Measurement Errors in EHR data

Rui Duan, Jing Huang and Yong Chen
ruiduan@; jing14@; ychen123@upenn.edu

Department of Biostatistics and Epidemiology
The University of Pennsylvania

Feb 16, 2017
EHR Based Phenotyping

Motivation

Simulation Study

Conclusion

Ongoing work

Tools

Phenotyping: Cohort Discovery Using EHR Data

Phenotyping is the practice of developing algorithms designed to identify specific phenomic traits within an individual. These algorithms are created using multiple variables enabling researchers to accurately identify traits and perform analyses. Best practice materials and data standardizations tools have been developed to aid with phenotyping protocols and collaboration. Phenotyping also seeks to advance the science of de-identification, transportable phenotyping methods, structure and standards, and portable components of algorithms and methods. Within the eMERGE network, phenotyping focuses both disease related and pharmacogenomic related phenotypes.

Tools

- eMERGE Record Counter
- PheKB
- PheWAS

Phenotyping-Related Publications

- Kho AN, Hayes MG, Rasmussen-Torvik L, Pacheco JA, Thompson WK, Lowe WL. Use of Diverse Electronic Medical Record Systems to Identify Genetic Risk for Type 2 Diabetes within a Genome-wide Association Study. J Am Med

<table>
<thead>
<tr>
<th>Institution</th>
<th>Repository size</th>
<th>% of genopool subjects reported</th>
<th>EMR description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHOP</td>
<td>80,367</td>
<td></td>
<td>Epic EMR since 2004</td>
</tr>
<tr>
<td>CCHMC</td>
<td>57,979</td>
<td>14.5% AA, 2.4% EL</td>
<td>Epic EMR since 2007</td>
</tr>
<tr>
<td>Columbia</td>
<td>26,000</td>
<td>66% EA, 10% AA, 4% bi/mixed, 10% other</td>
<td>a combination of home</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(18% female)</td>
<td>grown systems called &quot;SNP&quot;</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>and vender systems</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>provided by Aleregen.</td>
</tr>
<tr>
<td>Genentech</td>
<td>570,000</td>
<td>17.568</td>
<td>Epic EMR since 1996</td>
</tr>
</tbody>
</table>
EHR Based Phenotyping

Motivation

EHR Based Phenotyping

Simulation Study

Conclusion

Ongoing work

\[ \beta \approx \beta^* \]

Here the sensitivity is between **99%-100%**, and PPV is between **98%-100%**.
EHR Based Phenotyping: Scenario 1

Question 1: *If the phenotyping algorithm is of high accuracy, can we ignore the misclassification and treat the EHR-derived phenotype as the "true" disease status?*

Here the sensitivity is between 52%-63%, and the PPV is between 88%-94%. In general, phenotyping algorithm is highly disease dependent.
**Problem of Portability:** An algorithm may not perform equally well on different datasets since the structure of the data and characteristics of the study population can vary from study site to site.


<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Testing set</th>
<th></th>
<th></th>
<th></th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Partners</td>
<td>Northwestern</td>
<td>Vanderbilt</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PPV</td>
<td>Sensitivity</td>
<td>AUC</td>
<td>PPV</td>
<td>Sensitivity</td>
</tr>
<tr>
<td>Published algorithm</td>
<td>88%*</td>
<td>79%*</td>
<td>97%*</td>
<td>87%</td>
<td>60%</td>
</tr>
<tr>
<td>Retrained with</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Northwestern</td>
<td>79%</td>
<td>47%</td>
<td>89%</td>
<td>87%</td>
<td>73%</td>
</tr>
<tr>
<td>Vanderbilt</td>
<td>85%</td>
<td>74%</td>
<td>97%</td>
<td>82%</td>
<td>40%</td>
</tr>
<tr>
<td>Combined</td>
<td>86%</td>
<td>71%</td>
<td>97%</td>
<td>86%</td>
<td>65%</td>
</tr>
<tr>
<td>ICD-9 only†</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥1 RA code</td>
<td>22%</td>
<td>97%</td>
<td>N/A</td>
<td>26%</td>
<td>100%</td>
</tr>
<tr>
<td>&gt;3 RA code</td>
<td>55%</td>
<td>81%</td>
<td>N/A</td>
<td>42%</td>
<td>87%</td>
</tr>
<tr>
<td>97% Specificity</td>
<td>80%</td>
<td>49%</td>
<td>88%</td>
<td>80%</td>
<td>36%</td>
</tr>
<tr>
<td>Code count for 97% specificity</td>
<td>53</td>
<td>29</td>
<td>48</td>
<td>43.3</td>
<td></td>
</tr>
</tbody>
</table>

The PPV and sensitivity values reported represent model performance with a specificity set at 97% for logistic regression models.

