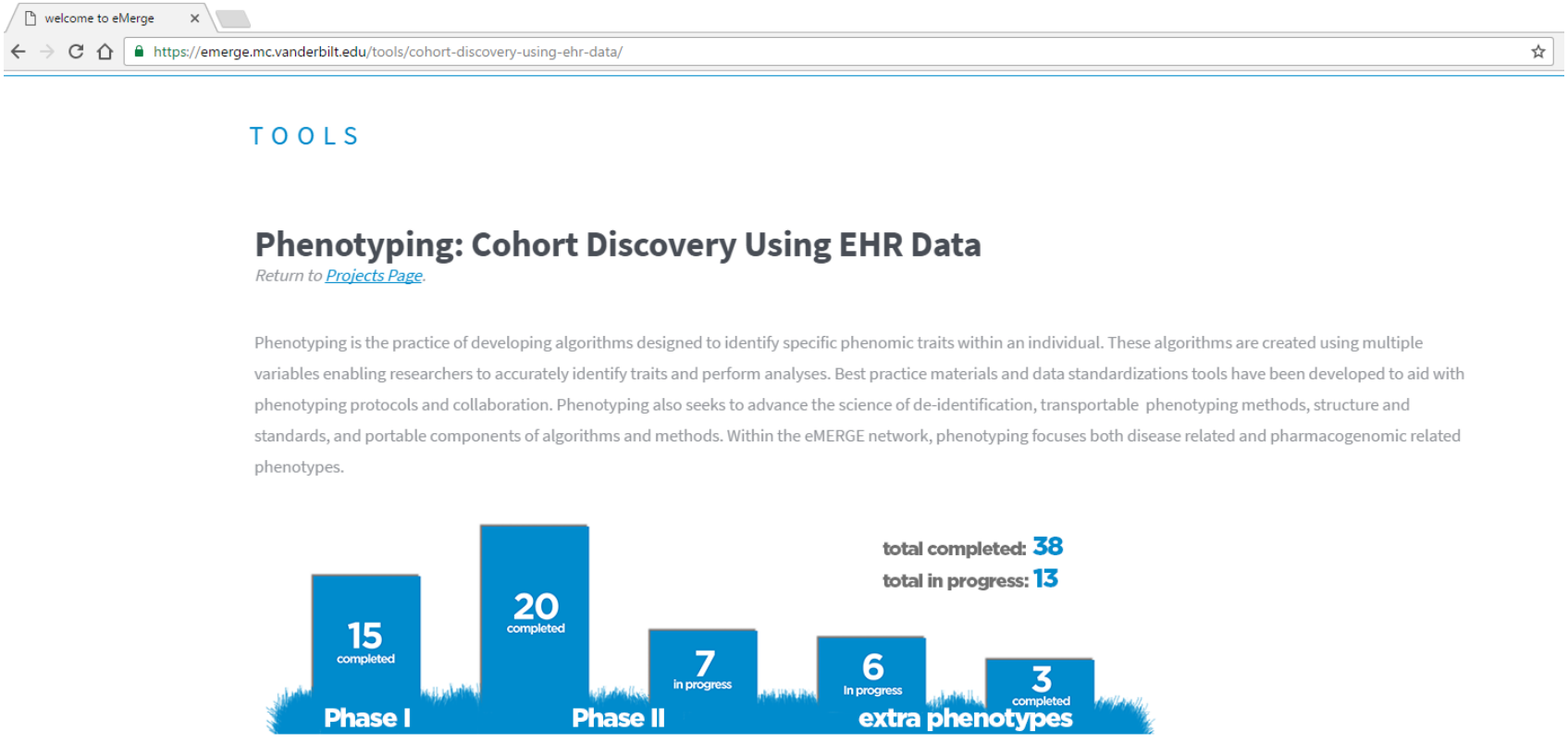
Measurement Errors in EHR data

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

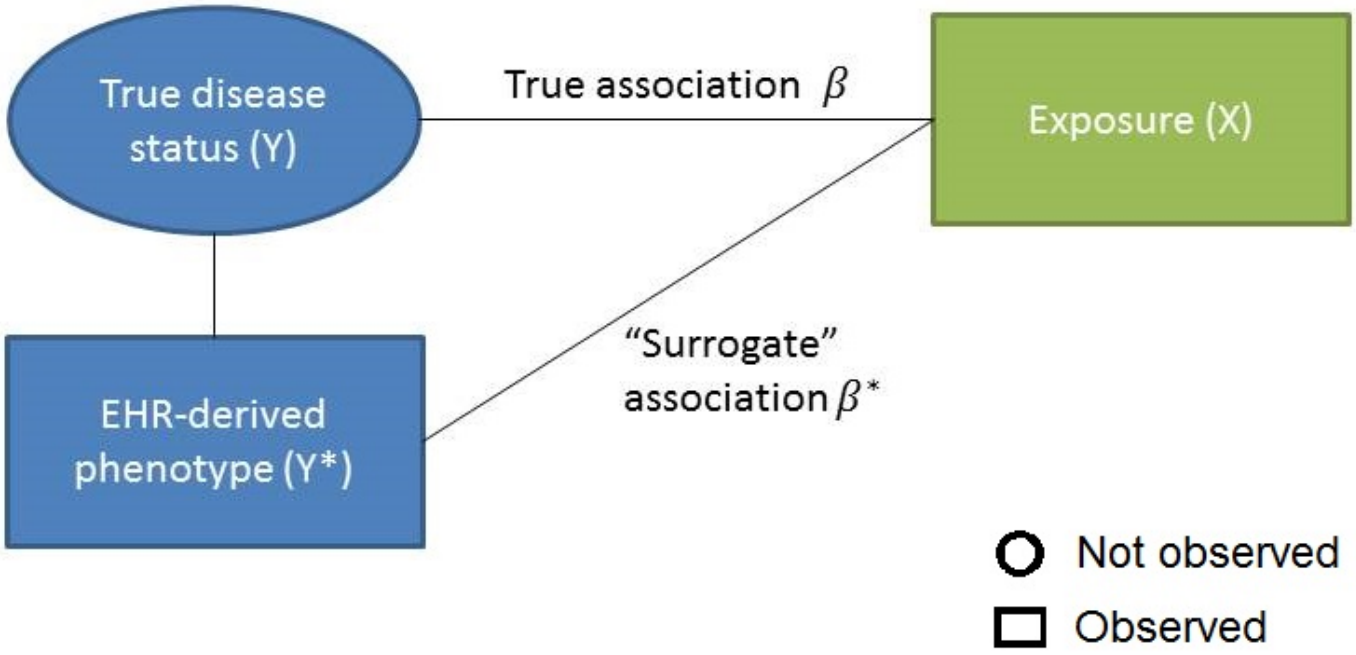
Department of Biostatistics and Epidemiology
The University of Pennsylvania

Feb 16, 2017

EHR Based Phenotyping



EHR Based Phenotyping



$$\beta \approx \beta^*?$$

EHR Based Phenotyping: Scenario 1

Kho, Abel N., et al. " Use of diverse electronic medical record systems to identify genetic risk for **type 2 diabetes** within a genome-wide association study". Journal of the American Medical Informatics Association (2012)

Table 3 Summary of chart review results at three participating sites

	Manual chart review								
	Northwestern University*			Vanderbilt University†			Marshfield Clinic‡		
	Case	Control	Total	Case	Control	Total	Case	Control	Total
EMR prediction									
Case	56	1	57	50	0	50	99	1	100
Control	0	43	43	0	50	50	1	49	50
Total	56	44	100	50	50	100	100	50	150

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?*

EHR Based Phenotyping: Scenario 2

Liao, Katherine P., et al. " Electronic medical records for discovery research in **rheumatoid arthritis**". Arthritis care & research (2010)

