and $Q(y^{(i)}) \in [0, 1)$ to determine which branch of Eq. 34 yields real solutions in $[0, 1]$. By Lemma 1, $D(Q(y^{(i)}, \hat{t}^{(i)}) \geq 0$, so the solutions are real. Then, by Lemma 2,

$$\frac{B(Q(y^{(i)}), \hat{t}^{(i)}) + \sqrt{D(Q(y^{(i)}, \hat{t}^{(i)})}}{2(\hat{t}^{(i)} + Q(y^{(i)})\hat{t}^{(i)})} \geq 1, \tag{35}$$

eliminating that branch. By elimination,

$$\hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)}) = \frac{B(Q(y^{(i)}), \hat{t}^{(i)}) - \sqrt{D(Q(y^{(i)}, \hat{t}^{(i)})}}{2(\hat{t}^{(i)} + Q(y^{(i)})\hat{t}^{(i)})}. \tag{36}$$

[5] We use letter $B(\cdot)$ because it corresponds to coefficient b in the conventional quadratic formula:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

for a quadratic polynomial $ax^2 + bx + c = 0$. We choose letter $D(\cdot)$ since it corresponds to the discriminant.

Lemma 3 verifies that the resulting $\hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)})$ are in [0, 1], as needed. To proceed, it suffices to show that $\hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)})$ is decreasing in $\hat{t}^{(i)}$ and attains a maximum of $Q(y^{(i)})$ as $\hat{t}^{(i)} \to 0^+$.

Applying Lemma 4, to prove that $\hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)})$ decreases in $\hat{t}^{(i)}$, it is sufficient to show

$$1 - 2\hat{t}^{(i)}Q(y^{(i)})^2 - \sqrt{D(Q(y^{(i)}), \hat{t}^{(i)}))} < 0 \tag{37}$$

because $1 - 2\hat{t}^{(i)}Q(y^{(i)})^2 - \sqrt{D(Q(y^{(i)}, \hat{t}^{(i)})}$ has the same sign as the derivative of $\hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)})$ for the values of $(\hat{t}^{(i)}, Q(y^{(i)}))$ of interest.

For values of $\hat{t}^{(i)} \in (0, 1]$ and $Q(y^{(i)}) \in [0, 1)$ such that $1 - 2\hat{t}^{(i)}Q(y^{(i)})^2 < 0$, Lemma 1 yields the desired result. For the remaining values of $(\hat{t}^{(i)}, Q(y^{(i)}))$, we can write

$$1 - 4Q(y^{(i)})^2\hat{t}^{(i)} + 4Q(y^{(i)})^4\hat{t}^{(i)2} < D(Q(y^{(i)}), \hat{t}^{(i)}) \tag{38}$$

$$\iff 1 - 4Q(y^{(i)})^2\hat{t}^{(i)} + 4Q(y^{(i)})^4\hat{t}^{(i)2} < 1 - 4Q(y^{(i)})^2\hat{t}^{(i)} + 4Q(y^{(i)})^2\hat{t}^{(i)2} \tag{39}$$

$$\iff Q(y^{(i)})^2 < 1 \tag{40}$$

which holds for all feasible values of $Q(y^{(i)}) \in [0, 1)$. Lastly, due to the monotonicity of $\hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)})$ in $\hat{t}^{(i)}$, the following one-sided limit is the maximum:

$$\lim_{\hat{t}^{(i)} \to 0^+} \hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)}) = \max_{\hat{t}^{(i)} \in (0,1]} \hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)}). \tag{41}$$

We want to show that the limit is $Q(y^{(i)})$. Note that, since $\mathcal{L}$ is finite and convex, it is continuous (Corollary 10.1.1, (Rockafellar, 1970)); hence, this limit exists. Since substituting $\hat{t}^{(i)} = 0$ yields the indeterminate form $0/0$, we appeal to L'Hôpital's rule:

$$\lim_{\hat{t}^{(i)} \to 0^+} \hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)}) = \lim_{\hat{t}^{(i)} \to 0^+} \frac{2Q(y^{(i)}) - \frac{4\hat{t}^{(i)}Q(y^{(i)})^2 - 2Q(y^{(i)})^2}{\sqrt{D(Q(y^{(i)}), \hat{t}^{(i)})}}}{2(Q(y^{(i)}) + 1)} = \frac{2Q(y^{(i)}) + 2Q(y^{(i)})^2}{2(Q(y^{(i)}) + 1)} = Q(y^{(i)}). \tag{42}$$

Note that $\sqrt{D(Q(y^{(i)}), \hat{t}^{(i)})}\Big|_{\hat{t}^{(i)}=0} = 1$. Since $Q(y^{(i)}) - \hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)}) > 0$,

$$R(J^{(i)}) = |Q(y^{(i)}) - \hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)})| = Q(y^{(i)}) - \hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)}). \tag{43}$$

Furthermore, since $\hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)})$ is decreasing in $\hat{t}^{(i)}$, $R(J^{(i)})$ must increase in $\hat{t}^{(i)}$, from which the theorem follows. □

Remark B.3. We comment on the potential for DCEM to improve robustness to low overlap. To do so, we analyze the sensitivity of the M-step optimum to extreme $\hat{t}^{(i)}$. While analyzing the asymptotic variance of consistent estimators is a common approach, asymptotic guarantees for DCEM are unclear due to the inherently non-convex (with respect to the parameters) objective function. Thus, we analyze the Lipschitzness of the M-step optimum versus other causal effect estimators. First, note that

$$\frac{d}{d\hat{t}}\hat{y}_{OPT}(Q(y), \hat{t}) = \frac{1 - 2Q(y)^2\hat{t}^2 - \sqrt{C}}{2(Q(y) + 1)\hat{t}^2\sqrt{C}} \tag{44}$$

where $C = 4Q(y)^2\hat{t}^2 - 4Q(y)^2\hat{t} + 1$. For all $Q(y^{(i)}) < 1$ and all $\hat{t^{(i)}}$, this derivative is *bounded* (*e.g.*, see Figure 5), and is $O(1)-$Lipschitz in $\hat{t}^{(i)}$. However, consider the expression for an inverse-propensity-weighted estimator, which sums terms of the form

$$\frac{y^{(i)}t^{(i)}}{\hat{t}^{(i)}} - \frac{y^{(i)}(1 - t^{(i)})}{1 - \hat{t}^{(i)}} \tag{45}$$

to obtain a final estimate. Eq. 45 has $O(\hat{t}^{(i)2})$-Lipschitz terms with respect to $\hat{t}^{(i)}$. Thus, in a Lipschitz sense, DCEM may be less sensitive to extreme propensity scores than causal effect estimators such as IPW.

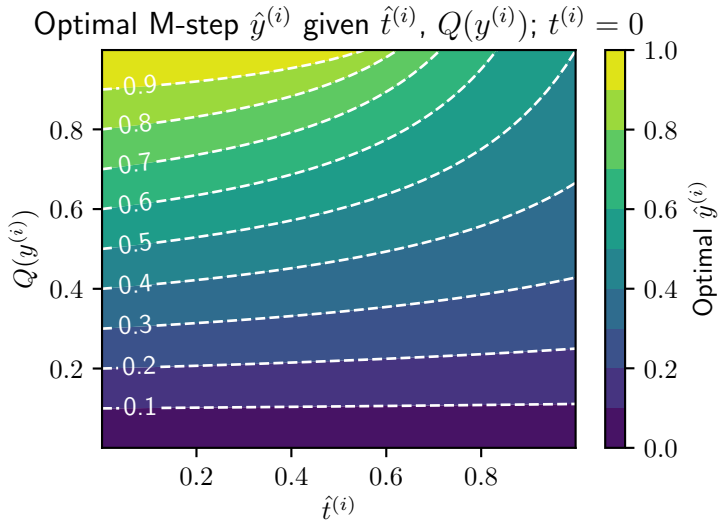

Figure 5. Contour plot of $\hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)})$ with respect to $\hat{t}^{(i)}$ (x-axis) and $Q(y^{(i)})$ (y-axis). $\hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)})$ scales with $Q(y^{(i)})$ but decreases in $\hat{t}^{(i)}$.

Corollary B.4. *For an example indexed by i, $Q(y^{(i)}) = 1$, and $J^{(i)}$ defined as in Eq. 88, $R(J^{(i)})$ is monotonically non-decreasing in $\hat{t}^{(i)}$ on $(0, 1]$.*

Proof. The proof is identical to that of Theorem 3.4, except we find that

$$\frac{\partial}{\partial \hat{t}^{(i)}} \hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)}) \leq 0, \tag{46}$$

instead of being strictly less than zero, from which the corollary follows. □

Remark B.5. For intuition, we show a contour plot of $\hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)})$ in Fig. 5. We verify the result in CVXPY.

B.4. Proposition 3.5

Proposition (Minimizer of M-step when $t^{(i)} = 1$)**.** *Suppose that $t^{(i)} = 1$, let $Q(y^{(i)}) \triangleq \mathbb{E}[y = 1 \mid t, \tilde{y}, a, \mathbf{x}]$, and let $\hat{y}^{(i)}$ be some estimate of $y^{(i)}$. Use $\mathcal{L} : [0, 1]^2 \to \mathbb{R}_+$ be shorthand for binary cross-entropy loss. Then, the minimization problem*

$$\min_{\hat{y}} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(Q(y^{(i)}), \hat{y}^{(i)}) + Q(y^{(i)}) \mathcal{L}(\tilde{y}^{(i)}, \hat{y}^{(i)} \hat{t}^{(i)}) \tag{47}$$

admits the solution $\hat{y}^{(i)} = y^{(i)}$ for all $i \in \{1, \ldots, N\}$.

Proof. We briefly verify the convexity of the objective, which follows from the convexity of binary cross-entropy loss and the closure of convexity under addition and positive scalar multiplication ($Q(y^{(i)}) \geq 0$). Thus, any minimizer of the objective is unique.

We proceed by cases. First, suppose that $y^{(i)} = 1$. Substituting the definition of $Q(y^{(i)})$, and using the fact that $t^{(i)} = 1 \implies y^{(i)} = \tilde{y}^{(i)}$), the objective function for a single example becomes

$$\mathcal{L}(1, \hat{y}^{(i)}) + \mathcal{L}(1, \hat{y}^{(i)} \hat{t}^{(i)}), \tag{48}$$

which, by inspection, is maximized for $\hat{y}^{(i)} = 1$. Similarly, for $y^{(i)} = 0$, the objective function for a single example is

$$\mathcal{L}(0, \hat{y}^{(i)}) \tag{49}$$

which reduces to binary cross-entropy loss, and $\hat{y}^{(i)} = 0$ minimizes the objective. Combining the two cases, the minimizer of the M-step objective when $t^{(i)} = 1$ is $\hat{y}^{(i)} = y^{(i)}$ as desired. □

B.5. Causal identifiability

For completeness, we provide the derivation of the causal identifiability results, though it follows directly from existing results (Imbens & Rubin, 2015).

Proposition. *Suppose that conditional exchangeability, or $\tilde{Y}(t) \perp\!\!\!\perp T \mid X$, holds. Then $\mathbb{E}[Y \mid X] = \mathbb{E}[\tilde{Y} \mid X, do(T=1)]$, which is identifiable as $\mathbb{E}[\tilde{Y} \mid X, T=1]$.*

Proof. We can write

$$\mathbb{E}[Y \mid X] = \mathbb{E}[Y \mid X, T=1] = \mathbb{E}[\tilde{Y} \mid X, T=1] = \mathbb{E}[\tilde{Y} \mid X, do(T=1)] \tag{50}$$

where the first equality is due to $Y \perp\!\!\!\perp T \mid X$, the second equality results from $T=1 \implies Y = \tilde{Y}$, and the final equality applies conditional exchangeability. Since $\mathbb{E}[Y \mid X] = \mathbb{E}[\tilde{Y} \mid X, do(T=1)] = \mathbb{E}[\tilde{Y} \mid X, T=1]$, the theorem follows. □

B.6. Lemmas used

Below are the lemmas and proofs referenced in the preceding theorem and proposition proofs.

