Surrogate-assisted Positive Unlabeled Learning on Electronic Health Record (EHR) data

Yiwen Li and Jessica Gronsbell; Department of Statistical Sciences, University of Toronto

Background

Motivating Problem

- Develop a machine learning (ML) model to identify patients with a particular disease based on their electronic health record data (EHR)
- The identified patients can be used for various applications such as disease surveillance and biomedical research

Challenge

- Gold-standard labeled data is obtained for time-consuming review of patient records by medical experts and can't be performed at scale

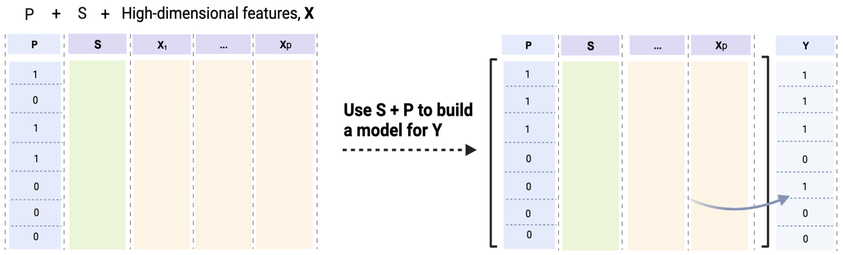
Objective

Develop the ML model without gold-standard labeled data by making use of **readily available:**

- **Positive labels (P)**: Labels that indicate if a patient has the disease with 100% certainty or if it is unclear if the patient has the disease (e.g., presence of a diagnostic billing code)
- **Surrogate labels (S)**: Labels that provide some, but imperfect**,** evidence a patient has the disease (e.g., mentions of the disease in clinical notes)

Surrogate-assisted positive-unlabeled (SAPUL) learning

P + S + High-dimensional features, **X**



Method

Challenges

- Existing positive-unlabeled learning methods can't accommodate surrogate label
- Most existing methods can't handle high-dimensional features

Our Proposal

- Make use of P and S with a 2-step estimation procedure called SAPUL to estimate a sparse logistic regression model for the gold-standard label

$$E(Y = 1 \mid \mathbf{X}, S) = \text{expit}(\beta_0 + \boldsymbol{\beta}_\mathbf{x}\mathbf{X} + \beta_s S)$$

STEP 1: Fit the sparse regression model

$$E(S \mid \mathbf{X}) = \gamma_0 + \boldsymbol{\gamma}_\mathbf{x}\mathbf{X}$$

to obtain estimates of the coefficients, $\hat{\gamma}_0$ and $\hat{\boldsymbol{\gamma}}_\mathbf{x}$.

WHY? When $S \perp \mathbf{X} \mid Y$, $\beta_x = c\,\boldsymbol{\gamma}_\mathbf{x}$ for some $c \neq 0$!

STEP 2: Fit the logistic regression model

$$E(P = 1 \mid \mathbf{X}, S) = \text{expit}(\beta_0 + c\,\boldsymbol{\gamma}_\mathbf{x}\mathbf{X} + \beta_s S)$$

using $\hat{\boldsymbol{\gamma}}_\mathbf{x}$ from step 1.

WHY? When the positive labels are selected completely at random (SCAR), we are guaranteed to recover the coefficients of the model for the gold-standard label!

Simulation Study

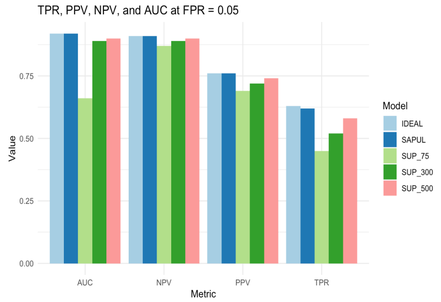
Generated data containing 20,000 examples when (i) SCAR holds or (ii) SCAR is violated with