Model	RA by algorithm or criteria, n	PPV (%) (95% CI)	Sens (%) (95% CI)	Difference in PPV* (95% CI)
Algorithms				
Narrative + codified (complete)	3585	94 (91, 96)	63 (51, 75)	reference
Codified only	3046	88 (84, 92)	51 (42, 60)	6 (2,9)**
NLP only	3341	89 (86, 93)	56 (46, 66)	5(1,8)**
Published administrative codified criteria				
≥ 3 ICD9 RA	7960	56 (47,64)	80 (72,88)	38 (29, 47)**
≥1 ICD9 RA + DMARD	7799	45 (37, 53)	66 (57,76)	49 (40, 57)**

* Difference in PPV= (PPV of complete algorithm) – (comparison algorithm or criteria)

** Significant difference in PPV compared to complete algorithm

Here the sensitivity is between 52%-63%, and the PPV is between 88%-94%.

In general, phenotyping algorithm is highly disease dependent.

EHR Based Phenotyping: Scenario 2

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.

Carroll, Robert J., et al. " Portability of an algorithm to identify rheumatoid arthritis in electronic health records". Journal of the American Medical Informatics Association (2012)

Table 3 Model performance

	Testing set											
	Partners			Northwestern			Vanderbilt			Average		
Algorithm	PPV	Sensitivity	AUC	PPV	Sensitivity	AUC	PPV	Sensitivity	AUC	PPV	Sensitivity	AUC
Published algorithm	88%*	79%*	97%*	87%	60%	92%	95%	57%	95%	90%	65%	95%
Retrained with												
Northwestern	79%	47%	89%	87%	73%	92%	93%	43%	89%	86%	54%	90%
Vanderbilt	85%	74%	97%	82%	40%	88%	97%	81%	97%	88%	65%	94%
Combined	86%	71%	97%	86%	65%	91%	97%	82%	96%	90%	72%	95%
ICD-9 only†												
≥1 RA code	22%	97%	N/A	26%	100%	N/A	49%	100%	N/A	33%	99%	N/A
≥3 RA code	55%	81%	N/A	42%	87%	N/A	73%	98%	N/A	57%	89%	N/A
97% Specificity	80%	49%	88%	80%	36%	84%	93%	43%	93%	84%	43%	88%
Code count for 97% specificity	53			29			48			43.3		

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.

EHR Based Phenotyping: Scenario 2

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:

Liu, Mei, et al. " A study of transportability of an existing smoking status detection module across institutions". American Medical Informatics Association (2012)

Table 4. Patient-level classification evaluation on the test set

	cTAKES Module			Customized Module		
	Precision	Recall	F-measure	Precision	Recall	F-measure
i2b2 Classification						
Current Smoker (C)	0.30	0.84	0.45	0.67	0.92	0.78
Past Smoker (P)	0.82	0.51	0.63	0.93	0.73	0.82
Non-smoker (N)	0.91	0.31	0.47	0.85	0.89	0.87
Macro Average	0.68	0.55	0.52	0.82	0.85	0.83
Micro Average	0.74	0.52	0.54	0.85	0.82	0.83
Ever/Never Classification						
Presence (current or past smokers)	0.74	0.98	0.84	0.95	0.92	0.93
Absence (non-smokers)	0.91	0.31	0.46	0.85	0.89	0.87

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?*

Statistical Issues for EHR based Association Study

The goal of our study is to investigate the impact of inaccurate phenotype on the association study through a simulation study.