Lemma 1. *Define $B(Q(y^{(i)}), \hat{t}^{(i)}) \triangleq 2Q(y^{(i)})\hat{t}^{(i)} + 1$ and $D(Q(y^{(i)}), \hat{t}^{(i)}) \triangleq (-B(Q(y^{(i)}), \hat{t}^{(i)}))^2 - 4Q(y^{(i)})(Q(y^{(i)})\hat{t}^{(i)} + \hat{t}^{(i)})$ on $Q(y^{(i)}) \in [0,1]$ and $\hat{t}^{(i)} \in (0,1]$. Then, $D(Q(y^{(i)}), \hat{t}^{(i)}) \geq 0$.*

Proof. Choose any $Q(y^{(i)}) \in [0,1]$ and $\hat{t}^{(i)} \in (0,1]$. We can write:

$$\begin{aligned}
& D(Q(y^{(i)}, \hat{t}^{(i)}) \geq 0 && (51)\\
\iff & B(Q(y^{(i)}), \hat{t}^{(i)})^2 \geq 4Q(y^{(i)})(Q(y^{(i)})\hat{t}^{(i)} + \hat{t}^{(i)}) && (52)\\
\iff & 4Q(y^{(i)})^2\hat{t}^{(i)2} + 4Q(y^{(i)})\hat{t}^{(i)} + 1 \geq 4Q(y^{(i)})(Q(y^{(i)})\hat{t}^{(i)} + \hat{t}^{(i)}) && (53)\\
\iff & 4Q(y^{(i)})^2\hat{t}^{(i)2} - 4Q(y^{(i)})^2\hat{t}^{(i)} + 1 \geq 0. && (54)
\end{aligned}$$

The final LHS is convex (by inspection) in $\hat{t}^{(i)}$, such that

$$\min_{\hat{t}^{(i)}} 4Q(y^{(i)})^2\hat{t}^{(i)2} - 4Q(y^{(i)})^2\hat{t}^{(i)} + 1 = 1 - Q(y^{(i)}) \geq 0 \tag{55}$$

where the minimum is attained at $\hat{t}^{(i)} = \frac{1}{2}$, and concave (by inspection) in $Q(y^{(i)})$, such that it suffices to evaluate the final LHS at $Q(y^{(i)}) \in \{0,1\}$:[6]

$$\begin{aligned}
& 4Q(y^{(i)})^2\hat{t}^{(i)2} - 4Q(y^{(i)})^2\hat{t}^{(i)} + 1\Big|_{Q(y^{(i)})=0} = 1 \geq 0 && (56)\\
& 4Q(y^{(i)})^2\hat{t}^{(i)2} - 4Q(y^{(i)})^2\hat{t}^{(i)} + 1\Big|_{Q(y^{(i)})=1} = 4(\hat{t}^{(i)2} - \hat{t}^{(i)}) + 1 \geq 4 \cdot \left(-\frac{1}{4}\right) + 1 = 0, && (57)
\end{aligned}$$

such that for all other $Q(y^{(i)}) \in (0,1)$, $4Q(y^{(i)})^2\hat{t}^{(i)2} - 4Q(y^{(i)})^2\hat{t}^{(i)} + 1 \geq 0$ as needed.

□

Lemma 2. *Define $B(Q(y^{(i)}), \hat{t}^{(i)})$ and $D(Q(y^{(i)}, \hat{t}^{(i)})$ as in Lemma 1. Then, for $Q(y^{(i)}) \in [0,1]$ and $\hat{t}^{(i)} \in (0,1]$,*

$$\frac{B(Q(y^{(i)}), \hat{t}^{(i)}) + \sqrt{D(Q(y^{(i)}, \hat{t}^{(i)})}}{2(\hat{t}^{(i)} + Q(y^{(i)})\hat{t}^{(i)})} \geq 1. \tag{58}$$

[6] Recall that, for a concave function f, $f(\alpha x + (1-\alpha)y) \geq \alpha f(x) + (1-\alpha)f(y)$ for $\alpha \in [0,1]$ with equality for $\alpha = 0$ or 1. Thus, via the extreme value theorem, the minimum of f on $[x, y]$ is achieved at x or y.

Proof. Choose any $Q(y^{(i)}) \in [0,1]$ and $\hat{t}^{(i)} \in (0,1]$. First, we rewrite

$$\frac{B(Q(y^{(i)}), \hat{t}^{(i)}) + \sqrt{D(Q(y^{(i)}, \hat{t}^{(i)})}}{2(\hat{t}^{(i)} + Q(y^{(i)})\hat{t}^{(i)})} \geq 1 \tag{59}$$

$$\iff 2Q(y^{(i)})\hat{t}^{(i)} + 1 + \sqrt{D(Q(y^{(i)}, \hat{t}^{(i)})} \geq 2(\hat{t}^{(i)} + Q(y^{(i)})\hat{t}^{(i)}) \tag{60}$$

$$\iff \sqrt{D(Q(y^{(i)}, \hat{t}^{(i)})} \geq 2\hat{t}^{(i)} - 1. \tag{61}$$

For $\hat{t}^{(i)} \in (0, \frac{1}{2})$, Lemma 1 yields the desired conclusion. For $\hat{t}^{(i)} \in [\frac{1}{2}, 1]$, we can write

$$\sqrt{D(Q(y^{(i)}, \hat{t}^{(i)})} \geq 2\hat{t}^{(i)} - 1 \tag{62}$$

$$\iff 4Q(y^{(i)2})\hat{t}^{(i)2} - 4Q(y^{(i)}2)\hat{t}^{(i)} + 1 \geq 4\hat{t}^{(i)2} - 4\hat{t}^{(i)} + 1 \tag{63}$$

$$\iff Q(y^{(i)2})(\hat{t}^{(i)} - 1) \geq \hat{t}^{(i)} - 1 \tag{64}$$

$$\iff Q(y^{(i)2}) \leq 1 \tag{65}$$

which all $Q(y^{(i)2}) \in [0,1]$ satisfy. This completes the proof. □

Lemma 3. *Define $B(Q(y^{(i)}), \hat{t}^{(i)})$ and $D(Q(y^{(i)}, \hat{t}^{(i)})$ as in Lemma 1. Then, for $Q(y^{(i)}) \in [0,1]$ and $\hat{t}^{(i)} \in (0,1]$,*

$$0 \leq \frac{B(Q(y^{(i)}), \hat{t}^{(i)}) - \sqrt{D(Q(y^{(i)}, \hat{t}^{(i)})}}{2(\hat{t}^{(i)} + Q(y^{(i)})\hat{t}^{(i)})} \leq 1. \tag{66}$$

Proof. Choose any $Q(y^{(i)}) \in [0,1]$ and $\hat{t}^{(i)} \in (0,1]$. Equivalently, we can show

$$0 \leq B(Q(y^{(i)}), \hat{t}^{(i)}) - \sqrt{D(Q(y^{(i)}, \hat{t}^{(i)})} \leq 2(\hat{t}^{(i)} + Q(y^{(i)})\hat{t}^{(i)}). \tag{67}$$

For the first inequality, note that

$$\sqrt{D(Q(y^{(i)}, \hat{t}^{(i)})} = \sqrt{(-B(Q(y^{(i)}), \hat{t}^{(i)}))^2 - 4Q(y^{(i)})(Q(y^{(i)})\hat{t}^{(i)} + \hat{t}^{(i)})} \leq \sqrt{(-B(Q(y^{(i)}), \hat{t}^{(i)}))^2} \tag{68}$$

$$= |B(Q(y^{(i)}), \hat{t}^{(i)}))| = B(Q(y^{(i)}), \hat{t}^{(i)})) \tag{69}$$

which rearranges to $0 \leq B(Q(y^{(i)}), \hat{t}^{(i)}) - \sqrt{D(Q(y^{(i)}, \hat{t}^{(i)})}$ as desired. For the second inequality, note that

$$B(Q(y^{(i)}), \hat{t}^{(i)}) - \sqrt{D(Q(y^{(i)}, \hat{t}^{(i)})} \leq 2(\hat{t}^{(i)} + Q(y^{(i)})\hat{t}^{(i)}) \tag{70}$$

$$\iff 1 - \sqrt{D(Q(y^{(i)}, \hat{t}^{(i)})} \leq 2\hat{t}^{(i)} \tag{71}$$

$$\iff 1 - 2\hat{t}^{(i)} \leq \sqrt{D(Q(y^{(i)}, \hat{t}^{(i)})}. \tag{72}$$

For $\hat{t}^{(i)} \in [\frac{1}{2}, 1]$, Lemma 1 yields the desired conclusion. For $\hat{t}^{(i)} \in (0, \frac{1}{2})$, the proof proceeds similarly to Lemma 2:

$$\sqrt{D(Q(y^{(i)}, \hat{t}^{(i)})} \geq 1 - 2\hat{t}^{(i)} \tag{73}$$

$$\iff 4Q(y^{(i)2})\hat{t}^{(i)2} - 4Q(y^{(i)})^2\hat{t}^{(i)} + 1 \geq 4\hat{t}^{(i)2} - 4\hat{t}^{(i)} + 1 \tag{74}$$

$$\iff Q(y^{(i)2})(\hat{t}^{(i)} - 1) \geq \hat{t}^{(i)} - 1 \tag{75}$$

$$\iff Q(y^{(i)2}) \leq 1 \tag{76}$$

which all $Q(y^{(i)2}) \in [0,1]$ satisfy. This completes the proof. □

Lemma 4. *Define $B(Q(y^{(i)}), \hat{t}^{(i)})$ and $D(Q(y^{(i)}, \hat{t}^{(i)})$ as in Lemma 1, and define $\hat{y}^{(i)}_{OPT}(Q(y^{(i)}), \hat{t}^{(i)})$ as in Definition 88. Then,*

$$\textit{sign}\left(\frac{\partial}{\partial \hat{t}^{(i)}}\, \hat{y}^{(i)}_{OPT}(Q(y^{(i)}), \hat{t}^{(i)})\right) = \textit{sign}\left(1 - 2\hat{t}^{(i)}Q(y^{(i)})^2 - \sqrt{D(Q(y^{(i)}), \hat{t}^{(i)}))}\right) \tag{77}$$

where sign$(x) : \mathbb{R} \to \{-1, 0, 1\}$ is the sign function:

$$\textit{sign}(x) \triangleq \begin{cases} -1 & x < 0 \\ 0 & x = 0 \\ 1 & x > 0 \end{cases}. \tag{78}$$

Proof. The proof is largely algebraic simplification based on sign-preserving operations. Taking derivatives:

$$\frac{\partial}{\partial \hat{t}^{(i)}}\, \hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)}) = \frac{Q(y^{(i)}) - \frac{2\hat{t}^{(i)}Q(y^{(i)})^2 - Q(y^{(i)})^2}{\sqrt{D(Q(y^{(i)}), \hat{t}^{(i)})}}}{\hat{t}^{(i)} + Q(y^{(i)})\hat{t}^{(i)}} - \frac{\left(B(Q(y^{(i)}), \hat{t}^{(i)}) - \sqrt{D(Q(y^{(i)}), \hat{t}^{(i)}))}\right)(Q(y^{(i)}) + 1)}{2(\hat{t}^{(i)} + Q(y^{(i)})\hat{t}^{(i)})^2} \tag{79}$$

via the quotient rule of derivatives and cancelling terms. We can apply sign-preserving operations, namely, positive scalar multiplication, canceling additive zeroes, and commuting additive terms, as follows:

$$\begin{aligned}
\frac{\partial}{\partial \hat{t}^{(i)}}\, \hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)}) &\propto (\hat{t}^{(i)} + Q(y^{(i)})\hat{t}^{(i)})\left(2Q(y^{(i)}) - \frac{4\hat{t}^{(i)}Q(y^{(i)})^2 - 2Q(y^{(i)})^2}{\sqrt{D(Q(y^{(i)}), \hat{t}^{(i)})}}\right) \\
&\quad - \left(B(Q(y^{(i)}), \hat{t}^{(i)}) - \sqrt{D(Q(y^{(i)}), \hat{t}^{(i)}))}\right)(Q(y^{(i)}) + 1) && (80) \\
&\propto \hat{t}^{(i)}\left(2Q(y^{(i)}) - \frac{4\hat{t}^{(i)}Q(y^{(i)})^2 - 2Q(y^{(i)})^2}{\sqrt{D(Q(y^{(i)}), \hat{t}^{(i)})}}\right) - B(Q(y^{(i)}), \hat{t}^{(i)}) + \sqrt{D(Q(y^{(i)}), \hat{t}^{(i)}))} && (81) \\
&= \left(\frac{2\hat{t}^{(i)}Q(y^{(i)})^2 - 4\hat{t}^{(i)2}Q(y^{(i)})^2}{\sqrt{D(Q(y^{(i)}), \hat{t}^{(i)})}}\right) - 1 + \sqrt{D(Q(y^{(i)}), \hat{t}^{(i)}))} && (82) \\
&\propto 2\hat{t}^{(i)}Q(y^{(i)})^2 - 4\hat{t}^{(i)2} - \sqrt{D(Q(y^{(i)}), \hat{t}^{(i)}))} + D(Q(y^{(i)}), \hat{t}^{(i)}) && (83) \\
&= 1 - 2\hat{t}^{(i)}Q(y^{(i)})^2 - \sqrt{D(Q(y^{(i)}), \hat{t}^{(i)}))} && (84)
\end{aligned}$$

which completes the proof. □

B.7. Definition 3.3: causal regularization strength

We expand on our definition of causal regularization strength here. Conventionally, regularization strength is operationalized in terms of a regularization parameter $\lambda \in \mathbb{R}_+$, given a loss $\ell(\theta)$ and a regularizer $r(\theta)$ (*e.g.*, $r(\theta) = \|\theta\|_2^2$ for some objective $J(\theta)$ of the form

$$J(\theta) \triangleq \ell(\theta) + \lambda r(\theta). \tag{85}$$

Eq. 85 is an instance of *regularized risk minimization* (Shalev-Shwartz & Ben-David, 2014). It is also identical to *linear scalarization*, a technique for characterizing tradeoffs in multi-objective optimization. The equivalence between regularized risk minimization and linear scalarization simply reflects that regularization can impose tradeoffs in optimizing $J(\theta)$ between minimizing $\ell(\theta)$ versus $r(\theta)$. Regularized risk minimization treats $r(\theta)$ as a "penalty" term, while linear scalarization treats $r(\theta)$ as merely another objective. As λ increases, the tradeoff increasingly favors $r(\theta)$, and vice versa.

