https://doi.org/10.1038/s44294-024-00011-5

Quantifying disparities in intimate partner violence: a machine learning method to correct for underreporting

Check for updates

Divya Shanmugam[1] ✉, Kaihua Hou[2] & Emma Pierson[3,4]

The first step towards reducing the pervasive disparities in women's health is to quantify them. Accurate estimates of the *relative prevalence* across groups—capturing, for example, that a condition affects Black women more frequently than white women—facilitate effective and equitable health policy that prioritizes groups who are disproportionately affected by a condition. However, it is difficult to estimate relative prevalence when a health condition is underreported, as many women's health conditions are. In this work, we present `PURPLE`, a method for accurately estimating the relative prevalence of underreported health conditions which builds upon the literature in positive unlabeled learning. We show that under a commonly made assumption—that the probability of having a health condition given a set of 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 conduct experiments on synthetic and real health data which demonstrate `PURPLE`'s ability to recover the relative prevalence more accurately than do previous methods. We then use `PURPLE` to quantify the relative prevalence of intimate partner violence (IPV) in two large emergency department datasets. We find higher prevalences of IPV among patients who are on Medicaid, not legally married, and non-white, and among patients who live in lower-income zip codes or in metropolitan counties. We show that correcting for underreporting is important to accurately quantify these disparities and that failing to do so yields less plausible estimates. Our method is broadly applicable to underreported conditions in women's health, as well as to gender biases beyond healthcare.

There are enormous disparities in women's health across race, age, socioeconomic status, and other dimensions. Mitigating these disparities requires accurate estimates of the extent to which a medical condition disproportionately affects different groups. The *relative prevalence* does so by capturing 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 to increase funding, research, and resources. Public health decisions often rely on such estimates to develop, allocate, and advocate for interventions. For example, research revealing startling disparities in maternal mortality between Black and white women[1] led to Congressional policy that has invested billions in funding towards evidence-based interventions to improve Black maternal health[2].

However, it remains challenging to produce accurate relative prevalence estimates for many conditions in women's health and healthcare more generally due to widespread *underreporting* (Fig. 1A). With underreported health conditions, only a small percentage of true positives may be labeled as positive; worse, the probability of correctly diagnosing a positive case can vary by group[3]. This is especially relevant to intimate partner violence, a notoriously underreported condition: true cases are only correctly diagnosed an estimated ~25% of the time, and this probability varies across racial groups[4]. Underreporting in one group and not another can skew estimates of health disparities, making it appear that a condition is

[1]Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA 02139, USA. [2]Malone Center of Engineering in Healthcare, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [3]Department of Computer Science, Cornell Tech, New York, NY 10044, USA. [4]Department of Population Health Sciences, Weill Cornell Medical College, New York, NY 10021, USA. ✉e-mail: divyas@mit.edu

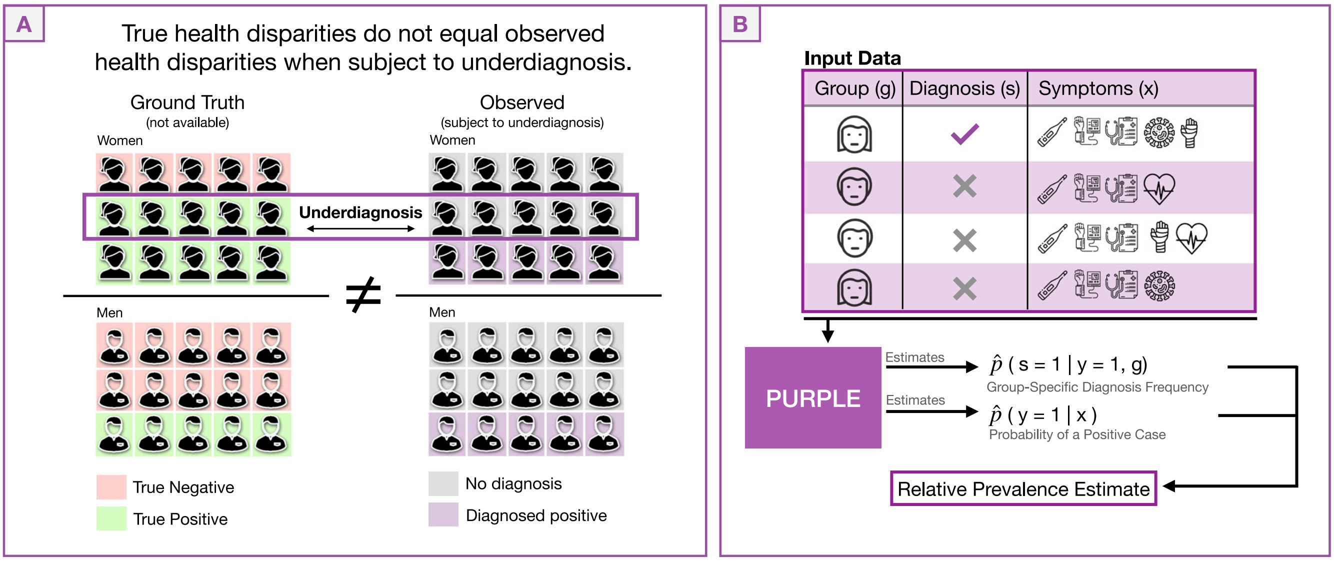

Fig. 1 | Underreporting can skew observed relative prevalences and conceal health disparities. PURPLE is designed to estimate the relative prevalence while correcting for underreporting. **A** Underreporting leads to inaccurate observed relative prevalences. Understanding the relative prevalence of a health condition between groups *g*—for example, men and women—is important to effective medical care. However, these estimates are often based on diagnoses *s* (i.e., positive diagnosis vs. no diagnosis) instead of the true patient state *y* (sick vs. not sick). Underreporting, which is known to vary by demographic groups, leads to inaccurate relative prevalence estimates that can hide the groups most affected by a condition. **B** PURPLE uses data on patient diagnoses *s*, symptoms *x*, and group membership *g* to accurately estimate the relative prevalence of a condition. PURPLE first estimates the group-specific diagnosis probability, $p(s=1|y=1,g)$, and disease likelihood, $p(y=1|x)$, up to constant multiplicative factors and then combines these estimates to compute the relative prevalence. We show this is possible under three widely-made assumptions: no false positives, random diagnosis within groups, and constant $p(y=1|x)$ between groups.

equally prevalent in two populations when it is not or more prevalent in one population than in another when it is not. These errors obscure where resources are most needed and consequently inhibit the development of effective health policy.

Efforts in both epidemiology and machine learning have addressed these challenges but often rely on either data that is unavailable or assumptions that are unrealistic in women's health contexts (refer to the Supplement for a detailed discussion of related work). Epidemiological work aims to quantify true prevalences in the context of imperfect diagnostic tests and commonly assumes the presence of information that is not always available (for example, ground truth annotations[5–9], multiple tests[10–12], or informative priors[13]). For conditions in women's health, we often have no access to ground truth, only a single diagnosis per patient, and little notion of how accurate that diagnosis truly is[14]. The machine learning literature has modeled underreporting using the positive-unlabeled (PU) learning framework, which assumes that only some positive cases are correctly labeled as positive, and the unlabeled examples consist of both true negatives and unlabeled true positives. In order to recover prevalence in the presence of underdiagnosis, many PU learning methods assume that there is a region of the feature space where cases are certain to be true positives. However, this is a restrictive assumption which, while potentially suitable in other PU settings, is unlikely to hold in health data[15] because it is rare that a set of symptoms corresponds to a health condition with 100% certainty. This is especially true in the context of intimate partner violence, where symptoms are frequently not specific to a particular condition (for example, pregnancy complications, which are well-known to occur at higher rates among IPV patients[16], could suggest a number of underlying conditions).

Here, we present PURPLE (**P**ositive **U**nlabeled **R**elative **P**reva**L**ence **E**stimator), a method that is complementary to the prior epidemiology and PU learning literature. In contrast to epidemiological approaches, it requires no external information (for example, the sensitivity and specificity of a test) to recover the relative prevalence of an underreported disease. In contrast to the prior PU learning literature, PURPLE relies on no assumptions about the overlap in symptom distribution of positive and negative cases. PURPLE is designed to address underreporting in intimate partner violence and women's health more broadly by estimating the relative prevalence of the condition given three assumptions: (1) no false-positive diagnoses; (2) random diagnosis within the group; and (3) constant $p(y=1|x)$ 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 PU assumptions about the separability of the positive and negative classes. 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. PURPLE does this by jointly estimating the conditional probability that a case is a true positive given a set of symptoms and the diagnosis probability (i.e., the probability a positive case is diagnosed as such; Fig. 1B). We demonstrate via experiments on synthetic and real health data that PURPLE recovers the relative prevalence more accurately than do existing methods. We provide procedures for checking whether PURPLE's underlying assumptions hold and show that even under a plausible violation of the assumptions, PURPLE still provides a useful lower bound on the magnitude of disparities.

Having validated PURPLE, we use it to estimate the relative prevalence of the condition that motivates this work—intimate partner violence (IPV)—in two widely-used datasets of electronic health records, which together describe millions of emergency department visits: MIMIC-IV[17] and NEDS[18]. Across both datasets, we find higher prevalences of IPV among patients who are on Medicaid than among those who are not. Relative prevalences are higher among non-white patients (though these disparities are noisily estimated in the MIMIC dataset). We also quantify the relative prevalence of IPV across income quartiles, marital statuses, and the rural–urban spectrum, finding that IPV is more prevalent among patients who are lower-income, not legally married, and in metropolitan counties. Finally, we show that PURPLE's corrections for underreporting are important: they yield more plausible estimates of how relative prevalence varies with income than estimation methods that do not correct for underreporting. Specifically, PURPLE estimates that the relative prevalence of IPV decreases with income, consistent with prior work[19–21]. In contrast, failing to correct for underdiagnosis (i.e., computing relative prevalence estimates using observed diagnoses) yields estimates that do not show any consistent trend with respect to income and which are harder to explain. Overall, this analysis contributes to the literature on IPV disparities in several ways: it uses some of the largest and most recent samples; evaluates robustness across multiple datasets; and corrects for underreporting.

