Quantifying Inequality in Underreported Medical Conditions

Divya Shanmugam* DIVYAS@MIT.EDU
Massachusetts Institute of Technology, USA

Emma Pierson* EMMA.PIERSON@CORNELL.EDU
Cornell Tech and Weill Cornell Medical College, USA

Abstract

Estimating the *prevalence* of a medical condition, or the proportion of the population in which it occurs, is a fundamental problem in healthcare and public health. Accurate estimates of the *relative prevalence* across groups — capturing, for example, that a condition affects women more frequently than men — facilitate effective and equitable health policy which prioritizes groups who are disproportionately affected by a condition. However, it is difficult to estimate relative prevalence when a medical condition is underreported. In this work, we provide a method for accurately estimating the relative prevalence of underreported medical conditions, building upon the positive unlabeled learning framework. We show that under the commonly made *covariate shift* assumption — i.e., that the probability of having a disease conditional on symptoms remains constant across groups — we can recover the relative prevalence, even without restrictive assumptions commonly made in positive unlabeled learning and even if it is impossible to recover the absolute prevalence. We provide a suite of experiments on synthetic and real health data that demonstrate our method's ability to recover the relative prevalence more accurately than do baselines, and the method's robustness to plausible violations of the covariate shift assumption.

1. Introduction

Reducing health disparities requires quantifying how a disease disproportionately affects different groups. The *relative prevalence* captures how much more frequently a condition occurs in one group compared to another — $\frac{\text{prevalence in group A}}{\text{prevalence in group B}}$ — with high relative prevalence estimates suggesting concrete areas for funding, outreach, and research (Penman-Aguilar et al., 2016). As one example, consider the diabetes epidemic among Native Americans in the early 1990s. Hundreds of papers were written about the high relative prevalence of diabetes for Native populations in the United States (Edwards and Patchell, 2009). In response, the U.S. Congress established the Special Diabetes Program for Indians, which developed community-based interventions in collaboration with over 40 tribes. Over the next 16 years, Native American tribes saw a massive decrease in diabetes related complications (30-50%) (IHS, 2017).

However, it is difficult to replicate this success for underreported medical conditions for several reasons. First, a small percentage of true positives are labeled as positive, producing inaccurate prevalence estimates. Second, the probability of a positive diagnosis can vary by group (Geiger, 2003). Consider intimate partner violence, a notoriously underdiagnosed condition: not only is the diagnosis probability of true cases estimated to be only $\sim$25%, but this probability varies across racial groups (Schafer et al., 2008).

The difficulty of estimating prevalence of underreported conditions has been formalized in the positive unlabeled (PU) learning literature, which assumes that only some positive cases are correctly labeled as positive. Past work in PU learning has proven that without assumptions about the separability of the positive and negative classes, it is impossible to estimate the prevalence of an underdiagnosed condition (Scott, 2015). Many PU learning methods consequently introduce restrictive assumptions in order to recover the prevalence, which do not accurately reflect healthcare data.

In this work, we present `PURPLE` (**P**ositive **U**nlabeled **R**elative **P**reva**L**ence **E**stimator), a method that can produce relative prevalence estimates for underreported medical conditions, given three assumptions: 1) per-group random diagnosis;

* Work done while at Microsoft Research

2) no false positives; and 3) covariate shift between groups, i.e., that the probability of having a disease conditional on symptoms remains constant across groups. The first two assumptions are standard in PU learning; the third, which is specific to our method, replaces restrictive PU assumptions like separability of positive and negative cases. We show that if these assumptions are satisfied, it is possible to recover the relative prevalence even if it is not possible to recover the absolute prevalence: that is, $\frac{\text{prevalence in group A}}{\text{prevalence in group B}}$ can be estimated even if neither the numerator nor denominator can. We demonstrate via experiments on synthetic and real health data that PURPLE recovers the relative prevalence more accurately than existing baselines, and illustrate that PURPLE still provides a useful lower bound on the magnitude of disparities even if the covariate shift assumption is violated.

We make the following contributions:

- We formalize *relative prevalence estimation* as a PU learning problem which is statistically feasible even in settings where absolute prevalence estimation is impossible.
- We outline three assumptions necessary to recover the relative prevalence and discuss the validity of these assumptions in the context of underreported health conditions.
- We present PURPLE, a method to estimate the relative prevalence of an underreported condition, and provide a set of experiments on both synthetic and real health data demonstrating that PURPLE outperforms common PU baselines at this task.

2. Related Work

Estimating prevalence is a central task in multiple fields. We describe relevant prior work in PU learning, epidemiology, and modelling underreported medical conditions below.

2.1. PU Learning

The area of work most closely related to our own is PU learning, a variant of semi-supervised learning which assumes access to a set of labeled positive examples and a typically larger set of unlabeled examples. PU learning is a natural framework for describing underdiagnosed conditions, where false positives are rare and false negatives are likely. For a detailed review of the PU learning literature, we refer the reader to Bekker and Davis (2020).

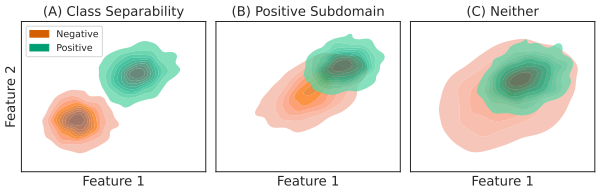

Figure 1: Pictured are three possible relationships between the positive and negative distributions. Most work in PU learning assumes (A), and while some methods can accommodate (B), none address (C). PURPLE can estimate relative prevalence in all three settings, assuming that patient subgroups have the same $p(y|x)$.

Estimating prevalence[1] is a critical step for many PU learning methods (Bekker and Davis, 2020). However, the prevalence is only identifiable under a restrictive set of assumptions about the structure of the positive and negative distributions (Jain et al., 2016). Many PU learning methods assume the distributions have no overlap (Elkan and Noto, 2008; Du Plessis and Sugiyama, 2014; Northcutt et al., 2017) (Fig. 1a), while some accommodate overlapping distributions with distinct support (Ramaswamy et al., 2016) (Fig. 1b). Łazęcka et al. (2021) introduce parametric assumptions and assume a specific functional form to estimate the class prior. These assumptions are rarely true for medical conditions because a set of symptoms is unlikely to correspond to a diagnosis with 100% probability.

Our work differs from past PU learning methods for inferring prevalence because 1) our goal is inferring the *relative* prevalence, not absolute prevalence and 2) our method is applicable even in instances when standard PU assumptions about class separability (Fig. 1a and 1b) are not satisfied (Fig. 1c).

2.2. Epidemiology

Epidemiological methods for prevalence estimation fall into two categories: *direct* and *indirect*. Direct methods rely on a well-known rule to separate positive cases from negative cases. Methods to estimate prevalence given imperfect diagnostic rules exist, but

1. The PU learning literature typically refers to the prevalence as the "class prior"; we use "prevalence" throughout this paper for consistency, but the terms are equivalent.

assume that the sensitivity and specificity of these rules are known (Diggle, 2011; Lewis and Torgerson, 2012; Haine et al., 2018). Direct methods are known to produce biased estimates in the context of underreported conditions (Hickman and Taylor, 2005).

