

Bias-corrected estimates for the fraction of unlabeled instances in the context of positive unlabeled learning for inferring health conditions

Kristan A. Schneider¹, Praveen Kumar¹, Fariha Moomtaheen¹, Rajesh Upadhayaya¹, Alexandria D Viszolay¹, Donald R Greer², Scott A. Malec¹, Jeremy J. Yang¹, Cristian G. Bologa¹, Yiliang Zhu¹, Mauricio Tohen³, Gerardo Villarreal^{3,4}, Douglas J. Perkins¹, Elliot M. Fielstein⁵, Sharon E. Davis⁵, Michael E. Matheny^{5,6}, David van der Goes^{4,7}, Christophe G. Lambert¹

¹University of New Mexico, Department of Internal Medicine, Albuquerque, NM, USA

²Greer Black Company, Bozeman, MT, USA

³University of New Mexico, Department of Psychiatry & Behavioral Sciences, Albuquerque, NM, USA

⁴VA New Mexico Healthcare System, Albuquerque, NM, USA

⁵Vanderbilt University Medical Center, Department of Biomedical Informatics, Nashville, TN, USA

⁶Tennessee Valley Healthcare System VA, Nashville, TN, USA

⁷Department of Economics, University of New Mexico, Albuquerque, NM, USA

Background

In many biomedical applications, including electronic health record (EHR) research, health conditions are only partially observed.¹⁻³ Specifically, some individuals with the condition of interest are explicitly labeled as positive, while the remaining individuals are unlabeled. Importantly, the absence of a diagnostic label does not imply the absence of the condition. This setting, commonly referred to as positive-unlabeled (PU) learning, presents fundamental challenges for estimation of the condition's prevalence and model validation. A representative example is post-traumatic stress disorder (PTSD), where underdiagnosis and incomplete coding lead to substantial numbers of unlabeled but truly positive cases.⁴

Formally, the problem of PU learning can be summarized as follows. Let N denote the total sample size of a dataset X . Let further N_1 be the number of labeled positive instances. The remaining observations are unlabeled and consist of an unknown number of unlabeled positives (N_2) and true negatives (N_3), such that $N = N_1 + N_2 + N_3$. While N and N_1 are observed, N_2 and N_3 are unknown. However, the number of unlabeled records ($N_2 + N_3 = N - N_1$) is known. Denoting the corresponding population proportions of labeled positives, unlabeled positives, and true negatives by p_1 , p_2 , and p_3 respectively, one seeks to estimate the proportion of unlabeled positives $\alpha = \frac{p_2}{p_2 + p_3}$.

Several methods have been proposed to estimate the fraction of positives among unlabeled, e.g. DEDPUL,⁵ TlCE,⁶ Elkan-Noto,⁷ PULSCAR, or PULSNAR.⁸ While TlCE exclusively estimates α , the other methods provide a posterior probability for each unlabeled instance to be positive.

An intrinsic problem of PU learning is the lack of a ground truth in the dataset X , i.e., it is unclear how biased a given estimate $\hat{\alpha}$ is. To resolve this problem, we introduce a method for bias correction in the context of PU learning, which is applicable to any PU classifier, which provides posterior probabilities of being positive for all unlabeled instances.