Together, our analyses illustrate how PURPLE is a general method for estimating relative prevalences in the presence of underreporting, allowing practitioners to discover and quantify group-specific disparities in a wide range of settings in which underreporting is common, including outcomes in women's health and beyond.

Results

Here, we introduce PURPLE, a method to quantify disparities in the prevalence of a health condition between groups given only positive and unlabeled data. A key idea underpinning our method is that knowing the exact prevalence in a group is not necessary to calculate the *relative* prevalence across groups: one can estimate the fraction $\frac{\text{prevalence in group A}}{\text{prevalence in group B}}$ without knowing its numerator or denominator. We adopt terminology standard in the PU learning literature and assume that we have access to three pieces of data for the *i*th example: a feature vector x_i; a group variable g_i; and a binary observed label s_i. We let y_i denote the true (unobserved) label. In healthcare, example i may correspond to a specific patient and their presenting symptoms (x_i), race (g_i), and observed diagnosis (s_i). Here, y_i corresponds to whether the patient truly *has* the medical condition. We first introduce PURPLE and then validate our approach on synthetic and semi-synthetic data.

Overview of PURPLE: Positive-Unlabeled Relative PrevaLence Estimation

We provide a conceptual overview of PURPLE here and describe the full details in the "Deriving the Relative Prevalence–Implementation" Section. PURPLE is designed to address how, for underreported women's health conditions, the ratio of *diagnosis rates* between demographic groups may not equal the ratio of *true prevalences* due to differential underreporting across groups. To address this, PURPLE first uses the observed data to learn a model of which symptoms correlate with having the condition; so long as this relationship between symptoms and the condition remains constant across groups, we will be able to estimate the relative prevalence. Mathematically, PURPLE first estimates $p(y = 1|x)$ up to a constant multiplicative factor; second, it uses this estimate to compute the relative prevalence between groups. We use p to denote the true probabilities in the underlying data distribution and $\hat{p}$ to denote PURPLE's estimates of these probabilities.

1. Estimate $p(y = 1|x)$ up to a constant factor. PURPLE fits the following model:

$$\underbrace{\hat{p}(s=1|g,x)}_{\text{probability patient is diagnosed}} = \underbrace{\hat{p}(y=1|x)}_{\text{probability patient truly has condition}} \cdot \underbrace{\hat{p}(s=1|y=1,g)}_{\text{probability true positives are correctly diagnosed}} \tag{1}$$

In other words, PURPLE models the probability a patient is diagnosed with a condition as the product of two terms: the probability the patient truly has the condition and the probability that true positives are correctly diagnosed. The first term is constant across groups g, while the second can vary, accounting for underdiagnosis. This decomposition is valid under three assumptions which we discuss below. To estimate the two terms on the right-hand side of Eq. (1), we parameterize the first term as a logistic regression and the second as a constant $c_g \in [0, 1]$ for each group g. We optimize these parameters by minimizing the cross-entropy loss between the predicted $\hat{p}(s = 1|g, x)$ and the empirical $p(s = 1|g, x)$, which is possible because s, g, and x are all observed ("Implementation" Section). Note that we can only estimate both terms on the right-hand side up to a constant multiplicative factor because multiplying the first term by a non-negative β and dividing the second term by β leaves $\hat{p}(s = 1|g, x)$ unchanged.

2. Estimate the relative prevalence using $\hat{p}(y = 1|x)$. Fortunately, even though $\hat{p}(y = 1|x)$ is only correct up to a constant multiplicative factor, this suffices to estimate the relative prevalence $\frac{p(y=1|g=a)}{p(y=1|g=b)}$, as we derive in the "Deriving the relative prevalence" Section. Specifically, our estimator of the relative prevalence is

$$\frac{\sum_x \hat{p}(y=1|x)\hat{p}(x|g=a)}{\sum_x \hat{p}(y=1|x)\hat{p}(x|g=b)} \tag{2}$$

In practice, this is simply the mean value of $\hat{p}(y = 1|x)$ for samples from group a divided by the mean value of $\hat{p}(y = 1|x)$ for samples from group b.

It is impossible to estimate outcome prevalence in PU settings without assumptions[22]. Our estimation procedure relies on three assumptions: (1) observed positives are true positives (the *positive-unlabeled* assumption common to all PU methods), (2) within each group, diagnosis s depends only on y (the *random diagnosis within group* assumption, commonly made in PU settings) and (3) the probability of having a disease conditional on symptoms remains constant across groups (the *constant* $p(y|x)$ assumption, common to work in both domain adaptation[23,24] and healthcare[25]). Details about the required assumptions can be found in the "Assumptions" Section. We also provide checks to assess whether the assumptions hold ("Assumption checks" Section) and show that even under a plausible violation of these assumptions, PURPLE is guaranteed to produce a lower bound on the true magnitude of disparities ("Robustness to violations of the Constant $p(y|x)$ assumption" Section). An illustration of PURPLE's behavior under violations of the PU assumption and the random-diagnosis-within-group assumption is available in the "Effect of positive-unlabeled assumption violations" and "Effect of random-diagnosis-within-groups assumption violations" Sections, respectively. We provide the full derivation of our estimation procedure in the "PURPLE: positive unlabeled prevalence estimator" Section.

PURPLE recovers the true relative prevalence in synthetic data

Prior to applying PURPLE to estimate the relative prevalence of IPV, we confirm that the method can correctly recover the true prevalence on synthetic data where the true relative prevalence is known, a standard machine learning check. We compare PURPLE to four previous machine learning methods ("Baselines" Section) drawn from the literature on PU learning, where estimating prevalence is a critical step[26]. We generate the synthetic data by simulating group-specific features ($p(x|g)$) and labels using a decision rule ($p(y|x)$). The two groups, a and b, correspond to 5D Gaussian distributions with different means (see "Gauss-Synth" Section for full data generation details).

Figure 2B compares PURPLE's performance to the performance of the other methods on purely synthetic data. We evaluate each approach in both separable (in which the datapoints with $y = 1$ and the datapoints with $y = 0$ can be perfectly separated in the feature space x) and non-separable settings. We perform this comparison because existing methods rely on separability assumptions which often do not hold in realistic health settings[15]. PURPLE is the only method that accurately recovers the relative prevalence in both the separable and non-separable settings. We also show that PURPLE maintains consistent performance regardless of the extent to which $p(x)$, or the distribution of symptoms, differs between groups (Fig. S1B).

PURPLE recovers the true relative prevalence in realistic semi-synthetic health data

Having established that PURPLE outperforms previous work on synthetic data, we investigate its performance on more realistic data: specifically, MIMIC-IV[27], a dataset of electronic health records that describes ~450,000 patient hospital visits between 2008 and 2018. We generate realistic semi-synthetic data based on these records to examine PURPLE's performance on the high-dimensional, sparse data common in clinical settings. Specifically, we use the patient symptoms x—encoded as a binary one-hot vector of ICD

