A Two-Stage Modeling Approach for Building and Evaluating Risk Prediction Models Using High-Dimensional Two-Phase Data

Perelman School of Medicine, University of Pennsylvania

Jill Hasler[1], Changcheng Li[2], Runze Li[2] , Ravi Parikh[3,4], Jinbo Chen[1]

University of Pennsylvania, Perelman School of Medicine Departments of Biostatistics[1], Medicine[3], and Medical Ethics[4]; Pennsylvania State University Department of Statistics[2]

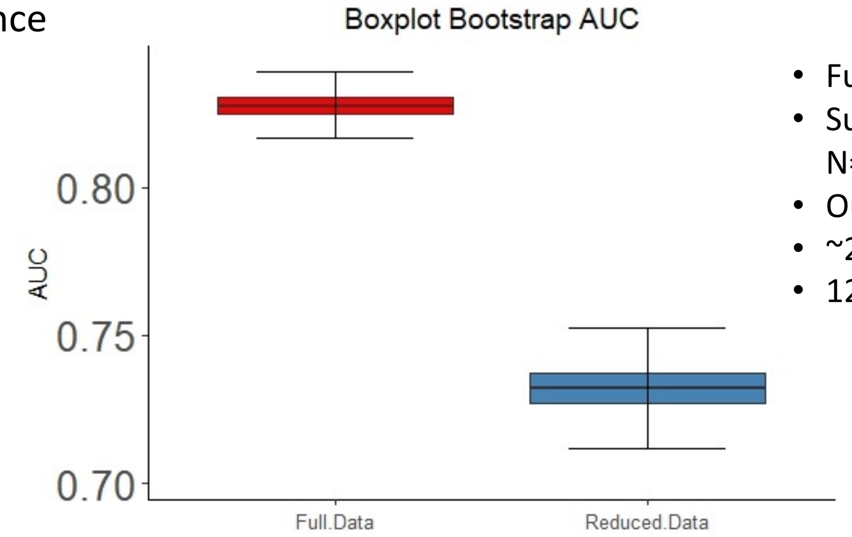
Introduction

- EHRs contain a wealth of information about a patient's health, making them valuable sources of data for building risk prediction models
- External data sources are often available to augment the EHR data
  - Surveys, Biobank, Wearable
- Incorporating external data has potential to improve predictive accuracy of model
- 2 Major challenges:

1. External data available for a subset of patients
2. EHR data often high dimensional

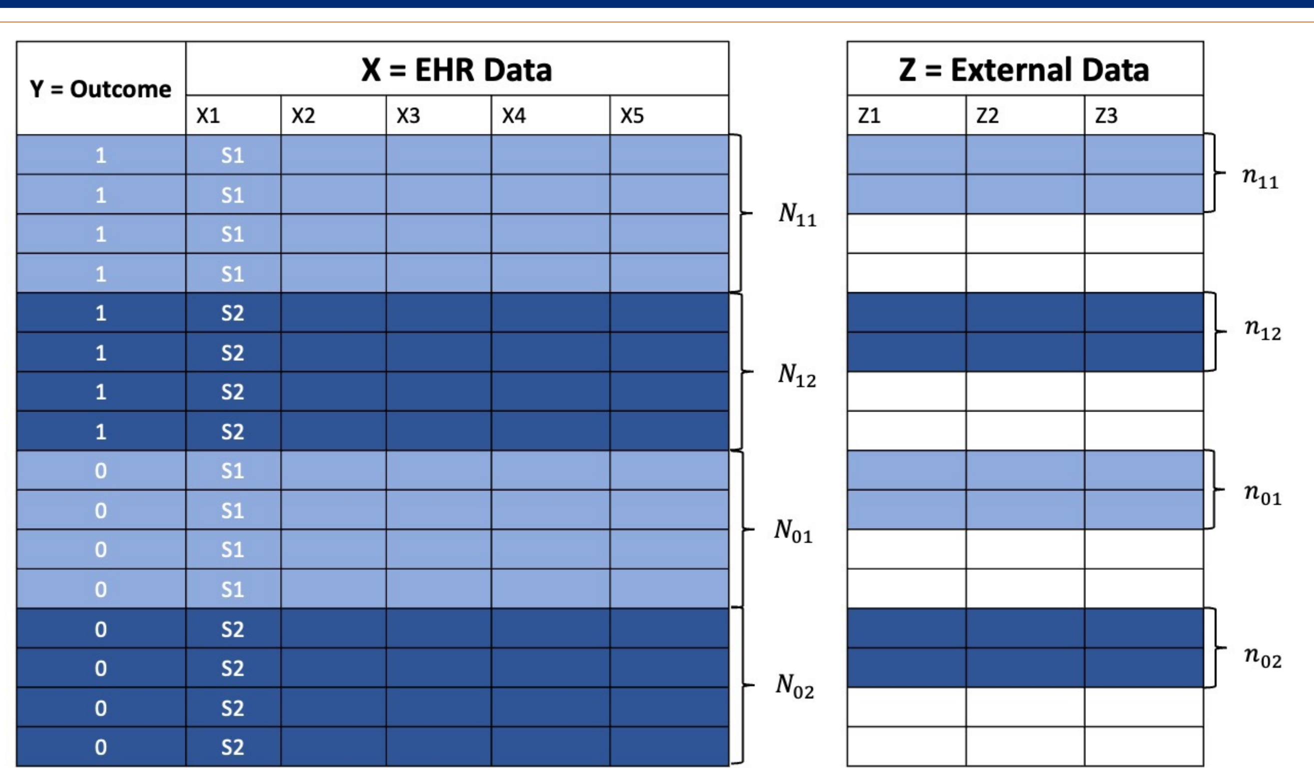
Motivating Example

- Life expectancy in oncology patients overestimated by physicians
- Missed opportunities for advanced care planning and palliative care
- Models built using EHR data can predict patients at risk of short-term mortality
- Patient Reported Outcomes (PROs) commonly collected
  - Not available for all patients
- Goal: Build a model to predict risk of short-term mortality using both EHR and PRO data
- Using only the subset of patients with both EHR and PRO data available impacts model performance



- Full data N=3,795
- Subset with both PRO available N=1,192
- Outcome=180-day mortality (3.4%)
- ~200 EHR variables,
- 12 PRO variables

Two-Phase Sampling Design



- Y: outcome of interest available for all patients
- X: EHR data available for N patients similar to phase I data
- Z: external data available for m patients (m<N) similar to phase II data
- R: indicator of whether or not Z is observed
- $S_1, S_2, \ldots, S_j$: strata defined by X, $P(R=1|X,Y) = N_{ys}/n_{ys}$
- Relationship between full data and subset of data with $R = 1$

$$\operatorname{logit}\{P(Y=1|X,Z,R=1)\} = \operatorname{logit}\{P(Y=1|X,Z)\} + o(X), \qquad o(X) = \log\frac{P(R=1|X,Y=1)}{P(R=1|X,Y=0)}$$

- Existing methods cannot accommodate high dimensionality of X

Our Method

Stage 1

- Use EHR data on to build preliminary model
- High dimensional methods
- Make use of all N patients

Reduce Dimension →

Stage 2

- Fit logistic regression model using one dimensional score and Z
- Model o(X) using more of the data

Step 1: Fit preliminary model between Y and X using EHR data for all N patients

$$\operatorname{logit}\{P^I(Y=1|\boldsymbol{X};\boldsymbol{\zeta})\} = \boldsymbol{\zeta}^T\boldsymbol{X}$$
$$\widehat{Y}_{\boldsymbol{x}}(\zeta) = \hat{\boldsymbol{\zeta}}\boldsymbol{X}$$

$$\operatorname{argmin}\left\{\frac{1}{N}\sum_{i=1}^{N}[\log\{1+\exp(\boldsymbol{X}_i\boldsymbol{\zeta}) - Y_i\boldsymbol{X}_i\boldsymbol{\zeta}\} + \lambda_n\sum_{j=1}^{p}\hat{\boldsymbol{w}}_j\left|\left|\boldsymbol{\zeta}_j\right|\right|_1\right\}$$

Working model defined as : $\operatorname{logit}\{P(Y=1|X,Z)\} = \beta_1^T\widehat{Y}_{\boldsymbol{x}}(\zeta) + \beta_1^T Z$

