
Positive-Unlabeled Regression: Learning from Partially Labeled Quantitative Outcomes

Paweł Teisseyre^{1,2}

Jan Mielniczuk^{1,2}

¹Faculty of Mathematics and Information Sciences, Warsaw University of Technology, Warsaw, Poland

²Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland

Abstract

We study a novel machine learning problem: Positive-Unlabeled Regression (PUR), which extends the classical Positive-Unlabeled (PU) learning framework to regression setting with a non-negative, discrete target variable. This setting arises naturally in applications where only some positive quantitative outcomes are reported, while the absence of a label may either indicate a true zero outcome or an unreported positive value. Applications include predicting the number of diseases a patient may have, the number of complications following an illness, or the advancement stage of a disease. We formalize the PUR problem and highlight the limitations of naive approaches that either use reported target variable or discard unlabeled data. To account for the inherent bias in such strategies, we propose two principled methods. The first is based on calibration of regression estimates using posterior probabilities from classical PU learning. The second builds on an empirical risk minimization framework, restating the target risk as a weighted function dependent on the instance-specific propensity score. We demonstrate both theoretically and empirically that the proposed approaches yield improved performance over standard baselines.

1 INTRODUCTION

1.1 MOTIVATION AND CONTRIBUTION

We consider a new learning problem for partially labeled data, hereinafter referred to as Positive-Unlabeled Regression (PUR), which generalizes the Positive-Unlabeled (PU) learning setting to the case of a quantitative target variable.

In classical PU learning [Elkan and Noto, 2008, Bekker

and Davis, 2020], the goal is to predict a *binary* target variable using training data consisting of labeled observations, which are guaranteed to be positive (i.e. belonging to a positive class), and unlabeled observations, which can be either positive or negative. PU learning has attracted significant interest in the ML community in recent years, resulting in the development of theoretical results regarding learning capabilities, necessary assumptions [Bekker and Davis, 2020, Gong et al., 2022], and the development of numerous algorithms for single-training sample and case-control setting [du Plessis et al., 2014, Kiryo et al., 2017, Sevetlidis et al., 2024, Jain et al., 2016, Chen et al., 2020, Bekker et al., 2019, Gerych et al., 2022, Gong et al., 2021a, Furmańczyk et al., 2023].

Unlike classical PU learning, in PU Regression, our goal is to predict a non-negative, quantitative, and usually discrete target variable $Y \in \{0, 1, 2, \dots\}$, based on a feature vector X . For example, we want to predict the number of complications that will arise for a patient after a myocardial infarction [Gong et al., 2021b], such as atrial fibrillation, pulmonary edema or Dressler syndrome, based on clinical, genetic, and lifestyle factors. In many practical situations, we do not have access to the ground-truth target variable Y , but only to the reported values, denoted by $\tilde{Y}$. If only such data is disponible, it should be taken into account that complications may not be reported for some patients. Then a value of $\tilde{Y} = 0$ could indicate either the absence of complications (negative case: $Y = 0$) or that they occurred (positive case: $Y > 0$) but were not reported. Therefore, a value of $\tilde{Y} = 0$ for a given observation is treated as an unlabeled case. Figure 1 shows an example of such data. When the number of unreported cases is large, a clear discrepancy between the true target variable and the reported variable occurs. In another setting, the variable Y may represent the number of attacks on a specific type of a computer network. This number may actually be equal to 0, but it will also be reported as 0 if the security systems fail to detect any attacks of such type. The probability of detection may depend, for example, on the type of security measures in place.

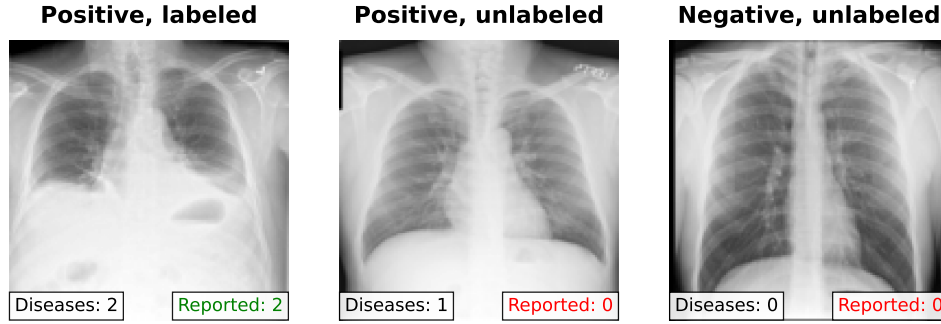


Figure 1: Illustration of the PU Regression problem using ChestMNIST data [Yang et al., 2023]. We distinguish three cases: positive-labeled ($Y, \tilde{Y} > 0$ and $\tilde{Y} = Y$), positive-unlabeled ($Y > 0$ and $\tilde{Y} = 0$), and negative-unlabeled ($Y = 0$ and $\tilde{Y} = 0$).

A final important example is the prediction of disease severity or progression (Y). This scenario is examined in our experiments, where, for the Parkinson’s disease dataset, we predict the UPDRS score (Unified Parkinson’s Disease Rating Scale), which reflects the severity of the disease. In this context, a zero value reported for a given indicator ($\tilde{Y} = 0$) may signify either the genuine absence of symptoms ($Y = 0$) or the absence of symptom reporting despite the symptoms actually being present ($Y > 0$).

A naive approach to this problem is to train a regression model using only the observed variable $\tilde{Y}$ as the target. However, this leads to a biased estimation of the true regression function corresponding to Y . Another natural strategy is to discard the unlabeled data and train the model solely on labeled instances, but this too results in biased estimates.

To overcome these limitations, we propose two alternative approaches. The first is based on calibrating the regression function estimated from labeled data. This calibration requires knowledge of the posterior probability of the positivity indicator, which can be estimated using techniques from classical PU learning. The second proposed approach is grounded in the empirical risk minimization (ERM) framework. We show that the theoretical risk associated with an unobserved response variable can be restated as a weighted risk function, where the weights depend on the propensity score, defined as the instance-specific probability that a positive observation is labeled. Given this propensity score or its estimators, the weighted risk function can be optimized using labeled data only. Propensity score estimation also relies on methods developed from PU learning.

Importantly, our methods assume that the labeling mechanism may depend on the feature vector, aligning with the Selected At Random (SAR) assumption commonly used in PU learning [Bekker et al., 2019, Gerych et al., 2022, Gong et al., 2021a, 2022, Furmańczyk et al., 2023]. However, effective learning from PU Regression data requires an additional assumption: the decision to assign a label to a positive instance of Y may depend on the corresponding

value of X , but not on the value of Y itself. This assumption is imposed for both proposed methods.

Our contributions are as follows. 1) We formalize a novel learning problem, Positive-Unlabeled Regression (PUR). 2) We introduce the assumption that enables effective learning from PUR data. 3) We establish a connection between PU regression and classical PU learning. 4) We propose two new methods, which are based on calibration and the ERM framework, and show experimentally that they outperform naive baselines.

1.2 PRIOR WORK

The Positive-Unlabeled Regression (PUR) scenario, despite its practical relevance, has not been studied before. However, key components of the methods proposed here, namely, calibration and empirical risk minimization, can be addressed in several competing ways. Both methods require the estimation of quantities within PU framework under the SAR assumption [Bekker et al., 2019, Gong et al., 2022]: posterior probability estimation is needed for the calibration method, while the propensity score is used in ERM (see Bekker and Davis [2020] for a review of PU learning and existing solutions). The available methods for posterior estimation under the SAR assumption include the SAREM method [Bekker et al., 2019], LBE [Gong et al., 2021a], TM [Furmańczyk et al., 2023] and PGLIN [Gerych et al., 2022]. Notably, all above methods also provide estimators for the propensity score.

Both proposed PU Regression methods rely on estimators of the regression function [Hastie et al., 2015]. This can be done either parametrically or non-parametrically. Here, to fix ideas, we consider two well-established approaches: linear regression and the Poisson regression model (see, e.g. Tutz [2012], Cameron and Trivedi [2013]). However, we emphasize that the proposed methods are generic and can be combined with a wide range of regression techniques, including non-parametric methods [Fan and Gijbels, 1996],

ensemble approaches [Duffy and Helmbold, 2002, Friedman, 2001] and neural networks [Bishop, 2024].

PU regression is related to hurdle and zero-inflated models [Lambert, 1992, Zhou et al., 2024] in that all three are designed for data with many observed zeros and use various mechanisms to explain them. However, in hurdle and zero-inflated models zeros are genuine outcome realizations generated by one or two stochastic processes, whereas in PU regression many observed zeros are unobserved positive outcomes caused by labeling mechanism.

Finally, PUR scenario may be considered as an example of missing data framework. More specifically, it is a special case of Missing Not At Random (MNAR) model for which missingness variable depends on both predictors and response [Kim and Yu, 2011, Zhao and Shao, 2015, Tchetgen Tchetgen and Wirth, 2017, Zhao and Ma, 2022]. However, inference for such a case requires that additional information is available, usually in the form of instrumental variable which depends on Y but not on missingness mechanism. The special type of MNAR considered here allows us to construct direct and effective solutions to the problem relying on inference for PU case and regression estimation for fully observable data.

2 POSITIVE UNLABELED REGRESSION

2.1 PROBLEM STATEMENT

We consider a partially observable regression scenario which formalizes practical situations discussed in the Introduction. At the population level, let $P(X, Y, S)$ be a trivariate probability distribution such that $Y \in \{0, 1, 2, \dots\}$ is a target variable, $X \in R^p$ is a vector of features used to predict Y , and $S \in \{0, 1\}$ is a random variable indicating whether the observation is labeled or not. We assume that if an observation is labeled, then it is definitely positive, i.e., $S = 1 \implies Y > 0$. Unlabeled observations can be either positive or equal 0, that is, for $S = 0$, variable $Y = 0$ or $Y > 0$. The observed target variable $\tilde{Y}$ can be conveniently written as $\tilde{Y} = Y \cdot S$. Thus, instead of samples from the distribution $P(X, Y)$, we observe samples from the distribution $P(X, \tilde{Y})$. We note that due to the imposed assumptions we have equivalence: $\tilde{Y} > 0 \iff S = 1$.

It will follow from the development, that assumption that Y is non-negative and discrete is not essential and can be replaced by assumption that Y is non-negative. The goal in PU Regression (PUR) is to estimate regression function $f(x) = E(Y|X = x)$ based on single training sample data $\mathcal{D}_{PUR} = \{(x_i, \tilde{y}_i) : i = 1, \dots, n\}$, which is an iid sample from the distribution $P(X, \tilde{Y})$.

2.2 PU REGRESSION VS PU LEARNING

Suppose we are interested in an easier task, which consists in predicting whether the variable Y takes positive values or the value 0, i.e., predicting the variable $W = \mathbb{1}\{Y > 0\}$. Note that $\mathcal{D}_{PU} = \{(x_i, s_i), i = 1, \dots, n\}$, where $s_i = \mathbb{1}\{\tilde{y}_i > 0\}$ is a Positive Unlabeled classification data corresponding to distribution $P(X, W)$. More precisely, these are Positive-Unlabeled data corresponding to the single-training sample scheme described in [Elkan and Noto, 2008, Bekker and Davis, 2020]. The observation follows immediately after noting that $S = 1$ implies that $W = 1$, however $S = 0$ corresponds to either $W = 1$ or $W = 0$.

It's important to emphasize, that the goals of PU learning and PU regression are different. Adopting a probabilistic perspective, the goal of PU learning is to estimate the posterior probability $P(W = 1|X = x)$, which then allows to define a classification rule. In contrast, the goal of PU regression is to estimate $f(x) = E(Y|X = x)$. Both tasks coincide only when Y is binary, as in this case $P(W = 1|X = x) = E(Y|X = x)$. However, as will be demonstrated in the following sections, PU learning methods will be used as necessary ingredients to solve the PU regression problems. More precisely, we will use the propensity score estimation methods developed for PU learning, as well as the estimation methods of $P(W = 1|X = x)$, in the proposed calibration method.

2.3 LABELING MECHANISM

A key element in PUR is the modeling of the random labeling mechanism, i.e. the mechanism that determines which positive observations are assigned a label. The fundamental concept is the propensity score function, which for scenario discussed above is defined as

$$e(x) = P(S = 1|Y > 0, X = x),$$

i.e. it is the observation-specific probability that a positive observation is assigned a label. Furthermore, the labeling frequency, $c = E_X e(X) = P(S = 1|Y > 0)$, is the probability of being labeled among all positive observations. We assume throughout that probability of labeling $e(x)$ is positive for any x such that $P(Y > 0|X = x) > 0$. This means that labeling of Y is possible for any x such that positive values of Y occur for such x . Since the estimation of the propensity score is demanding, some simplifying assumptions are usually made. The most restrictive is the SCAR (Selected Completely at Random) assumption, which states that the propensity score function is constant $e(x) = c$ or, equivalently, that the labeling mechanism does not depend on the feature vector x . Note that this scenario corresponds to usual SCAR assumption for PU data sampled from $P(X, W)$. The following lemma, useful for further consideration, shows the relationship between the feature

distributions for positive and labeled observations and the relationship between the posterior probabilities corresponding to S and W .

Lemma 1. *For the PU Regression scenario, the following properties hold:*

$$P(X = x|W = 1) = P(X = x|S = 1)\frac{c}{e(x)} \quad (1)$$

$$P(S = 1|X = x) = P(W = 1|X = x)e(x) \quad (2)$$

The first equality follows from equation (3) in [Bekker and Davis, 2020] as $P(X, S)$ was identified above as the underlying PU distribution associated with $P(X, W)$. In addition, equality (2) is a basic property valid in the PU single-sample framework [Bekker and Davis, 2020].

We now introduce an additional assumption, which will be used for learning from the PUR data.

Assumption (A). *The random variables Y and S are conditionally independent, given X and $W = 1$:*

$$Y \perp S|X, W = 1$$

The above assumption means that the decision on whether the positive value of Y will be labeled depends on the corresponding value of X , but not on the value of Y . This means that Missing At Random (MAR) assumption is satisfied on strata $W = 1$. Condition (A) is equivalent to

$$P(Y = y|S = 1, W = 1, X) = P(Y = y|W = 1, X),$$

as the above equality implies the analogous equality with $S = 1$ replaced by $S = 0$. We also have

$$\begin{aligned} P(Y = y|W = 1, X = x) \\ &= P(Y = y|S = 1, W = 1, X = x) \\ &= P(Y = y|S = 1, X = x) \end{aligned} \quad (3)$$

where the first equality is due to assumption (A) and the second to the fact that $S = 1$ implies $W = 1$. We note that when Y is binary, the assumption (A) is always satisfied. It is worthwhile to note what type of restrictions is imposed on the framework by assuming (A). E.g., when Y corresponds to the number of actual migraine attacks experienced and $\tilde{Y}$ to the number reported using dedicated smartphone application, (A) is satisfied when reporting depends solely on patient's characteristics such as proficiency in using the app or age but not on the number of attacks actually experienced.

As mentioned above the considered model with (A) assumption turns out to be a special case of MNAR model. Indeed, we have $P(S = 1|X, Y = 0) = 0$ and $P(S = 1|X, Y > 0)$ is a function of X only. Finally, we state the theorem, which shows the relationship between the distributions of (X, Y) for labeled and positive data. This generalizes Lemma 1 provided Assumption (A) holds. The theorem is used to justify a method based on the weighted empirical risk function minimization approach described in Section 3.3.