Indirect methods instead assume access to external information. Capture-Recapture Analysis (Sekar and Deming, 1949) requires multiple independent samples from a closed population and has produced prevalence estimates for muscular dystrophy (Smith et al., 2017), multiple sclerosis (Salhofer-Polanyi et al., 2017), and diabetes (LaPorte et al., 1993). Another approach is to use longitudinal data to model the distribution of visits per patient (Hay and Smit, 2003; Law et al., 2001; McKeganey et al., 1992). Data may also be aggregated over multiple disease indicators and sources to produce a more comprehensive prevalence estimate than any one dataset or indicator can provide (Simeone et al., 1995).

In contrast, we assume no knowledge about the sensitivity or specificity of the observed diagnoses. Furthermore, we assume no access to external data (e.g., disease registries, multiple hospital samples, or longitudinal data).

2.3. Modelling Underreported Conditions

A majority of machine learning approaches applied to rare conditions aim to correctly diagnose positive cases rather than produce prevalence estimates (Schaefer et al., 2020). Cui et al. (2020) propose a new method to address class imbalance in rare disease detection by generating synthetic positive cases using a generative adversarial network and note the dearth of machine learning work on low-prevalence diseases. Rigg et al. (2015) argue that SVMs are preferable to regression-based approaches for low-prevalence diseases. Applications of traditional ML approaches to diagnosing rare gynecologic (Tejera et al., 2011), neurological (Artoni et al., 2020), and cardiac diseases (Sengupta et al., 2016) have also been explored. Our work differs from this literature because our goal is relative prevalence estimation, not diagnosis of individual cases. In principle, methodological innovations developed by this literature to improve diagnostic accuracy, including new approaches to model training or data augmentation, may be applied in conjunction with the ideas presented here.

3. Problem Setting

We adopt terminology standard in the PU learning literature and assume that we have access to three pieces of data for the ith patient: a symptom vector x_i; their group membership g_i; and their binary observed diagnosis s_i, which is 0 if the patient is not diagnosed with the medical condition of interest and 1 if they are. We use y_i to denote whether the patient truly *has* the medical condition. This is an unobserved binary variable and because the medical condition is underreported, not all patients who truly have the condition are diagnosed with it, so $p(s_i = 1|y_i = 1) < 1$. Because we are interested in health disparities, we focus on groups g defined by sensitive attributes (e.g., gender, race, or socioeconomic status) but in principle our method is applicable to any set of groups. We use α_a to denote the prevalence $p(y|a)$ of a condition in group a.

Task For any two groups a and b, we aim to estimate the *relative prevalence* $\rho_{a,b}$ between group a and group b:

$$\rho_{a,b} = \frac{\alpha_a}{\alpha_b} = \frac{p(y = 1|g = a)}{p(y = 1|g = b)} \tag{1}$$

Estimating $\rho_{a,b}$ is easier than estimating α_a or α_b. We elaborate on why in the next section and show how restrictive assumptions of class separability (common to estimation approaches for α_a or α_b) are not necessary to estimate the relative prevalence.

Assumptions We make three assumptions:

1. *No False Positives*: We assume that examples labeled as positive ($s = 1$) are truly positive ($y = 1$): i.e., $p(y = 1|s = 0) = 0$. This is the positive unlabeled assumption and is the foundational assumption of PU learning methods.
2. *Random Diagnosis within Groups*: We assume that positive patients within a specific group are equally likely to receive a positive diagnosis: $p(s = 1|y = 1, g = a) = c_a$. This equates to the Selected-Completely-at-Random (SCAR) assumption within groups which is common in PU learning, where c_a represents the labeling frequency of group a.
3. *Covariate Shift between Groups*: We assume that while $p(x)$ varies across groups, $p(y|x)$ remains constant across groups: patients in different groups with the same symptoms are equally likely to have a medical condition. This is

the commonly made *covariate shift* assumption (Quiñonero-Candela et al., 2009), but it is not a standard assumption in PU learning setups.

Note that we make no assumptions about the separability of the positive and negative distributions, as much existing PU learning work does. Instead, we exchange these for the assumption of covariate shift, which past work has often used in medical settings (Nestor et al., 2019; Singh et al., 2021). While this assumption may not always apply—for example, the probability a patient has a medical condition may vary depending on their age group, even conditional on their symptoms—the assumption remains realistic in many cases, and it is not possible to estimate prevalence in the PU setting without making an assumption beyond (1) and (2) (Scott, 2015).

4. Method

The key idea underpinning our work is that knowing the exact prevalence of a condition in a specific group is not necessary to calculate the *relative* prevalence between groups: one can estimate a fraction without knowing its numerator or denominator. We will first show that one can recover the relative prevalence by estimating $p(y|x)$ up to a constant multiplicative factor. We then show that it is possible to conduct this estimation and conclude with implementation details.

4.1. Deriving the Relative Prevalence

We can compute the relative prevalence as follows:

$$\rho_{a,b} = \frac{p(y=1|g=a)}{p(y=1|g=b)} \tag{2}$$

$$= \frac{\sum_x p(y=1|x,g=a)p(x|g=a)}{\sum_x p(y=1|x,g=b)p(x|g=b)} \tag{3}$$

$$= \frac{\sum_x p(y=1|x)p(x|g=a)}{\sum_x p(y=1|x)p(x|g=b)} \tag{4}$$

$$= \frac{\sum_x \hat{p}(y=1|x)p(x|g=a)}{\sum_x \hat{p}(y=1|x)p(x|g=b)} \tag{5}$$

$$\text{for all } \hat{p}(y=1|x) \propto p(y=1|x)$$

where (4) follows from the covariate shift assumption and (5) follows because estimates of $p(y=1|x)$ up to a constant multiplicative factor will yield a constant term in the numerator and denominator which cancels. Thus, estimates of $p(y=1|x)$ up to a constant multiplicative factor suffice to recover the relative prevalence.

4.2. Estimating $\hat{p}(y=1|x)$

We have shown that, if we can estimate $\hat{p}(y=1|x) \propto p(y=1|x)$, we can use it to compute the relative prevalence $\rho_{a,b}$. Now we show how to estimate $\hat{p}(y=1|x)$. We do so by applying our three assumptions:

$$p(s=1|x,g) = p(y=1|x,g)p(s=1|y=1,x,g) \tag{6}$$

$$+ \; p(y=0|x,g)p(s=1|y=0,x,g)$$

$$= p(y=1|x,g)p(s=1|y=1,x,g) \tag{7}$$

$$= p(y=1|x,g)p(s=1|y=1,g) \tag{8}$$

$$= p(y=1|x)p(s=1|y=1,g) \tag{9}$$

Applying the *No False Positives* assumption allows us to remove the second term in (6). The *Random Labeling within Groups* assumption removes the dependence of the diagnosis probability on x, leading to (8). The *Covariate Shift* assumption leads to (9).

