Positive unlabeled learning imputation of undiagnosed and uncoded PTSD and self-harm among US Veterans

Praveen Kumar, MS[1,2]; Nicolas R. Lauve, MS[1,2]; Sharon E. Davis, PhD[3]; Sharidan K. Parr, MD[3]; Daniel Park, MS[3]; Michael E. Matheny, MD, PhD[3,4]; Gerardo Villarreal, MD[1,6]; George Uhl, MD[1]; Yiliang Zhu, PhD[1]; Mauricio Tohen, MD, DrPH, MBA[1]; Douglas J. Perkins, PhD[1]; Christophe G. Lambert, PhD[1,2,5*]

[1]Center for Global Health, Department of Internal Medicine, University of New Mexico Health Sciences Center, Albuquerque, New Mexico, USA; [2]Department of Computer Science, University of New Mexico, Albuquerque, New Mexico, USA; [3]Vanderbilt University Medical Center, Nashville, USA; [4]VA Tennessee Valley Healthcare System, Nashville, TN, USA; [5]Translational Informatics Division, Department of Internal Medicine, University of New Mexico Health Sciences Center, Albuquerque, New Mexico, USA; [6]VA New Mexico Healthcare System, Albuquerque, New Mexico, USA

*Corresponding author's email: cglambert[at]unm.edu.

Background

Noisy label machine learning (ML) can be used to rank-order patients by the probability of mental health conditions.[1-3] However, due to under-coding and under-diagnosis, it is hard to calibrate a probability threshold for a desired positive predictive value (or other ML metrics) without a large representative "gold-standard" sample of people who have been clinically assessed as both positive and negative for a given condition. Mental health phenotypes are under-reported in structured electronic health record (EHR) data and administrative claims data[4], limiting OHDSI cohort characterization, comparative effectiveness research (CER)[2], and patient-level prediction. We innovate and evaluate a positive and unlabeled learning (PU-learning) algorithm for estimating the true proportion of positives among patients with uncoded or undiagnosed post-traumatic stress disorder (PTSD) and self-harm.

Methods

Data sources: Simulated data. Veterans Health Administration (VHA) database mapped to OMOP common data model (v5.3).

1. Simulated data imputation (Figure 1):

- Data comprised 100,000 positives (label 1) and 100,000 unlabeled examples (label 0) with different fractions of positives among the unlabeled (1%, 5%, 10%, 20%, 30%) and 250 covariates.
- Data were generated using the scikit-learn make_classification() function with class_sep=0.3 (a difficult classification task).[5]
- XGboost[6] ML models were trained and tested with 20 iterations of 5-fold cross-validation.
- Each iteration's ML predicted probabilities were input to a leading PU method, CleanLab,[7] and our PU algorithm to estimate the true fraction of positives in the unlabeled set, with results compared against the known ground truth.

2. VHA PTSD data imputation. 255,643 coded PTSD cases and 934,754 controls observed ≥ 2 years during 2000-2020, with the last year blinded to assess "future" diagnostic conversion.

PTSD covariates: All condition and observation concepts (OMOP-mapped) present in the covariate windows of cases and controls (Figure 2)

ML approach and evaluation:

- Random 50% of data were used to train a model, applied to 50% test. Then a second model was built with test and train swapped.
- Given class imbalance (#controls>#cases), we built and averaged models over floor(#controls/#cases) balanced case/control sets.
- The probabilities from the models were used in the PU algorithm (below) to estimate the fraction of controls with PTSD.
- We assessed the distribution of predictions for the individuals who converted to PTSD in the final year versus those who did not.
- We reviewed charts of 50 probable but uncoded PTSD cases.

