Phenotyping Issues for Exploiting EHRs to Design Clinical Trials

Jinbo Chen, Ph.D.
Professor of Biostatistics
Dept. of Biostatistics, Epidemiology and Informatics
University of Pennsylvania Perelman School of Medicine

April 17, 2019

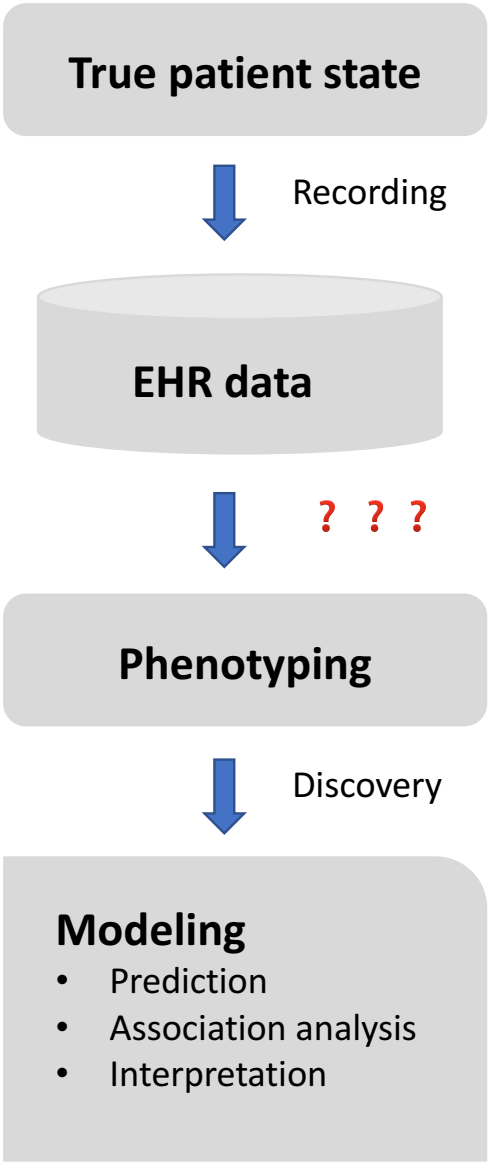
Electronic Health Records and Clinical Trials

- EHRs are expected to play increasingly important roles
  - To generate a list of potentially eligible patients
  - To generate RWE of comparative effectiveness
  - To generate evidence to support initiation of clinical trials
- Accurate EHR phenotyping is essential
  - Study efficiency: representativeness of "eligible" patients
  - Generalizability: high risk?
  - Unbiasedness of the RWE
- Inaccuracy in EHR phenotyping needs to be addressed in statistical analyses

Outline of the Talk

- An anchor-variable framework for EHR-phenotyping
  - Cost effective: minimum effort for chart review
  - High transferability across multiple EHRs
  - Part of student Lingjiao Zhang's dissertation work
- Estimating equation apporaches to correcting bias due to phenotyping inaccuracy
  - Case contamination for EHR-based case-control studies
  - Inaccuracy in cohort identification for EHR-based prospective studies
  - Part of student Lu Wang's dissertation work

EHR Phenotyping



- To identify eligible study subjects from EHR
  - Presence or absence of ICD billing codes
    - Low accuracy
  - Algorithms developed using structured and unstructured data
    - Significant expert involvement
    - Highly iterative process
    - Time-consuming medical chart review
    - Specific to phenotypes
- Need semi-automatic approach to utilizing error-prone EHR information for research

Typical Workflow for EHR-based Phenotyping

- Rule-based algorithms
  - Iterative process based on Clinical experts's knowledge
- Statistical classification methods
  - Identification of a set of "gold standard" cases and controls
  - Extraction of potential predictors from structured data: ICD-9 codes condition of interest, symptoms, comorbidities, common treatments
  - Extraction of useful information from unstructured data via NLP
  - Statitstical modeling: logistic regression, machine learning, AI..