Now, consider our M-step objective example-wise:

$$\mathcal{L}(Q(y^{(i)}), \hat{y}^{(i)}) + Q(y^{(i)})\mathcal{L}(\tilde{y}^{(i)}, \hat{y}^{(i)}\hat{t}^{(i)}). \tag{86}$$

The M-step objective can similarly be interpreted as variation of a regularized risk minimziation problem, where $\ell(\theta) = \mathcal{L}(Q(y^{(i)}, \hat{y}^{(i)})$, and $Q(y^{(i)})\mathcal{L}(\tilde{y}^{(i)}, \hat{y}^{(i)}, \hat{t}^{(i)}) = r(\theta), \lambda = 1$. However, $\hat{t}^{(i)}$ is a constant that can affect regularization strength, but is not a multiplier like λ. The purpose of our result is to characterize the impact of $\hat{t}^{(i)}$ on the tradeoff between the two terms of the M-step objective.

Thus, motivated by the tradeoff/multi-objective perspective of regularization, we define regularization strength in terms of a tradeoff between optimizing $\mathcal{L}(Q(y^{(i)}, \hat{y}^{(i)})$ and $Q(y^{(i)})\mathcal{L}(\tilde{y}^{(i)}, \hat{y}^{(i)}\hat{t}^{(i)})$. We observe that

$$Q(y^{(i)}) = \arg\min_{\hat{y}^{(i)}} \mathcal{L}(Q(y^{(i)}), \hat{y}^{(i)}) \tag{87}$$

and define causal regularization strength as the absolute distance between $Q(y^{(i)})$, the minimizer of $\mathcal{L}(Q(y^{(i)}), \hat{y}^{(i)})$, and the optimum of the example-wise M-step objective.

Definition B.6 (Causal regularization strength)**.** Given an example indexed by i, and a finite loss function $\mathcal{L} : [0, 1]^2 \to \mathbb{R}$ convex in $\hat{y}^{(i)}$ on [0, 1] for all i, define

$$\hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)}) \triangleq \arg\min_{\hat{y}^{(i)}} J^{(i)}(\hat{y}^{(i)}, \dots) \triangleq \arg\min_{\hat{y}^{(i)}} \mathcal{L}(Q(y^{(i)}), \hat{y}^{(i)}) + Q(y^{(i)})\mathcal{L}(\tilde{y}^{(i)}, \hat{y}^{(i)}\hat{t}^{(i)}). \tag{88}$$

The causal regularization strength of objective $J^{(i)}$ is defined as $R(J^{(i)}) = |Q(y^{(i)}) - \hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)})|$.

Intuitively, we define causal regularization strength in terms of the absolute distance between the optimum of each term of the M-step objective, which captures some notion of a tradeoff between the two terms. Note that this definition does not relate to convergence to $\hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)})$; we are largely interested in how much the solution to $\min\ \mathcal{L}(Q(y^{(i)}), \hat{y}^{(i)})$ shifts after adding the causal regularization term.

C. Additional experimental setup

For both settings, we set random seeds to 42 to facilitate reproducibility.

C.1. Additional details for fully synthetic dataset

We choose s_Y as follows:

$$S_Y(x) = (s_{Y1} \circ s_{Y2})(\mathbf{x})$$
$$s_{Y1}(\mathbf{x}) = x_1 - \frac{1}{4}\sin(8\pi x_0 + \psi) \quad s_{Y2}(\mathbf{x}) = \mathbf{R}_{\pi/6}\mathbf{x} + 0.5$$

where ψ is a simulation parameter, and $\mathbf{R}_{\pi/6}$ is a 2D rotation matrix. Intuitively, $s_Y(\mathbf{x})$ rotates and translates $\mathbf{x}$, then applies a sinewave-based function that yields a similarly rotated, sinewave-shaped decision boundary.

We choose s_T as follows:

$$s_T(\mathbf{x}^{(i)}, a^{(i)}) = \mathbf{1}^\top \mathbf{x}^{(i)} - \tau_{a^{(i)}}$$

where τ_a is a simulation parameter. For demonstration, we set c_y such that $P(Y = 1) = 0.25$ to allow for sufficiently-sized performance gaps across groups to emerge.[7] As a sensitivity analysis, we also replicate all experiments on fully synthetic data across $\psi \in [0, \pi/6, \pi/3, \dots, 11\pi/6]$, representing the "phase" of the decision boundary.

Computing simulation parameters. We discuss how we find simulation parameters for each value of q_y, q_t, and k. Given q_y and $P(A = 0) = P(A = 1) = 0.5$, we have:

$$q_y = \frac{P(Y = 1 \mid A = 0)}{P(Y = 1 \mid A = 1)}$$
$$P(Y = 1) = \frac{1}{4} = \frac{P(Y = 1 \mid A = 0) + P(Y = 1 \mid A = 1)}{2}$$

[7]Empirically, at extreme values of $P(Y = 1)$, we found artificially small performance gaps. This is because model errors tend to concentrate near the true decision boundary, which lies in the tails of the covariate distributions defining $X \mid A = a$. In those tail regions, the difference between the densities $X \mid A = a$ across values of A is smaller in our two-Gaussian simulation design.

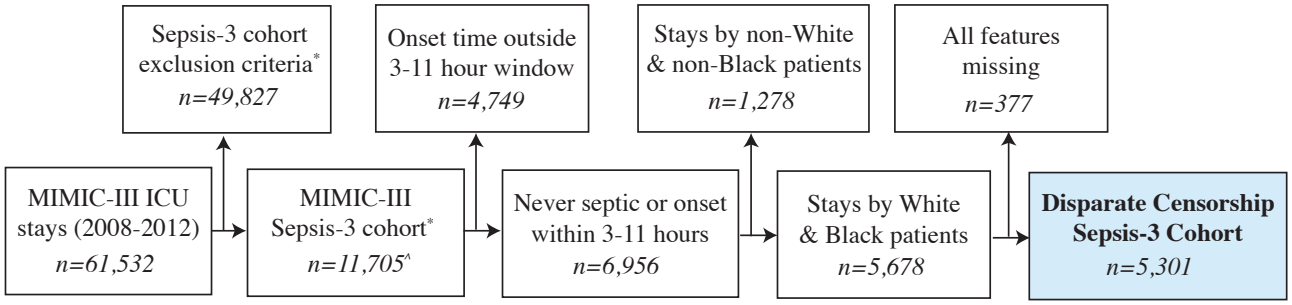

Figure 6. Cohort diagram for our Sepsis-3 cohort ($N = 5301$). *: Further details are provided in (Johnson et al., 2018). ^: Our cohort size differs slightly from that reported in (Johnson et al., 2018) due to an apparent pre-processing error in defining Sepsis-3 (Singer et al., 2016); our reported cohort size applies the relevant correction.

which yields, by substitution,

$$\frac{q_y}{2(q_y+1)} = P(Y=1 \mid A=0)$$
$$\frac{1}{2(q_y+1)} = P(Y=1 \mid A=1),$$

from which we use a binary search algorithm (bisection) evaluated using simulated versions of $X \mid A = a$ with the current estimate of the mean μ_a to solve for the requisite values of μ_a. Given values of μ_a, we can then solve for τ_a using q_t and k identically:

$$q_t = \frac{P(T=1 \mid A=0)}{P(T=1 \mid A=1)}$$
$$P(T=1) = kP(Y=1) = \frac{k}{4} = \frac{P(T=1 \mid A=0) + P(T=1 \mid A=1)}{2}$$

which yields, again by substitution,

$$\frac{q_t k}{2(q_t+1)} = P(T=1 \mid A=0)$$
$$\frac{k}{2(q_t+1)} = P(T=1 \mid A=1),$$

and we can again use binary search to solve for τ_a.

C.2. Additional details for pseudo-synthetic sepsis risk-stratification task

Cohort description. Our cohort follows from the MIMIC-III Sepsis-3 cohort (Johnson et al., 2018). Their cohort exclusion criteria are publicly available.[8] We corrected an apparent Sepsis-3 definition bug that erroneously labeled individuals with suspicion of infection if they received a blood culture at any time after an antibiotic before re-running their pipeline. In contrast, the Sepsis-3 (Singer et al., 2016) definition requires the blood culture to occur *within 24 hours* of the antibiotic time for suspicion of infection.[9] In practice, this stricter condition affects $< 1\%$ of rows in their original cohort: their cohort size is $N = 11,791$, while ours is $N = 11,705$.

Feature extraction. Following the Risk of Sepsis model (Delahanty et al., 2019), we extract the following 13 summary statistics over the initial 3-hour observation period:

1. Maximum lactic acid measurement,

[8] https://github.com/alistairewj/sepsis3-mimic

[9] There are multiple "paths" for meeting the criteria for suspicion of infection; for a full enumeration, see Table 2 of (Singer et al., 2016).

2. first shock index times age (years),
3. last shock index times age (years),
4. maximum white blood cell count,
5. change in lactic acid (last - first),
6. maximum neutrophil count,
7. maximum blood glucose,
8. maximum blood urea nitrogen,
9. maximum respiratory rate,
10. last albumin measurement,
11. minimum systolic blood pressure,
12. maximum creatinine, and
13. maximum body temperature (Fahrenheit).

The shock index is defined as the ratio of heart rate (beats per minute) and systolic blood pressure. Missing features are replaced with -9999 following the original manuscript.

Testing decision boundary. We define s_T as

$$\beta \cdot \frac{RR_{\max} - 22}{RR_\sigma} + (1-\beta) \cdot \frac{SBP_{\min} - 100}{SBP_\sigma} - \tau_a, \tag{89}$$

where $RR_{\max}$ and $SBP_{\min}$ are maximum respiratory rate and minimum systolic blood pressure, respectively, and RR_σ, SBP_σ are their corresponding standard deviations on the training split ($RR_\sigma = 9.8, SBP_\sigma = 21.8$). The parameter β allows us to examine different testing decisions. Thus, we replicate all experiments over $\beta \in \{0, 0.1, \ldots, 1\}$.

D. Hyperparameters & additional model details

All hyperparameters were selected using a validation set of examples. Hyperparameters for the sepsis simulation task were chosen such that all approaches attained similar performance when using y. We reimplement all existing methods following the original papers, using the code repository as a reference if applicable. We set random seeds to 42 for all models (used for initialization), unless otherwise noted.

D.1. Default hyperparameters

Fully synthetic. All models use a two-layer neural network with layer sizes $(64, 64)$, trained for 1000 epochs via Adam (Kingma & Ba, 2015) with learning rate 10^{-3} and no weight decay unless specified. EM approaches are trained up to 50 iterations with early stopping on validation loss (patience 3) and warm starts (initialized with solution from the previous iteration).

Sepsis classification. All predictors are three-layer neural networks with sizes $(128, 128, 16)$ trained for 10000 epochs using Adam with learning rate 10^{-5} and weight decay 10^{-3}. The DCEM propensity model (g_ζ) is trained for 20000 epochs with learning rate 10^{-5} and early stopping with patience 1000, and the DCEM model (f_θ) uses learning rate and weight decay 5×10^{-7} and 10^{-6}, respectively.