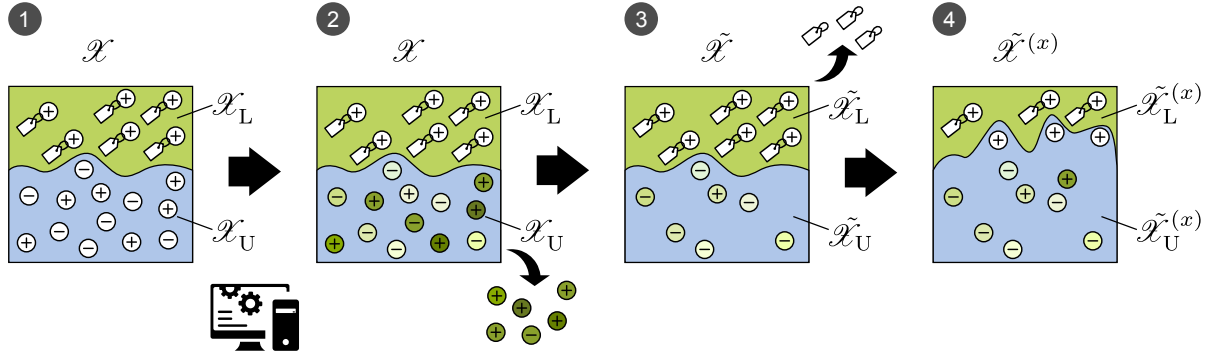


Figure 1. Illustration of the construction of reduced and modified data sets for bias correction. A dataset $\tilde{X}^{(x)}$ is constructed from $\tilde{X}$ by removing x labels. Hence, the number of unlabeled positives in $\tilde{X}^{(x)}$ increases exactly by x compared to $\tilde{X}$.

Methods

Let X_L and X_U denote, respectively, the sets of positive and unlabeled instances in the dataset X , and s a classifier for PU learning. First, the fraction of positives among unlabeled instances ($\hat{\alpha}$) is estimated (Fig. 1). Second, the posterior probabilities of instances in the set of unlabeled are calculated (Fig. 1). Third, those instances with the highest $2\hat{\alpha}$ probabilities of being positive are removed from the set X_U , yielding a reduced dataset $\tilde{X}_U$, which approximately contains no positive instances (Fig. 1). Hence, one obtains a new dataset $\tilde{X}$ consisting of the sets X_L of labelled and $\tilde{X}_U$ of unlabeled instances of size $\tilde{N}$, for which the ground truth is approximately known: $\tilde{p}_1 \approx \frac{|X_L|}{|X_L| + |\tilde{X}_U|}$ and $\tilde{p}_2 \approx 0$. Fourth, the dataset $\tilde{X}$ is manipulated by randomly removing the labels of x instances (Fig. 1), i.e., these instances are moved from the set X_L into the set $\tilde{X}_U$, yielding a new dataset $\tilde{X}^{(x)}$ of the same size $\tilde{N}$ consisting of the sets $\tilde{X}_L^{(x)}$ of labeled and $\tilde{X}_U^{(x)}$ unlabeled instances. In the new dataset, the fraction of labeled positives is $\tilde{p}_1^{(x)} = \tilde{p}_1 - \frac{x}{\tilde{N}}$ and the fraction of unlabeled positives is $\tilde{p}_2^{(x)} = \frac{x}{\tilde{N}}$. Fifth, step four is repeated for a range of values of x and a linear regression is fit to the classifier's estimates of $\tilde{p}_2^{(x)}$. If the classifier s was unbiased, the intercept of the regression equals 0 and the slope equals 1. If not, the estimated slope and intercept of the regression are used to correct the PU classifier s .

To illustrate the performance of this method for bias correction, it is applied to the three PU classifiers (DEDPUL, Elkan-Noto, and PULSCAR) using both synthetic and benchmark datasets from Kaggle and University of California, Irvine (UCI), including Wilt remote sensing ($N=4,839$, 5 features, 4,576 labelled positive, 236 unlabeled),⁹ Magic gamma telescope ($N=19,020$; 10 features, 12,332 labeled positive, 6,688 unlabeled),¹⁰ and Mouse protein expression ($N=1,080$, 77 features, 570 labeled positive, 510 unlabeled).¹¹

Results

All three methods, DEDPUL, Elkan-Noto, and PULSCAR, can be reasonably bias corrected with the proposed method (cf. Fig. 2). The bias correction fails to perform if the size of the dataset is too small, so that (i) there is high initial uncertainty in the original estimate $\hat{\alpha}$ and (ii) the reduced dataset $\tilde{X}_U$ is too small to represent the underlying population. The bias correction also fails if the original estimate $\hat{\alpha}$ is either so large or so small that removing the unlabeled instances, with the $2\hat{\alpha}$ highest probabilities of being positive, yields a reduced dataset $\tilde{X}_U$, which is either too small, or too similar the original set of unlabeled instances. In these cases, a different fraction than the top $2\hat{\alpha}$ likely positive instance should be