  Model valiation: PPV/NPV; Calibration largely ignored
- Available methods all required annovation of "gold standard" cases and controls
  - Anchor variable framework is an exception (Hapern et al. 2016)

An Anchor Variable Approach to Phenotyping (Our Work Under Review)

- Motivation: Phenotyping primary aldosteronism (PA) with positive-only data
- Our framework: An anchor variable framework
- Our proposed statistical methods
  - Maximum likelihood approach to model development
  - Nonparametric methods for model validation
- Development of a preliminary model for predicting PA
- Conclusion and future work

Motivating Example

- Primary Aldosteronism (PA)
  - PA is the most common cause of secondary hypertension, accounting for 5-10% of hypertensive patients
  - PA can be cured by adrenalectomy or administration of mineralocorticoid receptor antagonists
  - PA has been seriously underdiagnosed
  - To develop a phenotyping model for PA
  - "Positive-only" training data for PA
    - A retrospectively curated database composed of patients with PA referred to UPHS for evaluation (Wachtel et al., 2016)
    - No annotated controls
  - Traditional phenotyping techniques do not apply because of absence of labeled controls

Objectives for Analyzing Positive-Only Data

- Develop a model for predicting phenotype presence
  - Analyzing positive-only data
- Estimate phenotype prevalence
- Validate the trained classifier
  - Calibration
  - Predictive accuracy

The Concept of Anchor Variable

- An anchor is a binary variable summarizing domain expertise on patients' phenotype statuses (Halpern et al., 2014)
- High positive predictive value (PPV)
  - Anchor being positive indicates cases
  - Anchor being negative is non-deterministic of the true phenotype status
- Invariant anchor sensitivity
  - Anchor-positive cases are selected completely at random from all cases
- Example
  - A pathologic diagnosis of cancer
- Upon specification of an anchor variable
  - EHR = Anchor-positive cases + Unlabeled patients

Anchor Variable

- Notation
  - Y: True phenotype status ($Y = 1 : case, Y = 0 : control$)
  - $\mathbf{X}$: A vector of covariates predictive of Y, with density $f(\mathbf{X})$
  - S: Anchor variable ($S = 1 : presence, S = 0 : absence$)
  - q: Phenotype prevalence, $q = p(Y = 1)$
  - h: Anchor prevalence, $h = p(S = 1)$
  - $(\mathbf{X}, Y, S)$: Random variables, with joint distribution $p(\mathbf{X}, Y, S)$
- High PPV
  - $p(Y = 1|S = 1) = 1$
- Conditional independence
  - $p(S = 1|Y = 1, X) = p(S = 1|Y = 1) = c$
  - Bayes rule: $c = h/q$

Likelihood Approach

- Working model
  - logit $p(Y = 1|X) = X^T \beta$
- Likelihood function

$$
\begin{aligned}
L(\boldsymbol{\eta}, c) &= \prod_{i=1}^{N} p(\mathbf{X}_i, S_i = 1)^{S_i} \times p(\mathbf{X}_i, S_i = 0)^{1-S_i} \\
&\propto \prod_{i=1}^{N} \{cP(\mathbf{X}_i; \boldsymbol{\eta})\}^{S_i} \times \{1 - cP(\mathbf{X}_i; \boldsymbol{\eta})\}^{1-S_i}
\end{aligned}
$$

- $(\boldsymbol{\eta}, c)$ identifiable with positive-only data
- $(\hat{\boldsymbol{\eta}}, \hat{c})$: standard maximum likelihood estimation
- phenotype prevalence: $\hat{q} = \hat{h}/\hat{c}$, where $\hat{h} = N^{-1} \sum_{i=1}^{N} S_i$

Model Calibration Among the Unlabeled

- Nonparametric estimate of number of cases in interval $a < p(x; \hat{\eta}) < b$:

$$n_{nonpara} = \frac{n_{ab}\hat{p}_0 N_{S=1}^{-1} \sum_{i=1}^{N} I\{a < p(x_i; \hat{\beta}) < b\} I\{S_i = 1\}}{N_{S=0}^{-1} \sum_{i=1}^{N} I\{a < p(x_i; \hat{\beta}) < b\} I\{S_i = 0\}}$$

  - n_{ab}: total number of unlabeled patients in interval $a < p(x; \hat{\beta}) < b$
  - $N_{S=0}$: total number of unlabeled patients
  - $N_{S=1}$: total number of anchor-positive patients
  - $\hat{p}_0 = \{q^* - N^{-1} \sum_{i=1}^{N} S_i\} / \{1 - N^{-1} \sum_{i=1}^{N} S_i\}$
  - q^*: an educated guess of q
- Model predicted number of cases in interval $a < p(x; \hat{\beta}) < b$:

$$n_{para} = \sum_{i=1}^{N} \frac{I\{a < p(x_i; \hat{\beta}) < b\} I\{S_i = 0\}(1 - \hat{c})p(x_i; \hat{\beta})}{1 - \hat{c}p(x_i; \hat{\beta})}$$