*These results are from a fivefold cross-validation on the Partners training set. The PPV and sensitivity as published in Liao et al was calculated from a separate Partners validation set (PPV 94%, sensitivity 63%).

†ICD-9 cut-off used the count of 714.* codes, excluding codes for juvenile RA (714.3*).

AUC, area under the receiver operating characteristic curve; ICD-9, International Classification of Diseases, version 9 CM; PPV, positive predictive value; RA, rheumatoid arthritis.
Question 2: When a phenotyping algorithm does not perform well, what is the consequence of ignoring the misclassification in subsequent association studies?
EHR-based phenotyping: scenario 3

EHR-derived phenotype for exposure variables, e.g., smoking status:


<table>
<thead>
<tr>
<th>Table 4. Patient-level classification evaluation on the test set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>---------------------------</td>
</tr>
<tr>
<td><strong>i2b2 Classification</strong></td>
</tr>
<tr>
<td>Current Smoker (C)</td>
</tr>
<tr>
<td>Past Smoker (P)</td>
</tr>
<tr>
<td>Non-smoker (N)</td>
</tr>
<tr>
<td>Macro Average</td>
</tr>
<tr>
<td>Micro Average</td>
</tr>
<tr>
<td><strong>Ever/Never Classification</strong></td>
</tr>
<tr>
<td>Presence (current or past smokers)</td>
</tr>
<tr>
<td>Absence (non-smokers)</td>
</tr>
</tbody>
</table>
EHR Based Phenotyping: Scenario 3

**Question 3:** When the exposure variable is also subject to phenotyping error, what is the further impact on subsequent association studies?
The goal of our study is to investigate the impact of inaccurate phenotype on the association study through a simulation study.
Motivation
EHR Based Phenotyping
Simulation Study
Conclusion
Ongoing work

STATISTICAL ISSUES FOR EHR BASED ASSOCIATION STUDY

Phenotyping Errors may lead to:

1. For hypothesis testing:
   conservative or inflated Type I error, loss of power.

2. For estimation:
   biased estimation of the association.

We will consider:

1. Misclassified binary outcome (disease status)
   - When algorithms perform well
     (low misclassification rates)–Scenario 1
   - When algorithms lack portability
     (high misclassification rates)–Scenario 2

2. Misclassified covariate variables –Scenario 3
Simulation Settings

The impact of misclassification on power of association studies

Only outcome misclassification

- GWAS (covariate: genotype)
  - Common disease (Type II diabetes)
    - Low misclassification
      - (sen=0.84, spec=0.96)
    - High misclassification
      - (sen=0.88, spec=0.82)
  - Rare disease (multiple sclerosis)
    - Low misclassification
      - (sen=0.857, spec=0.997)
    - High misclassification
      - (sen=0.707, spec=0.988)

Average variant
- (MAF=0.2)

Rare variant
- (MAF=0.03)

Common variant
- (randomly sample from the MAF distribution)

Both outcome and covariate misclassification

- Epi study (covariate: smoking status)
  - Common disease (Type II diabetes)
    - Low misclassification
      - (sen=0.84, spec=0.96)
    - High misclassification
      - (sen=0.65, spec=0.82)
  - Rare disease (multiple sclerosis)
    - Low misclassification
      - (sen=0.857, spec=0.997)
    - High misclassification
      - (sen=0.707, spec=0.988)

- Number of cigarettes per day
  - (X~ZIP(0.85, 15))

- Current smoker, past smoker, non-smoker
  - (X~Multinomial(0.15, 0.183, 0.683, n=1))

Small measurement error
- (classifier one: sen=0.98, spec=0.97)
- (classifier two: sen=0.94, spec=0.92)