Thus, $p(s=1|x,g)$, which can be estimated from the observed data, is the product of two terms: the probability of a condition given a set of symptoms, $p(y=1|x)$, and the group-specific probability of diagnosis for a positive case, $p(s=1|y=1,g)$. Neither of these terms is identifiable because we can always multiply one by a constant factor while dividing the other by the same factor. However, as shown in §4.1, estimating $p(y|x)$ up to a constant multiplicative factor is sufficient to estimate the relative prevalence. Thus, we estimate $p(y|x)$ and $p(s=1|y=1,g)$ up to constant multiplicative factors by fitting to $p(s=1|x,g)$; we then use our constant-factor estimate of $p(y|x)$ as described in §4.1 to estimate the relative prevalence. Note that the method extends to any number of groups, where each group entails an additional parameter for the group-specific labeling frequency.

4.3. Implementation

We implement the model in PyTorch (Paszke et al., 2017) using a single-layer neural network to represent $p(y|x)$ and group-specific parameters $c_g = p(s=1|y=1,g)$ for each group g. We train the model using the Adam optimizer with default parameters. We use L1 regularization because our experiments are conducted on high-dimensional symptom vectors, most of which we expect to be unrelated to the medical condition, and select the regularization parameter $\lambda \in [10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}]$ using a held-out validation set by maximizing the AUC with respect to the diagnosis labels s. While we use a single-layer

neural network because our symptoms x are one-hot encoded and we do not anticipate interactions between symptoms, our same approach could be applied with deeper neural network architectures to accommodate interactions and nonlinearities.

5. Experimental Set-up

5.1. Datasets

We make use of two datasets: *Gauss-Synth*, a completely synthetic dataset, and *MIMIC-Semi-Synth*, a semi-synthetic dataset based on real health data. We rely on synthetic and semi-synthetic data so that the ground truth labels y are known, as is standard in the PU learning literature (Bekker and Davis, 2020), enabling us to assess how well methods recover the relative prevalence. Code to reproduce all experiments can be found at https://github.com/epierson9/invisible-conditions.

5.1.1. Gauss-Synth

We simulate data with two groups, a and b, which are both 5D Gaussian distributions. The likelihood function is a logistic function of the signed distance to a hyperplane through the origin. The generative model is:

$$x_i \sim \begin{cases} \mathcal{N}_5(-\mathbf{1}, 16 \cdot \mathbb{1}) \text{ if } g_i = a \\ \mathcal{N}_5(\mathbf{1}, 16 \cdot \mathbb{1}) \text{ if } g_i = b \end{cases} \quad (10)$$

$$y_i \sim \text{Bernoulli}(\sigma((\mathbf{1}^T x_i) \,/\, \|\mathbf{1}\|)) \quad (11)$$

$$s_i \sim \text{Bernoulli}(c_{g_i} y_i) \quad (12)$$

where σ represents a logistic function ($\sigma(x) = \frac{1}{1+e^{-x}}$) and c_{g_i} is the group-specific labeling frequency for g_i. We generate data with 10,000 observations for group a and 20,000 for group b.

5.1.2. MIMIC-Semi-Synth

We generate semi-synthetic data using MIMIC-IV, a public dataset of real patient visits to a Boston-area hospital over the course of 2008-2018 (Johnson et al., 2020). We filter out ICD codes that appear 10 or fewer times, leaving 5,544 unique ICD codes. Each feature vector x_i is a one-hot vector corresponding to the ICD codes assigned in a particular patient visit to the hospital. We generate true labels y based on a set of suspicious symptoms. Formally, this replaces (11) in our generative model with:

$$y_i \sim \text{Bernoulli}(\sigma(v_{sym}^T \mathbf{x_i}) \,/\, \|v_{sym}\|) \quad (13)$$

where v_{sym} is a one-hot encoding of the suspicious symptoms and $v_{sym}^T x_i$ corresponds to the number of suspicious symptoms present during a hospital visit. Thus, the probability a patient has a medical condition is a logistic function of the number of suspicious symptoms. As before, we have $s_i \sim \text{Bernoulli}(c_{g_i} y_i)$.

In all experiments, we compute the relative prevalence for Black (group a) versus white (group b) patients since these are the largest race groups in MIMIC data and provide each method with data from each group. However, our method can be applied to more than 2 groups, as described above. To assess how our method performs under diverse conditions, we experiment with selecting the suspicious symptoms in three different ways; details can be found in each experiment section below.

5.2. Baselines

Each of the baseline methods described below is designed to estimate the *absolute* prevalence. To obtain the relative prevalence, we apply the baseline to each group and take the ratio of the estimated absolute prevalences.

- *Negative*: Assigns all unlabeled examples a negative label. This approach replaces $p(y = 1|x)$ with $p(s = 1|x)$. Past work refers to this model as a *nontraditional classifier* (Bekker and Davis, 2020).
- *EM*: A PU learning method which learns the labeling frequency and likelihood function using an expectation-maximization approach (Bekker et al., 2019).
- *KM2*: A PU learning method which models the distribution of unlabeled examples as a mixture of the positive and negative distribution and estimates the proportion of positives using a kernel mean embedding approach (Ramaswamy et al., 2016). This method is recognized as state-of-the-art in prevalence estimation (Chen et al., 2019).
- *Supervised*: Uses the true label y to estimate $p(y|x)$; cannot actually be applied in real data since y is unobserved, but represents an upper bound on performance.

5.3. Metrics

We report the mean ratio of the estimated relative prevalence to the true relative prevalence over 5 random train/test splits of the dataset; values closer to 1 correspond to better performance. We use a paired t-test to compare PURPLE's performance to the performance of each baseline across the 5 train/test splits.

6. Synthetic Experiments

In our experiments on purely synthetic data, we compare PURPLE's performance to baseline performance as we alter three quantities: the separability of the positive and negative classes, the label frequency for each group, and the level of covariate shift.

6.1. Class Separability

Set-up We consider two separability conditions: separable data and non-separable data. The non-separable data is generated according to the default generative model of *Gauss-Synth* described earlier. To create the separable data, we start with the non-separable generative model and replace each $p(y = 1|x) > 0.5$ with $p(y = 1|x) = 1$ and each $p(y = 1|x) < 0.5$ with $p(y = 1|x) = 0$, removing the 40% of the data closest to the original decision boundary to ensure the classes are cleanly separable, as illustrated in Figure 1a.

Results Figure 2 compares PURPLE to existing work for prevalence estimation under the two different class separability settings. As expected, *Supervised* (green) is able to recover the true relative prevalence (black dotted line) exactly given access to the true labels y. The *Negative* baseline fails to recover the relative prevalence in either setting, which is unsurprising because it learns the diagnosis probability ($p(s = 1|x)$) rather than the medical condition probability ($p(y = 1|x)$). *EM* performs well in the separable setting but its performance degrades in the non-separable case because the use of a log-likelihood loss guides the model towards predicting probabilities of 1 or 0 for positive and negative examples, and ultimate produces the incorrect boundary. *KM2* fails to recover the relative prevalence in either setting. This result reflects a well-known weakness of kernel mean embedding approaches in higher dimensions (Scott and Sain, 2004). In a 1D setting, *KM2* is able to recover the relative prevalence exactly on the separable data, but is not able to on the non-separable data (results in the appendix). PURPLE is the only method to accurately recover the relative prevalence in both the separable and non-separable settings, significantly more accurately than existing work (p-value $< .0001$).