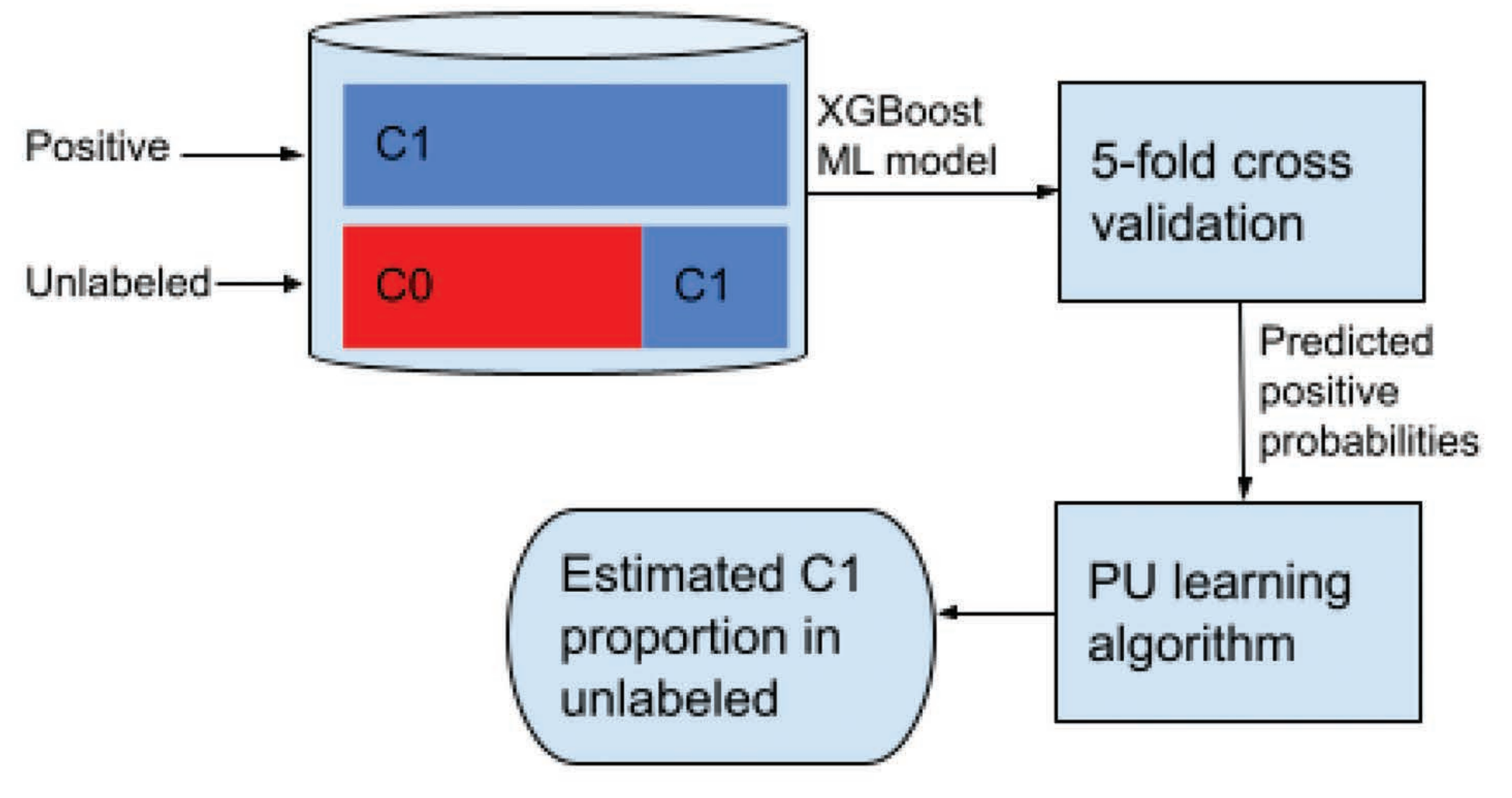

Figure 1. PU learning method on simulated data. C1= class 1 data, C0= class 0 data. Data were generated using difficult class separability (class_sep=0.3). For a given simulation with a given fraction of positives in unlabeled, 5-fold cross validation was executed 20 times. PU learning was applied on predicted probabilities from each iteration.