Theorem 2. *Assume that condition (A) holds. We then have:*

$$\begin{aligned} P(x = x, Y = y|W = 1) \\ &= P(X = x, Y = y|S = 1)\frac{c}{e(x)} \end{aligned}$$

Proof. We have

$$\begin{aligned} P(X = x, Y = y|W = 1) \\ &= P(X = x|W = 1)P(Y = y|X = x, W = 1) \\ &= \frac{c}{e(x)}P(X = x|S = 1)P(Y = y|X = x, S = 1) \\ &= \frac{c}{e(x)}P(X = x, Y = y|S = 1), \end{aligned}$$

where the second equality follows from the Lemma 1 and (3). $\square$

3 ESTIMATION OF THE REGRESSION FUNCTION

3.1 BASELINE ALGORITHMS: NAIVE AND LABELED SET METHODS

There are two direct and simple approaches to tackle the PUR problem: naive method and labeled set method. As shown in the following, both are inherently biased.

Naive method involves fitting a regression model to the observed data $\mathcal{D}_{PUR}$, which allows us to estimate the expected value $\tilde{f}(x) = E(\tilde{Y}|X = x)$. This will obviously lead to a biased estimate of $f(x)$ as shown by the following lemma (its proof is relegated to the supplement).

Lemma 3. *Let $\tilde{f}(x) = E(\tilde{Y}|X = x)$. Then $\tilde{f}(x) = f(x) - b(x)$, where*

$$\begin{aligned} b(x) &= E(Y|X = x, S = 0)P(S = 0|X = x) = \\ &= E(Y|X = x, S = 0)[1 - e(x)P(W = 1|X = x)] \end{aligned}$$

The above lemma indicates that the consistent estimator of $\tilde{f}(x)$ will inherently underestimate the true value of $f(x)$. When $e(x)$ is approximately 1, which means that most positive cases will be assigned a label, the bias becomes negligible.

Labeled Set (LS) method uses labeled data $\mathcal{D}_{PUR}^+ = \{(x_i, \tilde{y}_i) : s_i = 1\}$ to estimate $f_{LS}(x) = E(\tilde{Y}|X = x, S = 1) = E(Y|X = x, S = 1)$. Of course, $f_{LS}(x)$ usually differs from $f(x)$, so this method is also biased. It follows from Theorem 4 and Lemma 3 that $f_{LS}(x) \geq f(x) \geq \tilde{f}(x)$. Also, it is seen from (3) that $\mathcal{D}_{PUR}^+$ is an iid sample from the $P(X, Y|\tilde{Y} > 0)$ distribution and not from the distribution of $P(X, Y|Y > 0)$, which would correspond to inference from truncated data, what prevents direct application of methods designed for such a case.

3.2 REGRESSION FUNCTION CALIBRATION

Direct estimation of the regression function $f(x) = E(Y|X = x)$ based on training data $\mathcal{D}_{PUR} = \{(x_i, \tilde{y}_i) : i = 1 \dots, n\}$ is not possible because we do not observe the true target values y_i . However, the following result shows that the expected value of interest can be expressed in terms of quantities that we can estimate using the available data. This theorem gives rise to the proposed calibration method.

Theorem 4. *Assume that condition (A) holds. We then have the following equality*

$$E(Y|X = x) = \underbrace{E(Y|X = x, S = 1)}_{=f_{LS}(x)} P(W = 1|X = x)$$

Proof. The proof follows from noting that (3) implies that $P(Y = y|X = x)$ may be equivalently written as

$$\begin{aligned} P(Y = 0|X = x), & \text{ if } y = 0, \\ P(Y = y|X = x, S = 1) P(W = 1|X = x), & \text{ if } y > 0. \end{aligned}$$

and calculating $E(Y|X = x)$ using the above equality. $\square$

Theorem 4 gives rise to the natural estimator of $f(x)$, which is the calibration-based estimator of the regression function using labeled data, namely

$$\hat{f}(x) = \hat{E}(Y|X = x, S = 1) \hat{P}(W = 1|X = x).$$

This corresponds to two-part method with positive-part regression fitted on $S = 1$ and zero-part via PU-estimated $P(W = 1|X = x)$. Indeed, note that $f_{LS}(x) = E(Y|X = x, S = 1)$ can be estimated by any classical regression estimator (e.g. linear or Poisson regression, tree-based algorithms or neural networks) using labeled data $\mathcal{D}_{PUR}^+ = \{(x_i, \tilde{y}_i) : s_i = 1\}$. Secondly, $P(W = 1|X = x)$ is posterior probability in classification problem when classes are given by $W = 0, 1$ and can be estimated using any estimator of such probability for PU single sample data. In our experiments, $P(W = 1|X = x)$ is estimated using a parametric model and by minimizing the weighted empirical risk, as detailed in supplement B. This is a standard approach used in biased PU learning [Bekker and Davis, 2020, Gerych et al., 2022].

3.3 WEIGHTED EMPIRICAL RISK MINIMIZATION

Another approach is to consider an Empirical Risk Minimization (ERM) framework, in which $h_\theta(x)$ is the output of a parametric regression model (e.g. linear, Poisson, or neural network), θ is a vector of parameters, and $l(y, h_\theta(x))$ denotes the loss function. In ERM approach, the goal is to

minimize the empirical risk function, which corresponds to the theoretical risk

$$R(h_\theta) = E_{X,Y} l(Y, h_\theta(X)). \quad (4)$$

Again, direct optimization of the empirical equivalent of (4) is not possible because we do not have access to the true values of Y . The proposed approach generalizes ERM method described in Bekker and Davis [2020] and falls into category of Weighted Empirical Risk Minimisation [Vogel et al., 2020]. It is based on the theorem below (see the proof in the supplement A.2), which shows that, under the additional assumption (A), the risk function (4) can be expressed in terms of the expected value for the labeled data. The similar approach in semi-supervised setting is called Inverse Propensity Weighting (IWP) [Su et al., 2024] or, more generally, Weighted Scoring Rules (WSR) approach [Hernandez Orallo, 2014].

Theorem 5. *Assume that condition (A) is satisfied. Let $\pi_S = P(S = 1)$. Then*

$$\begin{aligned} R(h_\theta) &= E_{X,Y} l(Y, h_\theta(X)) = \\ &\pi_S E_{X,Y|S=1} \left\{ \frac{1}{e(X)} [l(Y, h_\theta(X)) - l(0, h_\theta(X))] \right\} \\ &+ E_X l(0, h_\theta(X)) \end{aligned}$$

Importantly, an empirical counterpart of the risk given in Theorem can be optimized provided the propensity score is known. This is possible because for labeled observations ($S = 1$), we observe both X and the true values of variable Y . Moreover, it follows that in the case of, e.g. quadratic loss or Poisson loss $l(y, h_\theta(x)) = \exp(x^T \theta) - yx^T \theta$ and when the model is well specified, minimisation of the expression in Theorem 4 yields the true regression $E(Y|X = x)$. The empirical unbiased risk function, corresponding to (4) is given as

$$\begin{aligned} \hat{R}(h_\theta) &= \frac{1}{n} \sum_{i:s_i=1} \left\{ \frac{1}{e(x_i)} [l(\tilde{y}_i, h_\theta(x_i)) - l(0, h_\theta(x_i))] \right\} \\ &+ \frac{1}{n} \sum_{i=1}^n l(0, h_\theta(x_i)). \end{aligned} \quad (5)$$

For simple models, such as the linear model or the Poisson model, the function can be easily optimized.

Consider first a linear model in which $h_\theta(x) = x^T \theta$ and a loss function $l(y, h_\theta(x)) = (y - h_\theta(x))^2$. Moreover, let $w_i = e(x_i)^{-1}$. Then, the empirical risk function (5) can be

written as follows

$$\begin{aligned}
\widehat{R}(h_\theta) &= \\
&= \frac{1}{n} \sum_{i:s_i=1} w_i [(\tilde{y}_i - x_i^T \theta)^2 - (x_i^T \theta)^2] + \frac{1}{n} \sum_{i=1}^n (x_i^T \theta)^2 \\
&\stackrel{(i)}{=} \frac{1}{n} \sum_{i=1}^n w_i [(\tilde{y}_i - x_i^T \theta)^2 - (x_i^T \theta)^2] + \frac{1}{n} \sum_{i=1}^n (x_i^T \theta)^2 \\
&= \frac{1}{n} \sum_{i=1}^n w_i \tilde{y}_i^2 - \frac{2}{n} \sum_{i=1}^n w_i \tilde{y}_i (x_i^T \theta) + \frac{1}{n} \sum_{i=1}^n (x_i^T \theta)^2
\end{aligned}$$

where (i) follows from the fact that for $s_i = 0$ we have $\tilde{y}_i = 0$. The gradient is given by

$$\begin{aligned}
\nabla_\theta \widehat{R}(h) &= -\frac{2}{n} \sum_{i=1}^n w_i \tilde{y}_i x_i + \frac{1}{n} \sum_{i=1}^n 2(x_i^T \theta) x_i \\
&= -\frac{2}{n} \mathbb{X}^T \mathbb{W} \tilde{\mathbb{Y}} + \frac{2}{n} \mathbb{X}^T \mathbb{X} \theta,
\end{aligned}$$

where $\mathbb{X}$ is a matrix containing feature vectors in rows, $\mathbb{W}$ is a diagonal matrix with elements $e(x_1)^{-1}, \dots, e(x_n)^{-1}$, and $\tilde{\mathbb{Y}} = (\tilde{y}_1, \dots, \tilde{y}_n)$. By equating the gradient to zero, we obtain the solution

$$\hat{\theta} = \arg \min_{\theta} \widehat{R}(h_\theta) = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbb{W} \tilde{\mathbb{Y}}. \quad (6)$$

We note that (6) is the standard least squares solution in which the response vector $\tilde{\mathbb{Y}}$ is replaced by a scaled vector $\mathbb{W} \tilde{\mathbb{Y}}$. The scaling mechanism works in such a way that observations reported as positive ($\tilde{Y} > 0$), for which the propensity score takes small values, are assigned a higher value of the scaled response $\mathbb{W} \tilde{\mathbb{Y}}$.

Consider now a Poisson model in which $h_\theta(x) = \exp(x^T \theta)$ and a loss function $l(y, h_\theta(x)) = \exp(x^T \theta) - y \cdot x^T \theta$. It is easily seen that such loss function is Fisher consistent, that is, if the true parameter is θ^* i.e. $E(Y|X = x) = e^{x^T \theta^*}$, then $\arg \min_{\theta} E(Y, h_\theta(X)) = \theta^*$. Note that in this case, $l(0, h_\theta(x)) = \exp(x^T \theta)$ and $l(y, h_\theta(x)) - l(0, h_\theta(x)) = -y \cdot (x^T \theta)$. Therefore, the empirical risk function can be written as

$$\widehat{R}(h_\theta) = \frac{1}{n} \sum_{i:s_i=1} (-1) \tilde{y}_i w_i (x_i^T \theta) + \frac{1}{n} \sum_{i=1}^n \exp(x_i^T \theta)$$

The gradient has the form

$$\nabla_\theta \widehat{R}(h_\theta) = \frac{1}{n} \sum_{i:s_i=1} (-1) \tilde{y}_i w_i x_i + \frac{1}{n} \sum_{i=1}^n \exp(x_i^T \theta) x_i.$$

Using the above formula, one can use gradient algorithms (GD, SGD, ADAM) to find the optimal solution.

4 EXPERIMENTS

The aim of the experiments is to compare the methods described above. We investigate the impact of parameters such as labeling frequency, class prior, data labeling scheme, choice of the underlying regression model, and propensity score estimation on prediction quality.

4.1 METHODS

We examine the effectiveness of the proposed methods: the calibration method (CAL) and the empirical risk minimization method (ERM). We compare them with baseline approaches: the NAIVE method and LS method. We consider two base regression models: a linear model and a Poisson model, the latter being particularly suitable for discrete response variable. In the latter case, for the LS and CAL methods, we use the truncated Poisson model to estimate $f_{LS}(x) = E(\tilde{Y}|X = x, S = 1) = E(\tilde{Y}|X = x, \tilde{Y} > 0)$, which is natural in this situation. The CAL and ERM require knowledge of the propensity score function $e(x)$ or its estimator. To address this, we examine two variants: one that assumes that $e(x)$ is known, and another that estimates it using the SAR-EM method [Bekker et al., 2019], a standard estimation approach discussed in the literature. The estimated variants are denoted as CAL (EST) and ERM (EST). An analysis of other propensity score estimation methods, demonstrating that SAREM is a reasonable choice (especially for low labeling frequencies), is presented in the supplement (Table 11). Technical details of the implementation are given in the supplement D.

4.2 DATASETS

We focus on datasets that are naturally aligned with the PU Regression task. Detailed descriptions of the datasets and preprocessing steps are provided in the Supplement C. In the MIMIC dataset [Johnson et al., 2016], the objective is to predict the number of diseases a patient will develop. In the Complications dataset [Golovenkin et al., 2020], the task is to predict the number of complications a patient will experience following a heart attack. The Obesity dataset [Palechor and De la Hoz Manotas, 2019] is used to predict obesity levels, while in the Parkinson dataset [Erdogdu Sakar et al., 2013], the target variable is the Unified Parkinson's Disease Rating Scale (UPDRS), which quantifies the severity of Parkinson's disease. In the Chest dataset [Yang et al., 2023] our objective is to predict the number of diseases based on X-ray scans of the lungs. Across all datasets, the target variable takes discrete, nonnegative values, with 0 indicating the absence of disease. Additionally, we consider two artificial datasets: Art1 and Art2; the method of generating them is given in the Supplement C.

4.3 EXPERIMENTAL FRAMEWORK

Each dataset is split into training and test subsets in the proportion of 0.5. From the fully labeled training set, we construct a PU regression dataset by assigning the label $S = 1$ to positive observations $Y > 0$, according to the probability specified by the propensity score function $e(x) = P(S = 1|X = x, Y > 0)$. The following labeling strategies are considered.

Strategy S1: $e(x) = c$, where c is a parameter representing the labeling frequency.

Strategy S2: $e(x) = \sigma(\theta_0 + x^T \theta)$, where θ_0, θ are parameters and σ is logistic sigmoid function.

Strategy S3: $e(x) = [\sigma(\theta_0 + x^T \theta)]^{10}$, where θ_0, θ are parameters.

Strategy S1 corresponds to the SCAR assumption in PU learning, whereas S2-S3 to SAR assumption. For S2-S3, the vector θ is estimated using a logistic model fitted to the variables W, X . Given θ , the intercept θ_0 is chosen to control the labeling frequency c . For example, in the case of S2, θ_0 is determined as an approximate solution of the equation $c = E_X \sigma(\theta_0 + X^T \theta)$, for fixed c and θ . Strategies S2 and S3 were used e.g. in Gong et al. [2021a]. In addition, we consider strategy S4, in which the labeling mechanism depends on the value of Y and thus assumption (A) is not met.

Strategy S4 assumes that $P(S = 1|X = x, Y = y > 0) = \sigma(\theta_0 + x^T \theta)(\frac{y}{M} + \epsilon)(1 + \epsilon)^{-1}$, where θ_0, θ are parameters, M is the maximum value achieved by Y for a given data set and $\epsilon = 0.001$.

An important aspect of the setup is to control the prior probability $\pi = P(Y > 0)$. To ensure a fixed value of π , we applied downsampling of the sample (X, Y) prior to the train-test split. For each method, we compute the Root Mean Square Error (RMSE) on the test set. The procedure is repeated 20 times, and the results are averaged.