- Similar values of $n_{nonpara}$ and n_{para} indicate good calibration

- Estimation with positive-only data

$$\widehat{TPR}_v = N_{S=1}^{-1} \sum_{i=1}^{N} I\{p(x_i;\hat{\beta}) > v\} I(S_i = 1)$$

$$\widehat{PPV}_v = \frac{N_{S=1}^{-1} \sum_{i=1}^{N} I\{p(x_i;\hat{\beta}) > v\} I(S_i = 1)}{N_{S=0}^{-1} \sum_{j=1}^{N} I\{p(x_j;\hat{\beta}) > v\} I(S_i = 0)} \hat{p}_0$$

$$\widehat{FPR}_v = \frac{N_{S=0}^{-1} \sum_{j=1}^{N} I\{p(x_j;\hat{\beta}) > v\} I(S_i = 0) - \hat{p}_0 N_{S=1}^{-1} \sum_{i=1}^{N} I\{p(x_i;\hat{\beta}) > v\} I(S_i = 1)}{1 - \hat{p}_0}$$

$$\widehat{NPV}_v = 1 - \frac{N_{S=1}^{-1} \sum_{i=1}^{N} I\{p(x_i;\hat{\beta}) < v\} I(S_i = 1)}{N_{S=0}^{-1} \sum_{i=1}^{N} I\{p(x_j;\hat{\beta}) < v\} I(S_i = 0)} \hat{p}_0$$

$$\widehat{AUC} = \int \widehat{TPR}_v \, d\widehat{FPR}_v$$

Development of a Preliminary Model for Predicting PA

- 6319 patients retrospectively extracted from UPHS EHRs
  - Underwent aldosterone screening test
  - Demographics, laboratory results, encounter meta data, diagnosis codes, clinical notes
- Data transformation
  - Highly skewed variables were log transformed
  - Continuous variables were standardized
- Assumed missing completely at random
  - Analyses were restricted to patients with complete observations on included variables
- Anchor variables for PA
  - Anchor 1: Being included in the retrospective PA research database
  - Anchor 2: Being included in the retrospective PA research database or underwent diagnostic adrenal vein sampling procedure

Selection of Candidate Predictors

- Univariate analyses: logit $p(S = 1|X; \theta) = X^T\theta$
- Candidate predictors chosen by domain expert considering both statistical and clinical significance

	VARIABLE	VARIABLE.DESCRIPTION
Demographics	age	Age when aldosterone or renin test was performed (year)
	gender	Gender
	race	Race
	hisp	Hispanic (Yes/No)
Pre-visit	dbp	Diastolic blood pressure, from office visit closest (<= 14 days) to aldosterone/renin testing
	sbp	Systolic blood pressure, from office visit closest (<= 14 days) to aldosterone/renin testing
	time_bp_to_1st_RAR_yr	Time interval (years) between first office visit with blood pressure recorded to aldosterone/renin test
	time_enc_to_1st_AVS_yr	Time interval (years) between first clinical encounter to aldosterone/renin test
Laboratory results	aldo	Serum aldosterone concentration (ng/dL)
	pra	Plasma renin activity (ng/mL/hr)
	aldo:pra	The aldosterone:renin ratio ((ng Aldosterone/dL)/(ng Angiotensin II/mL/hr))
	test_potassium	Blood potassium concentration (mmol/L)
	test_sodium	Blood sodium concentration (mmol/L)
	test_carbon_dioxide	Blood carbon dioxide concentration (mmol/L)
Encounter	enc_n	Number of clinical encounters
	enc_bp_n	Number of office visits with blood pressure recorded
	time_bp_after_1st_RAR_yr	Time interval (years) between aldosterone/renin test and last office visit with blood pressure
	time_enc_after_1st_AVS_yr	Time interval (years) between aldosterone/renin test and last clinical encounter
Diagnosis codes	Dx_h2_E26.0_9_n	Sum of the number of encounters with primary aldosteronism diagnosis codes (255.1, 255.10, 255.11, 255.12, E26.0, E26.01, E26.02, E26.09, E26.9)
	Dx_h2_E26.1_8_n	Sum of the number of encounters with other hyperaldosteronism diagnosis codes (255.13, 255.14, E26.1, E26.81, E26.89)
Clinical notes	re_hyperaldo	count of 'hyperaldo' mentions in clinical notes
	re_primaryaldo	count of 'primary aldo' mentioned in the clinical notes
	re_bah	count of 'bah' mentioned in the clinical notes
	re_adrenal_adenoma	count of 'adrenal_adenoma' mentioned in the clinical notes
	re_htn	count of 'hypertension' mentioned in the clinical notes
	re_adrenalectomy	count of 'adrenalectomy' mentioned in the clinical notes

