# PULSNAR: Positive Unlabeled Learning Selected Not At Random -- towards imputing undocumented conditions in EHRs and estimating their incidence

**Praveen Kumar[2]; Sharon E. Davis[1]; Michael E. Matheny[1,3]; Gerardo Villarreal[2,4]; Yiliang Zhu[2]; Mauricio Tohen[2]; Douglas J. Perkins[2]; Christophe G. Lambert[2]**

**[1]Vanderbilt University Medical Center, Nashville;[2] University of New Mexico, Albuquerque, NM; [3]VA Tennessee Valley Healthcare System, Nashville, TN; [4]VA New Mexico Healthcare System, Albuquerque, NM**

## Introduction

Deriving high-quality evidence from electronic health records (EHRs) is compromised by gaps between what is documented versus the true patient conditions, particularly in mental health.[1–3] *Noisy label learning* can rank order patients by the probability of uncoded or undiagnosed MH conditions,[4] but it has remained an unsolved problem to calibrate the predictions to the true disease incidence, absent a large representative sample of people who have been clinically assessed as *both* positive and negative. The algorithm we presented at the 2021 OHDSI Symposium showed promise for self-harm and PTSD imputation and estimation of the true fraction (α) of positives among imperfectly coded/diagnosed patients.[3] However, performance suffered when the *selected completely at random assumption* (SCAR)[5] did not hold, i.e., coded positives must be representative of the uncoded/undetected positives, which is unlikely in healthcare data (e.g. milder cases more likely undiagnosed). We introduce a new *positive-unlabeled (PU) learning algorithm*, PULSNAR, that can estimate α among patients with uncoded or undiagnosed conditions without "gold standard" assessment of positives and negatives even when the SCAR assumption fails. This technique opens up the possibility of estimating the incidence of undiagnosed conditions to inform public health, guide screening efforts for poorly captured conditions, and identify health equity issues where coding levels differ by sociodemographics.[6] We describe the method and our procedure for generating non-SCAR simulated data, then demonstrate that PULSNAR outperforms state-of-the-art algorithms.

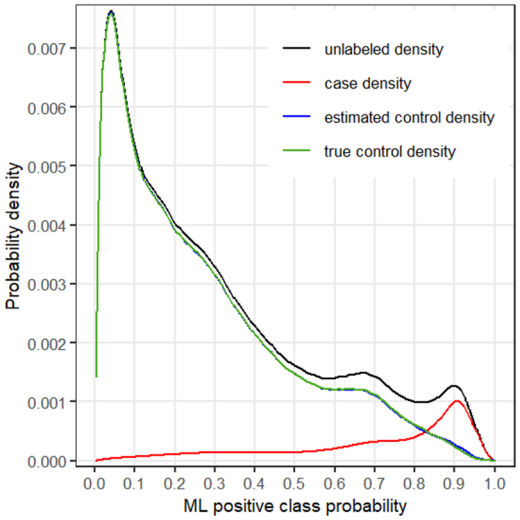


*Figure 1 Kernel density estimates for simulated data (SCAR) with α=10% cases in the unlabeled set.*

## Methods

We developed two new PU learning algorithms to estimate the proportion of cases (positives) among unlabeled samples. The first, PULSCAR, makes the SCAR assumption and is used as a subroutine in the second algorithm, PULSNAR, which overcomes the SCAR assumption.

**PU-Learning Selected Completely At Random (PULSCAR) algorithm**: Given any ML algorithm, A(x), that generates [0-1] probabilities to differentiate between unlabeled and positives using covariates x, let $f_p(x)$, $f_n(x)$, and $f_u(x)$ be probability density functions (PDFs) corresponding to cases, controls, and unlabeled distributions of A(x), respectively (**Figure 1**). Let α be the unknown proportion of cases among the unlabeled, then $f_u(x) \equiv \alpha f_p(x) + (1-\alpha) f_n(x)$. Our proposed PU learning method uses beta kernel estimates of the PDF of $f_p(x)$ and $f_u(x)$ to estimate α. A key observation is that $\alpha f_p(x)$ cannot exceed $f_u(x)$ anywhere, lest the PDF $f_n(x)$ have negative probabilities. We estimate α, by finding where the finite-difference slope of our error function ε(α) changes maximally: $\varepsilon(\alpha)=\log(\min(|f_u(x)-\alpha f_p(x)|))$. Our algorithm is similar to DEDPUL,[7] a PU-learning technique that also uses density estimates and makes the SCAR assumption, but differs in approach to density estimation and α estimation.