Statistical Issues for EHR based Association Study

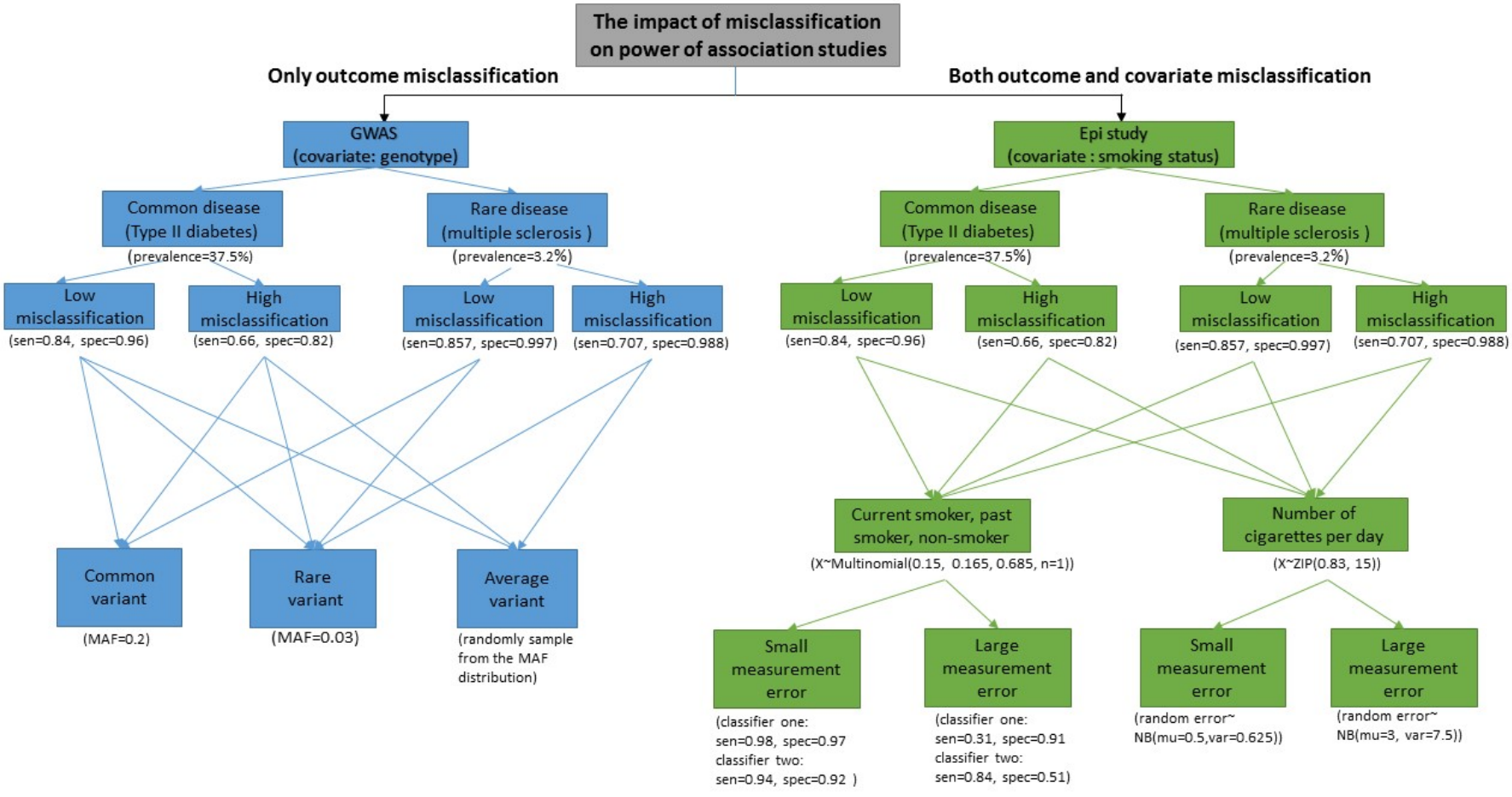
Phenotyping Errors may lead to:

1. For hypothesis testing:
   conservative or inflated Type I error, loss of power.
2. For estimation:
   biased estimation of the association.