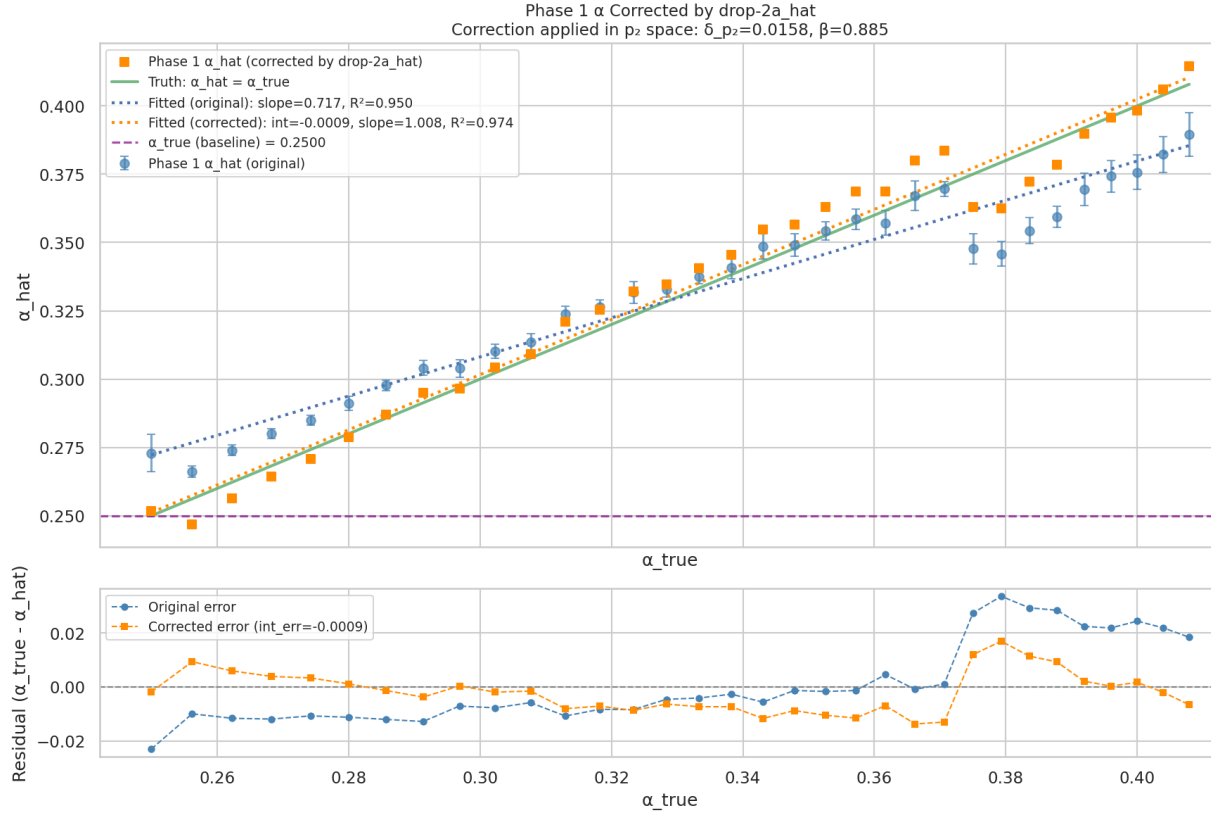


Figure 2. Bias correction of synthetic data. Illustration of bias correction for PULSCAR with synthetic data. Data were created with true $\alpha=0.25$. PULSCAR was used to estimate α with increasing fractions of positives spiked in, with the ground truth shown as the solid green line. The blue estimates show error bars for 40 independent alpha estimates, and the orange shows the results after the new bias correction algorithm was applied. Importantly, the orange bias-corrected y-intercept, $\alpha=0.2521$ is almost exactly the true α .

removed from the original data in the second step of the bias correction, i.e., when constructing $\tilde{X}$. Furthermore, the range of values x in the regression-step of the bias correction has to be carefully balanced to guarantee that the classifier can be trained on a balanced data set (cf. Table 1).