**PU-Learning Selected Not At Random (PULSNAR) Algorithm**: A key innovation was to first identify more homogenous subtypes of positives using unsupervised clustering and then to estimate the proportion of

each subtype among unlabeled observations using the PULSCAR algorithm, see **Figure 2**.

**Probability calibration:** The PDF of the coded and uncoded positives should follow the same pattern for SCAR data. Using this feature, we identify uncoded positives among the unlabeled examples. Then, the isotonic or sigmoid calibration method is applied to the data with coded and imputed positives to get calibrated probabilities.

**Evaluation of methods:** We tested Cleanlab,[8] DEDPUL with CatBoost, our PULSCAR method with XGBoost ML, and PULSNAR. Our simulations were as follows: a) used sklearn make_classification() with class_sep=0.3 (a difficult classification task) to generate 6 classes, defining one class as negative and the other 5 as positive; b) the labeled positives were assigned in equal proportions, but the unlabeled positives were assigned to be markedly non-SCAR with the 5 types comprising 1/31, 2/31, 4/31, 8/31, and 16/31 of the positives in the unlabeled data respectively; c) negatives were added to the unlabeled set to create different proportions for 5 datasets with α ranging from 1% to 50% positive among the unlabeled; d) the ratio of unlabeled samples to labeled positives was set to 10:1 for class imbalance. To create confidence intervals, the ML models were trained and tested with 20 iterations of 5-fold cross-validation. A simulation was also performed to assess the performance of the algorithms when the SCAR assumption holds, where the types (step b) had equal ratios. The PULSNAR algorithm was evaluated on VHA self-harm and PTSD data.

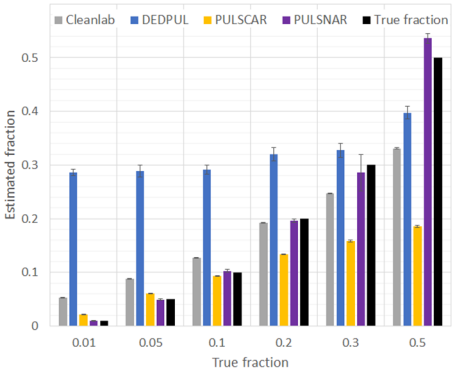


***Figure 2: Comparison of PU-Learning algorithms on non-SCAR data.*** *The 3 algorithms that make the SCAR assumption perform poorly at estimating the true fraction of positives (α) among unlabeled observations, but our new PULSNAR algorithm (purple) obtains close to the true answer over a broad range of α values, with no apparent systematic bias. Error bars are 2 standard deviations.*

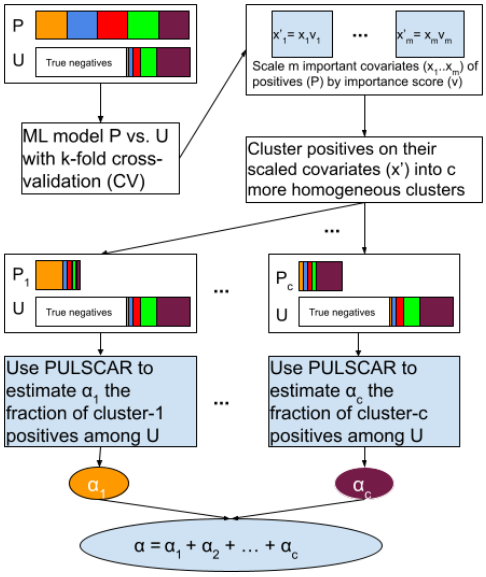


***Figure 3: Schematic of PULSNAR algorithm****. A ML model is trained and tested with 5-fold CV on all positive and unlabeled examples. The important covariates that the model used are scaled by their importance value. Positives are divided into c clusters using the scaled important covariates. c ML models are trained and tested with 5-fold CV on the records from a cluster and all unlabeled records. We estimate the proportions ($\alpha_1 \ldots \alpha_c$) of each subtype of positives in the unlabeled samples using PULSCAR. The sum of those estimates gives the overall fraction of positive samples in the unlabeled set. P = positive examples, U= Unlabeled examples.*