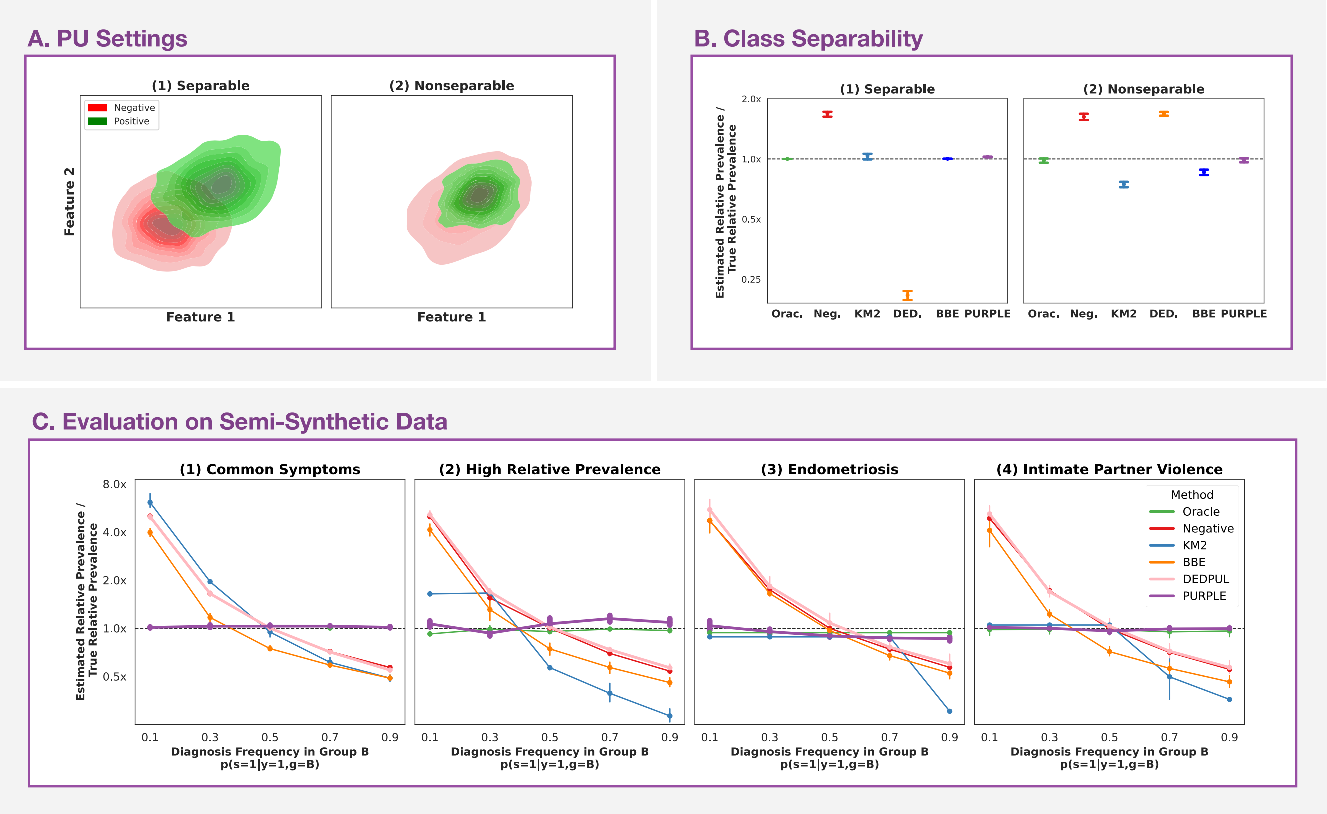

Fig. 2 | Validation of PURPLE on synthetic and semi-synthetic data. A Methods in positive-unlabeled learning commonly make assumptions about the separability of the positive and negative distributions. Settings in which underreporting occurs map directly to work in positive-unlabeled learning, in which learning algorithms have access to a set of positive labeled examples and an unlabeled mixture of positive and negative examples. Most works in positive-unlabeled learning assume A (left panel), or a positive subdomain. PURPLE makes no assumptions about the separability of the positive and negative distributions and instead assumes that $p(y = 1|x)$ remains constant across patient subgroups. **B** PURPLE accurately recovers the relative prevalence on both separable and nonseparable synthetic data. The vertical axis plots the ratio of estimated relative prevalence to true relative prevalence, with 1 (dotted line) indicating perfect performance. We report variation across 5 randomized train, validation, and test splits. *Negative*, *KM2*, *BBE*, and *DEDPUL* baselines do not always accurately estimate the relative prevalence, especially on nonseparable data. *Oracle* is impossible to implement in practice because it relies on ground truth labels y which are not available; it is provided as a metric for ideal performance. **C** PURPLE recovers the relative prevalence accurately in simulations based on real health data. We generate semi-synthetic data by using patient visits from MIMIC-IV[27] and simulating a disease label given a set of symptoms. This allows us to test PURPLE on a real, high-dimensional distribution of symptoms while retaining access to ground truth labels. Each dataset simulates disease likelihood on the basis of a different symptom set: (1) symptoms that appear most frequently, (2) symptoms that occur frequently in one group but not the other, (3) symptoms that co-occur frequently with endometriosis, and (4) symptoms known to indicate risk of intimate partner violence based on past literature. We define group A to be Black patients and group B to be white patients. Across symptom sets and a range of group-specific diagnosis frequencies, PURPLE produces more consistently accurate relative prevalence estimates than existing work. Two semisynthetic experiments involving real conditions in women's health—endometriosis and intimate partner violence—demonstrate the potential to apply PURPLE to conditions in women's health and produce accurate, actionable relative prevalence estimates.

codes—to simulate whether the patient truly has the medical condition, y. Using data in which we know y allows us to assess how accurately PURPLE recovers the relative prevalence; in contrast, if we did not simulate y, we would not have access to ground truth and could not assess relative prevalence estimates. We simulate y for four settings: (1) a condition with common symptoms, (2) a condition that is less common among Black patients, (3) endometriosis, and (4) intimate partner violence (see Section "MIMIC-semi-synth" for full details).

Across the semi-synthetic settings we consider, the estimation error of previous methods is large, with some methods producing relative prevalence estimates more than 4x the true value (Fig. 2C). Further, each previous method produces both overestimates and underestimates of the true relative prevalence depending on how underreported the medical condition is. In contrast, PURPLE remains accurate across the different settings.

Quantifying the relative prevalence of intimate partner violence

We have validated PURPLE's accuracy in recovering the relative prevalence by using synthetic and semi-synthetic datasets where the true relative prevalence is known. We now use PURPLE to estimate relative prevalence on two real datasets where the true relative prevalence is *unknown*. Specifically, we apply PURPLE to quantify the relative prevalence of the underdiagnosed condition motivating this work—intimate partner violence (IPV)—across different demographic groups.

Datasets. We conduct our study using two widely-used datasets of emergency department visits: MIMIC-IV ED[17] and the 2019 Nationwide Emergency Department Sample (NEDS)[18]. MIMIC-IV ED describes 293,297 emergency department visits to a single, Boston-area hospital; NEDS is a nationwide sample that is approximately one hundred times as large (it contains 33.1 million emergency department visits, which, when reweighted, represent the universe of 143 million US emergency department visits in 2019). We assess results across multiple datasets to verify the robustness of the disparities we observe. Because our sample consists of emergency department visits, we estimate the relative prevalence of IPV *conditional on going to the emergency department*—in particular, our data does not allow us to quantify disparities among populations who do not interact with the healthcare system at all[28]. Relative prevalence estimates among patients who visit emergency

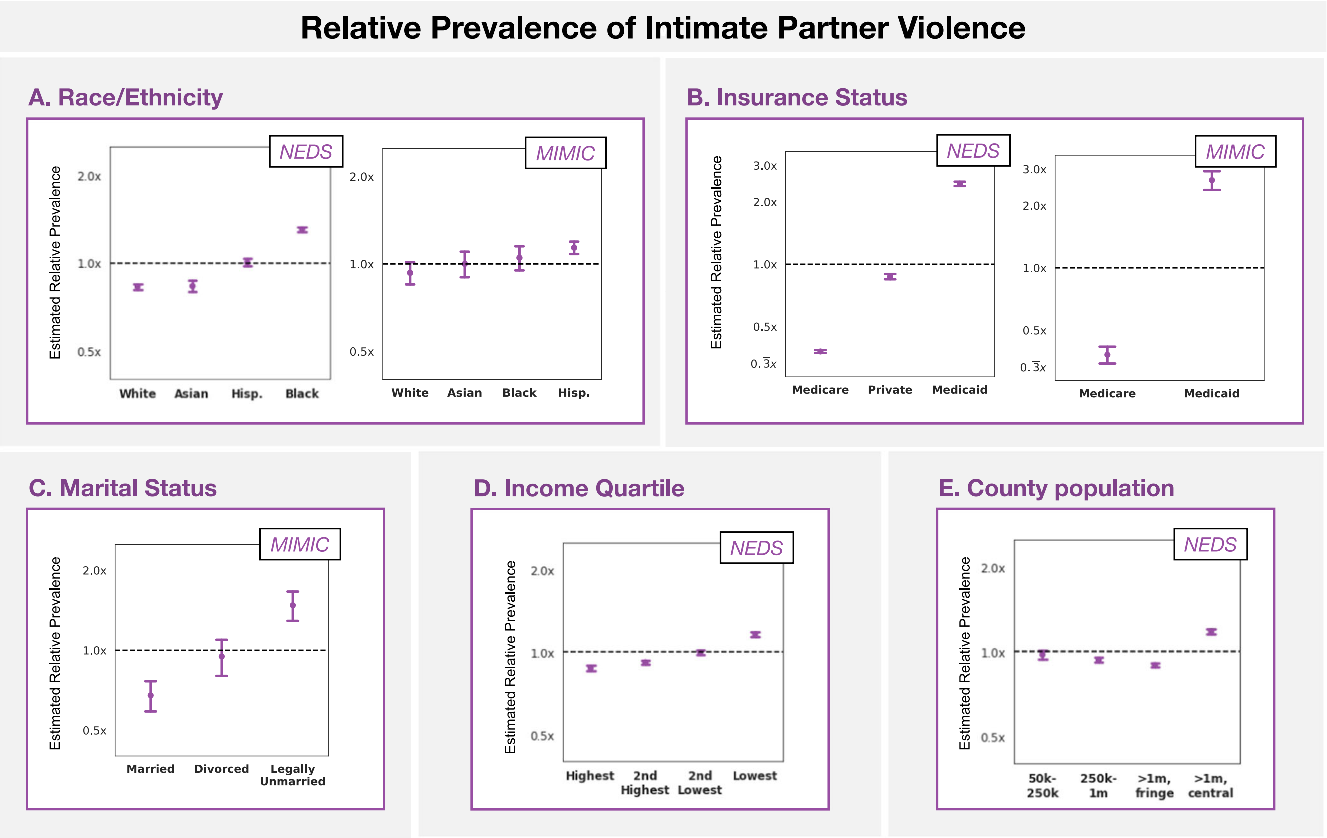

Fig. 3 | Estimates of the relative prevalence of intimate partner violence across demographic subgroups. We apply PURPLE to two large emergency department datasets: NEDS[18] and MIMIC-IV ED[17]. We compute each relative prevalence with respect to each group's complement (i.e., estimating the prevalence of IPV among married patients vs. non-married patients). Relative prevalences are higher among non-white patients overall (**A**) though these disparities are noisily estimated in the MIMIC-IV ED dataset, and the ordering across race groups is not completely consistent. We also find higher prevalences among patients who are on Medicaid (**B**), not legally married (**C**), in lower-income zip codes (**D**), and in metropolitan counties (**E**). Counties with more than 1 million residents are further differentiated by how urban they are (as "central" or "fringe"). Error bars report the standard deviations for each estimate across 5 randomized train/test splits of each dataset. Panels **C**–**E** are reported on only a single dataset due to different demographic feature availability in the two datasets; note the slightly larger *y*-axis range on Panel **B** due to the greater range of the estimates.

departments, however, remain of interest to IPV researchers due to the unique role emergency departments play as a point of care *and* intervention for patients who experience IPV[29].