6.2. Label Frequency

Set-up Using the *Gauss-Synth* generative model, we set the label frequency of group a, c_{g_a}, to 0.5. We then investigate performance for $c_{g_b} \in [0.1, 0.3, 0.5, 0.7, 0.9]$. This simulates a range of diagnosis probabilities between groups.

Results Across all settings, PURPLE remains able to recover the relative prevalence across different sets of labeling frequencies more accurately than are the baselines (Fig. 2, p-value $< .0001$).

6.3. Covariate Shift

Set-up We modify the *Gauss-Synth* generative model and vary the magnitude of covariate shift between groups a and b by modifying the difference in group means ($\mathbb{E}[x|g = a]$ - $\mathbb{E}[x|g = b]$). The mean of g_a is fixed to be $\mathbf{1}$, while $\mathbb{E}[x|g = b]$ is set to $v \cdot \mathbf{1}$, with $v \in [-1, 0, 0.5, 0.75, 1]$. This simulates patient subgroups that vary to different extents in their distribution of symptoms.

Results Figure 2 plots the effects of the magnitude of covariate shift on method performance. At a covariate shift of 0, the group distributions are identical. Each method follows a similar trend: smaller magnitudes of covariate shift translate to more accurate relative prevalence estimates. PURPLE is the only method to maintain consistent performance (p-value $< .0001$) regardless of the extent to which the group distributions differ.

7. Semi-synthetic Experiments

We have established PURPLE's ability to recover the relative prevalence more accurately than existing baselines in purely synthetic data. We now investigate whether these results hold in more realistic data: specifically, the high-dimensional, sparse data characteristic of clinical settings. To provide a diverse set of tests of our method, we create three different types of

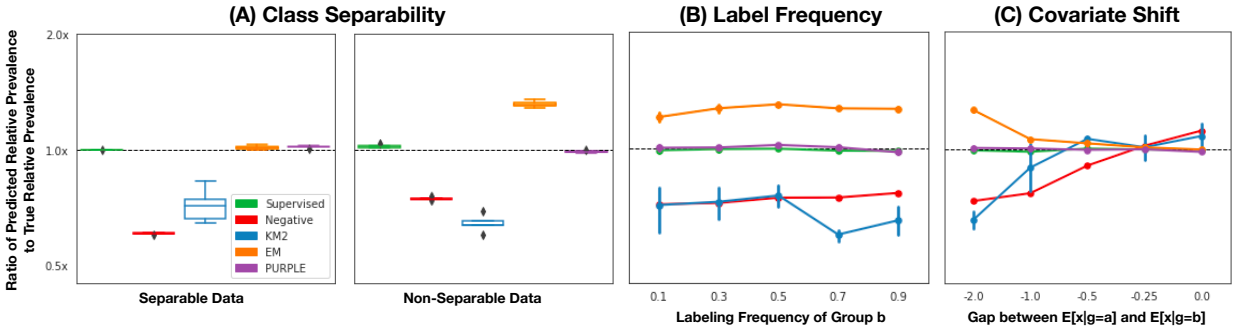

Figure 2: PURPLE recovers the relative prevalence more accurately and consistently than the *Negative*, *KM2*, and *EM* baselines across different synthetic settings for class separability, labeling frequency, and covariate shift, and comparably accurately to the supervised method which represents an upper bound on performance. Error bars are too small to appear for all methods except *KM2*. We include numerical results on error bars in the appendix.

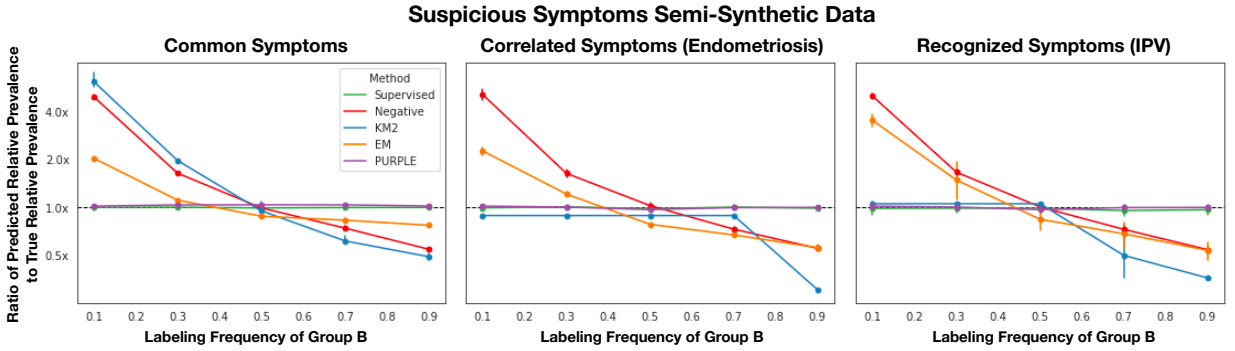

Figure 3: PURPLE recovers the relative prevalence accurately across three semi-synthetic datasets that draw suspicious symptoms from 1) symptoms that appear most frequently (left); 2) symptoms correlated with endometriosis (middle); and 3) symptoms known to correspond with intimate partner violence based on past literature (right). PURPLE produces more consistently accurate relative prevalence estimates across all label frequencies.

semi-synthetic data where the set of suspicious symptoms is drawn from: 1) the most common symptoms, 2) symptoms correlated with known diagnoses for a condition, and 3) symptoms known to be related to a condition from prior work. A full list of each suspicious symptom set can be found in the appendix.

7.1. Common Suspicious Symptoms

Symptom Definition We identify the 50 most common ICD codes in MIMIC-IV and randomly select 25 to be suspicious symptoms. Group a consists of 73,090 visits from Black patients ($p(y = 1|g = a) = 0.157$) and group b consists of 305,002 visits from White patients ($p(y = 1|g = b) = 0.185$).

Results The estimation error of the baselines on the real data far exceeds their error on *Gauss-Synth*, with some methods producing relative prevalence estimates more than 4x the true value. Further, each method produces both overestimates and underestimates the true relative prevalence at different label frequencies. As the labeling frequency of group b increases, the estimated positive prevalence in group b increases, which ultimately produces the downward trend the baselines exhibit across graphs. PURPLE remains accurate and robust to different labeling frequencies (p-value $< .0001$).

7.2. Correlated Suspicious Symptoms

We next consider endometriosis, a widely underdiagnosed condition in women's health (Moradi et al., 2014).

