Detecting PTSD and self-harm among US Veterans using positive unlabeled learning

Praveen Kumar; Nicolas R. Lauve; Sharon E. Davis; Sharidan K. Parr; Daniel Park; Michael E. Matheny; Gerardo Villarreal; George Uhl; Yiliang Zhu; Mauricio Tohen; Douglas J. Perkins; Christophe G. Lambert

Background

Cohort characterization, comparative effectiveness research (CER), and patient-level prediction are limited by patient conditions being incompletely recorded in structured electronic health records (EHRs) and administrative claims data. This is particularly true of mental health (MH) phenotypes. We recently used noisy label learning on administrative claims data in patients with major mental illness to impute uncoded self-harm,[1] and used it as an outcome to enhance statistical power in a comparative effectiveness study.[2] We estimated that only about 1 in 19 self-harm events were coded, but these estimates require evaluation and validation.

Noisy label machine learning (ML) can rank-order patients by the probability that they might have an MH condition.[1–3] Classical approaches calibrate probability thresholds for desired positive predictive values and other ML metrics using samples of people who have been clinically assessed as *both* positive and negative for a condition. We now report evaluation of methods for estimating the true proportion of positives among patients with uncoded or undiagnosed conditions without such "gold standard" assessments.[3–5] We employ positive and unlabeled learning (*PU-learning)*, a noisy label ML method that uses: a) a set of known positives [e.g. with diagnoses of post-traumatic stress disorder (PTSD)], b) an unlabeled set of individuals with an unknown proportion of positives and negatives, and c) an ML model to distinguish the two. PU-learning then estimates the *class prior* (proportion of positives) among the unknowns. We describe our PU-learning method, assess its performance on simulated data to detect rare and common phenotypes, and use it to detect self-harm and PTSD.

Methods

We received IRB approval (20-H317) to study Veterans Health Administration (VHA) EHR data (N=5.3M), including patient notes. PU methods were tested on simulated data, then applied to VHA PTSD and self-harm.

Simulated data imputation. We simulated a difficult classification task using 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, using sklearn make_classification() with class_sep=0.3.[8] We created confidence intervals on estimation. XGboost[9] ML models were trained and tested with 20 iterations of 5-fold cross-validation on each dataset with different random shuffles. Each iteration's ML predicted probabilities were input to an alternative PU method, CleanLab,[10] and our PU method to estimate the fraction of positives in the unlabeled set.

PTSD data imputation. We selected 255,643 coded PTSD cases and 934,754 controls from the VHA database mapped to the OMOP common data model (v5.3) observed ≥ 2 years during 2000-2020. Data from the last year were blinded to assess "future" diagnostic conversion. Case time windows before the first PTSD diagnosis were matched with controls to have similar enrollment dates and covariate window durations and to minimize dropped control years. Except for PTSD, all OMOP-mapped condition and observation concepts present in the covariate windows of cases and controls were used as ML covariates. An XGBoost model was built with 5-fold cross-validation. The fraction of uncoded/undiagnosed PTSD cases among controls was estimated using our PU method. We also 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 the 50 probable but uncoded PTSD cases.

Self-harm data imputation. We selected 36,962 coded self-harm cases and 2,621,278 controls. We evaluated the presence of condition, observation, and procedure covariates over the entire period of patient observation except for the last year. We employed the same XGBoost and PU method as for PTSD to estimate the fraction of uncoded self-harm. We reviewed charts of 50 individuals with probable but uncoded self-harm to assess whether the patient notes indicated any history of self-harm (including before VHA enrollment).

PU learning method. Our approach has similarities to a recently-described algorithm, DEDPUL.[6] Given an ML algorithm, A(x), that generates [0-1] probabilities to differentiate between positives and unknowns using covariates x, let $f_p(x)$, $f_n(x)$, and $f_u(x)$ be probability density functions (PDFs) corresponding to positives, negatives, and unlabeled distributions of A(x), respectively (Figure 1A). Let α be the proportion of positives in the unlabeled distribution, then $f_u(x) \equiv \alpha f_p(x) + (1 - \alpha) f_n(x)$. Our PU learning method uses 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 estimated slope of our error function ε(α) changes maximally: $\varepsilon(\alpha)=\log(\min(|f_u(x)-\alpha f_p(x)|))$, (Figure 1B, 1C). we note that this approach makes the *selected completely at random* (SCAR) assumption[7] that the coded positives are representative of all positives.