4.4 DISCUSSION

In the main part of the paper, we show the values of prediction error (RMSE) for schemes S1–S4, assuming a linear regression model as the base model (Table 1 and Figure 2). The complete results, including the Poisson model, the effect of labeling frequency, the impact of propensity score estimation, and the computation times, are provided in the Supplement.

Our primary goal was to compare the performance of the proposed methods (CAL and ERM) against the baselines (LS and NAIVE). First, for most datasets and labeling strategies, we observe that CAL and ERM outperform the NAIVE and LS methods. This advantage is particularly noticeable

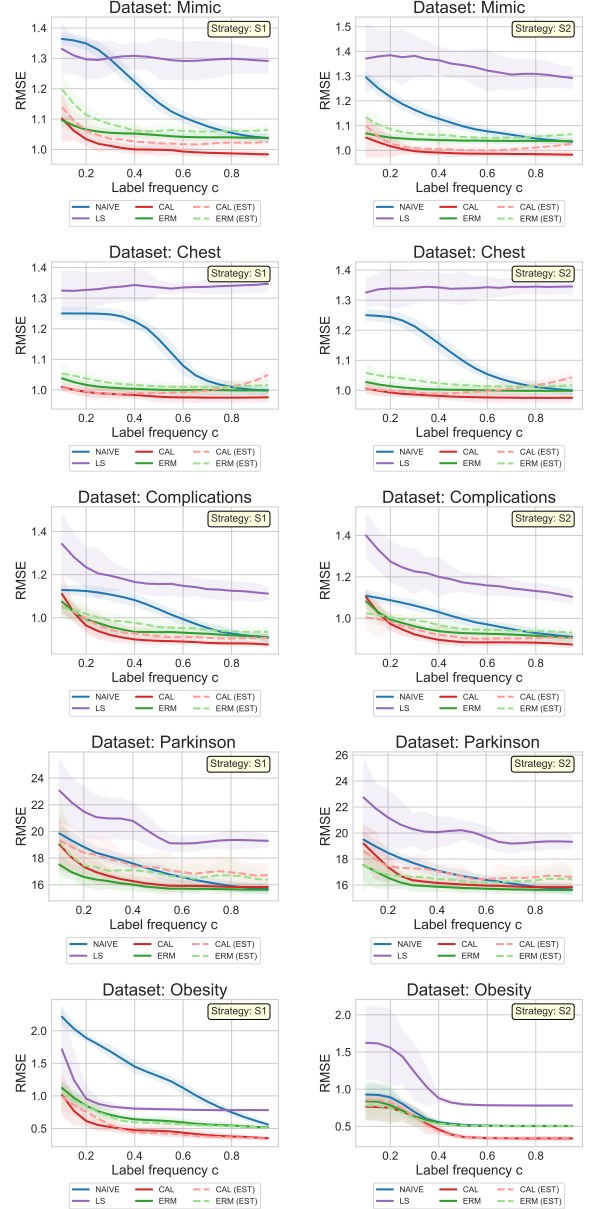


Figure 2: Root Mean Squared Errors with respect to label frequency $c = P(S = 1|Y > 0)$ for labeling strategies S1-S2. Parameter $\pi = P(Y > 0) = 0.5$ and linear model is used to estimate regression function.

at low labeling frequencies. When the linear model is used, CAL performs slightly better than ERM for most datasets, with the exception of the Parkinson dataset, where ERM turns out to be the better method (Figure 2 and Table 1). When using the Poisson regression, the situation changes, and for low label frequencies ERM performs noticeably better than CAL on three datasets (Complications, Parkinson and MIMIC), see Tables 7-10 in the Supplement.

In most of the settings considered, the NAIVE method performs better than LS, with the exception of the Obe-

Table 1: Root mean square error (RMSE) for the considered methods, labeling strategies S1-S4, $c = 10\%$ and for $\pi = 0.5$. A **linear model** was used as the basic regression model. The method with the lowest RMSE is highlighted in bold, and the second best is in light gray. The methods CAL and ERM are excluded from comparison as they assume known $e(x)$.

	Dataset	NAIVE	LS	CAL (EST)	ERM (EST)	CAL	ERM
S1	Art1	3.54 ± 0.11	0.73 ± 0.01	0.46 ± 0.18	0.49 ± 0.09	0.56 ± 0.17	0.64 ± 0.12
	Art2	1.86 ± 0.32	1.43 ± 0.2	1.41 ± 0.26	1.38 ± 0.14	1.26 ± 0.18	1.26 ± 0.18
	Mimic	1.36 ± 0.02	1.33 ± 0.08	1.14 ± 0.04	1.2 ± 0.05	1.1 ± 0.08	1.1 ± 0.02
	Parkinson	19.86 ± 0.5	23.07 ± 2.39	19.35 ± 2.1	19.22 ± 3.67	19.02 ± 1.89	17.51 ± 0.78
	Obesity	2.21 ± 0.13	1.71 ± 0.66	0.99 ± 0.22	1.04 ± 0.18	1.02 ± 0.3	1.12 ± 0.09
	Compl.	1.13 ± 0.02	1.34 ± 0.14	1.05 ± 0.05	1.05 ± 0.05	1.11 ± 0.09	1.07 ± 0.05
	Chest	1.25 ± 0.02	1.32 ± 0.06	1.01 ± 0.02	1.05 ± 0.03	1.01 ± 0.01	1.04 ± 0.01
S2	Art1	0.15 ± 0.02	0.71 ± 0.01	0.18 ± 0.02	0.12 ± 0.02	0.19 ± 0.03	0.17 ± 0.02
	Art2	1.5 ± 0.19	1.32 ± 0.11	1.12 ± 0.08	1.22 ± 0.17	1.11 ± 0.08	1.29 ± 0.17
	Mimic	1.3 ± 0.02	1.37 ± 0.14	1.1 ± 0.05	1.13 ± 0.03	1.05 ± 0.08	1.07 ± 0.02
	Parkinson	19.5 ± 0.56	22.74 ± 2.97	18.62 ± 1.25	17.57 ± 1.99	19.19 ± 1.41	17.54 ± 1.24
	Obesity	0.93 ± 0.14	1.62 ± 0.49	0.86 ± 0.09	0.78 ± 0.12	0.77 ± 0.17	0.84 ± 0.26
	Compl.	1.11 ± 0.02	1.4 ± 0.11	1.0 ± 0.06	1.02 ± 0.03	1.1 ± 0.04	1.08 ± 0.04
	Chest	1.25 ± 0.02	1.33 ± 0.05	1.01 ± 0.02	1.06 ± 0.03	1.01 ± 0.01	1.03 ± 0.01
S3	Art1	0.21 ± 0.06	0.86 ± 0.29	0.25 ± 0.02	0.15 ± 0.02	0.25 ± 0.06	0.17 ± 0.04
	Art2	1.41 ± 0.24	1.2 ± 0.08	1.22 ± 0.15	1.27 ± 0.14	1.18 ± 0.09	1.28 ± 0.15
	Mimic	1.25 ± 0.02	1.46 ± 0.15	1.11 ± 0.03	1.14 ± 0.02	1.14 ± 0.13	1.1 ± 0.02
	Parkinson	19.08 ± 0.47	21.81 ± 2.1	19.01 ± 1.27	17.34 ± 0.89	19.54 ± 1.63	17.63 ± 1.08
	Obesity	1.44 ± 0.22	3.19 ± 0.31	1.3 ± 0.14	1.29 ± 0.14	1.25 ± 0.12	1.27 ± 0.15
	Compl.	1.09 ± 0.02	1.35 ± 0.11	1.03 ± 0.05	1.03 ± 0.03	1.11 ± 0.07	1.14 ± 0.15
	Chest	1.25 ± 0.02	1.35 ± 0.07	1.04 ± 0.03	1.08 ± 0.02	1.05 ± 0.08	1.07 ± 0.04
S4	Art1	0.9 ± 0.07	0.72 ± 0.02	0.46 ± 0.22	0.3 ± 0.17	0.38 ± 0.13	0.4 ± 0.26
	Art2	1.69 ± 0.27	1.72 ± 0.21	1.23 ± 0.08	1.32 ± 0.28	1.42 ± 0.29	1.26 ± 0.23
	Mimic	1.25 ± 0.03	1.62 ± 0.02	1.06 ± 0.03	1.1 ± 0.02	1.06 ± 0.03	1.05 ± 0.01
	Parkinson	18.23 ± 0.53	26.46 ± 1.86	19.67 ± 2.9	18.19 ± 2.13	17.89 ± 0.84	16.25 ± 0.58
	Obesity	1.13 ± 0.13	2.49 ± 0.5	0.94 ± 0.1	0.9 ± 0.11	1.04 ± 0.24	1.01 ± 0.36
	Compl.	1.09 ± 0.03	1.53 ± 0.1	0.96 ± 0.04	0.99 ± 0.03	1.01 ± 0.05	0.97 ± 0.03
	Chest	1.25 ± 0.02	1.57 ± 0.03	0.98 ± 0.01	1.03 ± 0.01	1.01 ± 0.01	1.01 ± 0.01

sity data set under strategy S1, where the NAIVE method clearly performs the worst (Figure 2). Recall that for the NAIVE method, unlabeled instances with positive target values are incorrectly treated as if the response variable was zero. Therefore, the NAIVE method performs especially poorly when the labeling frequency c is low, as this results in a large number of such misclassified observations. This is reflected in the results.

It is also worth noting the superiority of the proposed CAL and ERM methods under strategy S4 in scenarios where assumption (A) is violated (Tables 1, 6 and 10). This suggests that CAL and ERM may be robust to violations of this assumption.

The frequency of labeling c significantly affects the performance of all methods. RMSE generally decreases with increasing c for most methods, regardless of other parameters (prior probability π , labeling strategy, or choice of base regression model). This is expected, as for low c , only a

small fraction of positive instances are labeled, leading to a substantial discrepancy between the true response variable Y and the observed variable $\tilde{Y}$. When the labeling frequency approaches 1, the NAIVE, CAL, and ERM methods perform very similarly. In some cases, the NAIVE method can even outperform the proposed methods (see e.g. Table 7). This is due to the fact that, for $c \approx 1$, the NAIVE method performs similarly to regression estimator with known Y , whereas proposed methods suffer from additional variability due to estimation of propensity score, which in this case is unnecessary as it is approximately 1. These conclusions are consistent with similar findings in the context of classification methods used in traditional PU learning [Bekker and Davis, 2020].

The impact of estimating propensity score $e(x)$ on the performance of the CAL and ERM methods is analysed. As expected, the CAL (EST) and ERM (EST) methods, in which $e(x)$ is estimated, perform slightly worse than their coun-

terparts with known $e(x)$ (CAL and ERM), see Figure 2. However, the degradation in performance is not substantial. Moreover, for some datasets (e.g., Obesity), the CAL and CAL (EST) methods perform almost identically. Similarly, for the Obesity dataset, the ERM and ERM (EST) methods yield very similar results.

It is worth noting that, under labeling strategies S2–S4, there exists a functional positive dependence between the propensity score function $e(x)$ and the true posterior probability $P(W = 1|X = x)$, since both are determined by the same linear combination $x^T\theta$. This assumption is natural in many practical applications. Consider, for example, the tagging of web pages: certain keywords may simultaneously indicate that a page is likely to be relevant to a particular topic and increase the probability that the page will be tagged. Consequently, concepts such as Invariance of Order (IOO), which assumes that the propensity score and the true posterior probability are co-monotonic, are of particular interest [Kato et al., 2019, Gerych et al., 2022]. Nevertheless, scenarios in which the variables influencing the response differ from those governing the labeling mechanism are also relevant in practice. To investigate such situations, we conducted an additional experiment using the MIMIC dataset. Specifically, we modified the S2 labeling scheme by defining the propensity function as $e(x) = \sigma(\theta_0 + X_k)$, where the covariate X_k used in the propensity model was the one exhibiting the weakest correlation with the target variable Y and θ_0 is chosen to control the value of c . The results obtained under this modified setting were very similar to those observed for the original S2 scheme. For example, when $c = 10\%$, the RMSE values obtained for the modified labeling scheme were 1.36 (NAIVE), 1.33 (LS), 1.11 (CAL (EST)), and 1.16 (ERM (EST)). These values are very close to those obtained under sampling scheme S2, namely 1.30 (NAIVE), 1.33 (LS), 1.10 (CAL (EST)), and 1.13 (ERM (EST)). The superiority of the proposed ERM and CAL methods over the NAIVE and LS approaches remains evident. It is also worth noting that, among all methods considered for estimating $P(W = 1|X = x)$, only the PGLIN method relies on the co-monotonicity assumption. Importantly, our final method of choice, SAR-EM [Bekker et al., 2019], does not require this assumption.

An analysis of the influence of the prior probability π on the results shows that it has the greatest impact on the RMSE of the LS method (Figures 3 and 4). This is consistent with expectations, as the LS relies solely on labeled observations. For small values of π , the training set contains very few positive observations, resulting in a limited number of labeled instances, especially when c is also small. The results for the largest dataset (MIMIC) show that LS performs clearly the worst for $\pi = 0.3$ and $\pi = 0.5$, while for $\pi = 0.7$, it outperforms the NAIVE method. In the case of $\pi = 0.7$ and low c , a more pronounced gap is observed between the NAIVE method and the proposed methods (CAL and

ERM).

We also report computational times (Table 12 in the Supplement) indicating that for CAL and ERM methods they are determined by times of estimation of propensity score.

Finally, we compared the performance of two base regression models: linear and Poisson. The results indicate that the main conclusions outlined above hold similarly for both base models.

5 CONCLUSIONS AND FUTURE WORK

In this work, we introduced Positive-Unlabeled Regression (PUR), a novel extension of the classical PU learning problem. We proposed two methods for learning in the PUR setting: CAL and ERM. Both approaches rely on the conditional independence assumption, under which the probability of observing a label depends on the feature vector X , but not on the actual value of the target variable Y . Extensive empirical evaluations on a variety of datasets and labeling mechanisms confirmed the effectiveness of the proposed methods. CAL and ERM consistently outperformed baseline methods (NAIVE and LS), especially under low labeling frequency, where incorrect assumptions made by the baselines led to substantial degradation of performance. Estimating the propensity score function $e(x)$ resulted in only a minor performance loss compared to scenarios where $e(x)$ is known, which supports the feasibility of deploying these methods in practice. An important component of this work is the theoretical analysis of the PUR framework. In particular, we demonstrated that the regression function for the true response variable can be expressed as a suitably rescaled regression function based on labeled data, which forms the foundation of the CAL method. The second key result shows that the risk of a regression estimator for unobserved target variable can be represented as a weighted risk function based on labeled data. These two results may serve as a foundation for developing many new learning methods in the PUR problem.

In this study, we combined CAL and ERM with relatively simple regression models (linear and Poisson). In future research, it would be of interest to explore more complex regression models based on ensembles or neural networks. Also, we believe it would be worthwhile to apply the developed methodology to other tasks, such as quantile regression [Koenker, 2005] for PU setting, where the use of alternative loss functions enables the estimation of conditional quantiles rather than regression.