For both datasets, we filter for female patients because the symptoms associated with IPV in male patients are less well understood, and the constant $p(y=1|x)$ assumption may not hold[30]; we also filter out patients younger than 18 years old because symptoms that indicate intimate partner violence could be instances of child abuse in this patient subgroup[31,32]. We describe all preprocessing steps in the "Datasets" Section. All point estimates and uncertainties reported below represent the mean and standard deviation, respectively, across five randomized train/test splits of each dataset.

Analysis. Results are plotted in Fig. 3 (we verify that PURPLE passes the assumption checks detailed in the "Assumption checks" Section in SI Figs. 4–6). We find, in both datasets, that intimate partner violence is more common among patients on Medicaid (NEDS relative prevalence 2.44 ± 0.07 in Medicaid patients vs non-Medicaid patients; MIMIC-IV relative prevalence 2.65 ± 0.31) and less common among patients on Medicare (NEDS relative prevalence of 0.37 ± 0.01; MIMIC-IV relative prevalence of 0.38 ± 0.04). Of course, Medicaid is likely not the *causal* factor underlying IPV risk; rather, it acts as a proxy that identifies populations who are disproportionately affected by IPV.

Examining racial differences reveals disparities that are smaller and less consistent than disparities by insurance status. In both datasets, white patients have the lowest relative prevalence of the four race groups, and in the NEDS dataset, white patients have significantly lower prevalence than non-white patients overall (relative prevalence for white patients vs. non-white patients equal to 0.82 ± 0.02). However, in MIMIC-IV, racial disparities are more noisily estimated due to the smaller size of the dataset, yielding an ordering of race groups that is similar but not completely consistent across the two datasets. This attests to the importance of using large samples and assessing results across multiple datasets.

The MIMIC-IV dataset provides information on patient marital status, allowing us to estimate that IPV is more common among patients who are "Legally Unmarried", or not officially married but may still be in relationships (relative prevalence equal to 1.48 ± 0.21). The NEDS dataset provides information on the population density and estimated median household income of areas where patients live. We estimate higher rates of IPV among patients living in central metropolitan counties with a population > 1 million (relative prevalence equal to 1.18 ± 0.02). We also find that IPV prevalence decreases with income (relative prevalence equal to 1.16 ± 0.02 in the bottom income quartile versus 0.87 ± 0.03 in the top income quartile).

In Fig. S3, we report the prevalence of observed IPV diagnoses—i.e., $p(s=1|g)$—without correcting for underdiagnosis. While often the trends are qualitatively similar, in some cases correcting for underdiagnosis is important to yield plausible trends. For example, failing to correct for underdiagnosis produces an inconsistent relationship between IPV prevalence and income which is difficult to reconcile with past work, which consistently documents that IPV prevalence decreases with income[19–21]. This suggests the importance of using methods like PURPLE, which attempts to correct for underdiagnosis. Our income results also suggest that IPV is less likely to be correctly diagnosed in lower-income women, a finding that reflects the broader phenomenon of underdiagnosis among lower-income patients, as has been shown in the context of dementia[33,34], asthma[35,36], and depression[37–39].

Discussion

In this work, we provide a method for estimating relative prevalence even in the presence of underreporting, a difficult but essential task in healthcare and public health. We show that we can estimate the relative prevalence even in settings where absolute prevalence estimation is impossible, by exchanging the restrictive separability assumptions typical in the PU learning literature for the constant $p(y=1|x)$ assumption, which is arguably more appropriate in clinical settings. Although this assumption may not hold for all settings—for example, the conditional probability of intimate partner violence is known to be dependent on a patient's age group[40]—it is realistic in many settings, and we provide methods for checking its validity and a lower-bound guarantee even when it fails to hold. Based on these assumptions, we present a method for relative prevalence estimation, PURPLE, a complementary approach to those in the epidemiology and PU learning literature: it works when one does not have the external information that epidemiological methods generally require and cannot make the separability assumptions PU learning methods rely on. We show PURPLE outperforms previous methods in terms of its ability to recover the relative prevalence on both synthetic and real health data.

We apply PURPLE to estimate the relative prevalence of intimate partner violence in two widely-used, large-scale datasets of emergency department visits. We find that IPV is more prevalent among patients who are on Medicaid, non-white, not legally married, in lower-income zip codes, and in metropolitan counties. We also show that correcting for underdiagnosis produces estimates of IPV prevalence across income groups which are more plausible in light of prior work[19–21], highlighting the importance of modeling underdiagnosis. In general, past work on IPV disparities corroborates the plausibility of our findings. Our finding that intimate partner violence is more common among patients on Medicaid compared to patients who are not is consistent with earlier results that show that IPV is [40–42] more common among patients who live below the poverty line[20,43,44]. Past work documenting higher IPV prevalences among unmarried women[45–47] and in metropolitan areas[19,48,49] also corroborates the plausibility of our findings. Our finding that IPV is more common among non-white patients is corroborated by some past work[4,19,50]. However, the fact that we find that racial disparities are smaller and not completely consistent across datasets is also concordant with past work documenting inconsistent racial differences across samples[44,51–53]. This suggests the importance of using large samples and multiple datasets to assess how consistently and robustly racial disparities emerge. Overall, our analysis contributes to the literature on IPV disparities by using large samples; evaluating robustness across multiple datasets; and correcting for underreporting.

Our work is motivated by the widespread underreporting of women's health, and we foresee numerous opportunities for future work. PURPLE could be applied to obtain relative prevalence estimates for many other health conditions that are known to be underreported, including polycystic ovarian syndrome[54], endometriosis[55], and traumatic brain injuries[56]. Additionally, quantifying relative prevalence in the presence of underreporting is a problem of interest in many domains beyond healthcare and public health: for example, quantifying the relative prevalence of underreported police misconduct across precincts or quantifying the relative prevalence of underreported hate speech across demographic groups. We believe that PURPLE can also yield useful insight into disparities in these non-healthcare settings.

Methods

PURPLE: Positive Unlabeled Relative PrevaLence Estimator

PURPLE estimates $p(y=1|x)$ up to a constant multiplicative factor in order to estimate the relative prevalence of a condition y. Underlying this procedure are two insights: first, that estimating $p(y=1|x)$ up to a constant factor suffices to estimate the relative prevalence, and second, that it is possible to produce this estimate using the observed labels s, symptoms x, and group statuses g. We first describe the three assumptions underlying PURPLE and show how these statements follow from them in the "Deriving the relative prevalence" Section and M1.3. We describe implementation details in the "Implementation" Section. We provide checks to determine whether PURPLE's assumptions hold true ("Assumption checks" Section) and show that even under a plausible violation of our assumptions, PURPLE produces a lower bound on the true magnitude of disparities ("Robustness to violations of the Constant $p(y|x)$ assumption" Section).

Assumptions. Neither the exact prevalence nor the relative prevalence can be recovered without making assumptions about the data-generating process: intuitively, without further assumptions, it is impossible to distinguish between whether a medical condition is truly rare or merely rarely *diagnosed*. We adopt terminology standard in the PU learning literature and assume that we have access to three pieces of data for the ith example: a feature vector x_i; a group variable g_i; and a binary observed label s_i. We let y_i denote the true (unobserved) label. In healthcare, example i may correspond to a specific patient and their presenting symptoms (x_i), race (g_i), and observed diagnosis (s_i). Here, y_i corresponds to 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 our method is applicable to any set of groups for which our assumptions hold. We make three assumptions:

1. *No false positives*: We assume that examples labeled as positive ($s=1$) are truly positive ($y=1$): i.e., $p(y=0|s=1)=0$ (and thus, by Bayes' rule, $p(s=1|y=0)=0$). This is the positive unlabeled assumption and is the foundational assumption of PU learning methods[57].
2. *Random diagnosis within groups*: We assume that positive examples within a specific group are equally likely to be labeled as positive: $p(s=1|y=1,g=a)=c_a$, where c_a represents the diagnosis frequency of group a (i.e., the probability that a positive case is diagnosed as such). Random diagnosis within groups amounts to the commonly made Selected-Completely-at-Random assumption[58] within each group. We allow c_a to vary across groups to allow for group-specific underdiagnosis rates.
3. *Constant $p(y=1|x)$ between Groups*: We assume that $p(y=1|x)$ remains constant across groups: examples in different groups with the same features are equally likely to be true positives. In the medical setting, this means that patients in different groups with the same symptoms have the same probability of truly having a condition. This is equivalent to assuming only a *covariate shift* between groups, a commonly made assumption in the literature on domain adaptation[23,24] and healthcare[25].

Notably, we make no assumptions about the separability of the positive and negative distributions. Past work in PU learning has shown that the true prevalence $p(y=1)$ can be recovered under a restrictive set of assumptions about the structure of the positive and negative distributions[59], which we refer to as "separability assumptions". Many PU learning methods assume that the positive distribution is not completely contained within the negative distribution: in healthcare, this means there is a region of the feature space where all examples are true positives[57,60–62] (Fig. 2A). This assumption is unrealistic in medical settings because it is unlikely that a set of symptoms maps to a diagnosis with 100% probability[63], and as a result, PURPLE makes no such assumption.

While it is necessary to make assumptions to infer the relative prevalence, no assumptions will hold on all datasets, a point we consider in the Discussion. To ensure PURPLE is applied to appropriate datasets, we provide two checks for violations of PURPLE's assumptions ("Assumption checks" Section). We also show that even under a plausible violation of the *Constant* $p(y = 1|x)$ assumption, PURPLE provides a useful *lower bound* on the magnitude of health disparities ("Robustness to violations of the Constant $p(y|x)$ assumption" Section).

