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 approaches 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

- True patient state
- Recording
- EHR data
- Phenotyping
- Discovery
- Modeling
  - Prediction
  - Association analysis
  - Interpretation
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
  - Statistical modeling: logistic regression, machine learning, AI..
Model validation: PPV/NPV; Calibration largely ignored
- Available methods all required annotation of “gold standard” cases and controls
  - Anchor variable framework is an exception (Hapern et al. 2016)
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
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
Notation

- **Y**: True phenotype status \( (Y = 1: \text{case}, Y = 0: \text{control}) \)
- **X**: A vector of covariates predictive of \( Y \), with density \( f(X) \)
- **S**: Anchor variable \( (S = 1: \text{presence}, S = 0: \text{absence}) \)
- **q**: Phenotype prevalence, \( q = p(Y = 1) \)
- **h**: Anchor prevalence, \( h = p(S = 1) \)
- \( (X, Y, S) \): Random variables, with joint distribution \( p(X, Y, S) \)

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

\[
L(\eta, c) = \prod_{i=1}^{N} p(X_i, S_i = 1)^{S_i} \times p(X_i, S_i = 0)^{1-S_i}
\]

\[
\propto \prod_{i=1}^{N} \{cP(X_i; \eta)\}^{S_i} \times \{1 - cP(X_i; \eta)\}^{1-S_i}
\]

- \((\eta, c)\) identifiable with positive-only data
- \((\hat{\eta}, \hat{c})\): 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 $a < p(x; \hat{\eta}) < b$:

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

- $n_{ab}$: 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 $a < p(x; \hat{\beta}) < b$:

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

- Similar values of $n_{\text{nonpara}}$ and $n_{\text{para}}$ indicate good calibration
Estimation with positive-only data

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

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

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

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

\[ \hat{AUC} = \int \hat{TPR}_v d\hat{FPR}_v \]
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**

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>VARIABLE.DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics</td>
<td></td>
</tr>
<tr>
<td>age</td>
<td>Age when aldosterone or renin test was performed (year)</td>
</tr>
<tr>
<td>gender</td>
<td>Gender</td>
</tr>
<tr>
<td>race</td>
<td>Race</td>
</tr>
<tr>
<td>hisp</td>
<td>Hispanic (Yes/No)</td>
</tr>
<tr>
<td>Pre-visit</td>
<td></td>
</tr>
<tr>
<td>dbp</td>
<td>Diastolic blood pressure, from office visit closest (&lt;= 14 days) to aldosterone/renin testing</td>
</tr>
<tr>
<td>sbp</td>
<td>Systolic blood pressure, from office visit closest (&lt;= 14 days) to aldosterone/renin testing</td>
</tr>
<tr>
<td>time_bp_to_1st_RAR_yr</td>
<td>Time interval (years) between first office visit with blood pressure recorded to aldosterone/renin test</td>
</tr>
<tr>
<td>time_enc_to_1st_AVS_yr</td>
<td>Time interval (years) between first clinical encounter to aldosterone/renin test</td>
</tr>
<tr>
<td>Laboratory results</td>
<td></td>
</tr>
<tr>
<td>aldo</td>
<td>Serum aldosterone concentration (ng/dL)</td>
</tr>
<tr>
<td>pra</td>
<td>Plasma renin activity (ng/mL/hr)</td>
</tr>
<tr>
<td>aldo:pra</td>
<td>The aldosterone:renin ratio ((ng Aldosterone/dL)/(ng Angiotensin II/mL/hr))</td>
</tr>
<tr>
<td>test_potassium</td>
<td>Blood potassium concentration (mmol/L)</td>
</tr>
<tr>
<td>test_sodium</td>
<td>Blood sodium concentration (mmol/L)</td>
</tr>
<tr>
<td>test_carbon_dioxide</td>
<td>Blood carbon dioxide concentration (mmol/L)</td>
</tr>
<tr>
<td>Encounter</td>
<td></td>
</tr>
<tr>
<td>enc_n</td>
<td>Number of clinical encounters</td>
</tr>
<tr>
<td>enc_bp_n</td>
<td>Number of office visits with blood pressure recorded</td>
</tr>
<tr>
<td>time_bp_after_1st_RAR_yr</td>
<td>Time interval (years) between aldosterone/renin test and last office visit with blood pressure</td>
</tr>
<tr>
<td>time_enc_after_1st_AVS_yr</td>
<td>Time interval (years) between aldosterone/renin test and last clinical encounter</td>
</tr>
<tr>
<td>Diagnosis codes</td>
<td></td>
</tr>
<tr>
<td>DX_h2_E26.0_9_n</td>
<td>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)</td>
</tr>
<tr>
<td>DX_h2_E26.1_8_n</td>
<td>Sum of the number of encounters with other hyperaldosteronism diagnosis codes (255.13, 255.14, E26.1, E26.81, E26.89)</td>
</tr>
<tr>
<td>Clinical notes</td>
<td></td>
</tr>
<tr>
<td>re_hyperaldo</td>
<td>count of 'hyperaldo' mentions in clinical notes</td>
</tr>
<tr>
<td>re_primaryaldo</td>
<td>count of 'primary aldo' mentioned in the clinical notes</td>
</tr>
<tr>
<td>re_bah</td>
<td>count of 'bah' mentioned in the clinical notes</td>
</tr>
<tr>
<td>re_adrenal_adenoma</td>
<td>count of 'adrenal adenoma' mentioned in the clinical notes</td>
</tr>
<tr>
<td>re_hln</td>
<td>count of 'hypertension' mentioned in the clinical notes</td>
</tr>
<tr>
<td>re_adrenalectomy</td>
<td>count of 'adrenalectomy' mentioned in the clinical notes</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th></th>
<th>Anchor 1</th>
<th>Anchor 2</th>
<th>Anchor 1</th>
<th>Anchor 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline model</td>
<td>0.100</td>
<td>0.300</td>
<td>0.787</td>
<td>0.150</td>
</tr>
<tr>
<td>+ Laboratory results</td>
<td>0.570</td>
<td>0.047</td>
<td>0.897</td>
<td>0.740</td>
</tr>
<tr>
<td>+ Encounter meta data</td>
<td>0.640</td>
<td>0.054</td>
<td>0.919</td>
<td>0.770</td>
</tr>
<tr>
<td>+ Diagnosis codes</td>
<td>0.480</td>
<td>0.071</td>
<td>0.963</td>
<td>0.540</td>
</tr>
<tr>
<td>+ Clinical notes</td>
<td>0.450</td>
<td>0.076</td>
<td>0.990</td>
<td>0.560</td>
</tr>
</tbody>
</table>