Step 2: Fit models for missingness indicator, R

Cases: $\operatorname{logit}\{P(R=1|Y=1,\boldsymbol{X};\boldsymbol{\tau_1},\boldsymbol{\zeta})\} = \alpha_{11}\boldsymbol{X_r} + f_1\{P^I(Y=0|\boldsymbol{X};\hat{\boldsymbol{\zeta}})\}$

Controls: $\operatorname{logit}\{P(R=1|Y=0,\boldsymbol{X};\boldsymbol{\tau_0},\boldsymbol{\zeta})\} = \alpha_{01}\boldsymbol{X_r} + f_0\{P^I(Y=1|\boldsymbol{X};\hat{\boldsymbol{\zeta}})\}$

X_r

- Stratum Indicators
- Additional EHR variables with low efficiency

Factors driving availability of data

$f_y\{P^I(Y=(1-y)|\boldsymbol{X};\hat{\zeta})\}$

- f_y smooth function
- P^I predicted from stage 1 model

Complex pattern availability of data

Step 3: Solve Estimating Equation

$$U(\beta,\hat{\boldsymbol{\tau}}_1,\hat{\boldsymbol{\tau}}_0,\hat{\boldsymbol{\zeta}}) = \frac{1}{N}\sum_{i=1}^{N} R_i\left(\hat{Y}_x(\zeta)_i, Z_i^T\right)^T\left[Y_i - \operatorname{expit}\{\beta_1\hat{Y}_x(\zeta)_i + \beta_2 Z_i + o(X)\}\right]=0$$

- Define $p_1(Y_X, Z;\beta)$ as predicted probability from our model
- For a probability threshold ν

$$TPR_\nu(\beta,F) = P\{p_1(Y_X,Z;\beta) > \nu|Y=1\} = \frac{\int I\{p_1(Y_X,Z;\beta)>\nu\}p_1(Y_X,Z;\beta)dF(Y_X,Z)}{p_1(Y_X,Z;\beta)dF(Y_X,Z)}$$

- Let n_{+Y_XZ} denote the number of individuals having $Y_X = y_x$ and $Z = z$ simultaneously

$$\hat{\eta}_{y_{x_i}}, z_i(\hat{\beta},\hat{\tau}) = N^{-1}\frac{n_{+y_xz}}{\sum_{y=0}^{1}P(R=1|Y=y,Y_x;\hat{\tau}_y)P(Y=y|Y_x,Z;\hat{\beta})}$$

$$\hat{F}(y_x,z;\hat{\beta},\hat{\tau}) = \sum_{y_{x_i},z_i:y_{x_i}\le y_x,z_i\le z}\hat{\eta}_{y_{x_i}}, z_i(\hat{\beta},\hat{\tau})$$