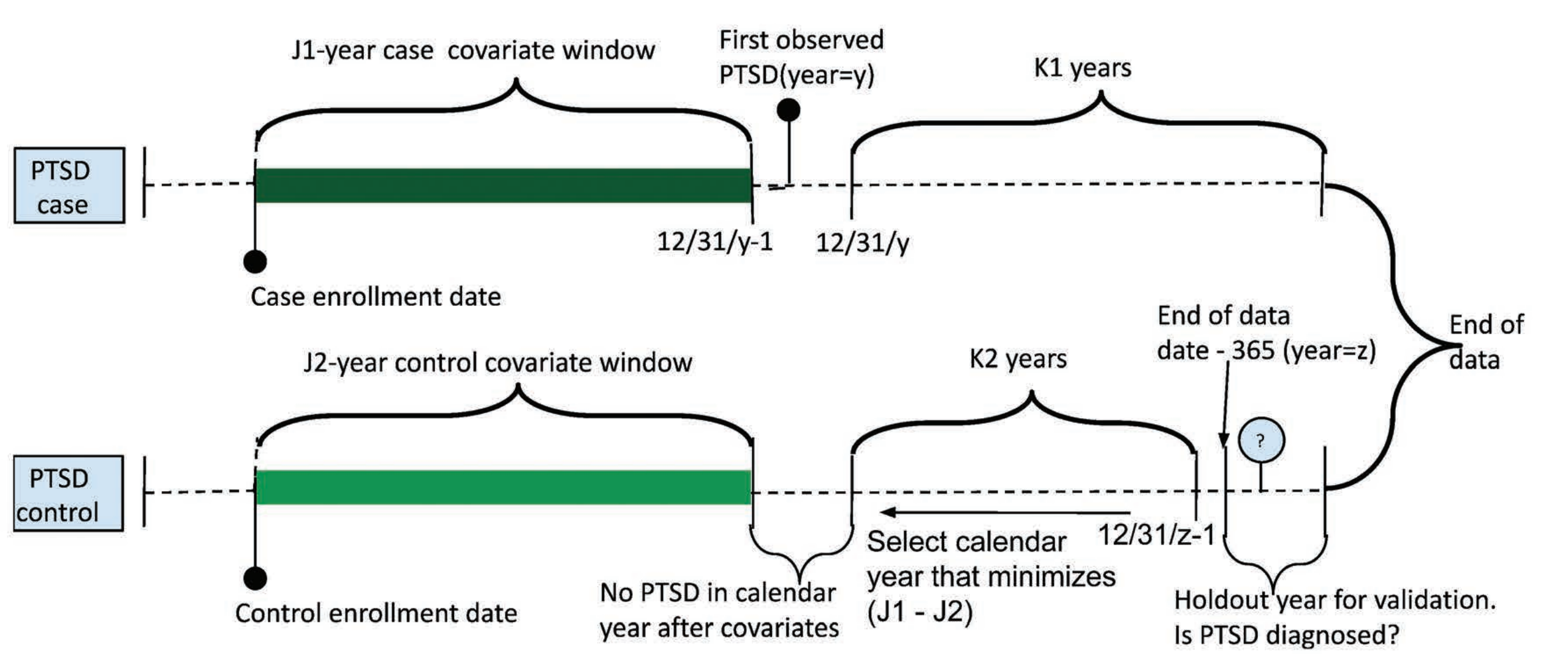
J1-year case covariate window
First observed PTSD(year=y)
K1 years
PTSD case
12/31/y-1
12/31/y
Case enrollment date
End of data date - 365 (year=z)
End of data
J2-year control covariate window
K2 years
PTSD control
?
Select calendar year that minimizes (J1 - J2)
12/31/z-1
Control enrollment date
No PTSD in calendar year after covariates
Holdout year for validation. Is PTSD diagnosed?

Figure 2. Time window approach to matching PTSD cases to controls. Good matches have: a) close enrollment dates; b) similar covariate windows; and c) minimize dropped control years, K2. The last year's data of all samples is blinded to allow future assessment of diagnostic conversion. All concepts (conditions and observations) present during J1 and J2 years are used as ML covariates.

3. VHA Self-harm data imputation:

- 36,962 coded self-harm cases (coded self-harm actions) and 2,621,278 controls (no history of self-harm coded)

Self-harm covariates: All condition, observation, and procedure covariates over the entire period of patient observation minus the last year.

ML approach and evaluation:

- Given class imbalance (#controls>#cases), we built and averaged models over floor(#controls/#cases) balanced case/control sets.
- The probabilities from the models were used in the PU algorithm (below) to estimate the fraction of controls with self-harm.
- We reviewed charts of 50 individuals with probable but uncoded self-harm.