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

Estimation of anchor sensitivity $c$ and PA prevalence $q$

<table>
<thead>
<tr>
<th></th>
<th>Anchor 1 (2.8%)</th>
<th>Anchor 2 (3.8%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{c}$ (95% CI)</td>
<td>0.374 (0.282, 0.466)</td>
<td>0.552 (0.476, 0.634)</td>
</tr>
<tr>
<td>$\hat{q}$ (95% CI)</td>
<td>0.076 (0.060, 0.092)</td>
<td>0.070 (0.058, 0.082)</td>
</tr>
</tbody>
</table>

$\hat{c}$ was sensitive to anchor selection

$\hat{q}$ was consistent regardless of anchor selection
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 valve 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
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) → biased
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
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

- True case
- Control
- Non-case
- Unknown

Who are they?

Random Validation Subset
(100 ~ 400)

Predict case status for those unknown

Contaminated Case pool

Control pool
Statistical Modeling

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

Model of interest:

$$\log \frac{P(D = 1|X; \beta_0, \beta_1)}{P(D = 0|X; \beta_0, \beta_1)} = \beta_0 + \beta_1^T X$$ (1)
Fit the logistic regression model to the case-control data as if the sampling were prospective (Prentice and Pyke, 1979)

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

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

where

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

- $\hat{\beta}_1$ is consistent
- The estimated intercept converges to a value different from $\beta_0$

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

$N_1/N_0$: numbers of cases/controls; $P(D = 1)$: phenotype prevalence
Weight the contribution of each non-validated candidate case by its probability of being a true case

\[
\sum_{i=1}^{N_1} \left( (1 - R_i) \mathbb{E}(S_i \mid Z_i) + R_i S_i \right) \tilde{X}_i P^* (D_i = 0 \mid X_i; \hat{\beta})
\]

\[
- \sum_{j=1}^{N_0} \tilde{X}_i P^* (D_j = 1 \mid X_j; \hat{\beta}) = 0
\]

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

\[
\text{logit } P^\nu(S_i = 1 \mid Z_i; \tau) = \tau_0 + \tau_1^T Z_i, \quad i = 1, \ldots, n_1
\]
The Estimating Equation Approach

- Develop $P^v(S = 1 \mid Z_i; \hat{\tau})$ using $n_1$ validated candidate cases
- Estimate probability of being a true case $P^v(S_j = 1 \mid Z; \hat{\tau})$ for non-validated candidate cases
- Plug $P^v(S = 1 \mid Z_j; \hat{\tau})$ back to the estimating equation to obtain $(\hat{\beta}_0^*, \hat{\beta}_1)$
- Large sample properties can be studied by applying standard M-estimation theory
  - Estimates $(\hat{\beta}_0^*, \hat{\beta}_1, \hat{\tau})$ are obtained by simultaneously solving
    \[
    \sum_{i=1}^{N_1} \left( (1 - R_i)P^v(S_i = 1 \mid Z_i; \hat{\tau}) + R_iS_i \right) \tilde{X}_i P^* (D_i = 0 \mid X_i; \hat{\beta}) \\
    - \sum_{j=1}^{N_0} \tilde{X}_i P^* (D_j = 1 \mid X_j; \hat{\beta}) = 0,
    \]
    and
    \[
    \sum_{i=1}^{N_1} R_i \tilde{Z}_i \left\{ S_i - P(S_i = 1 \mid Z_i; \hat{\tau}) \right\} = 0
    \]
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 $x$: age, gender (male: reference), race (EA, AA, other), hypertension status

- Phenotyping model
  - Outcome variable: AS status (case or non-case)
  - Predictors $z$: age, triglycerides (median value, indicator variable for availability)
AS Study Results

Race (AA) Race (other)

Age Gender (female) HTN

0.0
0.3
0.6
0.9
−0.6
−0.4
−0.2
0.0
−1.0
−0.5
0.0
0.00
0.05
0.10
0.15
0.20
−2
−1
0
Log−odds ratio estimator

method
Gold−standard
Naive
validation only
High PPV−1
High PPV−2
EE

Jinbo Chen
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 downstream 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!