## Results

In **Figure 3,** the results of random simulations are shown, demonstrating that the first 3 SCAR-assuming algorithms fail to estimate α accurately with non-SCAR data, but that PULSNAR closely estimates the true α. On SCAR data (not shown) Cleanlab performed poorly, DEDPUL performed better, and both PULSCAR and PULSNAR were extremely close to the true answer. DEDPUL struggled when the case fraction was low, as is true for many of our MH phenotypes. The performance of both Cleanlab and DEDPUL deteriorates if the dataset has a substantial class imbalance; which was not observed for either of our algorithms.

## Conclusion

Our PU learning algorithms outperformed existing methods on data with SCAR holding or failing. PULSNAR has the potential to estimate bounds on the incidence of under-coded conditions without time-consuming chart review, and generate highly calibrated classification models to support screening of undiagnosed patients. Future efforts will quantify performance on real-world healthcare data where the SCAR assumption cannot be made.

## References


1. Kumar P, Nestsiarovich A, Nelson SJ, Kerner B, Perkins DJ, Lambert CG. Imputation and characterization of uncoded self-harm in major mental illness using machine learning. J Am Med Inform Assoc. 2020 Jan 1;27(1):136–146. PMID: 31651956
2. Nestsiarovich A, Kumar P, Lauve NR, Hurwitz NG, Mazurie AJ, Cannon DC, Zhu Y, Nelson SJ, Crisanti AS, Kerner B, Tohen M, Perkins DJ, Lambert CG. Using Machine Learning Imputed Outcomes to Assess Drug-Dependent Risk of Self-Harm in Patients with Bipolar Disorder: A Comparative Effectiveness Study. JMIR Ment Health. 2021 Apr 21;8(4):e24522. PMID: 33688834
3. Kumar P, Lauve NR, Davis SE, Parr SK, Park D, Matheny ME, Villarreal G, Uhl G, Zhu Y, Tohen M, Perkins DJ, Lambert CG. Detecting PTSD and self-harm among US Veterans using positive unlabeled Learning. OHDSI 2021 Global Symposium [Internet]. Observational Health Data Sciences and Informatics; 2021. Available from: https://www.ohdsi.org/2021-global-symposium-showcase-103/
4. Agarwal V, Podchiyska T, Banda JM, Goel V, Leung TI, Minty EP, Sweeney TE, Gyang E, Shah NH. Learning statistical models of phenotypes using noisy labeled training data. J Am Med Inform Assoc. 2016 Nov;23(6):1166–1173. PMCID: PMC5070523
5. Elkan C, Noto K. Learning classifiers from only positive and unlabeled data. Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, NY, USA: Association for Computing Machinery; 2008. p. 213–220.
6. Davis SE, Kumar P, Lauve NR, Parr SK, Park D, Matheny ME, Villarreal G, Uhl G, Zhu Y, Tohen M, Perkins DJ, Lambert CG. Disparities in Coded and Imputed Post-Traumatic Stress Disorder and Self-Harm Among US Veterans. AMIA 2021 Annual Symposium. 2021.
7. Ivanov D. DEDPUL: Difference-of-Estimated-Densities-based Positive-Unlabeled Learning. 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA). ieeexplore.ieee.org; 2020. p. 782–790.
8. Northcutt C, Jiang L, Chuang I. Confident Learning: Estimating Uncertainty in Dataset Labels. J Artif Intell Res. jair.org; 2021 Apr 14;70:1373–1411.