We will consider:

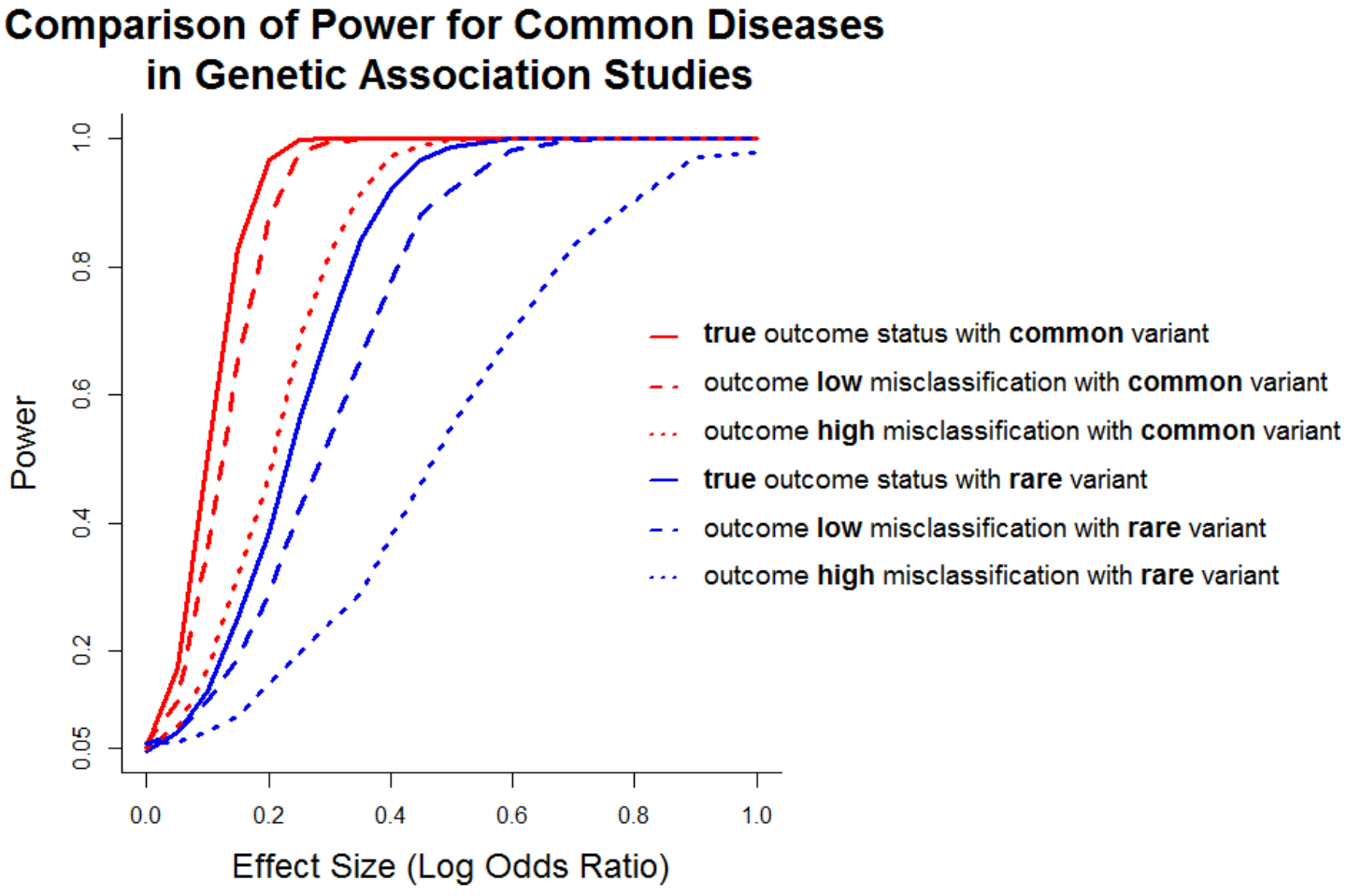
1. Misclassified binary outcome (disease status)
   - When algorithms perform well
     (low misclassification rates)–Scenario 1
   - When algorithms lack portability
     (high misclassification rates)–Scenario 2
2. Misclassified covariate variables –Scenario 3