References

J. Bekker and J. Davis. Learning from positive and unlabeled data: a survey. *Machine Learning*, 109:719–760, 2020.

- J. Bekker, P. Robberechts, and J. Davis. Beyond the Selected Completely At Random Assumption for Learning from Positive and Unlabeled Data. In *Proceedings of the 2019 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, ECML'19*, pages 71–85, 2019.
- Ch. Bishop. *Deep Learning*. Springer, 2024.
- A. C. Cameron and P. K. Trivedi. *Regression analysis of count data*. Cambridge University Press, 2013.
- X. Chen, W. Chen, T. Chen, Y. Yuan, C. Gong, K. Chen, and Z. Wang. Self-PU: Self boosted and calibrated positive-unlabeled training. In *Proceedings of the 37th International Conference on Machine Learning, ICML'20*, 2020.
- M. C. du Plessis, G. Niu, and M. Sugiyama. Analysis of learning from positive and unlabeled data. In *Proceedings of the International Conference on Neural Information Processing Systems, NIPS'14*, pages 703–711, 2014.
- N. Duffy and D. Helmbold. Boosting methods for regression. *Machine Learning*, 47(2):153–200, 2002.
- C. Elkan and K. Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '08*, pages 213–220, 2008.
- B. Erdogdu Sakar, M. Isenkul, C. O. Sakar, A. Sertbas, F. Gurgun, S. Delil, H. Apaydin, and O. Kursun. Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings. *IEEE Journal of Biomedical and Health Informatics*, 17:828–834, 2013.
- J. Fan and I. Gijbels. *Local Polynomial Modelling and its Applications*. Chapman and Hall, 1996.
- J. H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232, 2001.
- K. Furmańczyk, J. Mielniczuk, W. Rejchel, and P. Teisseyre. Double logistic regression approach to biased positive-unlabeled data. In *Proceedings of the European Conference on Artificial Intelligence, ECAI'23*, pages 764–771, 2023.
- W. Gerych, T. Hartvigsen, L. Buquicchio, E. Agu, and E. Rundensteiner. Recovering the propensity score from biased positive unlabeled data. In *Proceedings of the AAAI Conference on Artificial Intelligence, AAAI'22*, pages 6694–6702, 2022.
- S. E. Golovenkin, J. Bac, A. Chervov, E. M. Mirkes, Y. V. Orlova, E. Barillot, A. N. Gorban, and A. Zinovyev. Trajectories, bifurcations, and pseudo-time in large clinical datasets: applications to myocardial infarction and diabetes data. *GigaScience*, 9(11), 2020.
- C. Gong, Q. Wang, T. Liu, B. Han, J. You, J. Yang, and D. Tao. Instance-dependent positive and unlabeled learning with labeling bias estimation. *IEEE Trans Pattern Anal Mach Intell*, pages 1–16, 2021a.
- C. Gong, M. I. Zulfiqar, C. Zhang, S. Mahmood, and J. Yang. A recent survey on instance-dependent positive and unlabeled learning. *Fundamental Research*, 2022.
- F. F. Gong, I. Vaitenas, S. C. Malaisrie, and K. Maganti. Mechanical complications of acute myocardial infarction: A review. *JAMA Cardiology*, 6(3):341–349, 2021b.
- T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC, London, 2015.
- J. Hernandez Orallo. Probabilistic reframing for cost-sensitive regression. *ACM transactions on Knowledge Discovery from Data (TKDD)*, 8(4):1–55, 2014.
- S. Jain, M. White, and P. Radivojac. Estimating the class prior and posterior from noisy positives and unlabeled data. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, page 2693–2701, 2016.
- A. E. W. Johnson, T. J. Pollard, L. Shen, L.-W. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, Anthony C. L., and R. G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:1–9, 2016.
- M. Kato, T. Teshima, and J. Honda. Learning from positive and unlabeled data with a selection bias. In *Proceedings of the 7th International Conference on Learning Representations*, pages 1–12, 2019.
- J. Kim and C. Yu. A semiparametric estimation of mean functional with nonignorable missing data. *Journal of the American Statistical Association*, 106:157–165, 2011.
- R. Kiryo, G. Niu, M. C. du Plessis, and M. Sugiyama. Positive-unlabeled learning with non-negative risk estimator. In *Proceedings of the International Conference on Neural Information Processing Systems, NIPS'17*, pages 1674–1684, 2017.
- R. Koenker. *Quantile Regression*. Cambridge University Press, 2005.
- D. Lambert. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14, 1992.
- F. M. Palechor and A. De la Hoz Manotas. Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico. *Data in Brief*, 25, 2019.

- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- S. Seabold and J. Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.
- V. Sevetlidis, G. Pavlidis, S. Mouroutsos, and A. Gasteratos. Dense-PU: Learning a Density-Based Boundary for Positive and Unlabeled Learning. *IEEE Access*, 12:1–12, 2024.
- J. Su, S. Zhang, and Y. Zhou. Solving the missing at random problem in semi-supervised learning: An inverse probability weighting method. *Stat*, 13:e707, 2024.
- E. Tchetgen Tchetgen and E.J. Writh. A general instrumental variable framework for regression analysis with outcome missing not at random. *Biometrics*, 73:1123–1131, 2017.
- G. Tutz. *Regression for categorical data*. Cambridge University Press, 2012.
- R. Vogel, M. Achab, S. Cl  men  on, and C. Tillier. Weighted empirical risk minimization: sample selection bias correction based on importance sampling. In *Proceedings of International Conference on Machine Learning, Artificial Intelligence and Applications, ICMA2020*, 2020.
- Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.
- J. Zhao and Y. Ma. A versatile estimation procedure without estimating the nonignorable missingness mechanism. *Journal of the American Statistical Association*, 117:1916–1930, 2022.
- J. Zhao and J. Shao. Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data. *Journal of the American Statistical Association*, 110:1577–1590, 2015.
- Z. Zhou, D. Li, D. Huh, M. Xie, and E-Y. Mun. A simulation study of the performance of statistical models for count outcomes with excessive zeros. *Statistics in Medicine*, 43(24):4752–4767, 2024.

Positive-Unlabeled Regression: Learning from Partially Labeled Quantitative Outcomes (Supplementary Material)

Paweł Teisseyre^{1,2}

Jan Mielniczuk^{1,2}

¹Faculty of Mathematics and Information Sciences, Warsaw University of Technology, Warsaw, Poland

²Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland

A PROOFS

A.1 LEMMA 3 AND ITS PROOF

Let $\tilde{f}(x) = E(\tilde{Y}|X = x)$. Then $\tilde{f}(x) = f(x) - b(x)$, where

$$b(x) = E(Y|X = x, S = 0)P(S = 0|X = x) = E(Y|X = x, S = 0)[1 - e(x)P(W = 1|X = x)]$$

Proof. First we have

$$\begin{aligned}\tilde{f}(x) &= E(\tilde{Y}|X = x) = E(\tilde{Y}|X = x, S = 1)P(S = 1|X = x) + E(\tilde{Y}|X = x, S = 0)P(S = 0|X = x) \\ &= E(\tilde{Y}|X = x, S = 1)P(S = 1|X = x) = E(Y|X = x, S = 1)P(S = 1|X = x),\end{aligned}$$

which follows from the fact that $\tilde{Y} = 0$ for $S = 0$, and also $\tilde{Y} = Y$, for $S = 1$. Using the above property, we can write

$$\begin{aligned}f(x) &= E(Y|X = x) = E(Y|X = x, S = 1)P(S = 1|X = x) + E(Y|X = x, S = 0)P(S = 0|X = x) \\ &= \tilde{f}(x) + E(Y|X = x, S = 0)P(S = 0|X = x),\end{aligned}$$

which yields the first equality in the Lemma. The second equality follows from Lemma 1 in the main text. □

A.2 THEOREM 5 AND ITS PROOF

Assume that condition (A) is satisfied. Then

$$R(h_\theta) = P(S = 1)E_{X,Y|S=1} \left\{ \frac{1}{e(X)} [l(Y, h_\theta(X)) - l(0, h_\theta(X))] \right\} + E_X l(0, h_\theta(X))$$

Proof. Let us denote by $\pi = P(Y > 0)$. Then we can write

$$\begin{aligned}R(h) &= \pi E_{X,Y|Y>0} l(Y, h_\theta(X)) + (1 - \pi) E_{X,Y|Y=0} l(0, h_\theta(X)) \\ &\stackrel{(i)}{=} \pi E_{X,Y|Y>0} l(Y, h_\theta(X)) + E_X l(0, h_\theta(X)) - \pi E_{X|Y>0} l(0, h_\theta(X)) \\ &\stackrel{(ii)}{=} \pi E_{X,Y|S=1} \frac{c}{e(X)} l(Y, h_\theta(X)) + E_X l(0, h_\theta(X)) - \pi E_{X|S=1} \frac{c}{e(X)} l(0, h_\theta(X)) \\ &\stackrel{(iii)}{=} P(S = 1) E_{X,Y|S=1} \left\{ \frac{1}{e(X)} [l(Y, h_\theta(X)) - l(0, h_\theta(X))] \right\} + E_X l(0, h_\theta(X))\end{aligned}$$

Equality (i) uses the fact that $E_X l(0, h_\theta(X)) = \pi E_{X|Y>0} l(0, h_\theta(X)) + (1 - \pi) E_{X|Y=0} l(0, h_\theta(X))$. Equality (ii) follows directly from Theorem 2 and equality (iii) follows from the fact that $c\pi = P(S = 1)$. □

B EMPIRICAL RISK MINIMIZATION FOR TRADITIONAL PU LEARNING

The calibrated method (CAL) requires the estimation of $P(W = 1|X = x)$ based on PU data $\mathcal{D}_{PU} = \{(x_i, s_i) : i = 1, \dots, n\}$. Below, we show how one can estimate $P(W = 1|X = x)$ using a parametric function $g_\theta(x)$ (logistic regression, neural networks) and knowledge of the propensity score function $e(x)$. A standard approach, used in classical PU learning, is to apply an empirical risk minimization scheme [Bekker and Davis, 2020]. Let $l(w, g_\theta(x))$ denotes the loss function, e.g., the logistic loss $l(w, g_\theta(x)) = -[w \log g_\theta(x) + (1 - w) \log(1 - g_\theta(x))]$. In the ERM approach, the goal is to minimize the empirical risk function, which corresponds to the theoretical risk

$$R(g_\theta) = E_{X,W} l(W, g_\theta(X)). \quad (7)$$

Direct optimization of the empirical equivalent of (7) is not possible, because we do not have access to the true values of w_i , where w_i corresponds to $W = \mathbb{1}\{Y > 0\}$. However, the risk function (7) can be written as a weighted risk function

$$R(g_\theta) = P(S = 1) E_{X|S=1} \left[\frac{1}{e(X)} l(1, g_\theta(X)) + (1 - \frac{1}{e(X)}) l(0, g_\theta(X)) \right] + P(S = 0) E_{X|S=0} l(0, g_\theta(X))$$

The proof of the above fact can be found in [Bekker and Davis, 2020]. Importantly, the empirical counterpart of the above risk,

$$\hat{R}(g_\theta) = \hat{P}(S = 1) \frac{1}{n_1} \sum_{i:s_i=1} \left[\frac{1}{e(x_i)} l(1, g_\theta(x_i)) + (1 - \frac{1}{e(x_i)}) l(0, g_\theta(x_i)) \right] + \hat{P}(S = 0) \frac{1}{n_0} \sum_{i:s_i=0} l(0, g_\theta(x_i)),$$

where $\hat{P}(S = 1) = n_1/n$ and $n_0 = n - n_1$, can be optimized with respect to θ , provided that $e(x)$ is known. This approach is used in our CAL method. When $e(x)$ is unknown, its estimator is used.

C DATASETS

C.1 REAL DATASETS

We conducted experiments on five real-world datasets, with basic summary statistics provided in Table 2. In the MIMIC dataset [Johnson et al., 2016], each patient record includes information on the occurrence of eight diseases: coronary artery disease, diabetes, hypotension, kidney disease, lipid arthritis, liver disease, thrombosis, and thyroid disease. The target variable is defined as the number of these diseases a patient suffers from. While some patients have none of the listed conditions, others may suffer from several, with the maximum observed count being six. The MIMIC dataset includes 310 features, encompassing laboratory test results, clinical scores, and demographic or administrative information such as age, gender, and marital status. In the Complications dataset [Golovenkin et al., 2020], the task is to predict the number of complications a patient experiences following a heart attack, based on various clinical variables. In the Obesity dataset [Palechor and De la Hoz Manotas, 2019], obesity levels are predicted from patients' eating habits and physical condition. For the Parkinson dataset [Erdogdu Sakar et al., 2013], multiple types of voice recordings are collected from each patient, from which 26 linear and time-frequency features are extracted. The target variable for this dataset is the UPDRS (Unified Parkinson's Disease Rating Scale) score, which quantifies the severity of Parkinson's disease. The Chest dataset [Yang et al., 2023] consists of X-ray scans of the lungs, and our objective is to predict the number of diseases based on a flattened representation of the images, where each pixel corresponds to a single feature. For all datasets, categorical variables were transformed into quantitative features using one-hot encoding. For datasets containing more than 20 variables, we applied a variable selection step, retaining the 20 most informative features based on a simple mutual information filter.

C.2 ARTIFICIAL DATASETS

The datasets are generated in a way that allows easy control over the value of $\pi = P(Y > 0)$. The dataset Art1 is generated as follows. First, y_i is drawn from a multinomial distribution with 10 mutually exclusive events, with probabilities

$$1 - \pi, \frac{\pi}{9}, \frac{\pi}{9}, \dots, \frac{\pi}{9}.$$

Next, the feature vector $x_i = (x_i^{(1)}, \dots, x_i^{(p)})$ is generated from a uniform distribution such that

$$x_i^{(1)} \sim U[y_i, y_i + 1], \quad \text{and} \quad x_i^{(j)} \sim U[0, 1] \text{ for } j = 2, \dots, p.$$

Table 2: Summary statistics for the considered data sets

Dataset	Observations	Features	Range of Y	Mean of Y	$P(Y > 0)$
Mimic	19773	310	$[0, 6]$	1.432	0.837
Complications	1700	111	$[0, 5]$	0.793	0.55
Obesity	2111	16	$[0, 5]$	2.241	0.735
Parkinson	1040	26	$[0, 54]$	12.0	0.5
Chest	78468	784	$[0, 9]$	0.72	0.46

This procedure allows direct control of $\pi = P(Y > 0)$. The regression function is thus stepwise function. The dataset Art2 is generated as follows. First, the features are drawn from a Gaussian distribution:

$$x_i^{(j)} \sim N(0, 1).$$

Then, the response variable is drawn from a Poisson distribution:

$$y_i \sim \text{Poisson}(\lambda_i), \quad \text{where } \lambda_i = \exp(\theta_0 + x_i^{(1)}).$$

The parameter θ_0 is chosen to control the value of $\pi = P(Y > 0)$.

D TECHNICAL DETAILS OF THE IMPLEMENTATION

In experiments, all methods are based on two basic models: the linear model and the Poisson model/truncated Poisson model. We used the scikit-learn [Pedregosa et al., 2011] implementation of the linear regression model and the statsmodel [Seabold and Perktold, 2010] implementation of the truncated Poisson model. Since we predict a discrete response variable for all datasets, we used the projection technique for the linear model. More precisely, given an estimated parameter vector $\hat{\theta}$ for observations x_i , the predicted value is defined as: $\hat{y}_i = \arg \min_{k=0,1,2,3,\dots} |x_i^T \hat{\theta} - k|$. In the case of propensity score estimation, we used implementations made publicly available by the authors of SAREM, PGLIN, TM and LBE [Bekker et al., 2019, Gerych et al., 2022, Furmańczyk et al., 2023, Gong et al., 2021a], with default parameters used by them.