Model Building

- Baseline model included demographics and variables available at the time of PA screening
- Variables were added in sets serially until all candidate predictors were included

	Anchor 1			Anchor 2		
	$\hat{c}$	$\hat{q}$	$\widehat{AUC}$	$\hat{c}$	$\hat{q}$	$\widehat{AUC}$
Baseline model	0.100	0.300	0.787	0.150	0.270	0.780
+ Laboratory results	0.570	0.047	0.897	0.740	0.049	0.897
+ Encounter meta data	0.640	0.054	0.919	0.770	0.058	0.914
+ Diagnosis codes	0.480	0.071	0.963	0.540	0.082	0.972
+ Clinical notes	0.450	0.076	0.990	0.560	0.079	0.990

- Backward stepwise variable selection were performed until all included predictors had $p < 0.1$

Results

- Estimation of anchor sensitivity c and PA prevalence q

	Anchor 1 (2.8%)	Anchor 2 (3.8%)
$\widehat{c}$ (95%CI)	0.374 (0.282, 0.466)	0.552 (0.476, 0.634)
$\widehat{q}$ (95%CI)	0.076 (0.060, 0.092)	0.070 (0.058, 0.082)

  - $\hat{c}$ was sensitive to anchor selection
  - $\hat{q}$ was consistent regardless of anchor selection

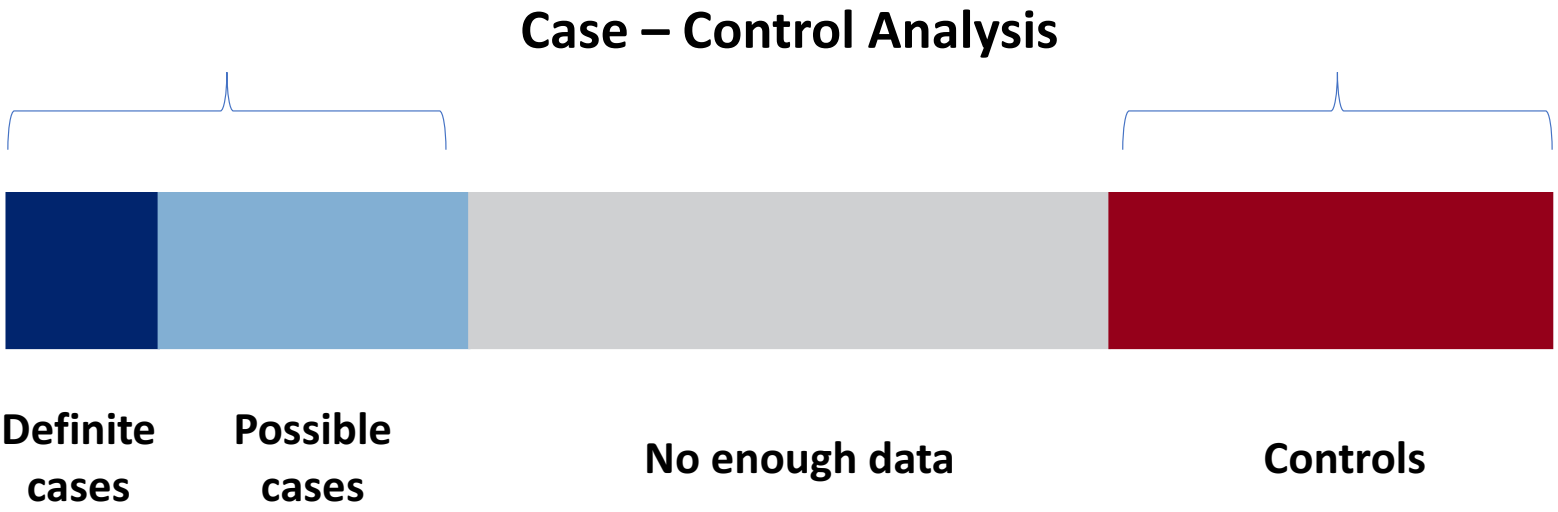
Error in EHR-Phenotyping: An Example EHR-based Study on Aortic Stenosis