- P with 0.5% prevalence
- S with AUC = 0.86 for Y
- **X** with p = 150

Compared performance of SAPUL with

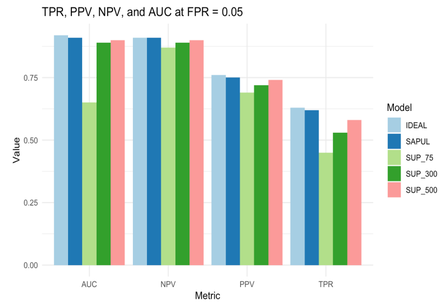
- Standard supervised learning with 75, 300,and 500 gold-standard labels (SUP_75, SUP_300, SUP_500)
- The ideal method with supervised learning with all 20,000 gold-standard labels (IDEAL)

Setting 1: SCAR HOLDS



- SAPUL **performs on par with supervised** learning with 500 labels
- Reliable performance for **high-dimensional** data

Setting 2: SCAR VIOLATED



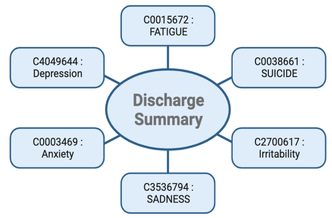
- Under setting 2, the labeling process is not independent of the covariates
- SAPUL is **robust to the SCAR** assumption

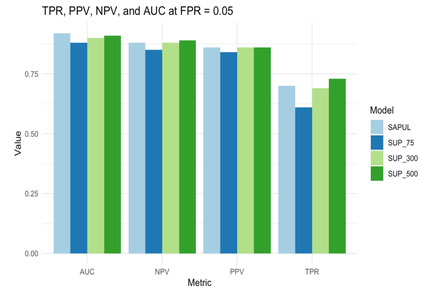
Real EHR Data Analysis

Analyzed EHR data from MIMIC to build a model for depression using 41,798 EHRS with

- Y: Presence of depression from medical chart review of 1,334 EHRs
- P**:** Presence of a depression billing code
- S: Total number of mentions of 'depression' in the discharge summary
- **X:** Contained p = 96 features from:
  - ✓ **Structured data**: Demographics, insurance, hospital stays, prescriptions
  - ✓ **Free-text data**: Clinical concepts related to depression extracted from an existing natural language processing (NLP) methods

Extraction of clinical concepts from discharge summaries





- SAPUL performs **on par with supervised** learning with 500 labels

Conclusions

Summary

- SAPUL performed as well as supervised learning with 500 labels in both real and simulated data
- SAPUL has potential to expedite ML model development

Next steps

- Apply SAPUL to more EHR datasets
- Develop methods to estimate model performance with positive-only labeled data

Acknowledgment

The project '***Surrogate-assisted Positive Unlabeled learning on Electronic Health Record (EHR) data***' is supported by the Data Sciences Institute, University of Toronto

References

[1] Gronsbell, J., Minnier, J., Yu, S., Liao, K., & Cai, T. (2019). Automated feature selection of predictors in electronic medical records data. *Biometrics*, *75*(1), 268–277. https://doi.org/10.1111/biom.12987

[2] Lee, S., Ma, Y., Wei, Y., & Chen, J. (2023). Optimal sampling for positive only electronic health record data. Biometrics. https://doi.org/10.1111/biom.13824

[3] Yu, S., Liao, K. P., Shaw, S. Y., Gainer, V. S., Churchill, S. E., Szolovits, P., Murphy, S. N., Kohane, I. S., & Cai, T. (2015). Toward high-throughput phenotyping: Unbiased automated feature extraction and selection from knowledge sources. *Journal of the American Medical Informatics Association*, *22*(5), 993–1000. https://doi.org/10.1093/jamia/ocv034

Contact

Yiwen Li
Department of Statistical Sciences
Email: even.li@mail.utoronto.ca

Want to know more about me? Scan the QR code to check out my Linkedin profile!

You can download the pdf version of this poster here.