For the ERM method and the quadratic loss function, we use the scikit-learn linear regression implementation, based on an appropriately modified response variable, as described in (6). The ERM method was implemented from scratch for the case where we use the loss function associated with the Poisson model. In this case we use standard gradient descent algorithm to optimize the parameters. We set the number of epochs to 500 and the learning rate to 0.01.

All experiments are executed on the same machine with access to 32Gb RAM and 8 cores of an Intel(R) Xeon(R) CPU E31270 3.40GHz.

E ADDITIONAL EXPERIMENTS

Below we show additional experimental results:

- Figures 3 and 4 show the impact of the prior probability $\pi = P(Y > 0)$ on the RMSE for MIMIC dataset.
- Tables 3-10 show complete results for strategies S1-S4 and both base regression models.
- Table 11 shows the effect of propensity score estimation on the RMSE of the CAL method.
- Table 12 contains computational times for the considered methods.

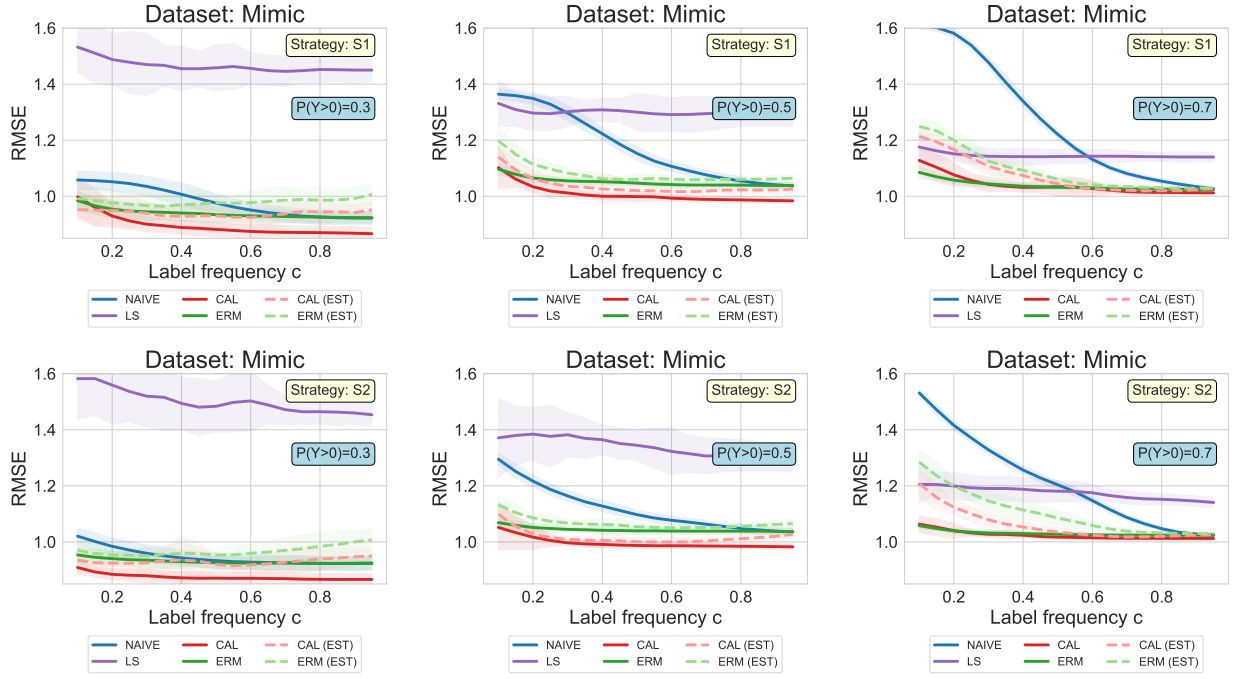


Figure 3: The impact of the prior probability $\pi = P(Y > 0)$ on the RMSE of the methods under consideration for MIMIC dataset. A **linear model** was used as the basic regression model.

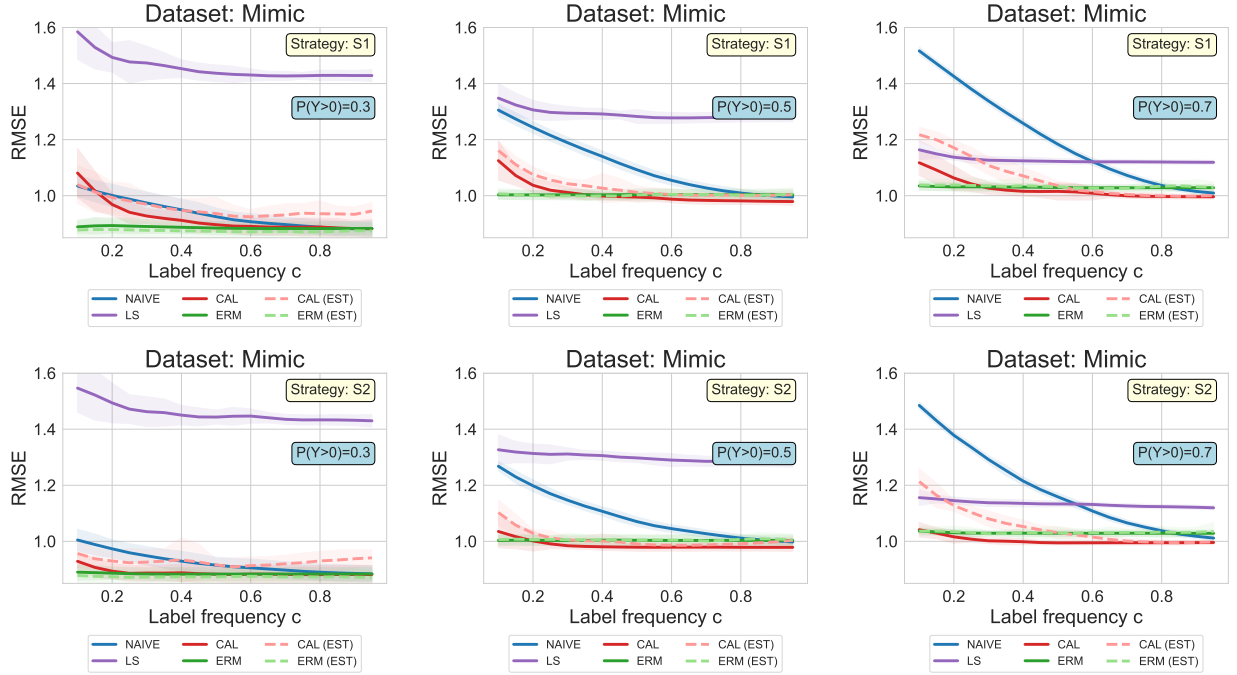


Figure 4: The impact of the prior probability $\pi = P(Y > 0)$ on the RMSE of the methods under consideration for MIMIC dataset. A **Poisson model** was used as the basic regression model.

Table 3: Root mean square error (RMSE) for the considered methods, **labeling strategy S1** and for $\pi = 0.5$. A **linear model** was used as the basic regression model. The method with the lowest RMSE is highlighted in bold, and the second best is in light gray. The methods CAL and ERM are excluded from comparison as they assume known $e(x)$.

c	Dataset	NAIVE	LS	CAL (EST)	ERM (EST)	CAL	ERM
10%	Art1	3.54 ± 0.11	0.73 ± 0.01	0.46 ± 0.18	0.49 ± 0.09	0.56 ± 0.17	0.64 ± 0.12
	Art2	1.86 ± 0.32	1.43 ± 0.2	1.41 ± 0.26	1.38 ± 0.14	1.26 ± 0.18	1.26 ± 0.18
	Mimic	1.36 ± 0.02	1.33 ± 0.08	1.14 ± 0.04	1.2 ± 0.05	1.1 ± 0.08	1.1 ± 0.02
	Parkinson	19.86 ± 0.5	23.07 ± 2.39	19.35 ± 2.1	19.22 ± 3.67	19.02 ± 1.89	17.51 ± 0.78
	Obesity	2.21 ± 0.13	1.71 ± 0.66	0.99 ± 0.22	1.04 ± 0.18	1.02 ± 0.3	1.12 ± 0.09
	Compl.	1.13 ± 0.02	1.34 ± 0.14	1.05 ± 0.05	1.05 ± 0.05	1.11 ± 0.09	1.07 ± 0.05
	Chest	1.25 ± 0.02	1.32 ± 0.06	1.01 ± 0.02	1.05 ± 0.03	1.01 ± 0.01	1.04 ± 0.01
40%	Art1	2.31 ± 0.12	0.72 ± 0.01	0.21 ± 0.04	0.32 ± 0.07	0.27 ± 0.05	0.51 ± 0.09
	Art2	1.51 ± 0.18	1.4 ± 0.11	1.14 ± 0.09	1.31 ± 0.27	1.2 ± 0.24	1.26 ± 0.19
	Mimic	1.22 ± 0.02	1.31 ± 0.04	1.03 ± 0.05	1.06 ± 0.02	1.0 ± 0.01	1.05 ± 0.01
	Parkinson	17.58 ± 0.4	20.93 ± 1.36	17.57 ± 0.87	17.27 ± 1.11	16.22 ± 0.41	15.96 ± 0.46
	Obesity	1.45 ± 0.1	0.81 ± 0.03	0.44 ± 0.04	0.59 ± 0.06	0.48 ± 0.05	0.64 ± 0.04
	Compl.	1.08 ± 0.03	1.16 ± 0.05	0.93 ± 0.03	0.98 ± 0.02	0.9 ± 0.02	0.94 ± 0.01
	Chest	1.23 ± 0.02	1.34 ± 0.05	0.99 ± 0.01	1.02 ± 0.01	0.98 ± 0.01	1.0 ± 0.01
70%	Art1	1.28 ± 0.09	0.72 ± 0.01	0.17 ± 0.01	0.25 ± 0.04	0.22 ± 0.09	0.36 ± 0.07
	Art2	1.3 ± 0.17	1.39 ± 0.1	1.1 ± 0.08	1.25 ± 0.15	1.11 ± 0.15	1.18 ± 0.12
	Mimic	1.08 ± 0.02	1.3 ± 0.06	1.02 ± 0.01	1.06 ± 0.01	0.99 ± 0.01	1.04 ± 0.01
	Parkinson	16.18 ± 0.37	19.22 ± 0.5	16.99 ± 1.1	16.52 ± 0.78	15.89 ± 0.31	15.68 ± 0.29
	Obesity	0.9 ± 0.08	0.78 ± 0.01	0.38 ± 0.03	0.54 ± 0.02	0.4 ± 0.05	0.55 ± 0.04
	Compl.	0.95 ± 0.02	1.14 ± 0.04	0.91 ± 0.02	0.95 ± 0.03	0.88 ± 0.01	0.93 ± 0.02
	Chest	1.03 ± 0.02	1.34 ± 0.02	1.0 ± 0.02	1.01 ± 0.01	0.98 ± 0.01	1.0 ± 0.01
100%	Art1	0.35 ± 0.02	0.72 ± 0.01	0.16 ± 0.02	0.19 ± 0.04	0.16 ± 0.02	0.22 ± 0.04
	Art2	1.21 ± 0.1	1.47 ± 0.15	1.18 ± 0.11	1.24 ± 0.19	1.15 ± 0.19	1.29 ± 0.15
	Mimic	1.04 ± 0.01	1.29 ± 0.04	1.03 ± 0.02	1.06 ± 0.01	0.98 ± 0.01	1.04 ± 0.01
	Parkinson	15.67 ± 0.32	19.29 ± 0.45	16.71 ± 0.91	16.37 ± 0.5	15.82 ± 0.34	15.62 ± 0.31
	Obesity	0.56 ± 0.02	0.78 ± 0.01	0.34 ± 0.02	0.51 ± 0.01	0.35 ± 0.02	0.52 ± 0.02
	Compl.	0.91 ± 0.02	1.11 ± 0.03	0.91 ± 0.02	0.94 ± 0.02	0.88 ± 0.01	0.91 ± 0.02
	Chest	1.0 ± 0.01	1.35 ± 0.02	1.05 ± 0.02	1.02 ± 0.02	0.98 ± 0.02	1.0 ± 0.01

Table 4: Root mean square error (RMSE) for the considered methods, **labeling strategy S2** and for $\pi = 0.5$. A **linear model** was used as the basic regression model. The method with the lowest RMSE is highlighted in bold, and the second best is in light gray. The methods CAL and ERM are excluded from comparison as they assume known $e(x)$.