α_{true}	α_{orig}	α_{corr}	err_orig	err_corr	improv %
0.01	0.0154	0.0107	0.0054	0.0007	0.862
0.02	0.0239	0.02	0.0039	0	0.994
0.05	0.0617	0.0557	0.0117	0.0057	0.513
0.1	0.11	0.1028	0.01	0.0028	0.717
0.15	0.1677	0.1546	0.0177	0.0046	0.739
0.2	0.2156	0.2012	0.0156	0.0012	0.924
0.25	0.2732	0.2521	0.0232	0.0021	0.91
0.3	0.3301	0.3042	0.0301	0.0042	0.862
0.35	0.363	0.3358	0.013	0.0142	-0.095

Table 1: Bias correction applied to PULSCAR. We observe over a large range of α values 0.01-0.30 that the bias correction improvement ranges from 51%-99%. The performance drops at 0.35 where too many probable positives are dropped.

Conclusion

The proposed bias correction performs well on synthetic data and benchmark datasets. It is also informative on the underlying bias of PU learning classifiers. However, the bias correction can only be applied to classifiers which provide posterior probabilities for unlabeled instances of being positive. Furthermore, the method is only applicable to dataset which are “reasonably balanced”) with regard to the number of labeled and unlabeled instances (i.e., the data must contain enough labelled and unlabeled instances. The method is computationally costly and is not feasible if running time for single estimates is too long. Besides these limitations, the agreement of theoretical predictions and observations in the selected data sets suggest that the proposed method is adequate for ascertaining the performance of PU learning algorithms, when applied to empirical data sets for which the status of unlabeled instances cannot be assessed, such as in the case of depression, bipolar disorder, or PTSD.

References

1. Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *J Am Med Inform Assoc*. 2018 Oct 1;25(10):1419-1428. doi: 10.1093/jamia/ocy068. PMID: 29893864; PMCID: PMC6188527.
2. Kumar P, Vizsolay AD, Upadhayaya R, Moomtaheen F, Greer DR, Bologa CG, Schneider KA, Davis SE, Matheny ME, van der Goes D, Villarreal G, Zhu Y, Tohen M, Malec SA, Yang JJ, Fielstein EM, Lambert CG. Detecting Uncoded Self-Harm in Veterans' Electronic Health Records Using Positive and Unlabeled Learning: Retrospective Cohort Study. *J Med Internet Res*. 2026 Jun 4;28:e89071. doi: 10.2196/89071. PMID: 42241701; PMCID: PMC13235979.
3. Kumar P, Moomtaheen F, Malec SA, Yang JJ, Bologa CG, Schneider KA, Zhu Y, Tohen M, Villarreal G, Perkins DJ, Fielstein EM, Davis SE, Matheny ME, Lambert CG. Detecting opioid use disorder in health claims data with positive Unlabeled learning. *IEEE J Biomed Health Inform [Internet]*. Institute of Electrical and Electronics Engineers (IEEE); 2025 Feb;29(2):750–757. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC11971012/> PMCID: PMC11971012
4. Wimalawansa S. Post-traumatic stress disorder: an under-diagnosed and under-treated entity. *Compr Res J Med Med Sci*. 2013;1(1):1-2.
5. Ivanov D. Dedpul: Difference-of-estimated-densities-based positive-unlabeled learning. In 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA) 2020 Dec 14 (pp. 782-790). IEEE.
6. Bekker, J., & Davis, J. (2018). Estimating the Class Prior in Positive and Unlabeled Data Through Decision Tree Induction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1). <https://doi.org/10.1609/aaai.v32i1.11715>
7. 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* 2008 Aug 24 (pp. 213-220).
8. Kumar P, Lambert CG. 2024. Positive Unlabeled Learning Selected Not At Random (PULSNAR): class proportion estimation without the selected completely at random assumption. *PeerJ Computer Science* 10:e2451 <https://doi.org/10.7717/peerj-cs.2451>
9. Johnson, B. (2013). Wilt [Dataset]. UCI Machine Learning Repository. <https://doi.org/10.24432/C5KS4M>.
10. Bock, R. (2007). MAGIC Gamma Telescope. UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C52C8B>.
11. Higuera, C., Gardiner, K., and Cios, K. (2015). Mice Protein Expression. UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C50S3Z.570>