- Study population
  - 44,191 patients having at least one echocardiogram recorded in Penn hospital electronic echocardiogram database between Jan 2009 and Oct 2015
- Aortic Stenosis (AS) cases identified by ICD-9 codes
  - At least one AS related codes: 424.1, 395.0, 395.2, 396.0, 396.2
  - Exclude those having bicuspid value disease: 746.3, 746.4
  - $N_1 = 6,525$
  - Chart-reviewed 327, 56.3% (184) had AS true cases
- AS controls identified by ICD-9 codes and NLP
  - Patients not having any relevant ICD-9 codes or specific key words in their echocardiography reports
  - $N_0 = 37,666$
  - Chart-reviewed 98, none had AS

An Example EHR-based Study on Aortic Stenosis

- PennSEEK algorithm for identifying "Gold-standard" AS cases (Small et al., 2018)
  - Used both ICD-9 codes and clinical notes in echocardiography reports
  - $N = 3,236$
  - Chart-reviewed 168, 166 had AS
- Estimated odds ratio parameters for Age (continuous)
  - Gold-standard cases: **1.12** (1.11, 1.12)
  - Validated cases: **1.12** (1.06, 1.14)
  - ICD-9 cases: **1.07** (1.07, 1.08) $\rightarrow$ **biased**

Design of EHR-based Case-Control Studies

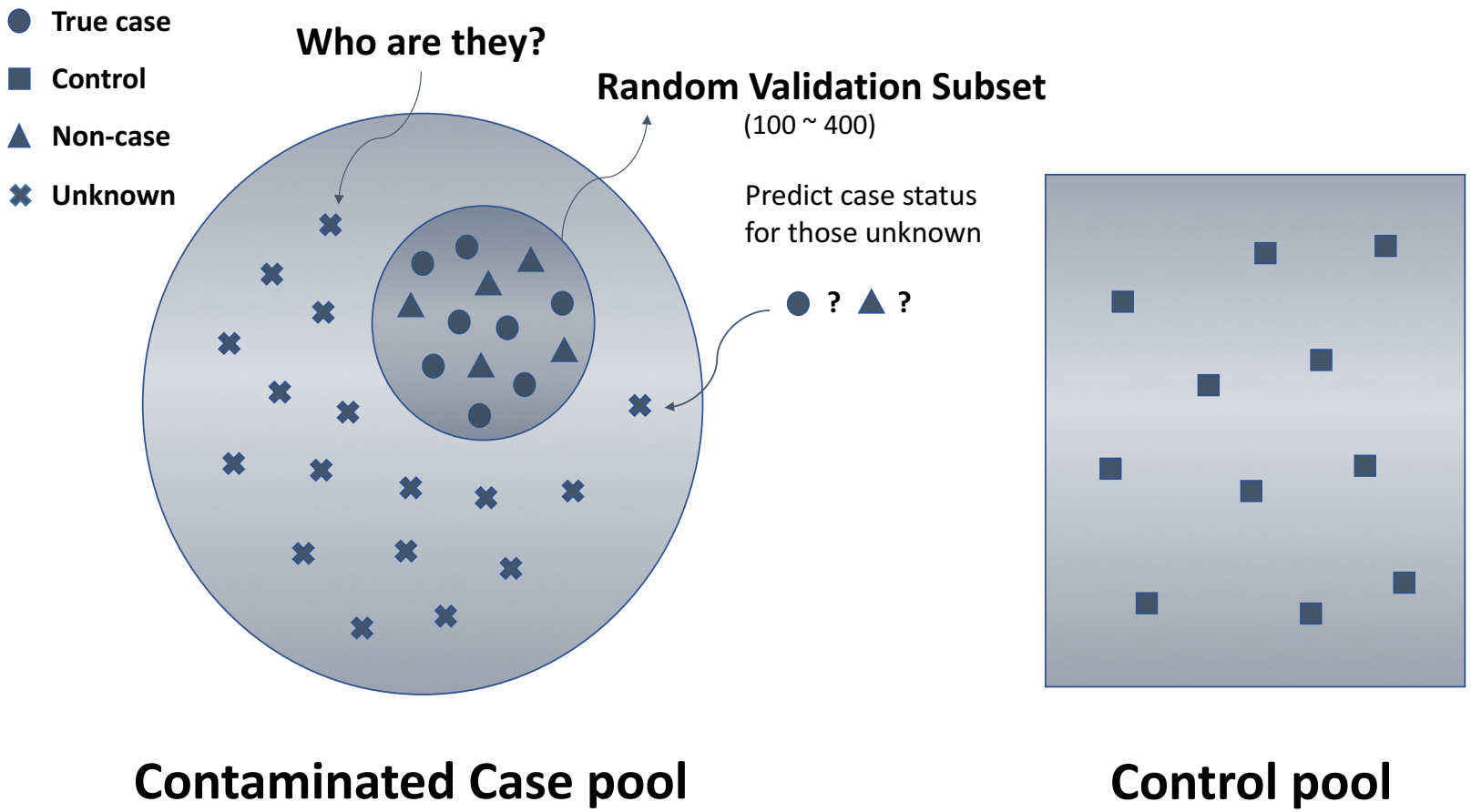


- IDENTIFY cases and controls from EHRs
- Perform standard logistic regression analysis
  - Stringent selection criteria in case identification ensures high accuracy at the price of low sample size
  - Relaxed criteria can lead to less accurate cases but larger numbers

Challenges for Analyzing EHR-based Case-Control Studies

- Ignoring inaccuracy in case identification can undermine statistical inference
  - Biased effect size estimates
  - Decreased power