c	Dataset	NAIVE	LS	CAL (EST)	ERM (EST)	CAL	ERM
10%	Art1	0.15 $\pm$ 0.02	0.71 $\pm$ 0.01	0.18 $\pm$ 0.02	0.12 $\pm$ 0.02	0.19 $\pm$ 0.03	0.17 $\pm$ 0.02
	Art2	1.5 $\pm$ 0.19	1.32 $\pm$ 0.11	1.12 $\pm$ 0.08	1.22 $\pm$ 0.17	1.11 $\pm$ 0.08	1.29 $\pm$ 0.17
	Mimic	1.3 $\pm$ 0.02	1.37 $\pm$ 0.14	1.1 $\pm$ 0.05	1.13 $\pm$ 0.03	1.05 $\pm$ 0.08	1.07 $\pm$ 0.02
	Parkinson	19.5 $\pm$ 0.56	22.74 $\pm$ 2.97	18.62 $\pm$ 1.25	17.57 $\pm$ 1.99	19.19 $\pm$ 1.41	17.54 $\pm$ 1.24
	Obesity	0.93 $\pm$ 0.14	1.62 $\pm$ 0.49	0.86 $\pm$ 0.09	0.78 $\pm$ 0.12	0.77 $\pm$ 0.17	0.84 $\pm$ 0.26
	Compl.	1.11 $\pm$ 0.02	1.4 $\pm$ 0.11	1.0 $\pm$ 0.06	1.02 $\pm$ 0.03	1.1 $\pm$ 0.04	1.08 $\pm$ 0.04
	Chest	1.25 $\pm$ 0.02	1.33 $\pm$ 0.05	1.01 $\pm$ 0.02	1.06 $\pm$ 0.03	1.01 $\pm$ 0.01	1.03 $\pm$ 0.01
40%	Art1	0.16 $\pm$ 0.02	0.71 $\pm$ 0.01	0.18 $\pm$ 0.02	0.15 $\pm$ 0.02	0.19 $\pm$ 0.06	0.17 $\pm$ 0.03
	Art2	1.26 $\pm$ 0.2	1.3 $\pm$ 0.07	1.1 $\pm$ 0.1	1.28 $\pm$ 0.14	1.12 $\pm$ 0.12	1.24 $\pm$ 0.19
	Mimic	1.13 $\pm$ 0.02	1.35 $\pm$ 0.05	1.01 $\pm$ 0.02	1.07 $\pm$ 0.02	0.99 $\pm$ 0.01	1.04 $\pm$ 0.01
	Parkinson	17.09 $\pm$ 0.41	20.06 $\pm$ 1.15	16.98 $\pm$ 0.47	16.57 $\pm$ 0.64	16.16 $\pm$ 0.26	15.92 $\pm$ 0.37
	Obesity	0.55 $\pm$ 0.02	0.86 $\pm$ 0.05	0.42 $\pm$ 0.04	0.53 $\pm$ 0.02	0.44 $\pm$ 0.06	0.55 $\pm$ 0.03
	Compl.	1.03 $\pm$ 0.02	1.21 $\pm$ 0.09	0.93 $\pm$ 0.05	0.97 $\pm$ 0.04	0.9 $\pm$ 0.02	0.94 $\pm$ 0.02
	Chest	1.16 $\pm$ 0.03	1.34 $\pm$ 0.05	0.99 $\pm$ 0.01	1.02 $\pm$ 0.01	0.98 $\pm$ 0.01	1.0 $\pm$ 0.01
70%	Art1	0.14 $\pm$ 0.03	0.72 $\pm$ 0.01	0.16 $\pm$ 0.02	0.18 $\pm$ 0.03	0.16 $\pm$ 0.01	0.15 $\pm$ 0.02
	Art2	1.21 $\pm$ 0.13	1.41 $\pm$ 0.21	1.16 $\pm$ 0.14	1.29 $\pm$ 0.14	1.12 $\pm$ 0.11	1.21 $\pm$ 0.11
	Mimic	1.06 $\pm$ 0.01	1.3 $\pm$ 0.06	1.0 $\pm$ 0.01	1.05 $\pm$ 0.01	0.99 $\pm$ 0.01	1.04 $\pm$ 0.01
	Parkinson	16.12 $\pm$ 0.33	19.18 $\pm$ 0.79	16.45 $\pm$ 0.78	16.22 $\pm$ 0.69	15.93 $\pm$ 0.29	15.67 $\pm$ 0.32
	Obesity	0.5 $\pm$ 0.01	0.78 $\pm$ 0.01	0.34 $\pm$ 0.02	0.5 $\pm$ 0.02	0.34 $\pm$ 0.02	0.5 $\pm$ 0.01
	Compl.	0.95 $\pm$ 0.02	1.14 $\pm$ 0.04	0.9 $\pm$ 0.02	0.94 $\pm$ 0.03	0.88 $\pm$ 0.01	0.92 $\pm$ 0.02
	Chest	1.03 $\pm$ 0.02	1.34 $\pm$ 0.03	1.01 $\pm$ 0.02	1.01 $\pm$ 0.02	0.98 $\pm$ 0.01	1.0 $\pm$ 0.01
100%	Art1	0.15 $\pm$ 0.02	0.72 $\pm$ 0.01	0.15 $\pm$ 0.02	0.19 $\pm$ 0.01	0.15 $\pm$ 0.02	0.15 $\pm$ 0.03
	Art2	1.25 $\pm$ 0.17	1.38 $\pm$ 0.13	1.14 $\pm$ 0.1	1.2 $\pm$ 0.06	1.12 $\pm$ 0.11	1.4 $\pm$ 0.32
	Mimic	1.04 $\pm$ 0.01	1.29 $\pm$ 0.04	1.03 $\pm$ 0.02	1.06 $\pm$ 0.02	0.98 $\pm$ 0.01	1.04 $\pm$ 0.01
	Parkinson	15.66 $\pm$ 0.32	19.33 $\pm$ 0.51	16.62 $\pm$ 0.8	16.44 $\pm$ 0.7	15.84 $\pm$ 0.31	15.62 $\pm$ 0.32
	Obesity	0.51 $\pm$ 0.01	0.78 $\pm$ 0.01	0.34 $\pm$ 0.02	0.51 $\pm$ 0.02	0.34 $\pm$ 0.02	0.51 $\pm$ 0.01
	Compl.	0.91 $\pm$ 0.02	1.1 $\pm$ 0.04	0.9 $\pm$ 0.01	0.93 $\pm$ 0.02	0.87 $\pm$ 0.01	0.91 $\pm$ 0.02
	Chest	1.0 $\pm$ 0.01	1.35 $\pm$ 0.03	1.04 $\pm$ 0.02	1.02 $\pm$ 0.02	0.98 $\pm$ 0.01	1.0 $\pm$ 0.01

Table 5: Root mean square error (RMSE) for the considered methods, **labeling strategy S3** and for $\pi = 0.5$. A **linear model** was used as the basic regression model. The method with the lowest RMSE is highlighted in bold, and the second best is in light gray. The methods CAL and ERM are excluded from comparison as they assume known $e(x)$.

c	Dataset	NAIVE	LS	CAL (EST)	ERM (EST)	CAL	ERM
10%	Art1	0.21 $\pm$ 0.06	0.86 $\pm$ 0.29	0.25 $\pm$ 0.02	0.15 $\pm$ 0.02	0.25 $\pm$ 0.06	0.17 $\pm$ 0.04
	Art2	1.41 $\pm$ 0.24	1.2 $\pm$ 0.08	1.22 $\pm$ 0.15	1.27 $\pm$ 0.14	1.18 $\pm$ 0.09	1.28 $\pm$ 0.15
	Mimic	1.25 $\pm$ 0.02	1.46 $\pm$ 0.15	1.11 $\pm$ 0.03	1.14 $\pm$ 0.02	1.14 $\pm$ 0.13	1.1 $\pm$ 0.02
	Parkinson	19.08 $\pm$ 0.47	21.81 $\pm$ 2.1	19.01 $\pm$ 1.27	17.34 $\pm$ 0.89	19.54 $\pm$ 1.63	17.63 $\pm$ 1.08
	Obesity	1.44 $\pm$ 0.22	3.19 $\pm$ 0.31	1.3 $\pm$ 0.14	1.29 $\pm$ 0.14	1.25 $\pm$ 0.12	1.27 $\pm$ 0.15
	Compl.	1.09 $\pm$ 0.02	1.35 $\pm$ 0.11	1.03 $\pm$ 0.05	1.03 $\pm$ 0.03	1.11 $\pm$ 0.07	1.14 $\pm$ 0.15
	Chest	1.25 $\pm$ 0.02	1.35 $\pm$ 0.07	1.04 $\pm$ 0.03	1.08 $\pm$ 0.02	1.05 $\pm$ 0.08	1.07 $\pm$ 0.04
40%	Art1	0.22 $\pm$ 0.06	0.72 $\pm$ 0.01	0.25 $\pm$ 0.02	0.19 $\pm$ 0.06	0.25 $\pm$ 0.03	0.17 $\pm$ 0.03
	Art2	1.36 $\pm$ 0.18	1.27 $\pm$ 0.11	1.1 $\pm$ 0.09	1.23 $\pm$ 0.12	1.08 $\pm$ 0.11	1.29 $\pm$ 0.28
	Mimic	1.12 $\pm$ 0.02	1.4 $\pm$ 0.07	1.0 $\pm$ 0.02	1.06 $\pm$ 0.02	0.99 $\pm$ 0.01	1.04 $\pm$ 0.01
	Parkinson	17.11 $\pm$ 0.45	20.29 $\pm$ 1.25	17.06 $\pm$ 0.7	16.65 $\pm$ 0.8	16.33 $\pm$ 0.65	16.18 $\pm$ 1.14
	Obesity	0.55 $\pm$ 0.02	0.87 $\pm$ 0.04	0.42 $\pm$ 0.03	0.53 $\pm$ 0.02	0.5 $\pm$ 0.11	0.56 $\pm$ 0.04
	Compl.	1.02 $\pm$ 0.02	1.22 $\pm$ 0.09	0.92 $\pm$ 0.02	0.96 $\pm$ 0.02	0.9 $\pm$ 0.01	0.94 $\pm$ 0.02
	Chest	1.14 $\pm$ 0.02	1.34 $\pm$ 0.05	1.0 $\pm$ 0.02	1.03 $\pm$ 0.01	0.98 $\pm$ 0.01	1.0 $\pm$ 0.01
70%	Art1	0.15 $\pm$ 0.02	0.72 $\pm$ 0.01	0.16 $\pm$ 0.02	0.18 $\pm$ 0.02	0.17 $\pm$ 0.01	0.15 $\pm$ 0.02
	Art2	1.22 $\pm$ 0.16	1.28 $\pm$ 0.06	1.1 $\pm$ 0.05	1.28 $\pm$ 0.32	1.07 $\pm$ 0.09	1.26 $\pm$ 0.17
	Mimic	1.06 $\pm$ 0.01	1.31 $\pm$ 0.06	1.01 $\pm$ 0.01	1.05 $\pm$ 0.01	0.98 $\pm$ 0.01	1.04 $\pm$ 0.01
	Parkinson	16.07 $\pm$ 0.32	19.19 $\pm$ 0.68	16.24 $\pm$ 0.28	16.04 $\pm$ 0.29	15.85 $\pm$ 0.29	15.65 $\pm$ 0.32
	Obesity	0.51 $\pm$ 0.01	0.78 $\pm$ 0.01	0.34 $\pm$ 0.02	0.5 $\pm$ 0.01	0.34 $\pm$ 0.02	0.51 $\pm$ 0.01
	Compl.	0.94 $\pm$ 0.02	1.15 $\pm$ 0.05	0.9 $\pm$ 0.02	0.94 $\pm$ 0.02	0.89 $\pm$ 0.01	0.92 $\pm$ 0.02
	Chest	1.03 $\pm$ 0.02	1.35 $\pm$ 0.03	1.01 $\pm$ 0.02	1.01 $\pm$ 0.02	0.97 $\pm$ 0.02	1.0 $\pm$ 0.01
100%	Art1	0.15 $\pm$ 0.04	0.71 $\pm$ 0.02	0.16 $\pm$ 0.02	0.19 $\pm$ 0.02	0.16 $\pm$ 0.02	0.15 $\pm$ 0.02
	Art2	1.18 $\pm$ 0.08	1.35 $\pm$ 0.07	1.14 $\pm$ 0.11	1.24 $\pm$ 0.14	1.07 $\pm$ 0.08	1.16 $\pm$ 0.12
	Mimic	1.03 $\pm$ 0.01	1.29 $\pm$ 0.04	1.03 $\pm$ 0.02	1.07 $\pm$ 0.02	0.98 $\pm$ 0.01	1.04 $\pm$ 0.01
	Parkinson	15.66 $\pm$ 0.33	19.3 $\pm$ 0.51	16.58 $\pm$ 0.73	16.4 $\pm$ 0.66	15.82 $\pm$ 0.33	15.63 $\pm$ 0.32
	Obesity	0.51 $\pm$ 0.01	0.78 $\pm$ 0.01	0.34 $\pm$ 0.02	0.51 $\pm$ 0.02	0.34 $\pm$ 0.02	0.51 $\pm$ 0.01
	Compl.	0.91 $\pm$ 0.02	1.11 $\pm$ 0.03	0.9 $\pm$ 0.01	0.93 $\pm$ 0.02	0.87 $\pm$ 0.01	0.91 $\pm$ 0.02
	Chest	1.0 $\pm$ 0.01	1.35 $\pm$ 0.03	1.04 $\pm$ 0.02	1.02 $\pm$ 0.01	0.98 $\pm$ 0.01	1.0 $\pm$ 0.01

Table 6: Root mean square error (RMSE) for the considered methods, **labeling strategy S4** and for $\pi = 0.5$. A **linear model** was used as the basic regression model. The method with the lowest RMSE is highlighted in bold, and the second best is in light gray. The methods CAL and ERM are excluded from comparison as they assume known $e(x)$.

c	Dataset	NAIVE	LS	CAL (EST)	ERM (EST)	CAL	ERM
10%	Art1	0.9 ± 0.07	0.72 ± 0.02	0.46 ± 0.22	0.3 ± 0.17	0.38 ± 0.13	0.4 ± 0.26
	Art2	1.69 ± 0.27	1.72 ± 0.21	1.23 ± 0.08	1.32 ± 0.28	1.42 ± 0.29	1.26 ± 0.23
	Mimic	1.25 ± 0.03	1.62 ± 0.02	1.06 ± 0.03	1.1 ± 0.02	1.06 ± 0.03	1.05 ± 0.01
	Parkinson	18.23 ± 0.53	26.46 ± 1.86	19.67 ± 2.9	18.19 ± 2.13	17.89 ± 0.84	16.25 ± 0.58
	Obesity	1.13 ± 0.13	2.49 ± 0.5	0.94 ± 0.1	0.9 ± 0.11	1.04 ± 0.24	1.01 ± 0.36
	Compl.	1.09 ± 0.03	1.53 ± 0.1	0.96 ± 0.04	0.99 ± 0.03	1.01 ± 0.05	0.97 ± 0.03
	Chest	1.25 ± 0.02	1.57 ± 0.03	0.98 ± 0.01	1.03 ± 0.01	1.01 ± 0.01	1.01 ± 0.01
40%	Art1	0.94 ± 0.12	0.74 ± 0.05	0.32 ± 0.15	0.19 ± 0.05	0.31 ± 0.09	0.36 ± 0.07
	Art2	1.48 ± 0.12	1.8 ± 0.25	1.4 ± 0.25	1.33 ± 0.24	1.25 ± 0.15	1.28 ± 0.16
	Mimic	1.21 ± 0.08	1.64 ± 0.06	1.05 ± 0.05	1.11 ± 0.06	1.04 ± 0.07	1.05 ± 0.08
	Parkinson	16.57 ± 0.48	26.42 ± 1.25	18.19 ± 1.49	17.17 ± 1.04	17.16 ± 0.7	15.78 ± 0.36
	Obesity	0.75 ± 0.08	0.89 ± 0.03	0.52 ± 0.07	0.57 ± 0.04	0.52 ± 0.06	0.56 ± 0.04
	Compl.	1.08 ± 0.06	1.51 ± 0.08	0.97 ± 0.07	0.98 ± 0.07	0.94 ± 0.06	0.94 ± 0.07
	Chest	1.25 ± 0.05	1.57 ± 0.03	1.0 ± 0.04	1.06 ± 0.04	1.04 ± 0.05	0.98 ± 0.05
70%	Art1	0.93 ± 0.07	0.75 ± 0.06	0.26 ± 0.07	0.29 ± 0.17	0.28 ± 0.06	0.34 ± 0.07
	Art2	1.63 ± 0.18	1.66 ± 0.11	1.26 ± 0.18	1.33 ± 0.07	1.76 ± 1.08	1.37 ± 0.2
	Mimic	1.21 ± 0.07	1.61 ± 0.04	1.08 ± 0.07	1.09 ± 0.06	1.05 ± 0.05	1.07 ± 0.06
	Parkinson	16.57 ± 0.46	26.41 ± 1.22	18.17 ± 1.51	17.2 ± 1.05	17.18 ± 0.72	15.79 ± 0.33
	Obesity	0.75 ± 0.06	0.85 ± 0.05	0.53 ± 0.06	0.57 ± 0.05	0.47 ± 0.08	0.56 ± 0.06
	Compl.	1.07 ± 0.06	1.51 ± 0.08	0.98 ± 0.09	0.99 ± 0.05	0.97 ± 0.06	0.97 ± 0.04
	Chest	1.27 ± 0.05	1.59 ± 0.06	0.99 ± 0.03	1.07 ± 0.04	1.02 ± 0.05	1.03 ± 0.04
100%	Art1	0.92 ± 0.07	0.74 ± 0.04	0.34 ± 0.12	0.31 ± 0.18	0.29 ± 0.08	0.36 ± 0.08
	Art2	1.59 ± 0.27	1.8 ± 0.43	1.3 ± 0.25	1.32 ± 0.09	1.23 ± 0.14	1.3 ± 0.18
	Mimic	1.19 ± 0.06	1.65 ± 0.02	1.05 ± 0.07	1.08 ± 0.05	1.07 ± 0.04	1.06 ± 0.08
	Parkinson	16.59 ± 0.47	26.42 ± 1.24	18.14 ± 1.49	17.18 ± 1.04	17.19 ± 0.7	15.78 ± 0.35
	Obesity	0.77 ± 0.09	0.87 ± 0.06	0.5 ± 0.05	0.59 ± 0.05	0.48 ± 0.06	0.59 ± 0.06
	Compl.	1.07 ± 0.08	1.51 ± 0.07	0.98 ± 0.08	1.0 ± 0.07	0.94 ± 0.07	0.95 ± 0.07
	Chest	1.26 ± 0.05	1.56 ± 0.05	0.98 ± 0.06	1.03 ± 0.04	1.02 ± 0.03	1.02 ± 0.05