Symptom Definition We define our suspicious symptoms as the symptoms most highly associated with known endometriosis codes. We first identify a set of patients who receive any one of 10 ICD endometriosis diagnosis codes (appendix). We then identify the ICD codes which are most highly associated with a diagnosis of endometriosis: for each ICD code, we compute the ratio $\frac{\text{prevalence of ICD code among endometriosis patients}}{\text{prevalence of ICD code among all patients}}$. We define our suspicious symptoms as the 25 ICD codes with the highest value of this ratio, which includes known endometriosis symptoms such as "Excessive and frequent menstruation with regular cycle" and "Pelvic and perineal pain". We use these suspicious symptoms to generate y and s as described in §5.1.2, resulting in a prevalence of 5.1%. We do not include the 10 ICD codes used to determine the 25 suspicious symptoms in x.

We filter for female patients because endometriosis is extremely rare among men (Jabr and Mani, 2014), leaving 47,138 unique hospital visits from Black patients ($p(y = 1|g = a) = 0.0534$) and 165,653 unique hospital visits from white patients ($p(y = 1|g = a) = 0.0495$).

Results This setting exhibits a much lower prevalence compared to the previous section. Despite the greater class imbalance, previous results hold and `PURPLE` delivers accurate relative prevalence estimates across labeling frequencies (p-value $< .0001$). While *KM2* produces the most accurate estimate out of the remaining methods, it is significantly less accurate than `PURPLE` (p-value $< .0001$).

7.3. Recognized Suspicious Symptoms

In other cases, one may use domain knowledge to define the set of suspicious symptoms. Intimate partner violence (IPV), a dramatically underreported condition which motivates this work, is one such case, which we consider below.

Symptom Definition Prior work has found that suspicious symptoms for IPV include head, neck and facial injuries (Wu et al. (2010)). The symptoms in this experiment consist of the 100 ICD codes corresponding to these injuries.

We filter for female patients because the symptoms associated with IPV in men are not well understood (Houry et al., 2008). We also filter out patients less than 18 years old because it is difficult to distinguish between intimate partner violence and child abuse in minors. The simulated prevalence is 5.5% where $p(y = 1|g = a) = 0.0541$ (25,546 unique patient visits) and $p(y = 1|g = a) = 0.0568$ (80,227 unique patient visits).

Results Similar to the results in the context of endometriosis, `PURPLE` provides significantly more accurate relative prevalence estimates than existing work (p-value $< .0001$). This setting differs from endometriosis in two ways. First, the sample size is smaller. Second, the suspicious symptoms rarely occur together. In the case of endometriosis, the number of suspicious symptoms during a particular visit can range from 0 to 5 (out of the possible 25). This is not the case with IPV; patient visits in MIMIC-IV can contain up to 2 suspicious IPV symptoms (out of the possible 100). The performance of `PURPLE` in this setting suggests the method generalizes to different symptom distributions.

8. Robustness to violations of the covariate shift assumption

Our method relies on the covariate shift assumption. In this section, we characterize performance of our method when this assumption fails to hold. Under a plausible violation of the assumption, `PURPLE` still provides a *lower bound* on the magnitude of disparities. This lower bound is useful because we can be confident that if `PURPLE` infers that a group suffers disproportionately from a condition, that is in fact the case.

Assume that group a has a higher prevalence of a condition than group b: i.e., $p(y = 1|g = a) > p(y = 1|g = b)$. Then as long as $p(y = 1|x, g = a) > p(y = 1|x, g = b)$, `PURPLE` will correctly infer that group a has a higher prevalence, and `PURPLE`'s relative prevalence estimate will provide a lower bound on the true relative prevalence $\frac{p(y=1|g=a)}{p(y=1|g=b)}$. This follows directly from Equations (3) and (4); if we assume that $p(y = 1|x, g = a) = p(y = 1|x, g = b)$ when in fact $p(y = 1|x, g = a) > p(y = 1|x, g = b)$, we will

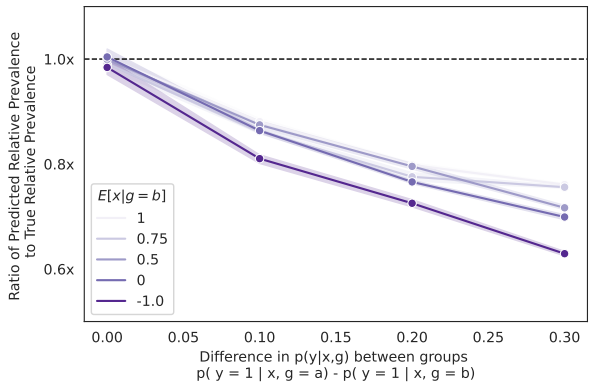

Figure 4: Under plausible violations of the covariate shift assumption, PURPLE will underestimate the true relative prevalence (dotted black line). Here, we plot this trend for different parameterizations of the Gaussian distributions each group is drawn from. Darker shades of purple correspond to distributions that are further from one another.

underestimate the fraction and thus the true relative prevalence.

The assumption that $p(y = 1|g = a) > p(y = 1|g = b) \rightarrow p(y = 1|x, g = a) > p(y = 1|x, g = b)$ is a reasonable one: when a condition is more prevalent in one group compared to another, the same symptoms plausibly correspond to higher posterior probabilities $p(y = 1|x)$ in the disproportionately affected group. For example, women are more likely than men to be victims of intimate partner violence overall (Houry et al., 2008), and if a woman and a man arrive in a hospital with the same injuries, doctors will be more likely to suspect intimate partner violence as the cause of the woman's injuries.

To demonstrate empirically that PURPLE provides a lower bound, we plot PURPLE's behavior over a range of values of $p(y = 1|x, g = a) - p(y = 1|x, g = b)$ (Figure 4), corresponding to a range of violations of the covariate shift assumption, using synthetic data as described in §5.1.1. In each case, PURPLE provides a lower bound on the true relative prevalence. We replicate this analysis over multiple parameterizations of the group-specific Gaussian distributions, where darker shades of purple correspond to distributions that are further from one another.

9. Discussion

In this work, we present *relative prevalence estimation*, a clinically important task previously unaddressed by the PU learning literature, which has focused on absolute prevalence estimation. We show that we can accomplish this task even in settings where absolute prevalence estimation is impossible, by exchanging the restrictive separability assumptions common in the PU learning literature for the covariate shift assumption, which is commonly used and arguably more appropriate in clinical settings. We present a new method, PURPLE, and show it outperforms baselines in terms of its ability to recover the relative prevalence on both synthetic and real health data.

We foresee opportunities for future work in two areas: PU learning methods and healthcare/public health. Methodologically, future work could explore other PU learning approaches to the task of relative prevalence estimation, as well as applications of PURPLE to relative prevalence estimation in the many other settings in which positive unlabeled data arises (Jaskie and Spanias, 2019). On the healthcare and public health side, PURPLE presents the opportunity for quantifying disparities in many underreported health conditions, including polycystic ovarian syndrome, endometriosis, premenstrual dysphoric disorder, and intimate partner violence.

Acknowledgments

We thank MSR seminar participants, Irene Chen, Jacquelyn Campbell, Karthik Chetty, John Guttag, Pang Wei Koh, Nat Roth, and Harini Suresh for helpful conversations. EP was supported by a Google Research Scholar award.