D.2. Simulation study

Peer loss & group peer loss: Both peer loss methods depend on a hyperparameter α, for which the optimal value depends on y. To show the peer loss methods in the best light, we manually calculate the optimal value for usage in training.

ITE corrected model (DragonNet): Our estimand of interest is the conditional average treatment effect (CATE) of the sensitive attribute A on testing T, which is identifiable via

$$\text{CATE}_{A\to T}(\mathbf{x}) \triangleq \mathbb{E}[T(1) \mid X = \mathbf{x}] - \mathbb{E}[T(0) \mid X = \mathbf{x}] = \mathbb{E}[T \mid X = \mathbf{x}, A = 1] - \mathbb{E}[T \mid X = \mathbf{x}, A = 0] \tag{90}$$

under assumptions of consistency ($T(a) = T$) and conditional exchangeability ($T(a) \perp\!\!\!\perp A \mid X$). We then apply the CATE as a correction factor to the default model:

$$\hat{y} \triangleq \hat{\tilde{y}} - \text{CATE}_{A\to T}(\mathbf{x}); \tag{91}$$

i.e., counterbalancing disparate censorship by "subracting out" the labeling bias. Note that this is an alternative to the counterbalancing approach of DCEM. We train and conduct inference with targeted regularization.

Truncated LQ: We searched across $k \in \{0.1, 0.2, \ldots, 1\}$ and $q \in \{0.1, 0.2, \ldots, 1\}$ (using the notation of the original paper), using $k = 0.1, q = 0.1$ for the final results.

SELF: We were unable to obtain convergence with Adam, so we used SGD with learning rate 0.01, momentum 0.9, noise parameter 0.05 (for input augmentation), consistency regularization parameter 1, and weight decay 1×10^{-6} as used for one of the experiments in the original paper. Weight decay was selected from $\{0, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$. The ensembling/mean teacher parameters were chosen from $\{0.9, 0.99, 0.999\}$. The noise was chosen from $\{0, 0.005, 0.01, 0.05, 0.1\}$. The regularization parameter was chosen from $\{1, 5, 10, 50\}$. SELF proceeds for a maximum of 50 iterations with patience 1 with respect to validation AUC. We set ensembling momentum to 0.9 and the mean teacher moving average parameter to 0.9. To show SELF in the best light, we prevented SELF from filtering tested positive individuals.

DivideMix: We use 20 warmup epochs, with $\alpha = 4$ as the Beta parameter for the MixMatch step, $T = 0.5$, $\lambda_u = 50$, $\lambda_r = 1$, and $\tau = 0.5$, and weight decay 5×10^{-4}. We also experimented with preventing DivideMix from filtering tested positive individuals, but DivideMix was unstable in both settings. Ultimately, we did not prevent DivideMix from filtering tested positive individuals.

EM-based methods (SAREM, DCEM): We tested SAREM and DCEM with and without the usage of warm starts in the M-step.

D.3. Sepsis risk-stratification

For all baselines, the setup matches the fully synthetic setting except as specified below.

DCEM: The learning rates under consideration were $[10^{-7}, 5 \times 10^{-7}, 10^{-6}, 5 \times 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}]$. The weight decay was selected from $[0, 10^{-6}, 2 \times 10^{-6}]$.

SELF: For the sepsis classification experiments, weused SGD and set the learning rate to 10^{-8}, the highest learning rate tested that did not result in NaN loss. We tested learning rates of the form $\{10^{-d}, 5 \times 10^{-d}\}$ for $d \in \{2, 3, 4, 5, 6, 7, 8\}$.

E. Additional empirical results and discussion

E.1. Full results

Here, we report empirical results for all baselines and settings. Due to the large number of empirical settings tested (simulation: 224, sepsis classification: 45), we include a representative subset of the figures, and report the raw numbers used for these results *and* results not shown in the Appendix via CSV files in the code appendix.

For the simulated task, we show results for $k \in [1/3, 3]$, $q_t \in [0.5, 2]$, and $q_y = 0.5$. Empirically, changing q_y did not affect the general trends, but amplified/dampened the scale. Increasing q_t beyond the selected range has similar impacts. For lower values of k, all methods perform poorly. For the sepsis classification task, we show results for $k \in [1/3, 5]$ and $q_t = 1.5$.

Summary of results. We summarize when our method (DCEM) empirically performed the best, when it performed similarly to baselines, and when it underperformed baselines.

DCEM is best where...

- (Both metrics) The higher-prevalance group is undertested ($q_y < 1$ but $q_t > 1$) *and*
- (Both metrics) testing rates are sufficiently high ($k \geq 0.5$).

DCEM is similar to baselines when...

- (Both metrics) Testing rates are moderately low ($1/3 \leq k \leq 1/2$), or sufficiently high that it is easier to extrapolate from labeled data ($k \geq 3$).

DCEM underperforms baselines when...

- (ROC gap only) when testing rates are low ($k \leq 1/2$) *and*
- (ROC gap only) the testing disparity aligns with the prevalence disparity (*e.g.*, q_t and $q_y < 1$ such that learning to predict $\tilde{y}$ preserves ranking in y), *or*
- (both metrics) testing rates are extremely low ($k < 1/3$).

The strongest alternatives to DCEM in our experiments were SELF (both datasets, bias mitigation), DragonNet (sepsis only, both metrics), and the tested-only model (simulation only, discriminative performance).

Index of figures. We provide here a list of all result figures in the Appendix, indexed by problem parameters k (overall testing rate multiplier), q_t (testing disparity), and q_y (prevalence disparity; simulation only).

Fully-synthetic data

- Figure 11: $q_y = 1/2, k = 1/3, q_t = 1/2$
- Figure 12: $q_y = 1/2, k = 1/3, q_t = 1$
- Figure 13: $q_y = 1/2, k = 1/3, q_t = 2$
- Figure 14: $q_y = 1/2, k = 1/2, q_t = 1/2$
- Figure 15: $q_y = 1/2, k = 1/2, q_t = 1$
- Figure 16: $q_y = 1/2, k = 1/2, q_t = 2$
- Figure 17: $q_y = 1/2, k = 1, q_t = 1/2$
- Figure 18: $q_y = 1/2, k = 1, q_t = 1$
- Figure 19: $q_y = 1/2, k = 1, q_t = 2$
- Figure 20: $q_y = 1/2, k = 2, q_t = 1/2$
- Figure 21: $q_y = 1/2, k = 2, q_t = 1$
- Figure 22: $q_y = 1/2, k = 2, q_t = 2$
- Figure 23: $q_y = 1/2, k = 3, q_t = 1/2$
- Figure 24: $q_y = 1/2, k = 3, q_t = 1$
- Figure 25: $q_y = 1/2, k = 3, q_t = 2$

Table 1. Sensitivity analysis of DCEM components with respect to AUC and ROC gap (min, max across s_T in parentheses) for $q_y = 0.5, k = 1, q_t = 2$. Maximum (minimum) median AUC (ROC gap) in bold.

Method	↑ AUC	↓ ROC gap
Imputation-only	.676 [.644, .715]	.063 [.036, .086]
No causal regularization	.767 [.733, .813]	.056 [.016, .086]
DCEM (ours)	**.791 [.763, .820]**	**.031 [.019, .072]**

Sepsis classification

- Figure 26: $k = 1/4, q_t = 3/2$
- Figure 27: $k = 1/3, q_t = 3/2$
- Figure 28: $k = 1/2, q_t = 3/2$
- Figure 29: $k = 1, q_t = 3/2$
- Figure 30: $k = 2, q_t = 3/2$
- Figure 31: $k = 3, q_t = 3/2$
- Figure 32: $k = 4, q_t = 3/2$
- Figure 33: $k = 5, q_t = 3/2$

E.2. DCEM ablation study

To understand how DCEM design choices impact performance, we conduct an ablation study of repeated iterations and causal regularization:

- **Imputation-only**: This approach trains a model on the tested-only (labeled) examples, imputes pseudo-labels for the remaining, then trains a model on both the pseudo-labeled and labeled data. This is equivalent to a single EM iteration without causal regularization.
- **No causal regularization**: This approach runs multiple EM iterations, but without causal regularization.

The results (Table 1) suggest that both repeated iterations and causal regularization are essential to the bias mitigation and discriminative capabilities of DCEM. The imputation-only approach fails due to low overlap between the tested vs. untested regions. Consequently, the imputed outcomes could be arbitrarily inaccurate. If we keep imputing and retraining (without causal regularization), we recover a form of pseudo-labeling (Lee, 2013). The empirical improvement in performance suggests that repeated supervision from reliably labeled examples helps improve discriminative performance. However, this approach does not adjust for labeling bias (e.g., by using A), and indeed the ROC gap does not improve. Incorporating causal regularization recovers the DCEM M-step. Adding causal regularization guarantees that DCEM locally maximizes log-likelihood, and allows it to mitigate labeling bias by incorporating A into a propensity score-like term (causal regularization; see Theorem B.3).

E.3. Sensitivity analysis of causally-motivated approaches

Here, we conduct a sensitivity analysis of causally-motivated approaches under disparate censorship. The causally-motivated approaches are theoretically consistent estimators of $P(Y \mid X)$, which we can interpret at the conditional average treatment effect of testing (T) on the observed outcome ($\tilde{Y}$; see Appendix B.5). We examine the following causally-motivated estimators:

- **Tested-only**: training models on tested individuals only, using X as covariates,
- **Tested-only + group**: training models on tested individuals only, using X and A as covariates,

Table 2. Sensitivity analysis of causal effect estimators for estimating $P(Y \mid X)$ compared to DCEM with respect to AUC and ROC gap (min, max across s_T in parentheses) for $q_y = 0.5, k = 1, q_t = 2$. Maximum (minimum) median AUC (ROC gap) in bold.

Method	↑ AUC	↓ ROCGap
Tested-only	.808 [.623, .876]	.052 [.020, .093]
Tested-only + group	.764 [.675, .863]	.078 [.025, .278]
IPW	**.829 [.598, .874]**	.048 [.020, .104]
DR-Learner	.643 [.558, .769]	.117 [.080, .216]
DCEM (ours)	.791 [.763, .820]	**.031 [.019, .072]**

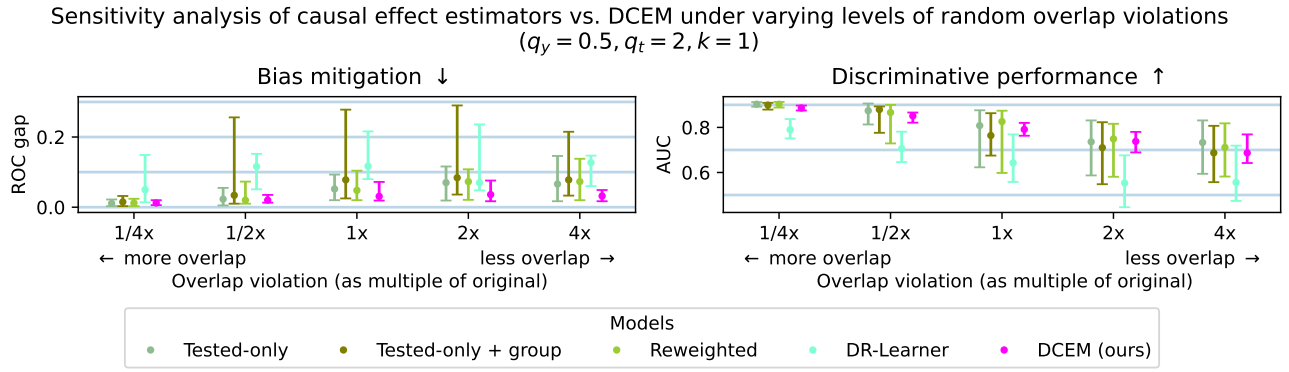

Figure 7. Sensitivity analysis of causal effect estimators with respect to AUC and ROC gap (min, max across s_T in parentheses) for $q_y = 0.5, k = 1, q_t = 2$ across varying levels of overlap. Causally-motivated methods are shown in green, while DCEM is shown in magenta. Empirically, DCEM improves robustness to overlap violations.

- **Inverse propensity weighting (IPW)**: an IPW-based (Rosenbaum & Rubin, 1983) version of the tested-only approach, and
- **Doubly-robust estimator (DR-Learner)**: a doubly-robust estimator of $P(Y \mid X)$ (Kennedy, 2023).