Table 7: Root mean square error (RMSE) for the considered methods, **labeling strategy S1** and for $\pi = 0.5$. A **Poisson model** was used as the basic regression model. The method with the lowest RMSE is highlighted in bold, and the second best is in light gray. The methods CAL and ERM are excluded from comparison as they assume known $e(x)$.

c	Dataset	NAIVE	LS	CAL (EST)	ERM (EST)	CAL	ERM
10%	Art1	3.59 ± 0.09	1.36 ± 0.07	0.59 ± 0.09	1.22 ± 0.13	0.83 ± 0.27	1.2 ± 0.21
	Art2	1.65 ± 0.1	1.37 ± 0.25	1.26 ± 0.14	1.14 ± 0.18	1.18 ± 0.14	1.13 ± 0.11
	Mimic	1.31 ± 0.02	1.35 ± 0.05	1.16 ± 0.03	1.01 ± 0.02	1.12 ± 0.07	1.0 ± 0.02
	Parkinson	19.5 ± 0.47	23.46 ± 2.23	19.9 ± 2.01	16.65 ± 0.63	20.16 ± 1.61	16.65 ± 0.62
	Obesity	2.12 ± 0.05	1.89 ± 0.59	1.47 ± 0.67	1.39 ± 0.38	1.17 ± 0.28	1.24 ± 0.17
	Compl.	1.08 ± 0.03	1.64 ± 0.17	1.12 ± 0.11	0.96 ± 0.07	1.28 ± 0.12	0.92 ± 0.05
	Chest	1.2 ± 0.02	1.29 ± 0.05	1.01 ± 0.03	1.21 ± 0.09	1.0 ± 0.02	1.22 ± 0.09
40%	Art1	2.48 ± 0.11	1.37 ± 0.04	0.43 ± 0.03	1.12 ± 0.06	0.64 ± 0.14	1.04 ± 0.11
	Art2	1.37 ± 0.23	1.25 ± 0.09	1.08 ± 0.15	1.05 ± 0.16	1.04 ± 0.11	1.04 ± 0.12
	Mimic	1.14 ± 0.02	1.29 ± 0.02	1.03 ± 0.05	0.99 ± 0.01	1.0 ± 0.01	1.0 ± 0.02
	Parkinson	17.59 ± 0.38	21.17 ± 1.42	17.69 ± 0.94	15.95 ± 0.34	16.39 ± 0.54	15.94 ± 0.36
	Obesity	1.47 ± 0.08	1.16 ± 0.07	0.45 ± 0.04	1.22 ± 0.4	0.6 ± 0.1	1.08 ± 0.08
	Compl.	0.97 ± 0.02	1.49 ± 0.05	0.97 ± 0.06	0.92 ± 0.06	0.94 ± 0.04	0.89 ± 0.03
	Chest	1.07 ± 0.02	1.24 ± 0.02	0.97 ± 0.01	1.21 ± 0.1	0.96 ± 0.01	1.22 ± 0.09
70%	Art1	1.58 ± 0.07	1.37 ± 0.03	0.43 ± 0.03	1.08 ± 0.09	0.47 ± 0.06	1.1 ± 0.07
	Art2	1.13 ± 0.18	1.19 ± 0.04	1.03 ± 0.08	0.98 ± 0.08	0.97 ± 0.05	0.96 ± 0.03
	Mimic	1.03 ± 0.02	1.28 ± 0.02	1.0 ± 0.01	1.01 ± 0.02	0.98 ± 0.01	1.0 ± 0.02
	Parkinson	16.3 ± 0.36	19.29 ± 0.46	17.02 ± 1.13	15.94 ± 0.3	15.92 ± 0.33	15.9 ± 0.32
	Obesity	0.96 ± 0.05	1.15 ± 0.03	0.42 ± 0.02	1.18 ± 0.41	0.47 ± 0.05	1.06 ± 0.09
	Compl.	0.89 ± 0.03	1.47 ± 0.02	0.96 ± 0.05	0.9 ± 0.03	0.91 ± 0.01	0.88 ± 0.02
	Chest	0.99 ± 0.02	1.23 ± 0.01	0.97 ± 0.01	1.22 ± 0.08	0.96 ± 0.02	1.22 ± 0.09
100%	Art1	1.07 ± 0.04	1.36 ± 0.02	0.43 ± 0.02	1.11 ± 0.05	0.46 ± 0.02	1.08 ± 0.03
	Art2	0.97 ± 0.06	1.3 ± 0.31	1.1 ± 0.15	1.01 ± 0.08	0.99 ± 0.07	1.01 ± 0.1
	Mimic	1.0 ± 0.01	1.28 ± 0.01	1.0 ± 0.01	1.01 ± 0.02	0.98 ± 0.01	1.0 ± 0.02
	Parkinson	15.81 ± 0.35	19.36 ± 0.46	16.73 ± 0.93	15.84 ± 0.32	15.85 ± 0.33	15.85 ± 0.33
	Obesity	0.65 ± 0.03	1.15 ± 0.02	0.41 ± 0.02	1.06 ± 0.1	0.41 ± 0.02	1.06 ± 0.08
	Compl.	0.86 ± 0.03	1.46 ± 0.02	0.97 ± 0.03	0.89 ± 0.04	0.89 ± 0.01	0.88 ± 0.02
	Chest	0.96 ± 0.02	1.23 ± 0.01	0.99 ± 0.01	1.2 ± 0.09	0.95 ± 0.01	1.22 ± 0.09

Table 8: Root mean square error (RMSE) for the considered methods, **labeling strategy S2** and for $\pi = 0.5$. A **Poisson model** was used as the basic regression model. The method with the lowest RMSE is highlighted in bold, and the second best is in light gray. The methods CAL and ERM are excluded from comparison as they assume known $e(x)$.

c	Dataset	NAIVE	LS	CAL (EST)	ERM (EST)	CAL	ERM
10%	Art1	1.01 $\pm$ 0.03	1.44 $\pm$ 0.04	0.39 $\pm$ 0.01	1.1 $\pm$ 0.02	0.41 $\pm$ 0.03	1.08 $\pm$ 0.04
	Art2	1.38 $\pm$ 0.14	1.22 $\pm$ 0.05	1.2 $\pm$ 0.17	0.99 $\pm$ 0.07	1.05 $\pm$ 0.1	1.07 $\pm$ 0.11
	Mimic	1.27 $\pm$ 0.02	1.33 $\pm$ 0.05	1.1 $\pm$ 0.05	1.01 $\pm$ 0.02	1.04 $\pm$ 0.06	1.0 $\pm$ 0.02
	Parkinson	19.35 $\pm$ 0.47	23.13 $\pm$ 2.9	18.93 $\pm$ 1.2	16.72 $\pm$ 0.6	20.17 $\pm$ 2.21	16.73 $\pm$ 0.56
	Obesity	1.14 $\pm$ 0.11	1.87 $\pm$ 0.36	0.84 $\pm$ 0.09	1.19 $\pm$ 0.18	0.84 $\pm$ 0.2	1.17 $\pm$ 0.11
	Compl.	1.08 $\pm$ 0.03	1.69 $\pm$ 0.13	1.04 $\pm$ 0.1	0.95 $\pm$ 0.07	1.26 $\pm$ 0.07	0.94 $\pm$ 0.06
	Chest	1.19 $\pm$ 0.02	1.26 $\pm$ 0.04	1.0 $\pm$ 0.02	1.25 $\pm$ 0.02	0.98 $\pm$ 0.01	1.22 $\pm$ 0.09
40%	Art1	1.03 $\pm$ 0.03	1.43 $\pm$ 0.02	0.4 $\pm$ 0.01	1.09 $\pm$ 0.04	0.44 $\pm$ 0.03	1.1 $\pm$ 0.03
	Art2	1.07 $\pm$ 0.09	1.26 $\pm$ 0.08	1.05 $\pm$ 0.12	1.01 $\pm$ 0.08	0.99 $\pm$ 0.08	1.01 $\pm$ 0.08
	Mimic	1.11 $\pm$ 0.02	1.3 $\pm$ 0.02	1.01 $\pm$ 0.03	1.0 $\pm$ 0.01	0.98 $\pm$ 0.01	1.0 $\pm$ 0.02
	Parkinson	17.3 $\pm$ 0.41	20.23 $\pm$ 1.2	16.99 $\pm$ 0.47	16.08 $\pm$ 0.37	16.19 $\pm$ 0.27	16.06 $\pm$ 0.39
	Obesity	0.72 $\pm$ 0.03	1.3 $\pm$ 0.06	0.42 $\pm$ 0.04	1.08 $\pm$ 0.07	0.53 $\pm$ 0.09	1.06 $\pm$ 0.07
	Compl.	0.96 $\pm$ 0.03	1.51 $\pm$ 0.08	0.95 $\pm$ 0.1	0.88 $\pm$ 0.04	0.94 $\pm$ 0.08	0.88 $\pm$ 0.03
	Chest	1.05 $\pm$ 0.02	1.24 $\pm$ 0.02	0.97 $\pm$ 0.01	1.23 $\pm$ 0.06	0.96 $\pm$ 0.01	1.22 $\pm$ 0.09
70%	Art1	0.99 $\pm$ 0.03	1.36 $\pm$ 0.02	0.47 $\pm$ 0.02	1.09 $\pm$ 0.02	0.43 $\pm$ 0.02	1.1 $\pm$ 0.04
	Art2	0.98 $\pm$ 0.07	1.22 $\pm$ 0.04	0.99 $\pm$ 0.07	0.97 $\pm$ 0.08	0.98 $\pm$ 0.09	0.99 $\pm$ 0.07
	Mimic	1.03 $\pm$ 0.01	1.29 $\pm$ 0.02	0.99 $\pm$ 0.01	1.0 $\pm$ 0.01	0.98 $\pm$ 0.01	1.0 $\pm$ 0.02
	Parkinson	16.34 $\pm$ 0.37	19.29 $\pm$ 0.78	16.47 $\pm$ 0.78	15.92 $\pm$ 0.32	15.97 $\pm$ 0.32	15.9 $\pm$ 0.34
	Obesity	0.61 $\pm$ 0.03	1.16 $\pm$ 0.02	0.43 $\pm$ 0.02	1.03 $\pm$ 0.08	0.41 $\pm$ 0.02	1.05 $\pm$ 0.08
	Compl.	0.89 $\pm$ 0.03	1.47 $\pm$ 0.02	0.93 $\pm$ 0.03	0.9 $\pm$ 0.02	0.9 $\pm$ 0.01	0.89 $\pm$ 0.03
	Chest	0.98 $\pm$ 0.02	1.23 $\pm$ 0.01	0.97 $\pm$ 0.02	1.2 $\pm$ 0.1	0.96 $\pm$ 0.02	1.22 $\pm$ 0.09
100%	Art1	0.98 $\pm$ 0.02	1.35 $\pm$ 0.02	0.48 $\pm$ 0.02	1.07 $\pm$ 0.03	0.44 $\pm$ 0.01	1.09 $\pm$ 0.04
	Art2	0.95 $\pm$ 0.06	1.23 $\pm$ 0.06	1.04 $\pm$ 0.1	0.98 $\pm$ 0.06	1.01 $\pm$ 0.1	0.98 $\pm$ 0.07
	Mimic	1.0 $\pm$ 0.01	1.28 $\pm$ 0.01	1.0 $\pm$ 0.01	1.01 $\pm$ 0.02	0.98 $\pm$ 0.01	1.0 $\pm$ 0.02
	Parkinson	15.82 $\pm$ 0.35	19.37 $\pm$ 0.51	16.65 $\pm$ 0.82	15.87 $\pm$ 0.31	15.86 $\pm$ 0.31	15.85 $\pm$ 0.33
	Obesity	0.6 $\pm$ 0.03	1.16 $\pm$ 0.02	0.43 $\pm$ 0.03	1.03 $\pm$ 0.08	0.41 $\pm$ 0.02	1.05 $\pm$ 0.08
	Compl.	0.87 $\pm$ 0.03	1.46 $\pm$ 0.02	0.97 $\pm$ 0.03	0.91 $\pm$ 0.05	0.89 $\pm$ 0.02	0.88 $\pm$ 0.02
	Chest	0.96 $\pm$ 0.02	1.23 $\pm$ 0.01	0.99 $\pm$ 0.02	1.13 $\pm$ 0.13	0.95 $\pm$ 0.02	1.22 $\pm$ 0.09

Table 9: Root mean square error (RMSE) for the considered methods, **labeling strategy S3** and for $\pi = 0.5$. A **Poisson model** was used as the basic regression model. The method with the lowest RMSE is highlighted in bold, and the second best is in light gray. The methods CAL and ERM are excluded from comparison as they assume known $e(x)$.