- EHR case identification error is a new analytical challenge
  - True cases are contaminated by non-cases who are not controls
  - EHR case-contamination is different from classical case-control label-switching (Magder and Hughes, 1997; Meuhaus, 1999)
- Novel statistical methods are needed for addressing case contamination
  - Contaminating subjects are "non-cases", but not controls
  - Non-cases may be more similar as cases than as controls
  - Desirable to honor consistency of control definition

Our Proposed Solution



- **Notation**:
  - D: True phenotype status ($D = 0$: control; $D = 1$: true case; $D = 2$: non-case)
  - $\mathbf{X}$: Covariates of interest
  - $\mathbf{Z}$: Predictors for discriminating true cases and non-cases
  - R: Binary indicator for case validation ($R = 1$: yes; $R = 0$: no)
- **Model of interest**:

$$\log \frac{P(D = 1|\mathbf{X}; \beta_0, \boldsymbol{\beta}_1)}{P(D = 0|\mathbf{X}; \beta_0, \boldsymbol{\beta}_1)} = \beta_0 + \boldsymbol{\beta}_1^\top \mathbf{X} \qquad (1)$$

Fit the logistic regression model to the case-control data as if the sampling were prospective (Prentice and Pyke, 1979)

- Estimates of $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0^*, \hat{\boldsymbol{\beta}}_1)$ are obtained by solving estimating equations

$$\sum_{i=1}^{N_1} \tilde{X}_i P^*(D_i = 0 | \mathbf{X}_i; \hat{\boldsymbol{\beta}}) - \sum_{j=1}^{N_0} \tilde{X}_i P^*(D_j = 1 | \mathbf{X}_j; \hat{\boldsymbol{\beta}}) = \mathbf{0},$$

  where

$$P^*(D = 1 \,|\, \mathbf{X}, \hat{\boldsymbol{\beta}}) = \exp(\hat{\beta}_0^* + \hat{\boldsymbol{\beta}}_1^T \mathbf{X}) / \{1 + \exp(\hat{\beta}_0^* + \hat{\boldsymbol{\beta}}_1^T \mathbf{X})\}$$