PU learning algorithm explained:

- Let $f_p(x)$, $f_n(x)$, and $f_u(x)$ be probability density functions (PDFs) corresponding to positives, negatives, and unlabeled samples' distributions, respectively (Figure 3A). Let α be the proportion of positives in the unlabeled distribution, then $\mathbf{f_u(x) \equiv \alpha f_p(x) + (1 - \alpha) f_n(x)}$.
- We estimate α, by finding where the finite-difference estimated slope of our error function ε(α) changes maximally: $\boldsymbol{\varepsilon(\alpha)=\log(\min(|f_u(x)-\alpha f_p(x)|))}$, (Figure 3B, 3C).
- The approach makes the selected completely at random (SCAR) assumption[8].

Results

PU learning algorithm estimated true class prior in simulated data that satisfies the SCAR assumption:

- Figure 3A shows that the estimated (10.68%) and true control density (10%) are almost identical when the error function (Figure 3B) was minimized.
- Our PU-learning estimates were extremely close to the truth (Figure 3C). The truth fell within the error bars, unlike the biased cleanlab estimates.

The PU learning algorithm worked on the VHA PTSD data where the SCAR assumption approximately holds:

- The PTSD model had positive predictive value=0.53, sensitivity=0.32, and specificity=0.92.
- 2.4% of those negative for PTSD were subsequently diagnosed in the holdout year. The probability of PTSD among these was similar to that of coded cases (mean 0.55 vs. 0.61) and higher than those who were not diagnosed in the holdout year (mean 0.39).
- Chart review of 50 probable patient cases without PTSD structured codes showed 18 with positive screens (with 3 subsequently diagnosed) and 32 had low evidence of PTSD.
- Figure 3D shows that the PU algorithm estimated that 7.55% of veterans without diagnosed PTSD have the condition.
- Top 10 important covariates: Age, Depressive disorder, Mood disorder, any mental disorder, anxiety, psychiatric care, adjustment disorder, abuse especially sexual, active duty injury, sleep disorder.

The SCAR assumption fails in self-harm data, and the PU learning algorithm underestimated uncoded self-harm.

- Chart review of a random 50 VHA patients with ML-imputed, but not coded self-harm, revealed 47 (94%) had clear evidence in their notes of suicide attempts and/or self-harm.
- The PU method estimated 0.43% of the uncoded patients had self-harm, which appears to be an underestimate (Figure 4).