Deriving the relative prevalence. Here, we show that an estimate of $p(y = 1|x)$ up to a constant multiplicative factor recovers the relative prevalence between groups a and b ($\rho_{a,b}$) exactly. The derivation is as follows:

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

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

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

$$= \frac{\sum_x \hat{p}(y=1|x)p(x|g=a)}{\sum_x \hat{p}(y=1|x)p(x|g=b)} \quad \text{for all } \hat{p}(y=1|x) \propto p(y=1|x) \tag{6}$$

where Eq. (5) follows from the constant $p(y = 1|x)$ assumption and Eq. (6) 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 compute the relative prevalence. $p(x|g)$ is directly observable from the data, so we can estimate the numerator as the mean of $\hat{p}(y=1|x)$ over all x in group a and similarly estimate the denominator as the mean of $\hat{p}(y=1|x)$ over all x in group b.

Estimating p(y = 1|x) up to a constant multiplicative factor. We have shown that if we can estimate $p(y = 1|x)$ up to a constant multiplicative factor, we can use this estimate to compute the relative prevalence $\rho_{a,b}$. Now we show how to estimate $p(y = 1|x)$ up to a constant multiplicative factor. We do so by applying our three assumptions to derive a decomposition for $p(s = 1|x, g)$:

$$\begin{aligned} p(s=1|x,g) = p(y=1|x,g)p(s=1|y=1,x,g) \\ + p(y=0|x,g)p(s=1|y=0,x,g) \end{aligned} \tag{7}$$

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

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

$$= \quad p(y=1|x)p(s=1|y=1,g) \tag{10}$$

Applying the *No False Positives* assumption allows us to remove the second term in Eq. (7), producing Eq. (8). The *Random Diagnosis within Groups* assumption removes the dependence of the diagnosis probability on x, leading to Eq. (9). The *Constant* $p(y = 1|x)$ assumption leads to Eq. (10).

Thus, $p(s = 1|x, g)$ can be decomposed as the product of two terms: the probability the patient truly has the condition given their symptoms, $p(y = 1|x)$, and the probability that true positives are correctly diagnosed, $p(s = 1|y = 1, g)$. The fact that the second term varies across groups accounts for group-specific underdiagnosis. This decomposition can be fit via maximum likelihood estimation with respect to the empirical $p(s = 1|x, g)$, since s, x, and g are observed. Note that this only allows estimation of the two terms on the right side of Eqn. (10) up to constant multiplicative factors, since we can multiply $p(y = 1|x)$ by a non-negative β and divide $p(s = 1|y = 1, g)$ by β while leaving our estimate of $p(s = 1|x, g)$ unchanged. However, constant-factor estimation of $p(y = 1|x)$ suffices to estimate the relative prevalence. Concretely, we estimate $p(y = 1|x)$ 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 $p(y = 1|x)$ to estimate the relative prevalence as described in Section "Deriving the relative prevalence".

We note that the probabilistic model described by Eqn. (10) has been previously applied to estimate absolute prevalence in PU settings[26]. Our novel contribution is to derive a precise set of assumptions in which this probabilistic model can be used to estimate *relative* prevalence and provide an estimation method to do so.

Implementation. Thus far, we have shown that it is possible to estimate the relative prevalence of an underreported condition by estimating $p(y = 1|x)$ up to a constant factor and provided a way to conduct this estimation given only the observed data. One can apply PURPLE to a new dataset in two steps:

1. Estimate $p(y = 1|x)$ up to a constant multiplicative factor using the observed diagnoses and the following probabilistic model:

$$\hat{p}(s=1|g,x) = \hat{p}(y=1|x)\hat{p}(s=1|y=1,g) \tag{11}$$

2. Plug our constant multiplicative factor estimate, $\hat{p}(y=1|x)$, into Eq. (6) to produce the relative prevalence estimate. Specifically, we estimate the relative prevalence $\rho_{a,b}$ as:

$$\frac{\sum_x \hat{p}(y=1|x)\hat{p}(x|g=a)}{\sum_x \hat{p}(y=1|x)\hat{p}(x|g=b)} \tag{12}$$

In practice, we can compute this fraction simply by taking the mean value of $\hat{p}(y=1|x)$ in each group to compute the numerator and denominator.

We implement the model in PyTorch[64] using a single-layer neural network to represent $\hat{p}(y=1|x)$ and group-specific parameters $c_g = \hat{p}(s=1|y=1,g)$ for each group g. Note that a single layer neural network, followed by a logistic activation, is functionally equivalent to a logistic regression, as they both learn a linear transformation of the input features followed by a logistic transformation to produce a predicted probability of the positive class. We train the model using the Adam optimizer with default parameters (i.e., a learning rate of 0.001, epsilon of 10^{-8}, and weight decay of 0) and implement early stopping based on the cross-entropy loss on the held-out validation set. For the semi-synthetic and real data, we use L1 regularization because these experiments are conducted on high-dimensional vectors describing thousands of symptoms, most of which we expect to be unrelated to the medical condition, and select the regularization parameter $\lambda \in [10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 0]$ using the 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 approach is general and could be applied with deeper neural network architectures to accommodate interactions and nonlinearities.

Assumption checks. Like all PU learning methods, PURPLE must rely on assumptions. To prevent users from applying PURPLE to datasets where these assumptions do not hold, we provide two empirical tests whose failure implies at least one of the underlying assumptions fails:

- *Compare the model fit of* PURPLE *to the fit of an unconstrained model.* If PURPLE's assumptions hold, the diagnosis likelihood $p(s = 1|x, g)$ decomposes as the product of two terms: $p(s = 1|x, g) = p(y = 1|x)$ $p(s = 1|y = 1, g)$. This is a constrained model of $p(s = 1|x, g)$: for example, it does not allow for interaction terms between group g and symptoms x. We can compare the performance of PURPLE to a fully unconstrained model for $p(s = 1|x, g)$ which allows these interaction terms. If the unconstrained model better fits the data, metrics, including the AUC and AUPRC, will be higher on a held-out set of patients. If the constrained and constrained models exhibit similar performance, it is still possible for one of the assumptions to not be true; however, if the models exhibit different performance, it is a sign that PURPLE's assumptions are not appropriate.

- *Compare calibration across groups.* PURPLE estimates a probabilistic model of diagnosis, $p(s=1|x,g)$, which means we can check how well the outputted probabilities reflect the real data by examining model calibration, a standard check[65]. Concretely, we expect that a proportion z of examples that our model gives a probability z of receiving a positive diagnosis truly receive a positive diagnosis, and we expect this to be true for each group. Violations of PURPLE's assumptions will often cause group-specific calibration to not hold. For example, if $p(y=1|x)$ differs between groups beyond a scalar constant factor, PURPLE's estimate of $p(s=1|x,g)$ cannot be correct for both groups (since PURPLE assumes $p(y=1|x)$ remains constant).

We note that these assumption checks cannot rule out all forms of model misspecification—and, indeed, no assumption checks can. Since only x, g, and s are observed, it is impossible to prove anything about the distribution of y. However, the assumption checks will rule out some forms of model misspecification and guide users away from datasets where applying PURPLE is clearly inappropriate.

Robustness to violations of the constant p(y|x) assumption. In this section, we show that under a plausible violation of the central new assumption of our work (constant $p(y=1|x)$ between groups), PURPLE produces 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, and we can be confident in targeting policy to that group.

Specifically, we relax the assumption of constant $p(y=1|x)$ across groups by assuming that if group A has a higher overall prevalence of a condition than group B—i.e., $p(y=1|g=A) > p(y=1|g=B)$—group A also has a higher prevalence by a constant factor given the same set of symptoms—i.e., $p(y=1|x,g=A)=\alpha\cdot p(y=1|x,g=B), \alpha>1$. This assumption is a reasonable one: when a condition is more prevalent in one group than another, the same symptoms plausibly correspond to higher posterior probabilities $p(y=1|x)$ in the disproportionately affected group. For example, female patients are more likely than male patients to be victims of intimate partner violence overall[30], and if a woman and a man arrive in a hospital with the same injuries, doctors are plausibly more likely to suspect intimate partner violence as the cause of the woman's injuries.

Proof 1. Without loss of generality, assume that group A is the group with higher overall disease prevalence: $p(y=1|g=A) > p(y=1|g=B)$. We make the following assumptions:

1. Disease prevalence in group A conditional on symptoms is higher by a constant multiplicative factor: $p(y=1|x,g=A)=\alpha\cdot p(y=1|x,g=B)$, $\alpha>1$.
2. The other PURPLE assumptions hold: that is, $p(y=1|s=1)=1$ (PU assumption) and $p(s=1|y=1,g,x)=p(s=1|y=1,g)$ (random diagnosis within groups).

Under these assumptions, we show that PURPLE's estimate provides a lower bound on the true relative prevalence. As before, we use p to denote the true probabilities in the underlying data distribution and $\hat{p}$ to denote PURPLE's estimates of these probabilities. We have

$$p(s=1|x,g=A)=p(s=1|y=1,g=A)\cdot p(y=1|x,g=B)\cdot\alpha$$
$$p(s=1|x,g=B)=p(s=1|y=1,g=B)\cdot p(y=1|x,g=B)$$

PURPLE estimates $\hat{p}(s=1|x,g)=\hat{p}(s=1|y=1,g)\cdot\hat{p}(y=1|x)$. It can minimize the cross-entropy loss by achieving an estimate $\hat{p}(s=1|x,g)$ which matches the true $p(s=1|x,g)$ by setting:

$$\begin{aligned}\hat{p}(y=1|x) &= \beta\cdot p(y=1|x,g=B)\\ \hat{p}(s=1|y=1,g=A) &= \tfrac{1}{\beta}\cdot\alpha\cdot p(s=1|y=1,g=A)\\ \hat{p}(s=1|y=1,g=B) &= \tfrac{1}{\beta}\cdot p(s=1|y=1,g=B)\end{aligned}$$

where β is a positive constant (this captures the fact that, as discussed previously, PURPLE only ever estimates $p(y=1|x)$ up to a constant multiplicative factor). In other words, PURPLE can perfectly match the true probabilities $p(s=1|g,x)$ by pushing the variation across groups in $p(y=1|x)$ into $\hat{p}(s=1|y=1,g)$ (since PURPLE assumes $\hat{p}(y=1|x)$ remains constant across groups.) Given its estimate $\hat{p}(y=1|x)$, PURPLE's estimate $\hat{\rho}_{A,B}$ of the relative prevalence is

$$\begin{aligned}\hat{\rho}_{A,B} &= \frac{\sum_{x\in\mathcal{A}}\hat{p}(y=1|x)}{\sum_{x\in\mathcal{B}}\hat{p}(y=1|x)}\\ &= \frac{\sum_{x\in\mathcal{A}}\beta\cdot p(y=1|x,g=B)}{\sum_{x\in\mathcal{B}}\beta\cdot p(y=1|x,g=B)}\\ &< \frac{\sum_{x\in\mathcal{A}}p(y=1|x,g=A)}{\sum_{x\in\mathcal{B}}p(y=1|x,g=B)}\\ &= \rho_{A,B}\end{aligned}$$

so PURPLE's estimate $\hat{\rho}_{A,B}$ provides a lower bound on the true relative prevalence $\rho_{A,B}$.