References

Pietro Artoni, Arianna Piffer, Viviana Vinci, Jocelyn LeBlanc, Charles A Nelson, Takao K Hensch, and Michela Fagiolini. Deep learning of spontaneous arousal fluctuations detects early cholinergic defects across neurodevelopmental mouse models and patients. *Proceedings of the National Academy of Sciences*, 117(38):23298–23303, 2020.

Jessa Bekker and Jesse Davis. Learning from positive and unlabeled data: A survey. *Machine Learning*, 109(4):719–760, 2020.

Jessa Bekker, Pieter Robberechts, and Jesse Davis. Beyond the selected completely at random assumption for learning from positive and unlabeled data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 71–85. Springer, 2019.

Hui Chen, Fangqing Liu, Yin Wang, Liyue Zhao, and Hao Wu. A variational approach for learning from positive and unlabeled data. *arXiv preprint arXiv:1906.00642*, 2019.

Limeng Cui, Siddharth Biswal, Lucas M Glass, Greg Lever, Jimeng Sun, and Cao Xiao. Conan: Complementary pattern augmentation for rare disease detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 614–621, 2020.

Peter J Diggle. Estimating prevalence using an imperfect test. *Epidemiology Research International*, 2011, 2011.

Marthinus Christoffel Du Plessis and Masashi Sugiyama. Class prior estimation from positive and unlabeled data. *IEICE TRANSACTIONS on Information and Systems*, 97(5):1358–1362, 2014.

Karethy Kay Edwards and Beverly Patchell. State of the science: A cultural view of native americans and diabetes prevention. *Journal of cultural diversity*, 16(1):32, 2009.

Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 213–220, 2008.

H Jack Geiger. Racial and ethnic disparities in diagnosis and treatment: a review of the evidence and a consideration of causes. *Unequal treatment: Confronting racial and ethnic disparities in health care*, 417, 2003.

Denis Haine, Ian Dohoo, and Simon Dufour. Selection and misclassification biases in longitudinal studies. *Frontiers in veterinary science*, 5:99, 2018.

Gordon Hay and Filip Smit. Estimating the number of drug injectors from needle exchange data. *Addiction Research & Theory*, 11(4):235–243, 2003.

Matthew Hickman and Colin Taylor. Indirect methods to estimate prevalence. In *Epidemiology of drug abuse*, pages 113–131. Springer, 2005.

JA Hirsch, G Nicola, G McGinty, RW Liu, RM Barr, MD Chittle, and L Manchikanti. Icd-10: history and context. *American Journal of Neuroradiology*, 37(4):596–599, 2016.

Debra Houry, Karin V Rhodes, Robin S Kemball, Lorie Click, Catherine Cerulli, Louise Anne McNutt, and Nadine J Kaslow. Differences in female and male victims and perpetrators of partner violence with respect to web scores. *Journal of interpersonal violence*, 23(8):1041–1055, 2008.

IHS. Changing the course of diabetes: Turning hope into reality. 2017.

Fadi I Jabr and Venk Mani. An unusual cause of abdominal pain in a male patient: Endometriosis. *Avicenna Journal of Medicine*, 4(4), 2014.

Shantanu Jain, Martha White, and Predrag Radivojac. Estimating the class prior and posterior from noisy positives and unlabeled data. *Advances in neural information processing systems*, 29:2693–2701, 2016.

Kristen Jaskie and Andreas Spanias. Positive and unlabeled learning algorithms and applications: A survey. In *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pages 1–8. IEEE, 2019.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and R Mark IV. Mimic-iv (version 0.4). *PhysioNet*, 2020.

Ronald E LaPorte, Daniel McCarty, Graziella Bruno, Naoko Tajima, and Shigeaki Baba. Counting diabetes in the next millennium: application of capture-recapture technology. *Diabetes care*, 16(2): 528–534, 1993.

Matthew G Law, Michael Lynskey, Joanne Ross, and Wayne Hall. Back-projection estimates of the number of dependent heroin users in australia. *Addiction*, 96(3):433–443, 2001.

Małgorzata Łazęcka, Jan Mielniczuk, and Paweł Teisseyre. Estimating the class prior for positive and unlabelled data via logistic regression. *Advances in Data Analysis and Classification*, pages 1–30, 2021.

Fraser I Lewis and Paul R Torgerson. A tutorial in estimating the prevalence of disease in humans and animals in the absence of a gold standard diagnostic. *Emerging themes in epidemiology*, 9(1):1–8, 2012.

Neil McKeganey, Marina Barnard, Alastair Leyland, Isobel Coote, and Edward Follet. Female streetworking prostitution and hiv infection in glasgow. *British Medical Journal*, 305(6857):801–804, 1992.

Maryam Moradi, Melissa Parker, Anne Sneddon, Violeta Lopez, and David Ellwood. Impact of endometriosis on women's lives: a qualitative study. *BMC women's health*, 14(1):1–12, 2014.

Bret Nestor, Matthew BA McDermott, Willie Boag, Gabriela Berner, Tristan Naumann, Michael C Hughes, Anna Goldenberg, and Marzyeh Ghassemi. Feature robustness in non-stationary health records: caveats to deployable model performance in common clinical machine learning tasks. In *Machine Learning for Healthcare Conference*, pages 381–405. PMLR, 2019.

Curtis G Northcutt, Tailin Wu, and Isaac L Chuang. Learning with confident examples: Rank pruning for robust classification with noisy labels. *arXiv preprint arXiv:1705.01936*, 2017.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

Ana Penman-Aguilar, Makram Talih, David Huang, Ramal Moonesinghe, Karen Bouye, and Gloria Beckles. Measurement of health disparities, health inequities, and social determinants of health to support the advancement of health equity. *Journal of public health management and practice: JPHMP*, 22(Suppl 1):S33, 2016.

Joaquin Quiñonero-Candela, Masashi Sugiyama, Neil D Lawrence, and Anton Schwaighofer. *Dataset shift in machine learning*. Mit Press, 2009.

Harish Ramaswamy, Clayton Scott, and Ambuj Tewari. Mixture proportion estimation via kernel embeddings of distributions. In *International conference on machine learning*, pages 2052–2060. PMLR, 2016.

J Rigg, H Lodhi, and P Nasuti. Using machine learning to detect patients with undiagnosed rare diseases: an application of support vector machines to a rare oncology disease. *Value in Health*, 18(7): A705, 2015.

Sabine Salhofer-Polanyi, Hakan Cetin, Fritz Leutmezer, Anna Baumgartner, Stephan Blechinger, Assunta Dal-Bianco, Patrick Altmann, Barbara Bajer-Kornek, Paulus Rommer, Michael Guger, et al. Epidemiology of multiple sclerosis in austria. *Neuroepidemiology*, 49(1-2):40–44, 2017.

Julia Schaefer, Moritz Lehne, Josef Schepers, Fabian Prasser, and Sylvia Thun. The use of machine learning in rare diseases: a scoping review. *Orphanet Journal of Rare Diseases*, 15(1):1–10, 2020.