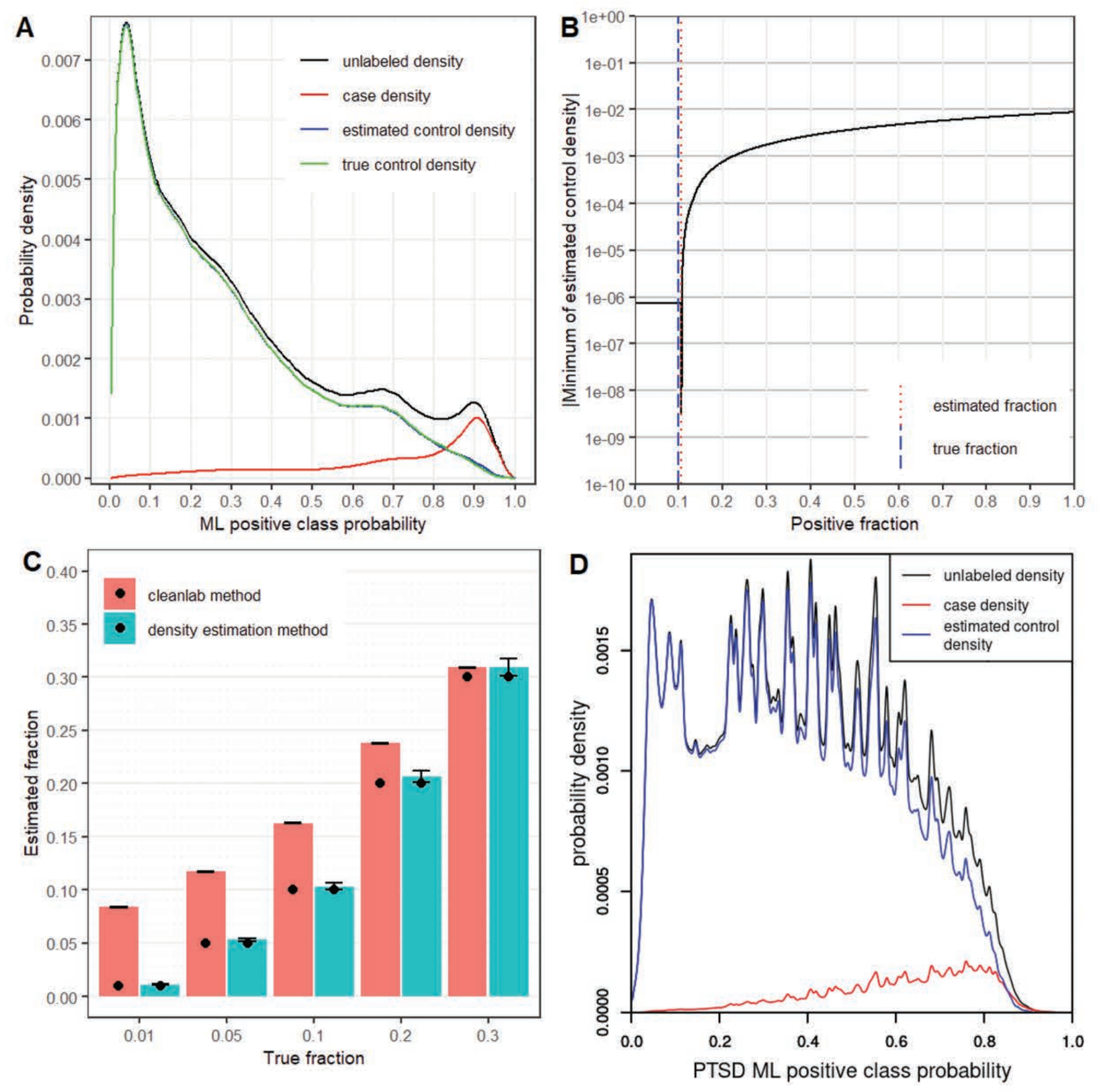

Figure 3: PU-learning evaluated in simulated data and VHA PTSD. A) kernel density estimates for simulated data with α=10% cases in the unlabeled set. B) Error function with α selected where the change is largest - close to α=10%. C) Comparison of CleanLab and our PU method for α=0.01-0.30; the bar represents the mean value of the estimated fraction, with 95% confidence intervals for estimated α, and black dot at true α. D) PU method applied to VHA data estimates 7.55% of veterans without diagnosed PTSD have the condition.