Empirical validation. We verify this behavior empirically under synthetic violations of the constant $p(y|x)$ assumption, again using the synthetic data described in Section "Gauss-Synth". Concretely, we vary the difference in the probability of a positive case between group a and group b ($p(y=1|x,g=a)-p(y=1|x,g=b)$), where $p(y=1|g=a)\geq p(y=1|g=b)$.

Figure 1B in the Supplement demonstrates how PURPLE consistently underestimates the true relative prevalence behavior empirically by plotting PURPLE's behavior over a range of values for α, where $\alpha\in[0,1]$ and $\alpha p(y=1|x,g=a)=p(y=1|x,g=b)$. In each case, PURPLE provides a lower bound on the true relative prevalence. We replicate this analysis for varying separations of the group-specific Gaussian distributions, where darker shades of purple correspond to group-specific distributions that are further from one another.

Effect of positive-unlabeled assumption violations. Here we demonstrate how PURPLE's performance varies when the positive-unlabeled assumption is violated, meaning that the sample of observed positives contains some number of negatives. This can arise if the diagnostic test used to create the set of observed positives exhibits a non-zero false positive rate. In SI Fig. 2A, we plot the behavior of PURPLE as we increase the extent to which the setting violates the PU assumption; specifically, we vary the percentage of the observed positives that truly are negative—$p(y=0|s=1)$—from 0% to 20%. When $p(y=0|s=1)=0$, this is equivalent to the positive-unlabeled assumption, and PURPLE recovers the relative prevalence exactly, as expected. At greater violations of the PU assumption, PURPLE's performance degrades somewhat, as expected.

Effect of random-diagnosis-within-groups assumption violations. We illustrate PURPLE's behavior under violations of the random-diagnosis-within-groups assumption in Fig. S2B. The assumption states that the probability that a true positive is diagnosed as such does not depend on x or that $p(s=1|y=1,g,x)=p(s=1|y=1,g)$. We simulate violations of the assumption by generating s according to $p(s=1|y=1,g,x)=c_g\cdot\sigma(\beta\cdot x_0)$. In other words, the diagnosis frequency for group g is scaled by the sigmoid function of parameter β multiplied by the first component of x. For $\beta=0$, the setting adheres to the random-diagnosis-within-groups assumption. Higher values of β translate to a higher correlation between the diagnosis probability and x_0. As Fig. S2B demonstrates, PURPLE recovers the relative prevalence under no violation of the assumption ($\beta = 0$) and incurs small errors in the estimated relative prevalence as β increases.

Datasets

We make use of five datasets. We begin with two synthetic datasets: *Gauss-Synth*, a completely synthetic dataset, and *MIMIC-Semi-Synth*, a semi-synthetic dataset based on real health data. We then apply PURPLE to three non-synthetic datasets: *MIMIC-IV ED*, a dataset of electronic health records collected from a single hospital in the Boston area, and *NEDS*, a dataset of emergency department visits occurring in the US in 2019.

We begin with synthetic and semi-synthetic data so that the ground truth labels y are known, as is standard in the PU learning literature[22], enabling us to assess how well methods recover the relative prevalence. To assess the performance of all methods in our synthetic and semi-synthetic experiments, 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. Code to reproduce all experiments can be found at https://github.com/epierson9/invisible-conditions.

Gauss-Synth. We generate completely synthetic data by simulating group-specific features ($p(x|g)$), and labels using a decision rule ($p(y|x)$). Formally, we simulate groups a and b using two 5D Gaussian distributions with different means:

$$x_i \sim \begin{cases} \mathcal{N}_5(-\mathbf{1}, 16 \cdot \mathbf{1}) \text{ if } g_i = a \\ \mathcal{N}_5(\mathbf{1}, 16 \cdot \mathbf{1}) \text{ if } g_i = b \end{cases} \tag{13}$$

The likelihood function ($p(y = 1|x)$) is a logistic function of the signed distance to a hyperplane through the origin. Observed labels s are drawn such that positive labels in group i are observed with a probability of c_{g_i}. The generative model for y and s is:

$$y_i \sim \text{Bernoulli}\,(\sigma((\mathbf{1}^T x_i)/\parallel \mathbf{1} \parallel)) \tag{14}$$

$$s_i \sim \text{Bernoulli}\,(c_{g_i} y_i) \tag{15}$$

where σ represents a logistic function ($\sigma(x) = \frac{1}{1+e^{-x}}$) and c_{g_i} is the group-specific diagnosis frequency for g_i. We draw 10,000 observations for group a and 20,000 for group b. We create the separable data by modifying the generative model described above, which does not generate separable data. Specifically, we replace each $p(y = 1|x) > 0.5$ with $p(y = 1|x) = 1$ and each $p(y = 1|x) < 0.5$ with $p(y = 1|x) = 0$, and remove the 40% of the data closest to the original decision boundary to ensure the classes are cleanly separable, as illustrated in Fig. 2A.

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[27]. We filter out ICD codes that appear 10 or fewer times, leaving 5544 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 Eq. (15) in our generative model with:

$$y_i \sim \text{Bernoulli}\,(\sigma(v_{sym}^T \mathbf{x_i})/\parallel v_{sym} \parallel) \tag{16}$$

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. We filter the dataset for patients belonging to 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 v_{sym} in four different ways:

Common symptoms. We identify the 50 most common ICD codes in MIMIC-IV and randomly select 25 to be suspicious symptoms (described fully in the Supplement). 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 ($p(y = 1|g = b) = 0.185$).

High relative prevalence symptoms. We filter out ICD codes that appear fewer than 50 times in each group and patients less than 18 years old. After ranking the ICD codes by relative prevalence—prevalence among visits by white patients, divided by prevalence among visits by Black patients—we select the top 10 ICD codes as our suspicious symptom set (described fully in the Supplement). Group a contains 14,618 visits from Black patients $p(y = 1|g = a) = 0.061$), and group b contains 61,000 visits from white patients ($p(y = 1|g = b) = 0.098$).

Correlated symptoms. We consider endometriosis, a widely under-diagnosed condition[66]. 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 gold-standard ICD endometriosis diagnosis codes (described fully in the Supplement). We then identify the ICD codes which are most highly associated with a gold-standard 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" (all codes are included in the Supplement). 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 male patients[67], 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 ($p(y = 1|g = a) = 0.0495$).

Recognized symptoms for IPV. Prior work has found that suspicious symptoms for IPV include head, neck, and facial injuries[68]. The symptoms in this experiment consist of the 100 ICD codes corresponding to these injuries (described fully in the Supplement). We filter for female patients because the symptoms associated with IPV in male patients are not well understood[30]. 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. This results in a dataset with $p(y = 1|g = a) = 0.0541$ (25,546 unique patient visits) and $p(y = 1|g = b) = 0.0568$ (80,227 unique patient visits).

In Sections "MIMIC-IV ED" and "NEDS", we describe the construction of the two non-synthetic datasets, MIMIC-IV and NEDS. In all these datasets, y is unknown, so we need only define the features x and known positive examples in which $s = 1$.

MIMIC-IV ED

Data filtering. MIMIC-IV contains two related databases: one representing diagnoses made in the hospital (which we will refer to as the hospital database) and one representing diagnoses made in the emergency department. There are slight inconsistencies between the two, as is to be expected; for example, one emergency department stay can be associated with multiple unique hospital admissions. We exclude 600 hospital admissions linked to multiple emergency department stays and exclude 59 emergency department stays associated with invalid hospital admissions. For emergency department stays that result in hospital admission, we include all diagnoses assigned in the emergency department or hospital. We also include patient visits that appear only in the hospital database but indicate admission through the emergency department via the "admission_location" field. We provide code to replicate these preprocessing steps.

We further filter for patients who are female and above 18. We do so because we are interested in the relative prevalence of intimate partner violence between subgroups of adult female patients. This leaves 293,297 individual hospital visits over 133,470 unique patients. For each demographic attribute we wish to analyze disparities over (ethnicity, insurance status, and marital status), we also filter out patients who are missing data for this attribute. This translates to 192,768 stays across marital statuses, 208,512 stays across ethnicities, and 108,948 stays across insurance statuses. We have significantly fewer stays in the insurance subgroups because we only have patients for whom the insurance status is known (i.e., Medicare and Medicaid recipients). We produce 5 randomized dataset splits, where we reserve 60% of the data for training, 20% for validation, and 20% for testing. Each patient appears in only one of these sets.

Defining features x. We represent each patient visit as a one-hot encoding of the ICD codes assigned. Concretely, 15,699 features represent each patient visit, where each feature corresponds to the presence or absence of one ICD code (across the ICD-9 and ICD-10 standards). Note that this is different from the semi-synthetic setup and we do *not* filter out codes that appeared fewer than 10 times. We do this because IPV itself is rare, and we do not want to exclude symptoms that are predictive of IPV and do not occur frequently.