Sean D Schafer, Linda L Drach, Katrina Hedberg, and Melvin A Kohn. Using diagnostic codes to screen for intimate partner violence in oregon emergency departments and hospitals. *Public Health Reports*, 123(5):628–635, 2008.

Clayton Scott. A rate of convergence for mixture proportion estimation, with application to learning from noisy labels. In *Artificial Intelligence and Statistics*, pages 838–846. PMLR, 2015.

DW Scott and SR Sain. Multi-dimensional density estimation handbook of statistics vol 23 data mining and computational statistics ed cr rao and ej wegman, 2004.

C Chandra Sekar and W Edwards Deming. On a method of estimating birth and death rates and the extent of registration. *Journal of the American statistical Association*, 44(245):101–115, 1949.

Partho P Sengupta, Yen-Min Huang, Manish Bansal, Ali Ashrafi, Matt Fisher, Khader Shameer, Walt Gall, and Joel T Dudley. Cognitive machine-learning algorithm for cardiac imaging: a pilot study for differentiating constrictive pericarditis from restrictive cardiomyopathy. *Circulation: Cardiovascular Imaging*, 9(6):e004330, 2016.

Ronald S Simeone, William M Rhodes, and Dana Eser Hunt. A plan for estimating the number of "hardcore" drug users in the united states. *International journal of the addictions*, 30(6):637–657, 1995.

Harvineet Singh, Rina Singh, Vishwali Mhasawade, and Rumi Chunara. Fairness violations and mitigation under covariate shift. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 3–13, 2021.

Michael G Smith, Julie Royer, Joshua Mann, Suzanne McDermott, and Rodolfo Valdez. Capture-recapture methodology to study rare conditions using surveillance data for fragile x syndrome and muscular dystrophy. *Orphanet journal of rare diseases*, 12(1):1–8, 2017.

Eduardo Tejera, Maria Jose areias, Ana Rodrigues, Ana Ramoa, Jose Manuel nieto villar, and Irene Rebelo. Artificial neural network for normal, hypertensive, and preeclamptic pregnancy classification using maternal heart rate variability indexes. *The Journal of Maternal-Fetal & Neonatal Medicine*, 24(9):1147–1151, 2011.

Victor Wu, Harold Huff, and Mohit Bhandari. Pattern of physical injury associated with intimate partner violence in women presenting to the emergency department: a systematic review and meta-analysis. *Trauma, Violence, & Abuse*, 11(2):71–82, 2010.

Appendix A. Suspicious Symptoms

We list the suspicious symptoms used to produce each of the semi-synthetic datasets in Tables 1 (common symptoms), 3 (correlated symptoms), and 4 (recognized symptoms). Note that the codes include both ICD-9 and ICD-10 codes since the patient visits span 2008-2018, and ICD-10 codes gained wider usage in 2013 (Hirsch et al., 2016).

Appendix B. KM2 Performance in 1D

As discussed in §6.1, *KM2* should theoretically be able to identify the relative prevalence in separable data. However, kernel mean embedding approaches are sensitive to the dimensionality of a dataset; replicating the experiment in the 1D setting (Fig. 5), we see that *KM2* does recover the relative prevalence correctly on average.

Appendix C. Synthetic Results

The numerical equivalents to the graphs in Fig. 2 can be found in Tables 5, 6, and 7. Each entry includes the mean estimated relative prevalence by each method, accompanied by the standard deviation over 5 random train/test splits. `PURPLE` delivers the most accurate estimates of the relative prevalence, with frequently smaller variation than existing work.

ICD Code	ICD Code Description
2724	Other and unspecified hyperlipidemia
2859	Anemia, unspecified
30000	Anxiety state, unspecified
412	Old myocardial infarction
41401	Coronary atherosclerosis of native coronary artery
42731	Atrial fibrillation
496	Chronic airway obstruction, not elsewhere class...
5849	Acute kidney failure, unspecified
E039	Hypothyroidism, unspecified
E669	Obesity, unspecified
F419	Anxiety disorder, unspecified
I10	Essential (primary) hypertension
I4891	Unspecified atrial fibrillation
N179	Acute kidney failure, unspecified
V5861	Long-term (current) use of anticoagulants
V5866	Long-term (current) use of aspirin
Y929	Unspecified place or not applicable
Z7901	Long term (current) use of anticoagulants
Z794	Long term (current) use of insulin

Table 1: **Common Suspicious Symptoms.** Suspicious symptoms for the semi-synthetic dataset created using common symptoms. We select these symptoms randomly from the 50 most common ICD codes in MIMIC-IV.

ICD Code	ICD Code Description
N80	Endometriosis
N800	Endometriosis of uterus
N801	Endometriosis of ovary
N802	Endometriosis of fallopian tube
N803	Endometriosis of pelvic peritoneum
N804	Endometriosis of rectovaginal septum and vagina
N805	Endometriosis of intestine
N806	Endometriosis in cutaneous scar
N808	Other endometriosis
N809	Endometriosis, unspecified
6179	Endometriosis, site unspecified

Table 2: **Endometriosis Diagnoses.** Diagnoses used to identify endometriosis cases. We use all symptoms containing reference to endometriosis.

Appendix D. Semi-Synthetic Results

Similarly, we include the numerical equivalent to Fig. 3 in Tables 8, 9, and 10. The estimated relative prevalences illuminate how *KM2* frequently estimates a relative prevalence of 1.00 (i.e., in Table 9), which is close to the true value and appears correct. However, this is an artifact of estimating each group's prevalence to be .5. Conditions where relative prevalences are close to 1 (as is the case with endometriosis and IPV in the semisynthetic datasets) merit extra

ICD Code	ICD Code Description
33819	Other acute pain
5951	Chronic interstitial cystitis
6205	Torsion of ovary, ovarian pedicle, or fallopian...
6260	Absence of menstruation
78902	Abdominal pain, left upper quadrant
78904	Abdominal pain, left lower quadrant
78905	Abdominal pain, periumbilic
7891	Hepatomegaly
C561	Malignant neoplasm of right ovary
D251	Intramural leiomyoma of uterus
D252	Subserosal leiomyoma of uterus
D259	Leiomyoma of uterus, unspecified
D270	Benign neoplasm of right ovary
E43	Unspecified severe protein-calorie malnutrition
F911	Conduct disorder, childhood-onset type
G8921	Chronic pain due to trauma
K661	Hemoperitoneum
N739	Female pelvic inflammatory disease, unspecified
N952	Postmenopausal atrophic vaginitis
O210	Mild hyperemesis gravidarum
O2341	Unspecified infection of urinary tract in pregn...
O26891	Other specified pregnancy related conditions, f...
Q600	Renal agenesis, unilateral
R1310	Dysphagia, unspecified
R17	Unspecified jaundice

Table 3: **Correlated Suspicious Symptoms.** Suspicious symptoms for the semi-synthetic dataset created using symptoms correlated with endometriosis. We select the 25 ICD codes with the highest relative proportion among endometriosis patients.