Models are evaluated for $q_y = 0.5, k = 1, q_t = 2$ (*i.e.*, same setting as Fig. 2). Under disparate censorship, low overlap is common due to the "sharpness" of the testing boundary. To validate this hypothesis, we also evaluate causal effect estimators versus DCEM at varying levels of overlap ($1/4$x, $1/2$x, 1x, 2x, and 4x of the original setting). Overlap is controlled by the coefficient inside the sigmoid for generating $t^{(i)}$ (*i.e.*, 30 in the original experiments).[10] For the DR-learner, we trimmed propensity scores (threshold: 0.05) to obtain estimates that were in [0, 1] (the possible values of $P(Y \mid X)$).

DCEM has better bias mitigation capabilities than causal approaches, and a tighter range of discriminative performance. Table 2 shows that, empirically, DCEM exhibits lower variance under overlap violations than causally-motivated approaches. Notably, DCEM achieves the lowest median ROC gap, and maintains competitive (but not necessarily best) median AUC. Causally-motivated methods generally have good median discriminative performance, but poor bias mitigation properties. Furthermore, the wide performance ranges of causally-motivated approaches may be unacceptable for safety-critical/high-stakes domains. We note that the DR-learner may underperform in this setting due if the propensity score trimming introduces sufficient bias: recall that, although double-robustness only requires one correctly-specified model, the asymptotic properties may still depend on the asymptotics of each model (*e.g.*, as shown in (Wager, 2020)).

DCEM is empirically more robust to overlap violations. Figure 7 shows that, empirically, as overlap violations increase, DCEM degrades more slowly than causally-motivated approaches in terms of both bias mitigation and discriminative performance. Furthermore, DCEM maintains similarly tight performance ranges across levels of overlap, while the

[10] Recall that $t^{(i)}$ is generated as a Bernoulli random variable with parameters of the form $\sigma(ax + b)$.

Table 3. Sensitivity analysis of softmax temperature scaling parameter (T) with respect to DCEM AUC and ROC gap (min, max across s_T in parentheses) for $q_y = 0.5, k = 1, q_t = 2$. Maximum (minimum) median AUC (ROC gap) in bold.

τ	↑ AUC	↓ ROC gap
0.01	.778 [.737, .815]	.051 [.020, .104]
0.1	.791 [.762, .818]	**.025 [.014, .057]**
1 (default)	.791 [.763, .820]	.031 [.019, .072]
10	**.800 [.730, .858]**	.051 [.021, .096]
100	.762 [.667, .835]	.071 [.032, .097]

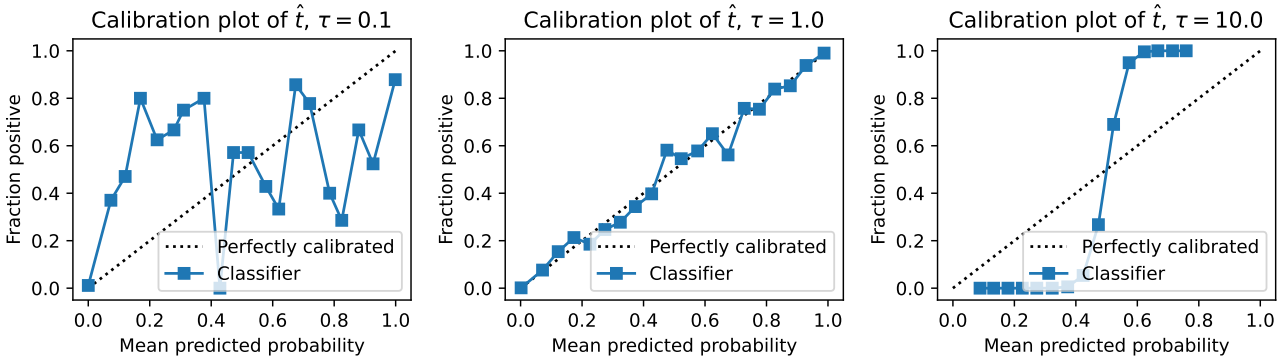

Figure 8. Calibration plot of $\hat{t}$ for (from left to right) $\tau \in \{0.1, 1, 10\}$. While $\hat{t}$ is well-calibrated for $\tau = 1$, changing τ in either direction (< 1 vs. > 1) induces miscalibration error.

performance ranges of causal approaches widens as overlap violations increase. At low overlap, causally-motivated approaches have similarly tight performance ranges as DCEM.

E.4. Sensitivity analysis of softmax temperature scaling

We can further tune the smoothness of $\hat{t}^{(i)}$ via the softmax temperature τ of the binary classifier for $\hat{t}^{(i)}$:

$$\hat{t}^{(i)} := \frac{\exp(z_t/\tau)}{\exp(z_t/\tau) + \exp(z_{1-t}/\tau)} \tag{92}$$

where z_t is the logit outputted by g_ζ for each $t \in \{0, 1\}$. Lower values of τ sharpen $\hat{t}^{(i)}$ towards $\{0, 1\}$, while larger values smooth $\hat{t}^{(i)}$ toward $\frac{1}{2}$. Note that $\tau = 1$ recovers the standard softmax function. Thus, adjusting τ allows us to control the smoothness of the $\tilde{y}^{(i)} = t^{(i)} y^{(i)}$ constraint.

Table 3 shows full results (median AUC and ROC gap, plus minima and maxima across s_Y) for DCEM across various values of temperature scaling parameter τ. Empirically, our results suggest that temperature scaling does not significantly change the AUC, and may trade off with bias mitigation since $\hat{t}^{(i)}$ may no longer be calibrated. Furthermore, even though median AUC improves in one case ($\tau = 10$), the range of AUC is much larger (0.057 vs. 0.128), and $\tau = 1$ still yields the maximum empirical worst-case AUC (0.763).

Values of τ away from 1 tend to yield larger ROC gaps. We find that $\hat{t}^{(i)}$ is well-calibrated for $\tau = 1$, but not so for values of τ (Figure 8). Since $\hat{t}^{(i)}$ is critical to counterbalancing disparate censorship, miscalibration error in $\hat{t}^{(i)}$ could result in larger ROC gaps by reducing the effectiveness/correctness of the causal regularization term. Thus, we opt to maintain $\tau = 1$.

E.5. Sensitivity analysis of E-step initialization

We compare random initialization to using a tested-only model as initialization (the final approach). Empirically, Table 4 shows trivial changes to performance when using a model trained on labeled data to initialize the E-step. This suggests that DCEM is able to overcome poor initialization in the settings studied; *i.e.*, the gains from tested-only initialization may be marginal, if nonzero.

Table 4. Sensitivity analysis of E-step initialization with respect to DCEM AUC and ROC gap (min, max across s_Y in parentheses) for $q_y = 0.5, k = 1, q_t = 2$. Maximum (minimum) median AUC (ROC gap) in bold.

Initialization scheme	AUC	ROC gap
random	.787 [.768, .822]	**.031 [.011, .060]**
tested-only	**.791 [.763, .820]**	**.031 [.019, .072]**

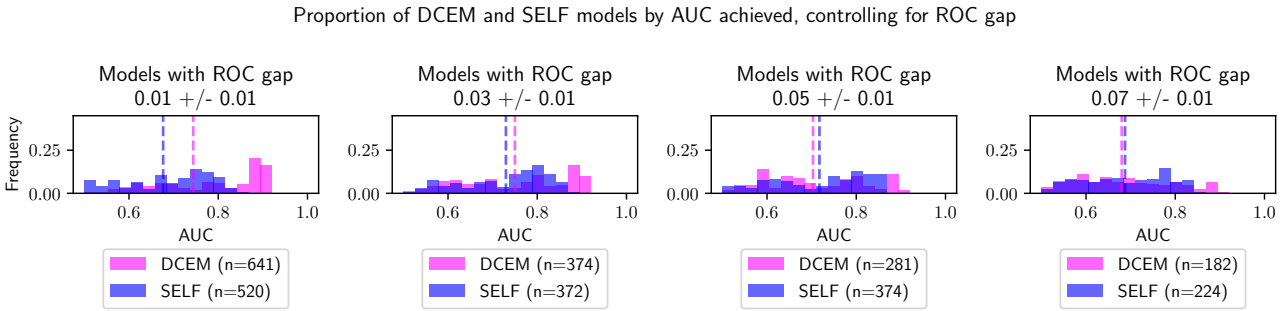

Figure 9. Relative frequencies of AUC for DCEM vs. SELF at similar ROC gaps, pooling models across all k, q_y, q_t tested. Dashed lines = mean AUC by model. DCEM improves AUC among models with similar ROC gaps when the ROC gap is below 0.04.

E.6. Tradeoffs between bias mitigation and discriminative performance: SELF

We compare instances of DCEM to SELF, controlling for ROC gap. We find that DCEM optimizes discriminative performance more effectively than SELF. Fig. 9 shows a histogram of AUC for SELF and DCEM models with similar ROC gaps across q_t, q_y, k and s_Y, increasing in ROC gap to the right. For models with ROC gaps < 0.04 (Fig. 9, 1st and 2nd from left), DCEM improves AUC compared to instances of SELF with similar ROC gaps. At larger ROC gaps, DCEM and SELF obtain similar AUCs (Fig. 9, 1st and 2nd from right). Similarly to the comparison with tested-only models, the results suggest that DCEM is not simply trading improved bias mitigation for performance, but is also able to optimize discriminative performance. Since SELF is a filtering approach that does not account for the causal structure of disparate censorship, its estimates of label bias are likely skewed. In contrast, DCEM explicitly uses the causal structure of disparate censorship to counterbalance label bias.

E.7. Sepsis classification and robustness to shifts in labeling decisions

Fig. 10 shows the performance of DCEM vs. models with bimodial behavior across different s_T, indexed by different feature weightings in s_T. Our results suggest that the baselines require specific s_T to perform above random. The baselines catastrophically underperform (AUC below 0.5) otherwise. Trends are analogous for the ROC gap.

Specifically, baseline performance improves when one feature is more heavily weighted than the other in the labeling decision (x-axis near 0 or 1). However, when both features feature in labeling decisions (x-axis near 0.5), the baselines catastrophically fail, while DCEM performance stays high. As seen in Fig. 4, DCEM AUC and ROC gap also exhibit less variation across the different s_T.

Determining which s_T is appropriate is a clinical problem that requires domain expertise, and we make no claims as to the clinical appropriateness of s_T. Thus, ML practitioners should not assume that their data will be representative of any particular decision-making pattern. DCEM is an alternative approach that is more robust than baselines to shifts in s_T, and thus warrants consideration when narrow assumptions about labeling biases are undesirable.

F. Computing Infrastructure

Hardware. We parallelize experiments across 4 A6000 GPUs and 256 AMD CPU cores (4x AMD EPYC 7763 64-Core processors), though the memory requirements of each model are under 2GB of VRAM.

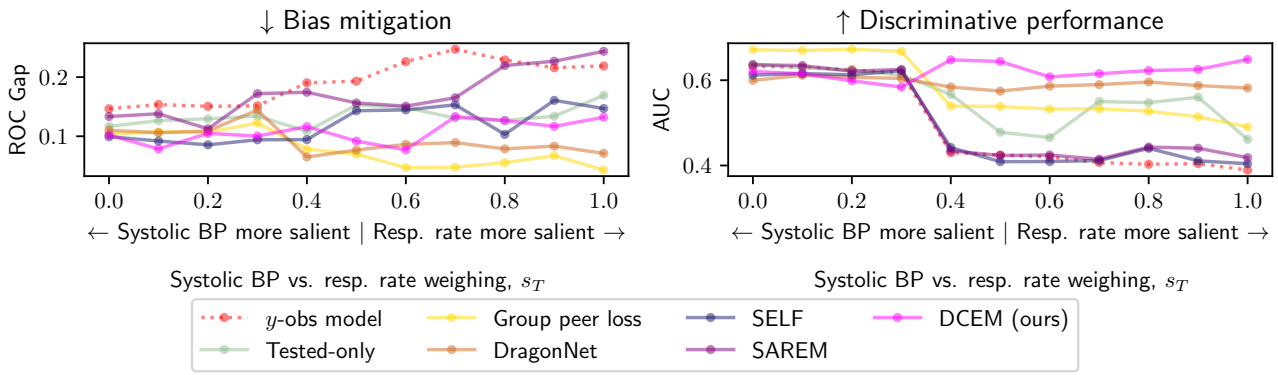

Figure 10. ROC gaps (left) and AUC (right) of selected models on sepsis classification task as weighting of systolic blood pressure (BP) and respiratory (resp.) rate (s_T) for testing changes ("0.0"/left: consider systolic BP only; "1.0"/right: consider resp. rate only) at $q_t = 1.5, k = 4$. If a feature is "more salient," it is weighted higher than the other in the testing decision function s_T.