Defining $s = 1$. To define examples where $s = 1$ (known positive examples), we use criteria for reported instances of intimate partner violence from prior work. The most specific code is E967.3, or "Battering by an intimate partner", drawn from the ICD-9 standard. Other codes include V6111, or "Counseling for a victim of spousal or partner abuse". The full code set can be found in Table S6. If a patient receives any one of the codes in the positive code set, the visit is deemed to be positively labeled for intimate partner violence.

Defining g. We define the groups over which we quantify disparities via demographic variables associated with each electronic health record. These include race/ethnicity (Black, white, Asian, Hispanic/Latino), insurance status (Medicare or Medicaid), and marital status (Legally Unmarried, Married, Divorced).

NEDS. The National Emergency Department Sample (NEDS) is the largest publicly available, all-payer database describing visits to hospital-owned emergency departments in the United States and is commonly used in studies of disease prevalence. In this work, we make use of NEDS 2019. The survey from this year produced a dataset containing 33.1 million visits, which represents the 143 million total visits occurring in hospital-owned EDs across the United States in 2019. To create nationally representative estimates, NEDS releases "discharge weights", which allow the analyst to reweight estimates to represent the universe of emergency department visits. We follow this procedure to reweight our relative prevalence estimates.

Data filtering. As with MIMIC-IV, we include visits from patients who are female and above the age of 18. The resulting dataset contains 15,357,528 visits.

Defining features x. As before, we treat the ICD codes logged during a visit as the input features, as a proxy for the symptoms a patient presents with. The feature set consists of 19,710 ICD-10 codes (excluding those used to identify positive cases, as described in the next section), which we one-hot encode to create features for each ED visit.

Defining $s = 1$. We use the same criteria to identify positive cases as with MIMIC-IV: we consider a case to be labeled positive if it is associated with if it has any of the codes described in the Supplement.

Defining g. We define groups according to race/ethnicity groups, insurance status, income quartile, and urban/rural designation. The income quartile is calculated using the estimated median household income in the patient's zipcode. In 2019, the first quartile corresponds to an estimated median household income between \$0 and \$48k, the second to \$48k to \$61k, the third to \$61k–\$82k, and the fourth to the remaining zip codes. The urban/rural designations are based on the population of the patient's home county. From most to least rural, the four categories are 50k residents to 250k residents, 250k residents to 1 million residents, 1 million residents in a fringe metropolitan county, and 1 million residents in a central metropolitan county.

Baselines

Each of the baseline methods described below is designed to estimate the *absolute* prevalence. To obtain the relative prevalence, we apply each baseline to groups a and b individually to obtain estimates of the absolute prevalence in each group; we then divide the resulting quantities to produce an estimate of the relative prevalence. To provide consistent comparisons, we constrain each baseline to use the same function class as `PURPLE`. We do not compare to baselines in the epidemiology literature because they assume access to external information (e.g., diagnostic accuracy) that is often not available; we also do not consider work that places parametric assumptions on $p(y = 1|x)$[69–72] because these assumptions will not hold in general.

- *Negative*: Assign all unlabeled examples a negative label. This approach replaces $p(y = 1|x)$ with $p(s = 1|x)$ and assumes no underreporting occurs. Past work refers to this model as a *nontraditional classifier*[57] (NTC). We use *sklearn*'s logistic regression implementation with no regularization and default settings for all other hyperparameters, trained with target s.
- *KM2*: 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[62]. This method is known to perform poorly on large datasets with many features[73]. KM2 assumes that there exists a function that only selects positive examples.
- *DEDPUL*: Uses a non-traditional classifier to map each example to a predicted probability of diagnosis and performs mixture proportion estimation using the classifier's outputs on the unlabeled examples[73]. Specifically, the method applies heuristics to the estimated densities of the positive and unlabeled distribution.
- *BBE*: Identifies a small subset of positive examples using the outputs of an non-traditional classifier on the positive and unlabeled sample[74]. The method uses this subset to infer the proportion of positive examples in the unlabeled sample.
- *Oracle*: Uses the true label y to estimate $p(y = 1|x)$. Importantly, this method cannot actually be applied in real data since y is unobserved, but it represents an upper bound on performance.

Data availability

Anonymized imaging and clinical data to reproduce the results of this study are available online. MIMIC-IV is a publicly available database of emergency department and hospital admissions occurring between 2008 and 2019 at the Beth Israel Deaconess Medical Center and can be found at: https://physionet.org/content/mimiciv/2.2/. We make use of the National Emergency Department Sample (2019), made available by the Agency for Healthcare Research and Quality, which is publicly available at https://www.hcup-us.ahrq.gov/nedsoverview.jsp. Code to preprocess all datasets and reproduce all experiments can be found at https://github.com/epierson9/invisible-conditions.

Received: 7 June 2023; Accepted: 23 February 2024;
Published online: 15 May 2024

References

1. MacDorman, M. F., Declercq, E. & Thoma, M. E. Trends in maternal mortality by socio-demographic characteristics and cause of death in 27 states and the district of columbia. *Obstet. Gynecol.* **129**, 811 (2017).
2. FACT SHEET: Biden-Harris Administration Announces Initial Actions to Address the Black Maternal Health Crisis. https://www.whitehouse.gov/briefing-room/statements-releases/2021/04/13/fact-sheet-biden-harris-administration-announces-initial-actions-to-address-the-black-maternal-health-crisis/. Accessed: 2022-10-16.
3. Geiger, H. J. Racial and ethnic disparities in diagnosis and treatment: a review of the evidence and a consideration of causes. *Unequal Treat.* **417**, 1–38 (2003).
4. Schafer, S. D., Drach, L. L., Hedberg, K. & Kohn, M. A. Using diagnostic codes to screen for intimate partner violence in oregon emergency departments and hospitals. *Public Health Rep.* **123**, 628–635 (2008).
5. Lyles, R. H. et al. Validation data-based adjustments for outcome misclassification in logistic regression: an illustration. *Epidemiology* **22**, 589 (2011).
6. Sekar, C. C. & Deming, W. E. On a method of estimating birth and death rates and the extent of registration. *J. Am. Stat. Assoc.* **44**, 101–115 (1949).
7. Simeone, R. S., Rhodes, W. M. & Hunt, D. E. A plan for estimating the number of "hardcore" drug users in the united states. *Int. J. Addict.* **30**, 637–657 (1995).
8. Hay, G. & Smit, F. Estimating the number of drug injectors from needle exchange data. *Addict. Res. Theory* **11**, 235–243 (2003).
9. McKeganey, N., Barnard, M., Leyland, A., Coote, I. & Follet, E. Female streetworking prostitution and hiv infection in glasgow. *Br. Med. J.* **305**, 801–804 (1992).
10. Hui, S. L. & Walter, S. D. Estimating the error rates of diagnostic tests. *Biometrics* **36**, 167–171 (1980).
11. Walter, S. D. & Irwig, L. M. Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. *J. Clin. Epidemiol.* **41**, 923–937 (1988).
12. Pepe, M. S. & Janes, H. Insights into latent class analysis of diagnostic test performance. *Biostatistics* **8**, 474–484 (2007).
13. Lewis, F., Sanchez-Vazquez, M. & Torgerson, P. Association between covariates and disease occurrence in the presence of diagnostic error. *Epidemiol. Infect.* **140**, 1515–1524 (2012).
14. Singh, H., Meyer, A. N. & Thomas, E. J. The frequency of diagnostic errors in outpatient care: estimations from three large observational studies involving US adult populations. *BMJ Qual. Saf.* **23**, 727–731 (2014).
15. Chen, J. H. & Asch, S. M. Machine learning and prediction in medicine—beyond the peak of inflated expectations. *N. Engl. J. Med.* **376**, 2507 (2017).
16. Davidov, D. M., Larrabee, H. & Davis, S. M. United states emergency department visits coded for intimate partner violence. *J. Emerg. Med.* **48**, 94–100 (2015).
17. Johnson, A. et al. Mimic-iv-ed (2021).
18. Cost, H. & (HCUP), U. P. The study examined emergency department visits for diabetes using discharge data from the nationwide emergency department sample (neds).
19. Rennison, C. & Welchans, S. Bureau of justice statistics special report: Intimate partner violence. retrieved november 12, 2007 (2000).
20. Bonomi, A. E., Trabert, B., Anderson, M. L., Kernic, M. A. & Holt, V. L. Intimate partner violence and neighborhood income: a longitudinal analysis. *Violence Women* **20**, 42–58 (2014).
21. Abramsky, T. et al. Women's income and risk of intimate partner violence: secondary findings from the maisha cluster randomised trial in north-western tanzania. *BMC Public Health* **19**, 1–15 (2019).
22. Bekker, J. & Davis, J. Learning from positive and unlabeled data: a survey. *Mach. Learn.* **109**, 719–760 (2020).
23. Sugiyama, M., Krauledat, M. & Müller, K.-R. Covariate shift adaptation by importance weighted cross validation. *J. Mach. Learn. Res.* **8**, 985–1005 (2007).
24. Quiñonero-Candela, J., Sugiyama, M., Lawrence, N. D. & Schwaighofer, A. *Dataset Shift in Machine Learning* (Mit Press, 2009).
25. Nestor, B. et al. Feature robustness in non-stationary health records: caveats to deployable model performance in common clinical machine learning tasks. In *Machine Learning for Healthcare Conference*, 381–405 (PMLR, 2019).
26. Bekker, J., Robberechts, P. & Davis, J. 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*, 71–85 (Springer, 2019).
27. Johnson, A. et al. Mimic-iv (version 0.4). *PhysioNet* (2020).
28. Riley, W. J. Health disparities: gaps in access, quality and affordability of medical care. *Trans. Am. Clin. Climatol. Assoc.* **123**, 167 (2012).
29. Alessandrino, F. et al. Intimate partner violence: a primer for radiologists to make the "invisible" visible. *Radiographics* **40**, 2080–2097 (2020).
30. Houry, D. et al. Differences in female and male victims and perpetrators of partner violence with respect to web scores. *J. Interpers. Violence* **23**, 1041–1055 (2008).
31. Louwers, E. C. et al. Detection of child abuse in emergency departments: a multi-centre study. *Arch. Dis. Child.* **96**, 422–425 (2011).
32. Loder, R. T. & Momper, L. Demographics and fracture patterns of patients presenting to us emergency departments for intimate partner violence. *JAAOS Glob. Res. Rev.* **4**, e20 (2020).
33. Amjad, H. et al. Underdiagnosis of dementia: an observational study of patterns in diagnosis and awareness in us older adults. *J. Gen. Intern. Med.* **33**, 1131–1138 (2018).
34. Bradford, A., Kunik, M. E., Schulz, P., Williams, S. P. & Singh, H. Missed and delayed diagnosis of dementia in primary care: prevalence and contributing factors. *Alzheimer Dis. Assoc. Disord.* **23**, 306–314 (2009).
35. Aaron, S. D., Boulet, L. P., Reddel, H. K. & Gershon, A. S. Underdiagnosis and overdiagnosis of asthma. *Am. J. Respir. Crit. Care Med.* **198**, 1012–1020 (2018).
36. Quinn, K., Shalowitz, M. U., Berry, C. A., Mijanovich, T. & Wolf, R. L. Racial and ethnic disparities in diagnosed and possible undiagnosed asthma among public-school children in chicago. *Am. J. Public Health* **96**, 1599–1603 (2006).
37. Swant, E. & Wyatt, L. 200 underdiagnosis of depression among low-income, predominantly latino, type 2 diabetics. (2007).
38. Lao, C.-K., Chan, Y.-M., Tong, H. H.-Y. & Chan, A. Underdiagnosis of depression in an economically deprived population in m acao, c hina. *Asia Pac. Psychiatry* **8**, 70–79 (2016).
39. Sorkin, D. H. et al. Underdiagnosed and undertreated depression among racially/ethnically diverse patients with type 2 diabetes. *Diabetes Care* **34**, 598–600 (2011).
40. Pathak, N., Dhairyawan, R. & Tariq, S. The experience of intimate partner violence among older women: a narrative review. *Maturitas* **121**, 63–75 (2019).
41. Evans, C. S., Hunold, K. M., Rosen, T. & Platts-Mills, T. F. Diagnosis of elder abuse in us emergency departments. *J. Am. Geriatrics Soc.* **65**, 91–97 (2017).
42. Gerino, E., Caldarera, A. M., Curti, L., Brustia, P. & Rollè, L. Intimate partner violence in the golden age: systematic review of risk and protective factors. *Front. Psychol.* **9**, 1595 (2018).
43. Cunradi, C. B., Caetano, R., Clark, C. & Schafer, J. Neighborhood poverty as a predictor of intimate partner violence among white, black, and hispanic couples in the united states: A multilevel analysis. *Ann. Epidemiol.* **10**, 297–308 (2000).
44. Mariscal, T. L., Hughes, C. M. & Modrek, S. Changes in incidents and payment methods for intimate partner violence related injuries in