Results

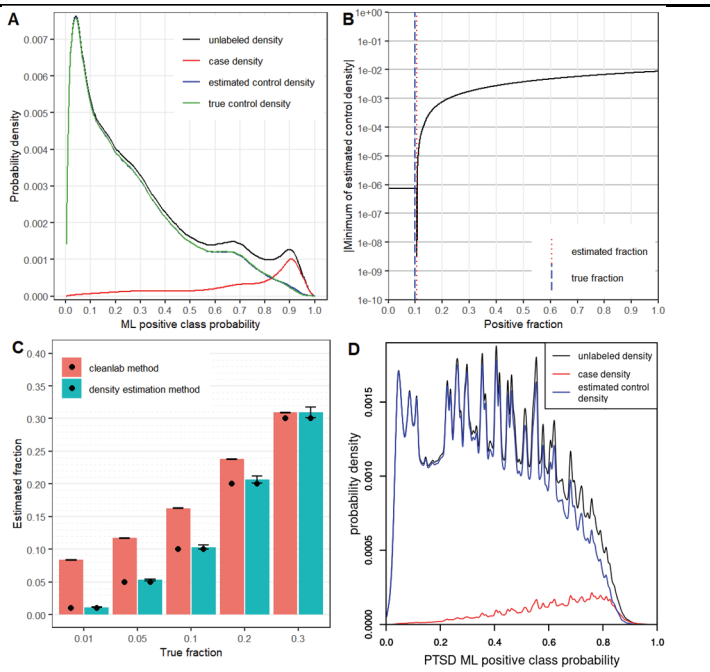

Figure 1: PU-learning evaluated in simulated data and VHA PTSD. A) kernel density estimates for simulated data with α=10% cases in the unlabeled set. B) Error function with α selected where the change is largest - close to α=10%. C) Comparison of CleanLab and our PU method for α=0.01-0.30; the bar represents 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.

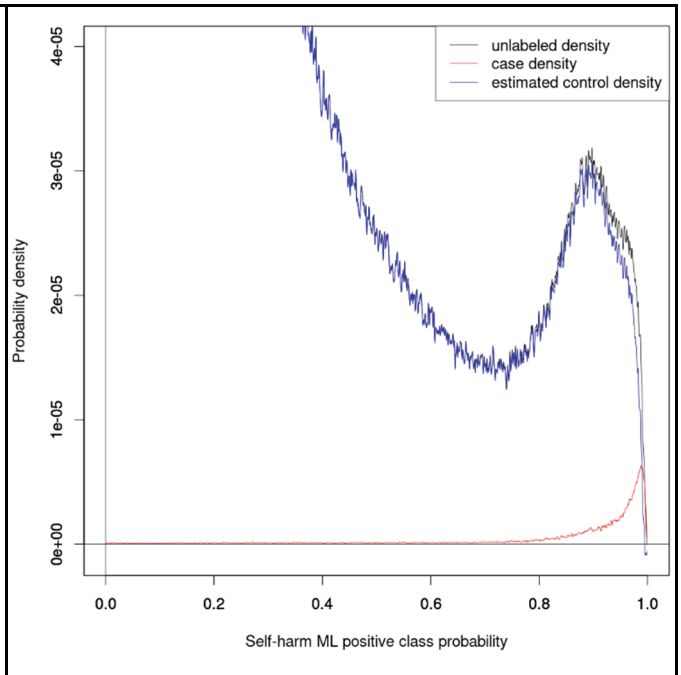

Figure 2: PU-learning evaluated in VHA self-harm estimates that 0.43% of veterans have uncoded self-harm. After subtracting the positive density (red) from the unlabeled density (black) to estimate the control density (blue), a large density remains above 0.8 among the controls. This, combined with chart review suggests that the SCAR assumption fails. That is, the coded positives are not representative of all positives: α=0.43% is an underestimate.

Simulated data. The estimated fraction (0.1068) of positives among the unlabeled cohort computed by our PU method was very close to the true fraction (0.1) (Figure 1B). For α=0.10, the true and estimated control densities follow the same pattern (Figure 1A). Our method was extremely close to the true answers across the full range of positive fractions - with confidence intervals enclosing the ground truth (Figure 1C). CleanLab was heavily biased upwards, performing most poorly when the positive fraction was low.

PTSD. Our PTSD model had *positive predictive value*=0.53, *sensitivity*=0.32, and specificity=0.92. Only 2.4% of those negative for PTSD were 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 (3 subsequently diagnosed) and 32 had low evidence of PTSD. Figure 1D shows that the algorithm estimated that 7.55% of veterans without diagnosed PTSD have the condition.

Self-harm. The PU method estimated 0.43% of the uncoded patients had self-harm, which appears to be an underestimate (Figure 2). Chart review of a random 50 VHA patients with ML-imputed, but not coded self-harm, revealed that 47 (94%) had clear evidence in their notes of suicide attempts and/or self-harm.

Discussion and Conclusion

Our PU-learning method displays outstanding performance in simulated data where the SCAR assumption is known to be valid. The SCAR assumption appears to hold for PTSD in the VHA where routine screening likely reduces observation bias. Despite this screening, our PU-learning estimate suggests that 21.6% of veterans with PTSD had not been diagnosed. For self-harm, on the other hand, the density estimates of positives in Figure 2 suggest that the SCAR assumption does not hold. The feature space of coded positives is different than all positives. The rising right tail in the unlabeled density suggests that our 0.43% estimate is low – it should taper towards zero if the positive (labeled) density was representative of all positives. Chart review of patient notes in the unlabeled observations in this tail showed that 94% had evidence of self-harm, further suggesting that our estimate is too low. Further work is needed to address the failure of the SCAR assumption in this condition. Nevertheless, our method makes advances towards estimating bounds on the incidence of under-coded conditions without time-consuming chart review, calibrating efforts to screen persons for undiagnosed conditions, and enhancing the statistical power of CER through inference about uncoded phenotypes.

References/Citations

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. Halpern Y, Horng S, Choi Y, Sontag D. Electronic medical record phenotyping using the anchor and learn framework. J Am Med Inform Assoc. jamia.oxfordjournals.org; 2016 Jul;23(4):731–740. PMCID: PMC4926745

5. 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

6. 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.

7. 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.

8. 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

9. 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.

10. 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.