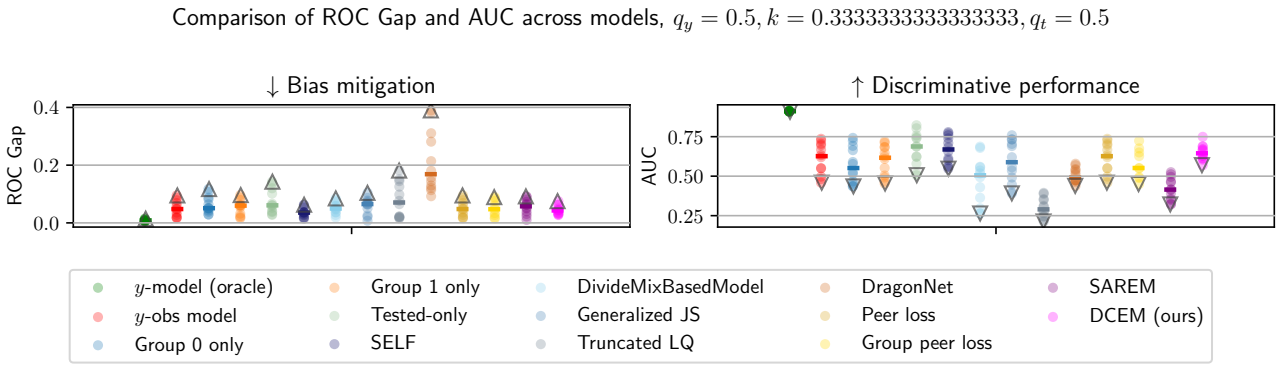

Figure 11. ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 1/3, q_t = 1/2$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

Software. All experiments are run on a distribution of Ubuntu 20.04.5 LTS (Focal Fossa) with Python 3.9.16 managed by conda 23.3.1. We use Pytorch 1.13.1 with CUDA 11.6 for all experiments (Paszke et al., 2019), with scikit-learn 1.2.2 (Pedregosa et al., 2011), scipy 1.10.1 (Virtanen et al., 2020), numpy 1.25.0 (Harris et al., 2020) and pandas 1.5.3 (The pandas development team, 2020) for data processing/analysis. Matplotlib 3.7.1 was used to generate figures. Additionally, torch_ema 0.3 was used in our implementation of SELF. For the simulation study, we use a modified version of the official disparate censorship repository at `https://github.com/MLD3/disparate_censorship` (Chang et al., 2022), which is included with our code repository.

G. Code

Code will be released at the MLD3 Github repository at `https://github.com/MLD3/DCEM`. We redact the data-processing code for the sepsis task only where necessary to ensure compliance with the terms of use for MIMIC-III (Johnson et al., 2016).

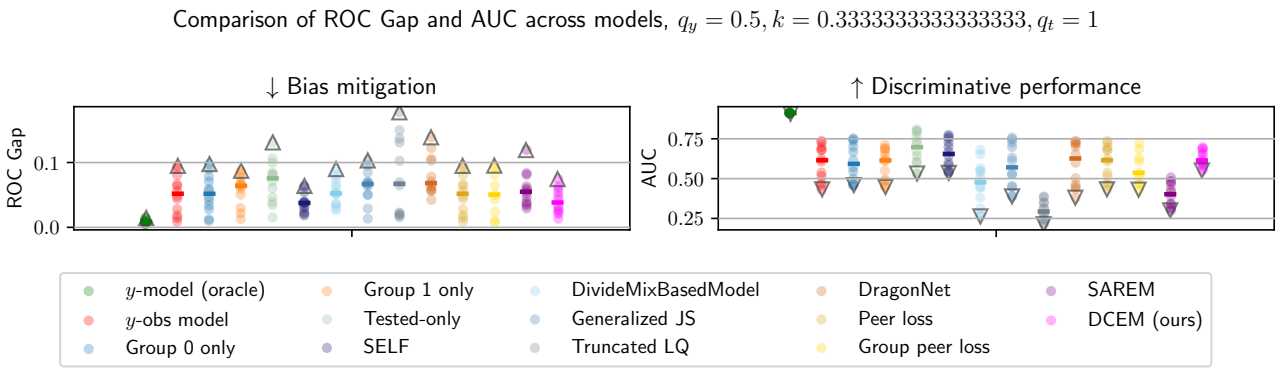

Figure 12. ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 1/3, q_t = 1$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

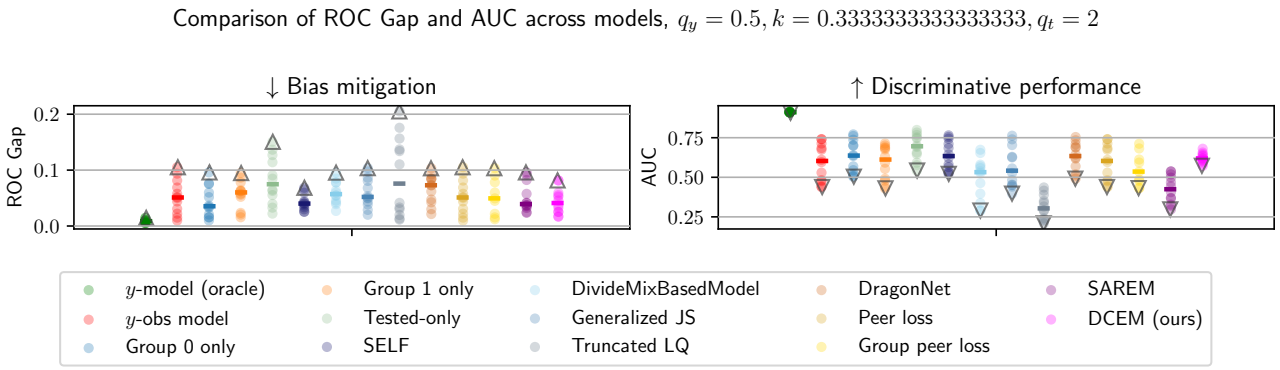

Figure 13. ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 1/3, q_t = 2$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

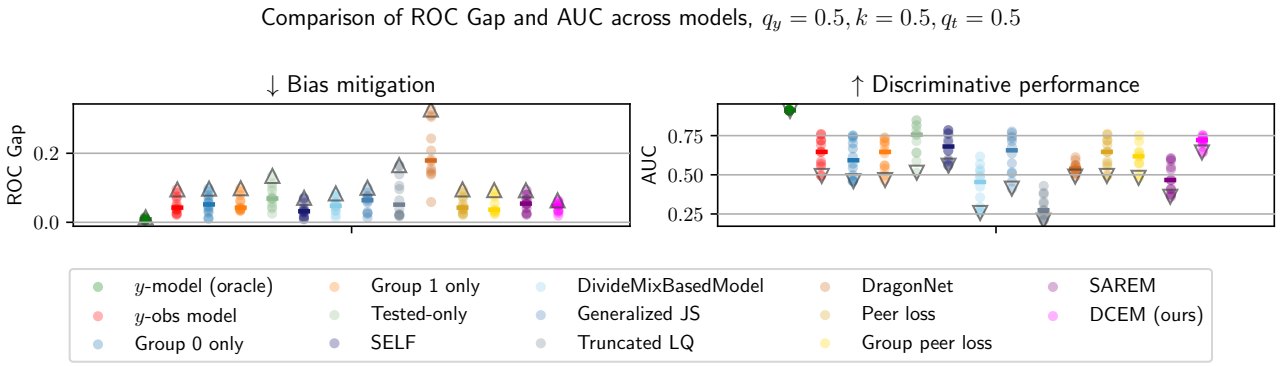

Figure 14. ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 1/2, q_t = 1/2$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

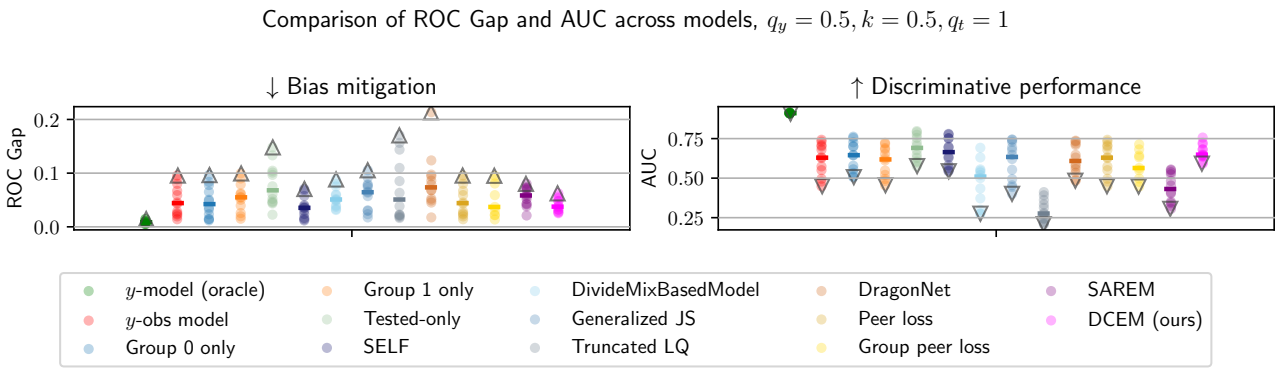

Figure 15. ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 1/2, q_t = 1$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

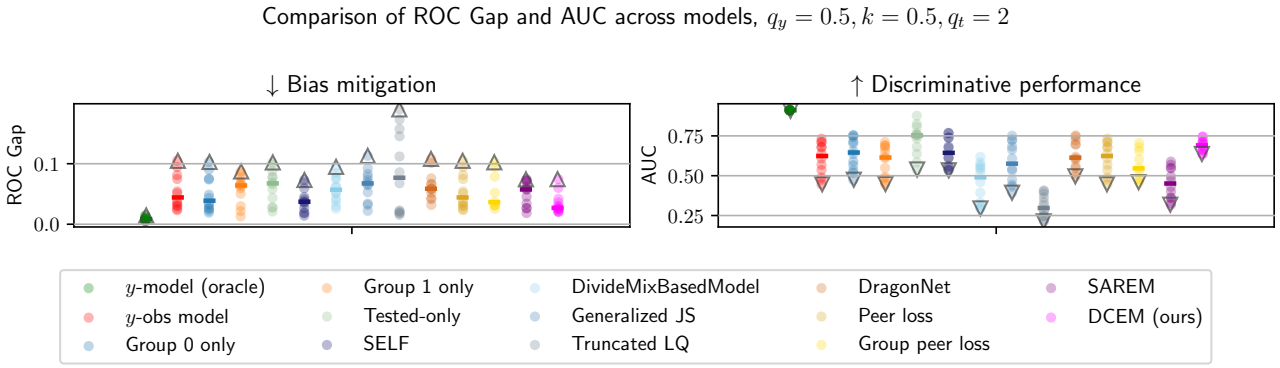

Figure 16. ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 1/2, q_t = 2$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

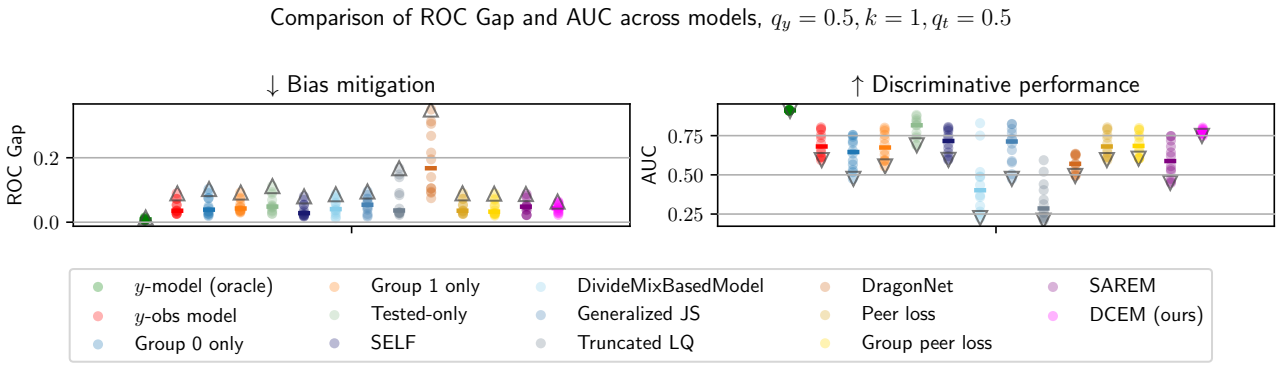

Figure 17. ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 1, q_t = 1/2$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