women residing in the united states, 2002 to 2015. *Women's Health Issues* **30**, 338–344 (2020).
45. Wong, J. Y.-H. et al. A comparison of intimate partner violence and associated physical injuries between cohabitating and married women: a 5-year medical chart review. *BMC Public Health* **16**, 1–9 (2016).
46. Abramsky, T. et al. What factors are associated with recent intimate partner violence? findings from the who multi-country study on women's health and domestic violence. *BMC Public Health* **11**, 1–17 (2011).
47. Capaldi, D. M., Knoble, N. B., Shortt, J. W. & Kim, H. K. A systematic review of risk factors for intimate partner violence. *Partn. Abus.* **3**, 231–280 (2012).
48. Ravi, K. E., Rai, A. & Schrag, R. V. Survivors' experiences of intimate partner violence and shelter utilization during covid-19. *J. Fam. Violence* **37**, 979–990 (2022).
49. DuBois, K. O., Rennison, C. M. & DeKeseredy, W. S. Intimate partner violence in small towns, dispersed rural areas, and other locations: Estimates using a reconception of settlement type. *Rural Sociol.* **84**, 826–852 (2019).
50. Lipsky, S., Caetano, R. & Roy-Byrne, P. Racial and ethnic disparities in police-reported intimate partner violence and risk of hospitalization among women. *Women's Health Issues* **19**, 109–118 (2009).
51. Cho, H. Racial differences in the prevalence of intimate partner violence against women and associated factors. *J. Interpers. Violence* **27**, 344–363 (2012).
52. Domestic, E. & Violence, G.-B. 2020 report on the intersection of domestic violence, race/ethnicity and sex (2020).
53. Hart JD, B. & Klein PhD, A. J. Practical implications of current intimate partner violence research for victim advocates and service providers (2013).
54. Hillman, S. C. & Dale, J. Polycystic ovarian syndrome: an under-recognised problem? *Br. J. Gen. Pract.* **68**, 244–244 (2018).
55. Agarwal, S. K. et al. Clinical diagnosis of endometriosis: a call to action. *Am. J. Obstet. Gynecol.* **220**, 354–e1 (2019).
56. Prince, C. & Bruhns, M. E. Evaluation and treatment of mild traumatic brain injury: the role of neuropsychology. *Brain Sci.* **7**, 105 (2017).
57. Elkan, C. & Noto, K. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 213–220 (2008).
58. Bekker, J. & Davis, J. Estimating the class prior in positive and unlabeled data through decision tree induction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32 (2018).
59. Jain, S., White, M. & Radivojac, P. Estimating the class prior and posterior from noisy positives and unlabeled data. *Adv. Neural Inf. Process. Syst.* **29**, 2693–2701 (2016).
60. Du Plessis, M. C. & Sugiyama, M. Class prior estimation from positive and unlabeled data. *IEICE Trans. Inf. Syst.* **97**, 1358–1362 (2014).
61. Northcutt, C. G., Wu, T. & Chuang, I. L. Learning with confident examples: rank pruning for robust classification with noisy labels. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1705.01936 (2017).
62. Ramaswamy, H., Scott, C. & Tewari, A. Mixture proportion estimation via kernel embeddings of distributions. In *International conference on machine learning*, 2052–2060 (PMLR, 2016).
63. Chen, I. Y., Joshi, S., Ghassemi, M. & Ranganath, R. Probabilistic machine learning for healthcare. *Annu. Rev. Biomed. Data Sci.* **4**, 393–415 (2021).
64. Paszke, A. et al. Automatic differentiation in pytorch (2017).
65. Steyerberg, E. W. et al. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology* **21**, 128 (2010).
66. Moradi, M., Parker, M., Sneddon, A., Lopez, V. & Ellwood, D. Impact of endometriosis on women's lives: a qualitative study. *BMC Women's Health* **14**, 1–12 (2014).
67. Jabr, F. I. & Mani, V. An unusual cause of abdominal pain in a male patient: endometriosis. *Avicenna J. Med.* **4**, 99–101 (2014).
68. Wu, V., Huff, H. & Bhandari, M. Pattern of physical injury associated with intimate partner violence in women presenting to the emergency department: a systematic review and meta-analysis. *Trauma Violence Abus.* **11**, 71–82 (2010).
69. Lazkecka, M., Mielniczuk, J. & Teisseyre, P. Estimating the class prior for positive and unlabelled data via logistic regression. *Adv. Data Anal. Classif.* **15**, 1039–1068 (2021).
70. Jaskie, K. & Spanias, A. Positive and unlabeled learning algorithms and applications: A survey. In *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, 1–8 (IEEE, 2019).
71. Teisseyre, P., Mielniczuk, J. & Lazkecka, M. Different strategies of fitting logistic regression for positive and unlabelled data. In *International Conference on Computational Science*, 3–17 (Springer, 2020).
72. Furmańczyk, K., Mielniczuk, J., Rejchel, W. & Teisseyre, P. Joint estimation of posterior probability and propensity score function for positive and unlabelled data. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2209.07787 (2022).
73. Ivanov, D. Dedpul: difference-of-estimated-densities-based positive-unlabeled learning. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 782–790 (IEEE, 2020).
74. Garg, S., Wu, Y., Smola, A. J., Balakrishnan, S. & Lipton, Z. Mixture proportion estimation and pu learning: a modern approach. *Adv. Neural Inf. Process. Syst.* **34**, 8532–8544 (2021).

Acknowledgements

We thank Microsoft Research seminar participants Irene Chen, Serina Chang, Jacquelyn Campbell, Karthik Chetty, Nikhil Garg, John Guttag, Pang Wei Koh, Allison Koenecke, Nat Roth, Priya Shanmugam, and Harini Suresh for their helpful conversations. DS was supported by the Wistron Corporation. EP was supported by a Google Research Scholar award, NSF CAREER #2142419, a CIFAR Azrieli Global scholarship, a LinkedIn Research Award, and the Abby Joseph Cohen Faculty Fund.

Author contributions

E.P. supervised the project. E.P. and D.S. conceived of the presented method, experiments, and theory and verified the results. E.P. and D.S. wrote the paper. D.S. conducted the synthetic and semi-synthetic experiments, carried out the case studies for intimate partner violence and content moderation, and created the figures. K.H. conducted analyses of the MIMIC-IV data and the NEDS data.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44294-024-00011-5.

Correspondence and requests for materials should be addressed to Divya Shanmugam.

Reprints and permissions information is available at http://www.nature.com/reprints