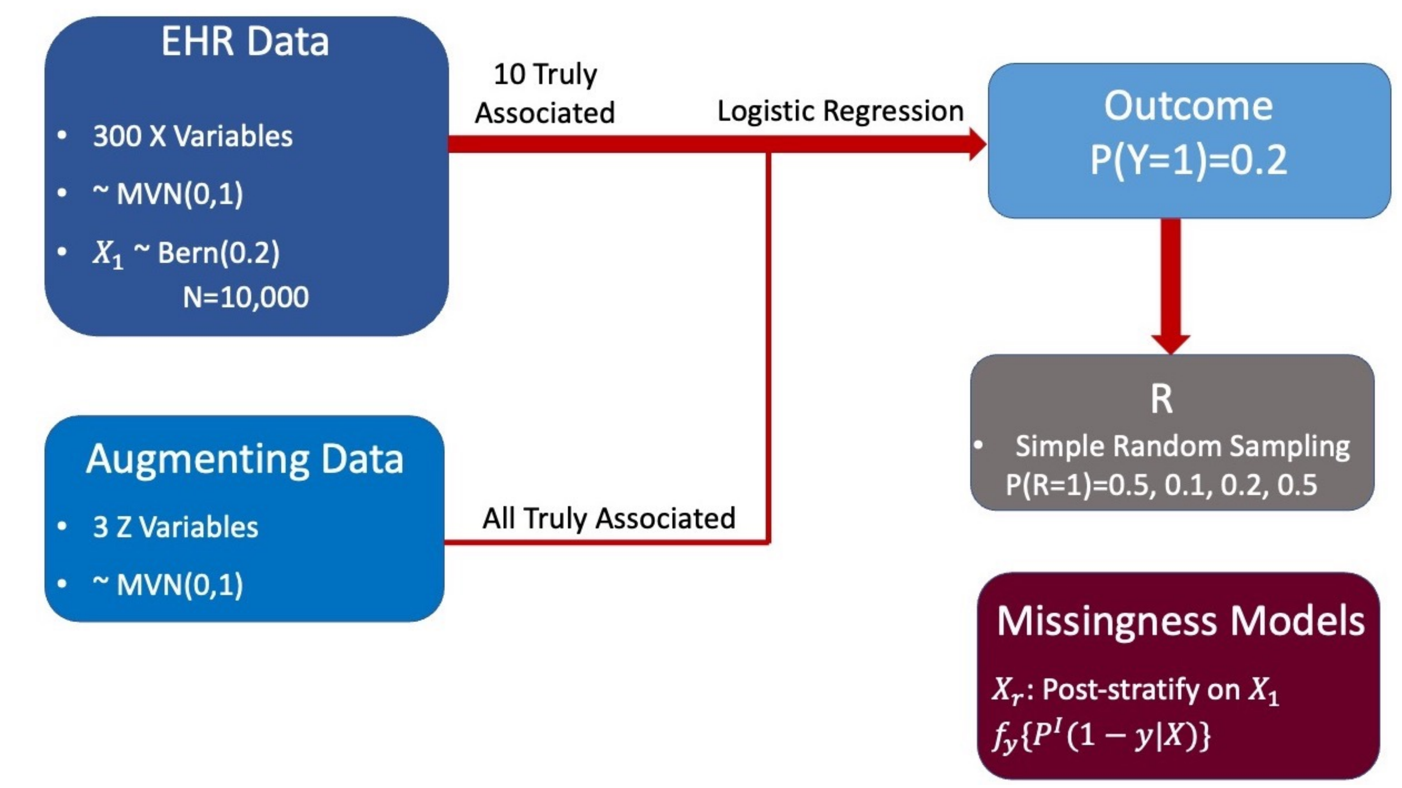
A1. $\frac{m}{N} \to \rho \in (0,1)$

A2. $\hat{\boldsymbol{\zeta}}$ has the oracle property

- Under Conditions A1 and A2

$$\sqrt{N}(\hat{\beta}-\beta) \xrightarrow{d} N(0,\Sigma)$$
$$\sqrt{N}(\widehat{AUC} - AUC) \xrightarrow{d} N(0,\Sigma)$$

Simulation Study



P(R=1)	Method	AUC
0.05	Logistic	0.804 (0.005)
	Ada-Lasso X Only	0.751 (0.009)
	Ada-Lasso R=1 Only	0.748 (0.044)
	Two-Phase	0.838 (0.015/0.016)
0.2	Logistic	0.839 (0.005)
	Ada-Lasso X Only	0.751 (0.009)
	Ada-Lasso R=1 Only	0.784 (0.018)
	Two-Phase	0.837 (0.008/0.008)

Data Analysis

	EHR Only	PRO Only	EHR+PRO
AUC	0.86 (0.85-0.86)	0.75 (0.76-0.79)	0.89 (0.88-0.90)
AUPRC	0.29 (0.25-0.32)	0.19 (0.16-0.22)	0.36 (0.31-0.40)
TPR	0.53 (0.49-0.56)	0.41 (0.35-0.47)	0.61 (0.56-0.65)
FPR	0.06 (0.05-0.07)	0.08 (0.06-0.09)	0.05 (0.04-0.06)

Conclusion and Discussion

- Proposed a pseudo-likelihood method for building risk prediction models using high dimensional two-phase data
- Reduce dimension of high dimensional X to make estimating joint distribution of predictors feasible
- Future Direction
  - Sampling individuals to collect external data
  - Which future patients should have external data collected

References

Breslow, N. E., & Cain, K. C. (1988). Logistic regression for two-stage case-control data. *Biometrika*, *75*(1), 11-20.

Cao, Y., Haneuse, S., Zheng, Y., & Chen, J. (2021). Two-phase stratified sampling and analysis for predicting binary outcomes. *Biostatistics*.

Schnall, J., Li, C., Li, R., Parikh, R., & Chen, J. Efficient Estimation of Prediction Models Using High-Dimensional Two-Phase Data. Under Preparation.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, *101*(476), 1418-1429.