- $\hat{\boldsymbol{\beta}}_1$ is consistent
- The estimated intercept converges to a value different from β_0

$$\beta_0^* = \beta_0 + \log(N_1/N_0) - \log\{P(D = 1)/P(D = 0)\},$$

  N_1/N_0: numbers of cases/controls; $P(D = 1)$: phenotype prevalence

A Novel Estimating Equation Approach

- Weight the contribution of each non-validated candidate case by its probability of being a true case

$$\sum_{i=1}^{N_1} \Big((1-R_i)\mathrm{E}(S_i \mid \mathbf{Z}_i) + R_i S_i \Big) \tilde{X}_i P^*(D_i = 0 \mid \mathbf{X}_i; \hat{\boldsymbol{\beta}}) - \sum_{j=1}^{N_0} \tilde{X}_i P^*(D_j = 1 \mid \mathbf{X}_j; \hat{\boldsymbol{\beta}}) = \mathbf{0}$$

  - $S_i = 1$: true case; $S_i = 0$: non-case
- Upon a valid model for $\mathrm{E}(S \mid \mathbf{Z})$, we show that
  - The estimating equation is unbiased
  - The estimates are expected to be consistent
- $\mathrm{E}(S \mid \mathbf{Z})$ is unknown
  - We develop a parametric model ("**phenotyping model**") using the validation data

$$\text{logit } P^v(S_i = 1 \mid \mathbf{Z}_i; \boldsymbol{\tau}) = \tau_0 + \boldsymbol{\tau}_1^T \mathbf{Z}_i, \quad i = 1, \ldots, n_1$$

- Develop $P^v(S = 1 \mid \mathbf{Z}_i; \hat{\boldsymbol{\tau}})$ using n_1 validated candidate cases
- Estimate probability of being a true case $P^v(S_j = 1 \mid \mathbf{Z}; \hat{\boldsymbol{\tau}})$ for non-validated candidate cases
- Plug $P^v(S = 1 \mid \mathbf{Z}_j; \hat{\boldsymbol{\tau}})$ back to the estimating equation to obtain $(\hat{\beta}_0^*, \hat{\boldsymbol{\beta}}_1)$
- Large sample properties can be studied by applying standard M-estimation theory
  - Estimates $(\hat{\beta}_0^*, \hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\tau}})$ are obtained by simutaneously solving

$$\sum_{i=1}^{N_1} \Big((1 - R_i) P^v(S_i = 1 \mid \mathbf{Z}_i; \hat{\boldsymbol{\tau}}) + R_i S_i \Big) \tilde{X}_i P^*(D_i = 0 \mid \mathbf{X}_i; \hat{\boldsymbol{\beta}}) - \sum_{j=1}^{N_0} \tilde{X}_i P^*(D_j = 1 \mid \mathbf{X}_j; \hat{\boldsymbol{\beta}}) = \mathbf{0},$$