c	Dataset	NAIVE	LS	CAL (EST)	ERM (EST)	CAL	ERM
10%	Art1	1.05 $\pm$ 0.03	1.54 $\pm$ 0.04	0.36 $\pm$ 0.01	1.1 $\pm$ 0.03	0.4 $\pm$ 0.07	1.07 $\pm$ 0.03
	Art2	1.31 $\pm$ 0.11	1.21 $\pm$ 0.05	1.17 $\pm$ 0.16	1.13 $\pm$ 0.19	1.09 $\pm$ 0.12	1.24 $\pm$ 0.48
	Mimic	1.26 $\pm$ 0.02	1.4 $\pm$ 0.09	1.12 $\pm$ 0.03	1.01 $\pm$ 0.02	1.11 $\pm$ 0.1	1.01 $\pm$ 0.02
	Parkinson	19.36 $\pm$ 0.47	22.34 $\pm$ 2.2	19.14 $\pm$ 1.3	17.22 $\pm$ 1.32	19.79 $\pm$ 1.69	17.28 $\pm$ 1.43
	Obesity	1.51 $\pm$ 0.15	3.19 $\pm$ 0.3	1.29 $\pm$ 0.14	1.47 $\pm$ 0.1	1.25 $\pm$ 0.12	1.45 $\pm$ 0.11
	Compl.	1.08 $\pm$ 0.03	1.6 $\pm$ 0.11	1.05 $\pm$ 0.09	1.01 $\pm$ 0.16	1.22 $\pm$ 0.13	0.97 $\pm$ 0.11
	Chest	1.19 $\pm$ 0.02	1.28 $\pm$ 0.04	1.04 $\pm$ 0.03	1.23 $\pm$ 0.07	1.02 $\pm$ 0.07	1.22 $\pm$ 0.09
40%	Art1	1.05 $\pm$ 0.03	1.52 $\pm$ 0.04	0.37 $\pm$ 0.01	1.1 $\pm$ 0.04	0.52 $\pm$ 0.4	1.09 $\pm$ 0.04
	Art2	1.11 $\pm$ 0.15	1.2 $\pm$ 0.05	1.01 $\pm$ 0.12	0.96 $\pm$ 0.05	1.0 $\pm$ 0.08	0.99 $\pm$ 0.08
	Mimic	1.11 $\pm$ 0.02	1.32 $\pm$ 0.02	1.0 $\pm$ 0.02	1.0 $\pm$ 0.01	0.98 $\pm$ 0.01	1.0 $\pm$ 0.02
	Parkinson	17.37 $\pm$ 0.46	20.43 $\pm$ 1.31	17.07 $\pm$ 0.71	16.16 $\pm$ 0.73	16.37 $\pm$ 0.67	16.15 $\pm$ 0.71
	Obesity	0.71 $\pm$ 0.03	1.32 $\pm$ 0.06	0.43 $\pm$ 0.03	1.06 $\pm$ 0.05	0.61 $\pm$ 0.15	1.06 $\pm$ 0.08
	Compl.	0.97 $\pm$ 0.03	1.51 $\pm$ 0.05	0.93 $\pm$ 0.04	0.91 $\pm$ 0.06	0.93 $\pm$ 0.03	0.89 $\pm$ 0.02
	Chest	1.05 $\pm$ 0.02	1.24 $\pm$ 0.02	0.98 $\pm$ 0.02	1.23 $\pm$ 0.08	0.96 $\pm$ 0.01	1.22 $\pm$ 0.09
70%	Art1	1.01 $\pm$ 0.04	1.36 $\pm$ 0.03	0.48 $\pm$ 0.02	1.07 $\pm$ 0.03	0.44 $\pm$ 0.02	1.08 $\pm$ 0.04
	Art2	1.02 $\pm$ 0.1	1.24 $\pm$ 0.02	1.02 $\pm$ 0.1	1.07 $\pm$ 0.2	1.04 $\pm$ 0.13	1.07 $\pm$ 0.26
	Mimic	1.02 $\pm$ 0.01	1.29 $\pm$ 0.02	0.99 $\pm$ 0.01	1.0 $\pm$ 0.02	0.98 $\pm$ 0.01	1.0 $\pm$ 0.02
	Parkinson	16.27 $\pm$ 0.35	19.27 $\pm$ 0.71	16.25 $\pm$ 0.27	15.92 $\pm$ 0.32	15.87 $\pm$ 0.29	15.88 $\pm$ 0.33
	Obesity	0.61 $\pm$ 0.03	1.16 $\pm$ 0.02	0.44 $\pm$ 0.02	1.06 $\pm$ 0.07	0.41 $\pm$ 0.02	1.05 $\pm$ 0.08
	Compl.	0.89 $\pm$ 0.03	1.48 $\pm$ 0.03	0.93 $\pm$ 0.04	0.89 $\pm$ 0.04	0.91 $\pm$ 0.02	0.89 $\pm$ 0.03
	Chest	0.98 $\pm$ 0.02	1.24 $\pm$ 0.01	0.97 $\pm$ 0.02	1.25 $\pm$ 0.02	0.95 $\pm$ 0.02	1.22 $\pm$ 0.09
100%	Art1	1.0 $\pm$ 0.02	1.37 $\pm$ 0.02	0.48 $\pm$ 0.01	1.1 $\pm$ 0.03	0.44 $\pm$ 0.02	1.08 $\pm$ 0.03
	Art2	0.97 $\pm$ 0.07	1.2 $\pm$ 0.06	0.99 $\pm$ 0.06	0.95 $\pm$ 0.06	1.0 $\pm$ 0.15	0.99 $\pm$ 0.06
	Mimic	1.0 $\pm$ 0.01	1.28 $\pm$ 0.01	1.0 $\pm$ 0.01	1.0 $\pm$ 0.01	0.98 $\pm$ 0.01	1.0 $\pm$ 0.02
	Parkinson	15.83 $\pm$ 0.36	19.36 $\pm$ 0.52	16.61 $\pm$ 0.74	15.84 $\pm$ 0.31	15.84 $\pm$ 0.32	15.85 $\pm$ 0.33
	Obesity	0.6 $\pm$ 0.03	1.16 $\pm$ 0.02	0.43 $\pm$ 0.03	1.03 $\pm$ 0.08	0.41 $\pm$ 0.02	1.05 $\pm$ 0.08
	Compl.	0.87 $\pm$ 0.03	1.46 $\pm$ 0.02	0.97 $\pm$ 0.03	0.9 $\pm$ 0.03	0.89 $\pm$ 0.02	0.88 $\pm$ 0.02
	Chest	0.96 $\pm$ 0.02	1.23 $\pm$ 0.01	0.99 $\pm$ 0.02	1.16 $\pm$ 0.12	0.95 $\pm$ 0.02	1.22 $\pm$ 0.09

Table 10: Root mean square error (RMSE) for the considered methods, **labeling strategy S4** and for $\pi = 0.5$. A **Poisson model** was used as the basic regression model. The method with the lowest RMSE is highlighted in bold, and the second best is in light gray. The methods CAL and ERM are excluded from comparison as they assume known $e(x)$.

c	Dataset	NAIVE	LS	CAL (EST)	ERM (EST)	CAL	ERM
10%	Art1	1.63 $\pm$ 0.06	1.65 $\pm$ 0.04	0.37 $\pm$ 0.04	1.11 $\pm$ 0.04	0.72 $\pm$ 0.42	1.08 $\pm$ 0.04
	Art2	1.34 $\pm$ 0.09	1.45 $\pm$ 0.12	1.13 $\pm$ 0.11	1.05 $\pm$ 0.08	1.07 $\pm$ 0.1	1.09 $\pm$ 0.1
	Mimic	1.2 $\pm$ 0.02	1.58 $\pm$ 0.04	1.07 $\pm$ 0.04	1.0 $\pm$ 0.02	1.05 $\pm$ 0.04	1.0 $\pm$ 0.02
	Parkinson	18.32 $\pm$ 0.45	26.54 $\pm$ 1.76	19.77 $\pm$ 2.81	16.13 $\pm$ 0.39	17.97 $\pm$ 0.84	16.14 $\pm$ 0.44
	Obesity	1.28 $\pm$ 0.1	2.45 $\pm$ 0.56	0.93 $\pm$ 0.1	1.37 $\pm$ 0.35	1.05 $\pm$ 0.23	1.27 $\pm$ 0.16
	Compl.	1.02 $\pm$ 0.03	1.79 $\pm$ 0.09	0.99 $\pm$ 0.06	0.9 $\pm$ 0.03	1.11 $\pm$ 0.07	0.89 $\pm$ 0.03
	Chest	1.12 $\pm$ 0.02	1.54 $\pm$ 0.05	0.99 $\pm$ 0.01	1.24 $\pm$ 0.04	1.0 $\pm$ 0.01	1.22 $\pm$ 0.09
40%	Art1	1.63 $\pm$ 0.1	1.62 $\pm$ 0.06	0.54 $\pm$ 0.3	1.09 $\pm$ 0.08	0.51 $\pm$ 0.16	1.11 $\pm$ 0.08
	Art2	1.37 $\pm$ 0.13	1.43 $\pm$ 0.1	1.16 $\pm$ 0.09	1.14 $\pm$ 0.19	1.11 $\pm$ 0.1	1.11 $\pm$ 0.09
	Mimic	1.17 $\pm$ 0.05	1.58 $\pm$ 0.06	1.06 $\pm$ 0.06	0.99 $\pm$ 0.07	1.06 $\pm$ 0.07	1.02 $\pm$ 0.05
	Parkinson	16.65 $\pm$ 0.46	26.58 $\pm$ 1.14	18.23 $\pm$ 1.55	15.98 $\pm$ 0.28	17.23 $\pm$ 0.73	15.94 $\pm$ 0.36
	Obesity	0.95 $\pm$ 0.07	1.28 $\pm$ 0.08	0.49 $\pm$ 0.07	1.24 $\pm$ 0.37	0.57 $\pm$ 0.11	1.08 $\pm$ 0.11
	Compl.	0.98 $\pm$ 0.06	1.74 $\pm$ 0.06	1.05 $\pm$ 0.11	0.92 $\pm$ 0.08	1.03 $\pm$ 0.08	0.88 $\pm$ 0.06
	Chest	1.12 $\pm$ 0.05	1.54 $\pm$ 0.06	0.98 $\pm$ 0.05	1.25 $\pm$ 0.06	1.01 $\pm$ 0.06	1.24 $\pm$ 0.11
70%	Art1	1.6 $\pm$ 0.06	1.61 $\pm$ 0.08	0.62 $\pm$ 0.2	1.06 $\pm$ 0.08	0.58 $\pm$ 0.1	1.16 $\pm$ 0.05
	Art2	1.39 $\pm$ 0.09	1.65 $\pm$ 0.43	1.16 $\pm$ 0.09	1.06 $\pm$ 0.07	1.07 $\pm$ 0.12	1.11 $\pm$ 0.18
	Mimic	1.18 $\pm$ 0.06	1.57 $\pm$ 0.04	1.09 $\pm$ 0.04	1.04 $\pm$ 0.04	1.06 $\pm$ 0.05	1.04 $\pm$ 0.05
	Parkinson	16.69 $\pm$ 0.43	26.59 $\pm$ 1.16	18.21 $\pm$ 1.54	15.97 $\pm$ 0.3	17.25 $\pm$ 0.74	15.93 $\pm$ 0.36
	Obesity	0.89 $\pm$ 0.06	1.3 $\pm$ 0.09	0.46 $\pm$ 0.06	1.24 $\pm$ 0.4	0.57 $\pm$ 0.11	1.11 $\pm$ 0.09
	Compl.	1.0 $\pm$ 0.05	1.73 $\pm$ 0.07	1.02 $\pm$ 0.13	0.93 $\pm$ 0.05	1.0 $\pm$ 0.06	0.92 $\pm$ 0.06
	Chest	1.13 $\pm$ 0.06	1.53 $\pm$ 0.05	0.99 $\pm$ 0.08	1.23 $\pm$ 0.05	1.01 $\pm$ 0.03	1.2 $\pm$ 0.08
100%	Art1	1.6 $\pm$ 0.04	1.64 $\pm$ 0.09	0.56 $\pm$ 0.26	1.08 $\pm$ 0.06	0.49 $\pm$ 0.09	1.12 $\pm$ 0.07
	Art2	1.47 $\pm$ 0.21	1.5 $\pm$ 0.13	1.15 $\pm$ 0.12	1.06 $\pm$ 0.13	1.01 $\pm$ 0.06	1.03 $\pm$ 0.08
	Mimic	1.15 $\pm$ 0.04	1.56 $\pm$ 0.04	1.08 $\pm$ 0.09	1.01 $\pm$ 0.04	1.07 $\pm$ 0.04	1.02 $\pm$ 0.05
	Parkinson	16.68 $\pm$ 0.42	26.57 $\pm$ 1.16	18.21 $\pm$ 1.54	16.0 $\pm$ 0.29	17.25 $\pm$ 0.74	15.95 $\pm$ 0.37
	Obesity	0.92 $\pm$ 0.12	1.28 $\pm$ 0.07	0.51 $\pm$ 0.06	1.23 $\pm$ 0.39	0.57 $\pm$ 0.1	1.07 $\pm$ 0.07
	Compl.	0.98 $\pm$ 0.06	1.77 $\pm$ 0.05	1.04 $\pm$ 0.14	0.92 $\pm$ 0.05	1.02 $\pm$ 0.06	0.87 $\pm$ 0.05
	Chest	1.11 $\pm$ 0.05	1.51 $\pm$ 0.04	1.0 $\pm$ 0.05	1.26 $\pm$ 0.04	1.05 $\pm$ 0.04	1.23 $\pm$ 0.07

Table 11: The effect of propensity score estimation on the RMSE of the CAL method. A **linear model** was used as the baseline regression model, with labeling strategy S2 and $\pi = 0.5$. The method with the lowest RMSE is highlighted in bold, and the second best is in light gray. (CAL which assumes known $e(x)$ is excluded from comparison). Four methods of estimating $e(x)$ are considered: SAREM [Bekker et al., 2019], PGLIN [Gerych et al., 2022], TM [Furmańczyk et al., 2023] and LBE [Gong et al., 2021a].

c	Dataset	CAL	CAL (SAREM)	CAL (PGLIN)	CAL (TM)	CAL (LBE)
10%	Mimic	1.05 ± 0.08	1.1 ± 0.05	1.19 ± 0.02	1.18 ± 0.1	1.37 ± 0.14
	Parkinson	19.19 ± 1.41	18.62 ± 1.25	18.62 ± 0.66	21.53 ± 3.0	22.74 ± 2.97
	Obesity	0.77 ± 0.17	0.86 ± 0.09	0.94 ± 0.11	0.99 ± 0.1	1.62 ± 0.49
	Complications	1.1 ± 0.04	1.0 ± 0.06	1.0 ± 0.02	1.2 ± 0.07	1.3 ± 0.17
	Chest	1.01 ± 0.01	1.01 ± 0.02	1.11 ± 0.02	1.21 ± 0.04	1.19 ± 0.1
40%	Mimic	0.99 ± 0.01	1.01 ± 0.02	1.02 ± 0.02	1.15 ± 0.05	1.35 ± 0.05
	Parkinson	16.16 ± 0.26	16.98 ± 0.47	16.51 ± 0.34	18.39 ± 0.9	20.06 ± 1.15
	Obesity	0.44 ± 0.06	0.42 ± 0.04	0.44 ± 0.04	0.47 ± 0.03	0.82 ± 0.12
	Complications	0.9 ± 0.02	0.93 ± 0.05	0.9 ± 0.02	1.04 ± 0.08	1.12 ± 0.15
	Chest	0.98 ± 0.01	0.99 ± 0.01	0.98 ± 0.02	1.27 ± 0.02	1.22 ± 0.11
70%	Mimic	0.99 ± 0.01	1.0 ± 0.01	0.98 ± 0.01	1.12 ± 0.01	1.3 ± 0.06
	Parkinson	15.93 ± 0.29	16.45 ± 0.78	15.96 ± 0.32	17.78 ± 0.49	19.18 ± 0.79
	Obesity	0.34 ± 0.02	0.34 ± 0.02	0.34 ± 0.02	0.34 ± 0.01	0.78 ± 0.01
	Complications	0.88 ± 0.01	0.9 ± 0.02	0.88 ± 0.01	0.99 ± 0.04	0.97 ± 0.12
	Chest	0.98 ± 0.01	1.01 ± 0.02	0.97 ± 0.02	1.25 ± 0.02	1.33 ± 0.03
100%	Mimic	0.98 ± 0.01	1.03 ± 0.02	0.99 ± 0.01	1.1 ± 0.01	1.29 ± 0.04
	Parkinson	15.84 ± 0.31	16.62 ± 0.8	15.85 ± 0.28	17.74 ± 0.63	19.33 ± 0.51
	Obesity	0.34 ± 0.02	0.34 ± 0.02	0.34 ± 0.02	0.34 ± 0.02	0.78 ± 0.01
	Complications	0.87 ± 0.01	0.9 ± 0.01	0.88 ± 0.01	0.95 ± 0.03	0.93 ± 0.11
	Chest	0.98 ± 0.01	1.04 ± 0.02	1.01 ± 0.01	1.17 ± 0.01	1.35 ± 0.03

Table 12: Computational times (averaged over 30 simulations) for the largest considered MIMIC dataset and example parameter setting: $\pi = 0.5$, $c = 0.5$, strategy S2. The estimation of the propensity score in the ERM (EST) and CAL (EST) methods is the most time-consuming component of the algorithms.

Model	NAIVE	LS	CAL	CAL (EST)	ERM	ERM (EST)
Linear	0.061	0.056	0.131	20.901	0.099	20.369
Poisson	0.107	1.093	0.915	20.181	0.333	18.865