Large measurement error
- (classifier one: sen=0.31, spec=0.91)
- (classifier two: sen=0.84, spec=0.51)

Small measurement error
- (random error~N[0, 0.5, var=0.625])

Large measurement error
- (random error~N[0, 8, var=7.5])
SIMULATION RESULTS

Genetic Association for one single SNP:

Comparison of Power for Common Diseases in Genetic Association Studies

- **true outcome status with common variant**
- **outcome low misclassification with common variant**
- **outcome high misclassification with common variant**
- **true outcome status with rare variant**
- **outcome low misclassification with rare variant**
- **outcome high misclassification with rare variant**
Epidemiological Association for Smoking Status:

Comparison of Power for Common Diseases in Epidemiological Studies

- **true outcome and exposure**
- **outcome low and exposure low** misclassification
- **outcome high and exposure low** misclassification
- **outcome low and exposure high** misclassification
- **outcome high and exposure high** misclassification
**SIMULATION RESULTS**

Epidemiological Association for Smoking Status:

![Comparison of Power for Common Diseases in Epidemiological Studies](image)

The statistical power you need
**Simulation Results**

Epidemiological Association for Smoking Status:

![Graph showing comparison of power for common diseases in epidemiological studies.](image)
CONCLUSIONS

1. Power loss depends on:
   - Effect size you want to detect.
   - Prevalence of disease.
   - Frequency of the exposure.
   - The accuracy of the phenotyping algorithm for outcome and exposure.

2. In most of the situations, the power loss are substantial if the misclassification is ignored.
1. From study design point of view:
   - Underestimation of the sample size to for a given power.
   - Correct sample size can be obtained through a similar simulation study.

2. From hypothesis testing point of view:
   - Methods need to be developed to account for phenotyping errors and improve statistical power.

3. From estimation point of view:
   - Methods need to be developed for bias correction.
**Future Works**

1. Investigate how the power loss is attributable to all the factors individually and jointly.

2. Develop new statistical methods and software for bias reduction and power improvement in EHR based association studies.

(This paper won the first prize of “Best of Student Papers in Knowledge Discovery and Data Mining (KDDM)” Awards)
Part II: Bias correction for measurement errors in EHR data
**Data Structure**

- True disease status ($Y$)
- EHR-derived phenotype ($Y^*$)
- Exposure ($X$)

True association $\beta$

"Surrogate" association $\beta^*$

$\beta \neq \beta^*$

- Not observed
- Observed
Existing work for Bias Reduction

- Without validation data: Irwin 1954; Barron 1977; Magder & Hughes 1997; Morrissey & Spiegelman 1999; Lyles 2002;
- With validation data: Greenland & Kleinbaum 1983; Copas 1988; Paulino et. al 2003; Lyles & Lin 2010; Shardell et. al 2015; Edwards et. al 2013.
Validation study as the gold standard

▶ Validation study by manual chart review is expensive
▶ How large is the chart review?
▶ How about the uncertainty of estimates from validation study?
The standard likelihood function: \( L(\alpha_1, \alpha_2, \beta) \)
- \( \alpha_1 \) and \( \alpha_2 \) are sensitivity and specificity.
- \( \beta \) is a vector of association parameters.

Commonly used methods:
- MLE method: directly maximize the joint likelihood
  i.e. maximize \( L(\alpha_1, \alpha_2, \beta) \).
- Naive method: ignore misclassification,
  i.e. maximize \( L(\alpha_1 = 1, \alpha_2 = 1, \beta) \).
- Fix value method: specify sensitivity and specificity,
  i.e. maximize \( L(\alpha_1 = \text{a fixed value}, \alpha_2 = \text{a fixed value}, \beta) \).
PERFORMANCE OF EXISTING METHODS

likelihood function

likelihood function with sensitivity and specificity incorrectly specified

likelihood function with sensitivity and specificity correctly specified

\( \beta_0 \)

\( \beta_1 \)

\( \hat{\beta}_{MLE} \)

\( \beta_2 \)
ALTERNATIVE: BAYESIAN APPROACH

- Bayesian method to use validation study as a prior
- Simulation studies to evaluate the performance
- Real application
More to come!