ICD Code	ICD Code Description
7842	Swelling, mass, or lump in head and neck
9100	Abrasion or friction burn of face, neck, and sc...
920	Contusion of face, scalp, and neck except eye(s)
95901	Head injury, unspecified
S0003XA	Contusion of scalp, initial encounter
S0011XA	Contusion of right eyelid and periocular area, ...
S0012XA	Contusion of left eyelid and periocular area, i...
S0990XA	Unspecified injury of head, initial encounter

Table 4: **Recognized Suspicious Symptoms.** Diagnoses associated with intimate partner violence. We first identify all ICD codes that refer to head, neck, and face injuries (100 ICD codes). We filter out codes that appear fewer than 10 times as a part of dataset preprocessing (§5.1.2), which leaves the 8 codes listed here.

caution when applying and evaluating prevalence estimation methods.

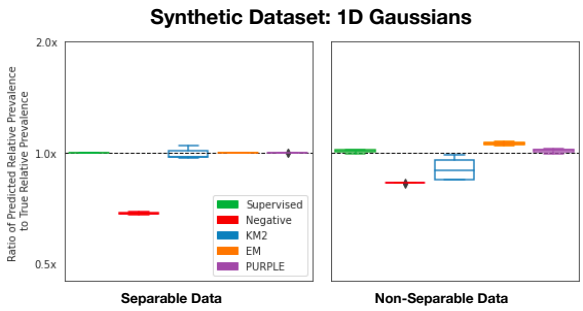

Figure 5: PURPLE continues to outperform baselines in the 1D setting, where we use the generative model for *Gauss-Synth* and change only the dimensionality of the group distributions to be 1 instead of 5. The key difference is the performance of *KM2*; in 1D, the method is able to recover the relative prevalence on average in the separable case.

Separability	Negative	KM2	EM	PURPLE	Supervised
overlap	1.40 ± 0.01	1.25 ± 0.07	2.47 ± 0.05	1.90 ± 0.03	1.92 ± 0.02
separable	1.86 ± 0.01	2.22 ± 0.23	3.12 ± 0.04	3.11 ± 0.03	3.08 ± 0.00

Table 5: **Class Separability.** Each row describes method performance in one of two separability conditions: “overlap" and “separable", corresponding to Fig. 1a and 1b respectively.

c_{g_b}	Negative	KM2	EM	PURPLE	Supervised
0.1	1.37 ± 0.01	1.36 ± 0.23	2.30 ± 0.08	1.92 ± 0.04	1.89 ± 0.03
0.3	1.40 ± 0.02	1.37 ± 0.16	2.46 ± 0.07	1.93 ± 0.03	1.93 ± 0.03
0.5	1.43 ± 0.00	1.43 ± 0.12	2.50 ± 0.03	1.93 ± 0.02	1.92 ± 0.01
0.7	1.47 ± 0.01	1.14 ± 0.05	2.50 ± 0.03	1.93 ± 0.03	1.94 ± 0.02
0.9	1.51 ± 0.01	1.25 ± 0.14	2.50 ± 0.04	1.92 ± 0.02	1.94 ± 0.02

Table 6: **Labeling Frequency.** Method performance across different labeling frequencies for group b. These results are the numerical equivalent to Fig. 2b.

Group Gap	Negative	KM2	EM	PURPLE	Supervised
-2.00	1.40 ± 0.00	1.25 ± 0.08	2.40 ± 0.05	1.92 ± 0.03	1.89 ± 0.03
-1.00	1.03 ± 0.01	1.17 ± 0.21	1.42 ± 0.02	1.32 ± 0.02	1.32 ± 0.01
-0.50	1.04 ± 0.00	1.19 ± 0.03	1.18 ± 0.02	1.12 ± 0.01	1.15 ± 0.01
-0.25	1.08 ± 0.00	1.06 ± 0.10	1.07 ± 0.02	1.05 ± 0.01	1.06 ± 0.01
0.00	1.12 ± 0.00	1.09 ± 0.11	1.00 ± 0.01	1.00 ± 0.00	0.99 ± 0.01

Table 7: **Covariate Shift.** Method performance across different shifts in $p(x|g = a)$ and $p(x|g = b)$. These results are the numerical equivalent to Fig. 2c.

c_{g_b}	Negative	KM2	EM	PURPLE	Supervised
0.1	4.02 ± 0.14	5.09 ± 0.77	1.67 ± 0.06	0.84 ± 0.01	0.85 ± 0.01
0.3	1.33 ± 0.05	1.62 ± 0.00	0.91 ± 0.02	0.85 ± 0.03	0.85 ± 0.02
0.5	0.81 ± 0.03	0.79 ± 0.12	0.72 ± 0.01	0.86 ± 0.02	0.86 ± 0.02
0.7	0.60 ± 0.02	0.51 ± 0.04	0.68 ± 0.02	0.85 ± 0.01	0.86 ± 0.03
0.9	0.44 ± 0.02	0.40 ± 0.03	0.63 ± 0.01	0.84 ± 0.01	0.85 ± 0.02

Table 8: **Common Symptoms Semi-Synthetic Dataset.** `PURPLE` outperforms existing baselines in estimating the relative prevalence of a condition simulated over a set of 25 common symptoms.

c_{g_b}	Negative	KM2	EM	PURPLE	Supervised
0.1	5.13 ± 0.54	1.00 ± 0.00	2.33 ± 0.19	NaN	1.05 ± 0.03
0.3	1.65 ± 0.12	1.00 ± 0.00	1.25 ± 0.04	1.89 ± 0.27	1.07 ± 0.03
0.5	1.03 ± 0.05	1.00 ± 0.00	0.81 ± 0.02	NaN	1.04 ± 0.04
0.7	0.73 ± 0.03	1.00 ± 0.00	0.69 ± 0.02	NaN	1.06 ± 0.04
0.9	0.56 ± 0.03	0.34 ± 0.00	0.58 ± 0.04	NaN	1.05 ± 0.03

Table 9: **Correlated Symptoms Semi-Synthetic Dataset.** `PURPLE` generalizes to conditions with lower prevalences, including endometriosis, as we have simulated in these experiments.

c_{g_b}	Negative	KM2	EM	PURPLE	Supervised
0.1	4.73 ± 0.30	1.00 ± 0.00	3.25 ± 0.43	0.97 ± 0.01	0.94 ± 0.12
0.3	1.58 ± 0.06	1.00 ± 0.00	1.37 ± 0.48	0.95 ± 0.01	0.94 ± 0.07
0.5	0.95 ± 0.06	1.00 ± 0.00	0.78 ± 0.15	0.95 ± 0.02	0.94 ± 0.08
0.7	0.69 ± 0.03	0.47 ± 0.29	0.64 ± 0.14	0.94 ± 0.01	0.91 ± 0.10
0.9	0.51 ± 0.01	0.34 ± 0.00	0.50 ± 0.08	0.95 ± 0.01	0.92 ± 0.10

Table 10: **Recognized Symptoms Semi-Synthetic Dataset.** Intimate partner violence exhibits a different distribution of symptoms compared to endometriosis, where there are fewer suspicious symptoms and fewer concrete diagnoses. `PURPLE` can deliver accurate estimates of relative prevalence in this case, too.