and

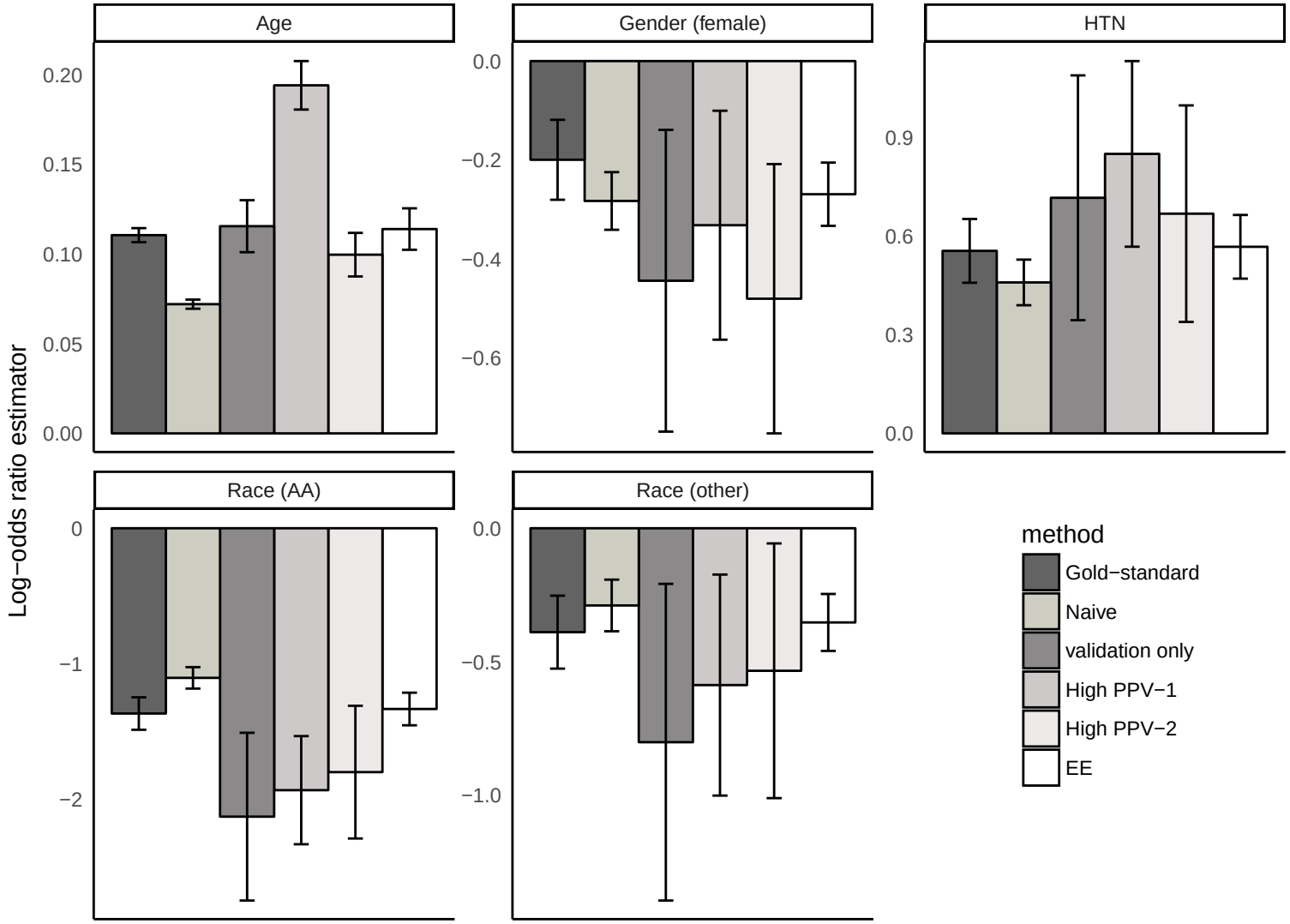
$$\sum^{N_1} R_i \tilde{Z}_i \Big\{ S_i - P(S_i = 1 \mid \mathbf{Z}_i; \hat{\boldsymbol{\tau}}) \Big\} = \mathbf{0}$$

Penn EHR-based Study on Aortic Stenosis

- Candidate cases identified by ICD-9 codes ($N_1 = 6,525$)
  - Chart-reviewed 327, 184(56.3%) had AS
- Controls identified by ICD-9 codes and NLP ($N_0 = 37,666$)
  - Chart-reviewed 98, none had AS
- True case status for this dataset was known for all $6,526$
  - $3,236$ AS cases were identified by a novel Penn algorithm
  - Chart-reviewed 168, 166(98.8%) had AS

- Association model of interest
  - Outcome variable: AS status (case or control)
  - Covariates $\mathbf{x}$: age, gender (male: reference), race (EA, AA, other), hypertension status
- Phenotyping model
  - Outcome variable: AS status (case or non-case)
  - Predictors $\mathbf{z}$: age, triglycerides (median value, indicator variable for availability)

AS Study Results



Bias Correction To address Inaccuracy in Cohort Identification

- Motivating example:
  - Investigate the development of cardiovascular diseases (e.g. CHD, PAD etc.) among individuals who have type II diabetes (T2D)
  - Study population: a cohort of individuals identified as having T2D in EHRs
- Challenge:
  - Cohort selected from EHRs might be mixed with those not having T2D, resulting in bias in down stream analysis
- The estimating equation approach can be easily extended

Acknowledgements

- Scott M. Damrauer, MD
  Upenn Assistant Professor of Surgery
- Aeron Small, MD
  Hospital resident, Yale Traditional Internal Medicine
- Daniel Herman, MD, PhD
  Upenn Assistant Professor of Laboratory Medicine and Pathology
- Rebecca A. Hubbard, PhD, Upenn
  Associate Professor of Biostatistics
- Jill Schnall, MS Biostatistics student
- Lu Wang, Ph.D. advisee
- Lingjiao Zhang, Ph.D. advisee

Thank you very much!