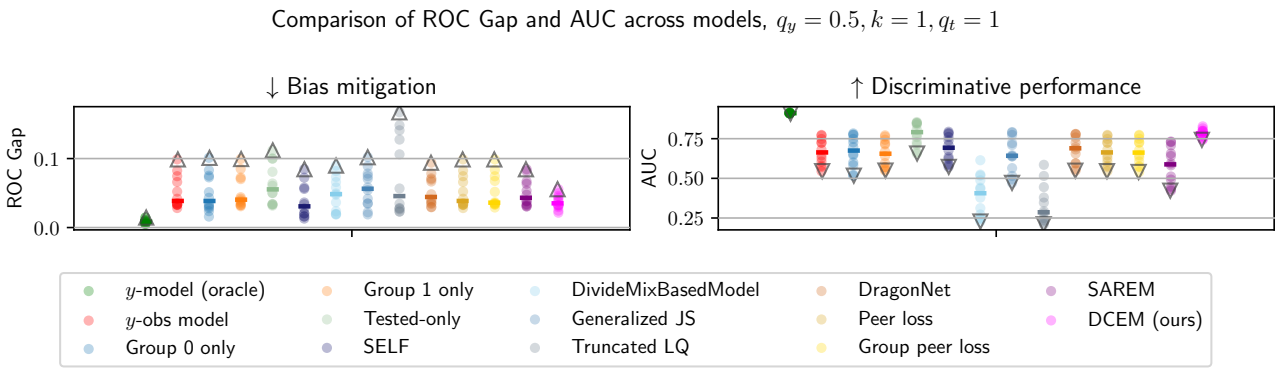

Figure 18. ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 1, q_t = 1$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

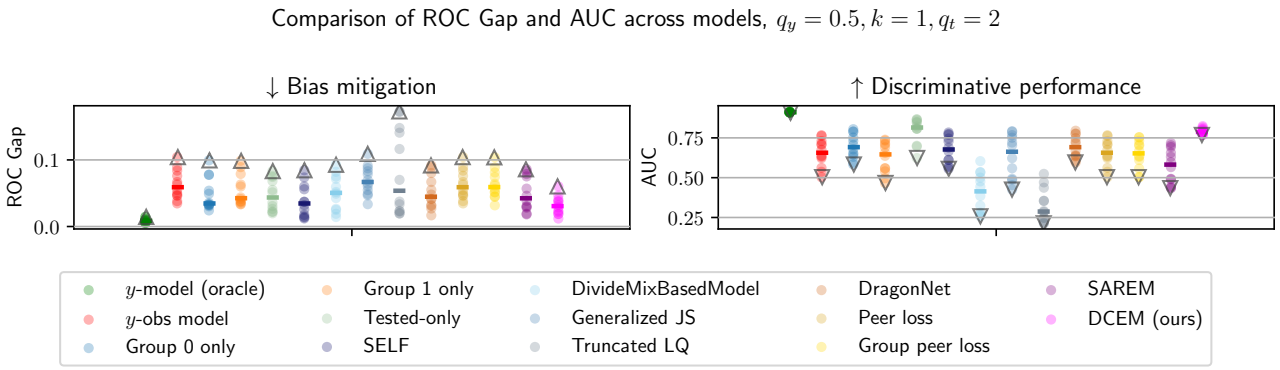

Figure 19. ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 1, q_t = 2$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

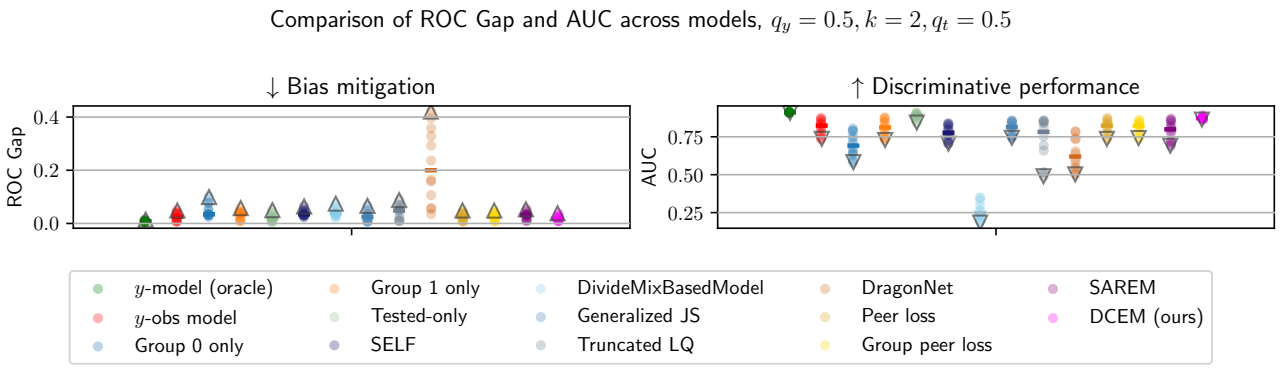

Figure 20. ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 2, q_t = 1/2$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

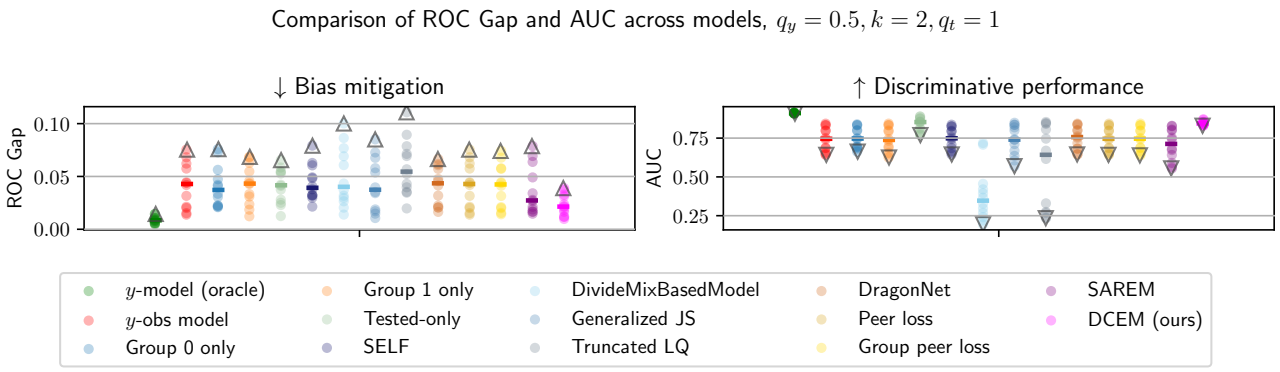

Figure 21. ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 2, q_t = 1$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

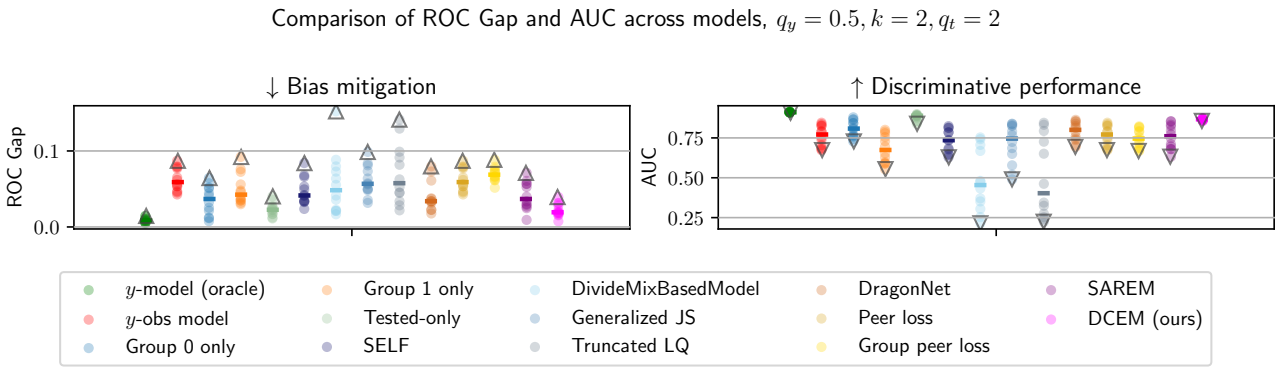

Figure 22. ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 2, q_t = 2$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

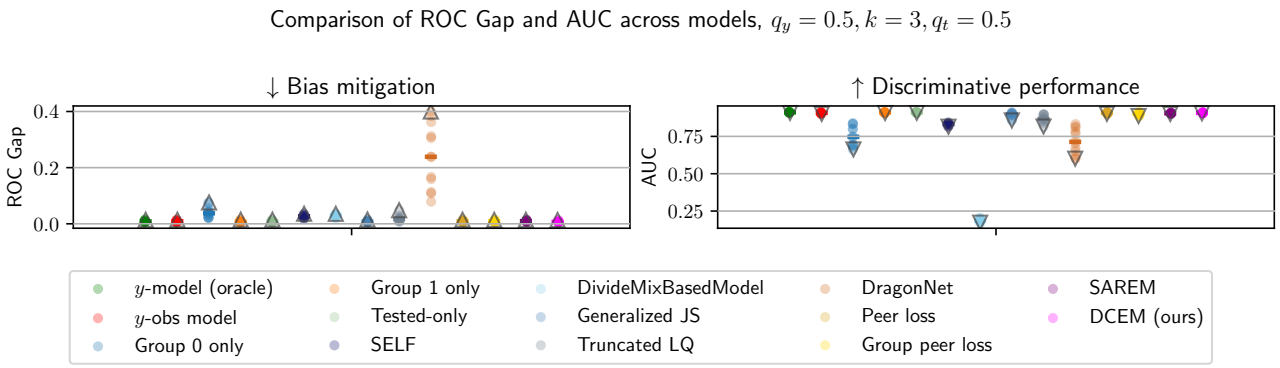

Figure 23. ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 3, q_t = 1/2$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

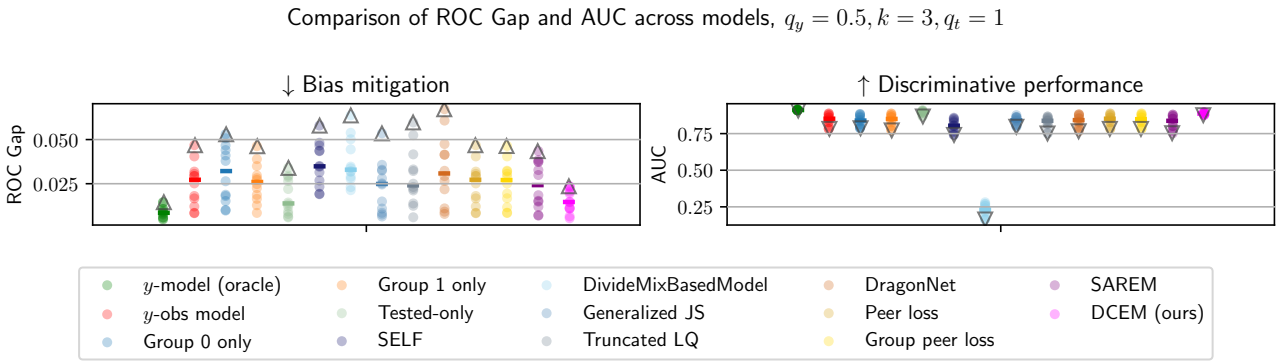

Figure 24. ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 3, q_t = 1$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

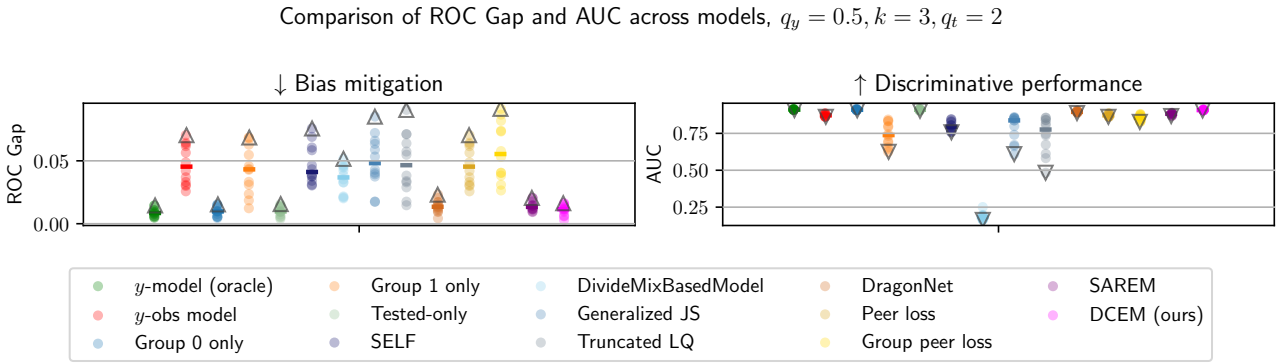

Figure 25. ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 3, q_t = 2$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