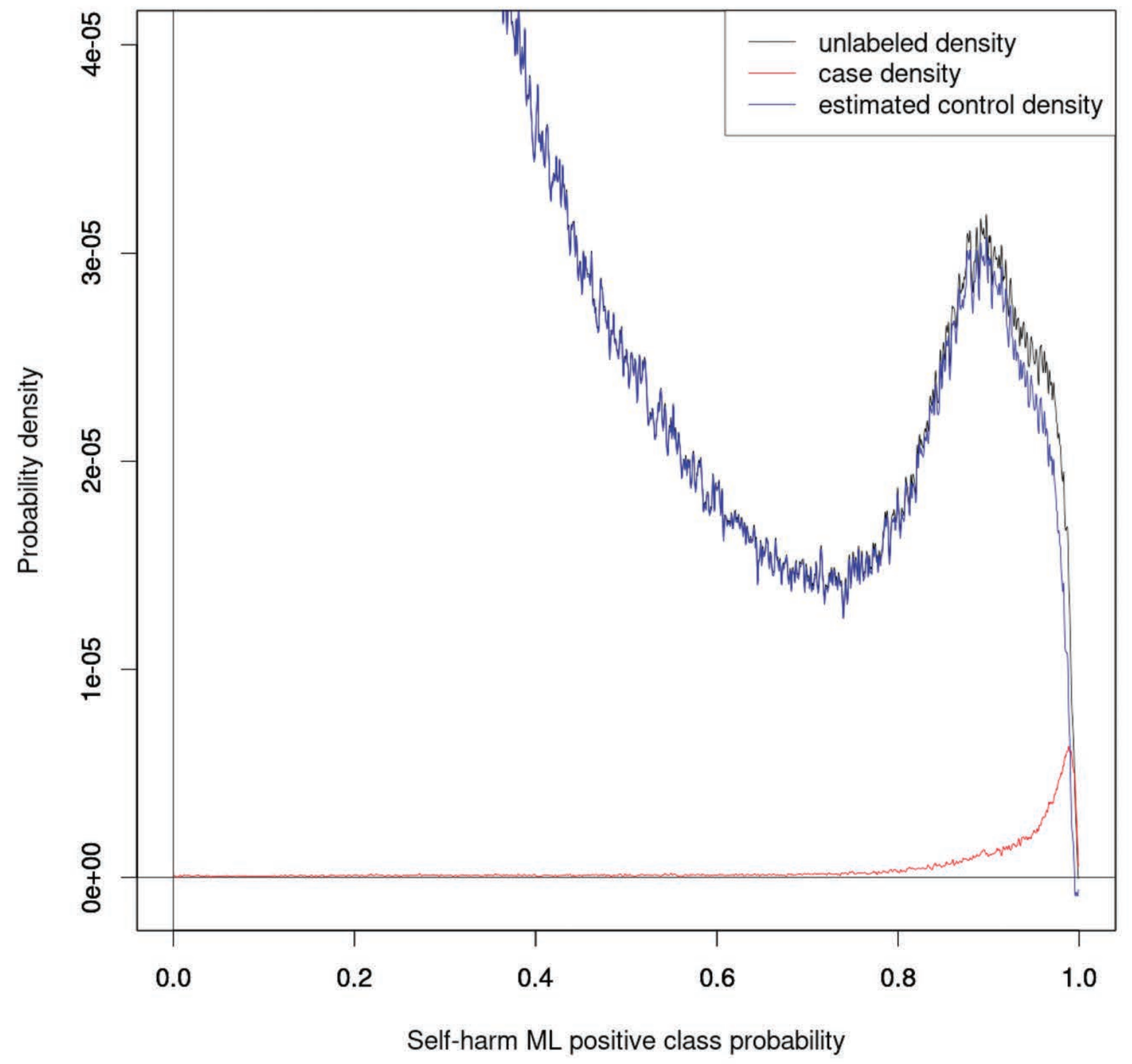

Figure 4: PU-learning evaluated in VHA self-harm estimates that 0.43% of veterans have uncoded self-harm. After subtracting the positive density (red) from the unlabeled density (black) to estimate the control density (blue), a large density remains above 0.8 among the controls. This, combined with chart review suggests that the SCAR assumption fails. That is, the coded positives are not representative of all positives: α=0.43% is an underestimate

Conclusions

- The PU method has the potential to estimate bounds on the incidence of under-coded conditions without time-consuming chart review, calibrating efforts for screening persons with undiagnosed conditions, and enhancing the statistical power of CER through inferring missing phenotypes.
- Our PU-learning method has outstanding performance on simulated data when the SCAR assumption is valid.
- For PTSD in the VHA, the SCAR assumption appears to roughly hold, likely due to an annual screening policy.
- For self-harm in VHA, the SCAR assumption appears to not hold, resulting in our method underestimating the uncoded self-harm.
- Further work is needed to address the SCAR assumption not holding.

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. Banda JM, Halpern Y, Sontag D, Shah NH. Electronic phenotyping with APHRODITE and the Observational Health Sciences and Informatics (OHDSI) data network. AMIA Jt Summits Transl Sci Proc. 2017 Jul 26;2017:48–57. PMCID: PMC5543379
4. Bharadwaj, P., Pai, M.M. and Suziedelyte, A., 2017. Mental health stigma. Economics Letters, 159, pp.57-60.
5. sklearn.datasets.make_classification — scikit-learn 0.24.2 documentation [Internet]. [cited 2021 May 27]. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_classification.html
6. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM; 2016. p. 785–794.
7. 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.
8. 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.