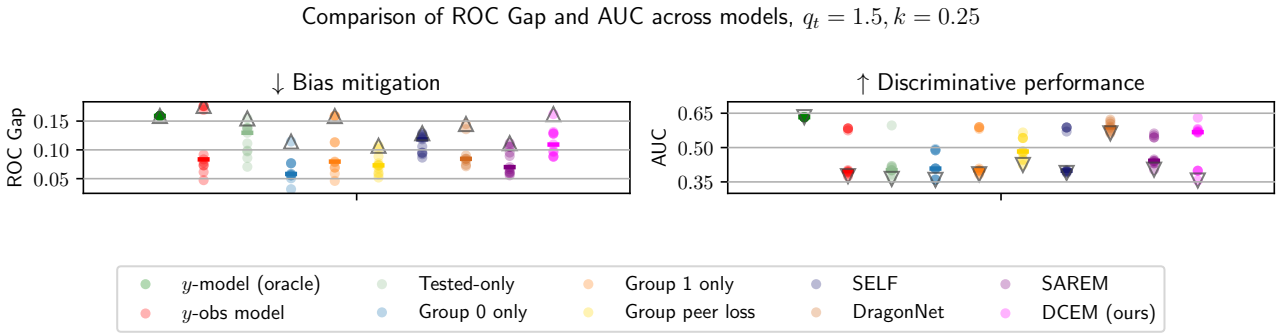

Figure 26. ROC gap (left) and AUC (right) of baselines on sepsis classification at $k = 1/4, q_t = 1.5$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

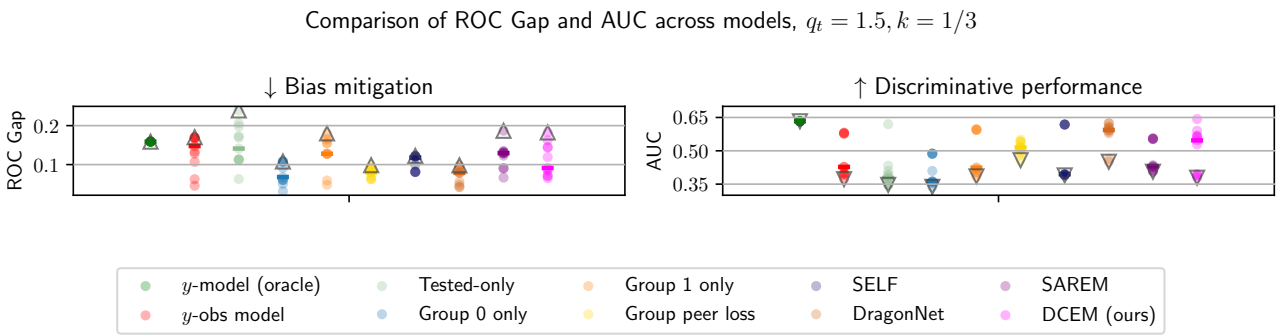

Figure 27. ROC gap (left) and AUC (right) of baselines on sepsis classification at $k = 1/3, q_t = 1.5$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

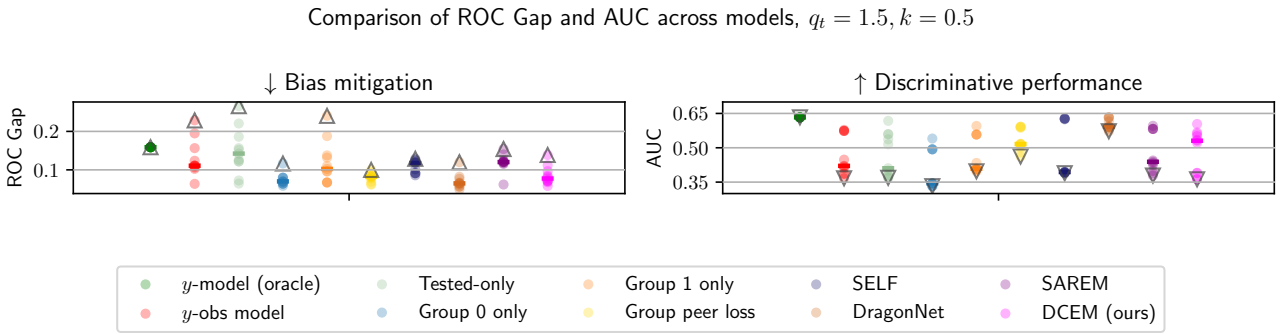

Figure 28. ROC gap (left) and AUC (right) of baselines on sepsis classification at $k = 1/2, q_t = 1.5$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

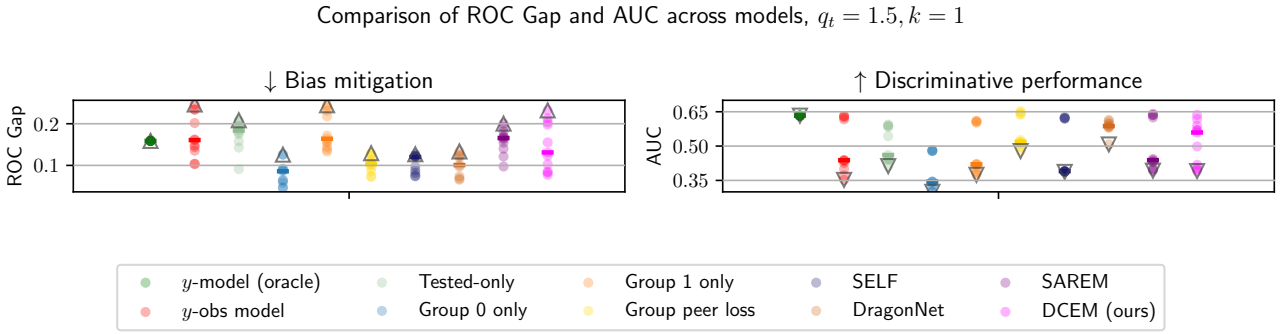

Figure 29. ROC gap (left) and AUC (right) of baselines on sepsis classification at $k = 1, q_t = 1.5$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

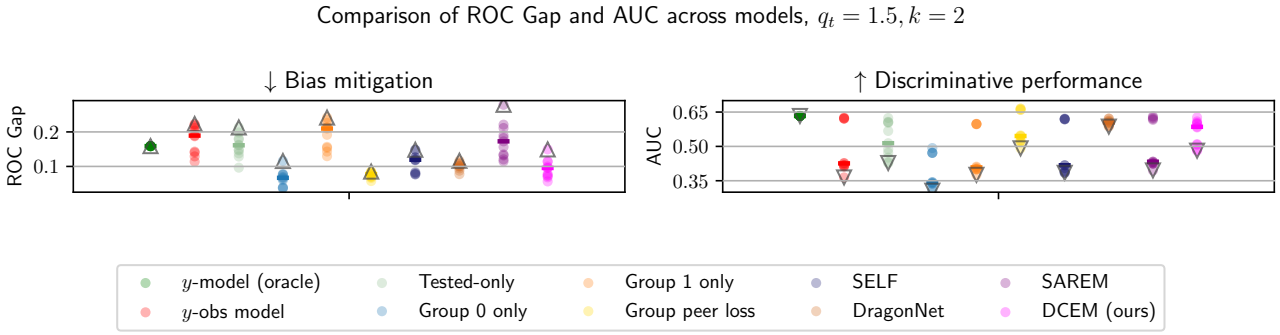

Figure 30. ROC gap (left) and AUC (right) of baselines on sepsis classification at $k = 2, q_t = 1.5$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

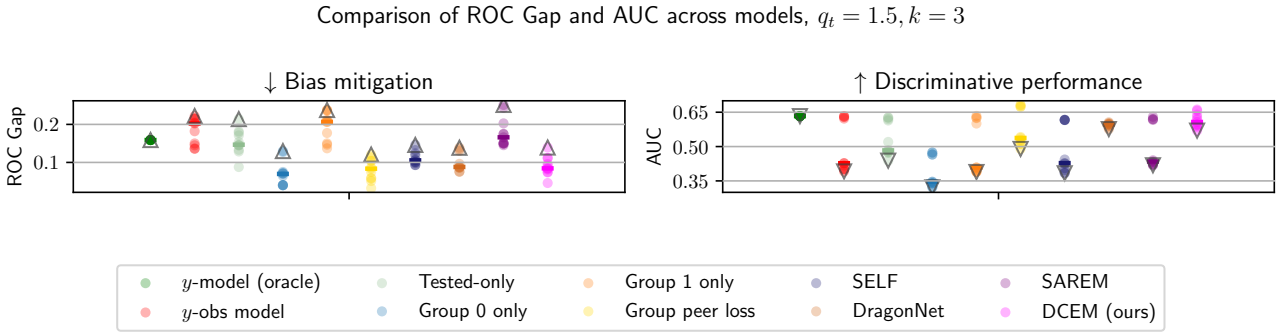

Figure 31. ROC gap (left) and AUC (right) of baselines on sepsis classification at $k = 3, q_t = 1.5$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

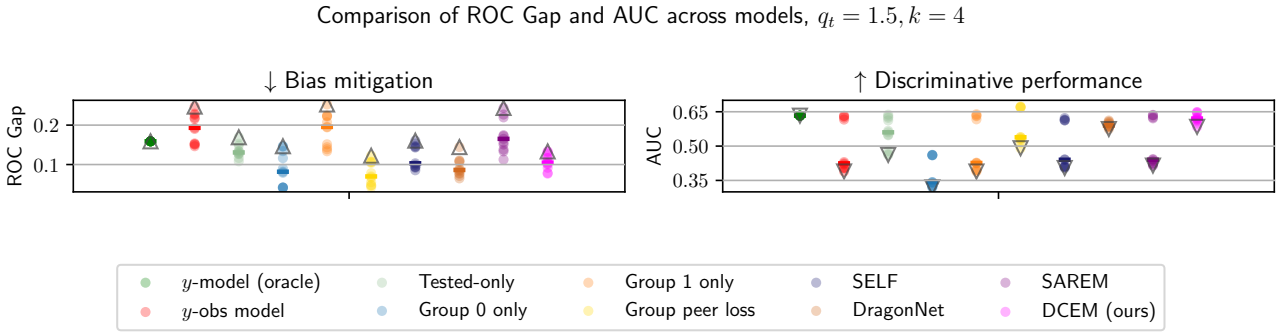

Figure 32. ROC gap (left) and AUC (right) of baselines on sepsis classification at $k = 4, q_t = 1.5$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

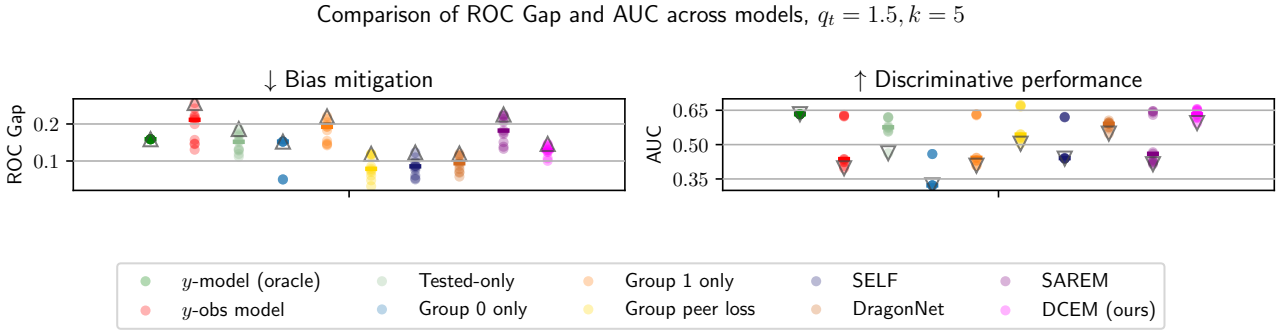

Figure 33. ROC gap (left) and AUC (right) of baselines on sepsis classification at $k = 5, q_t = 1.5$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.