Simulation Settings



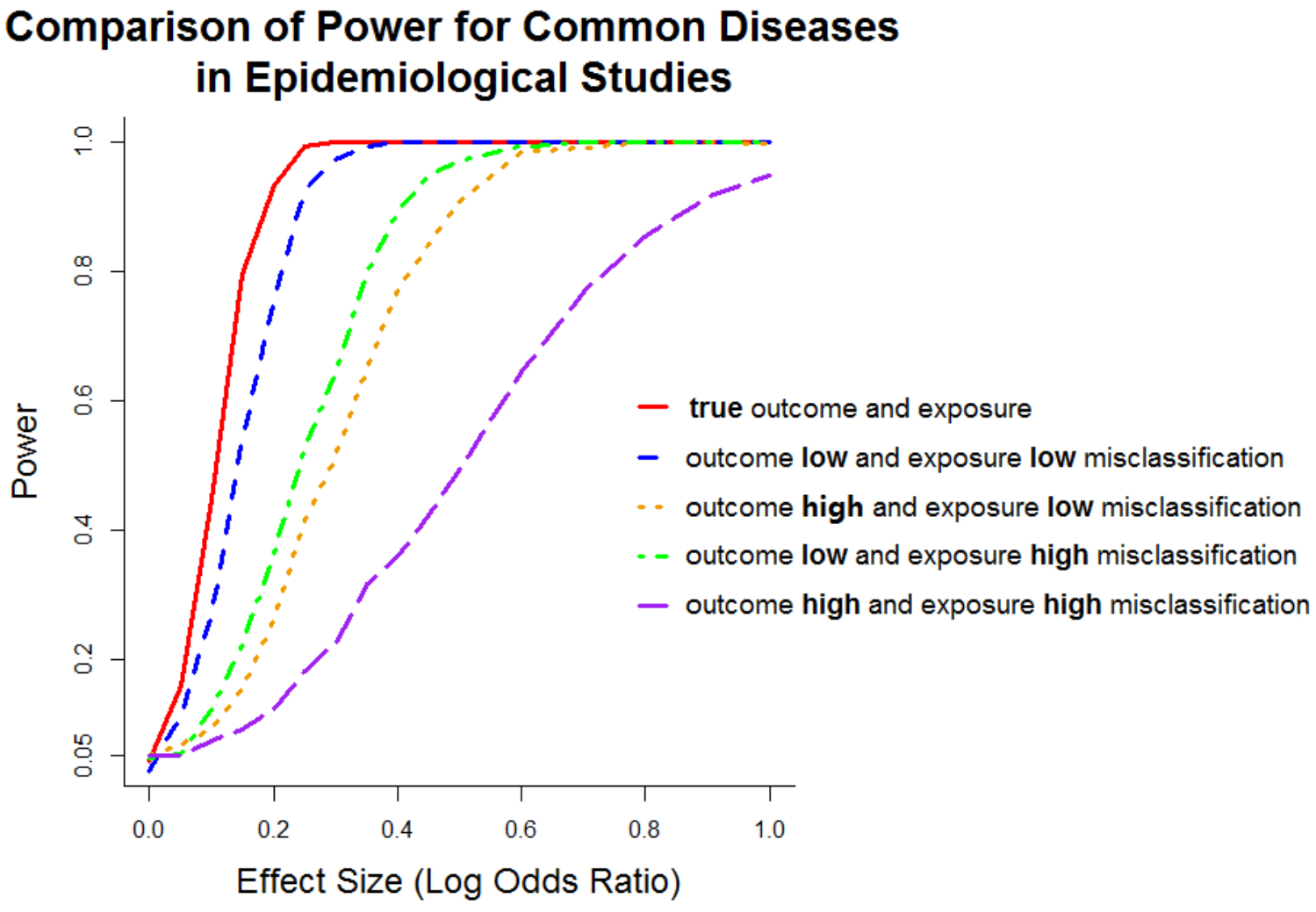
Simulation Results

Genetic Association for one single SNP:



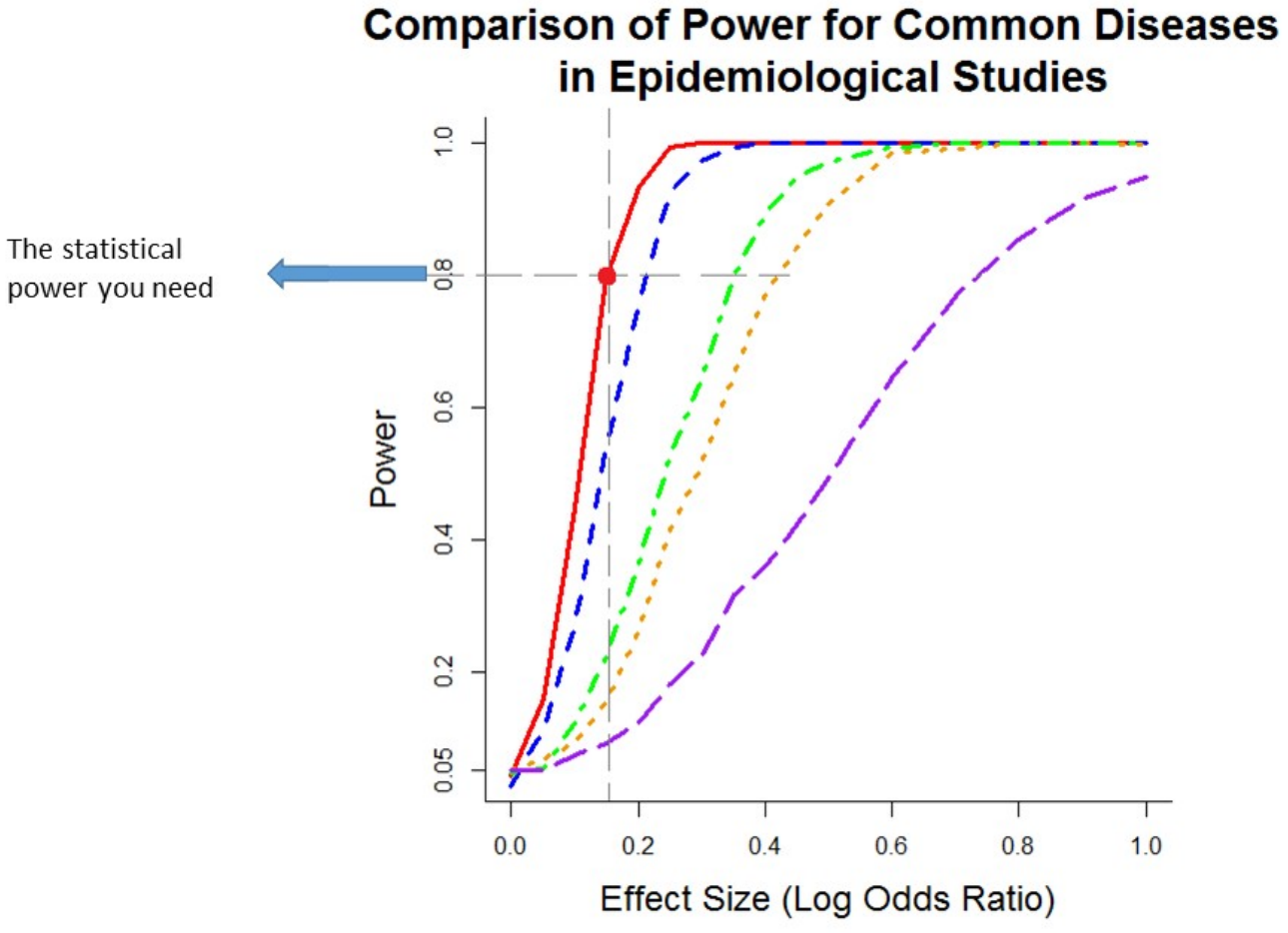
Simulation Results

Epidemiological Association for Smoking Status:



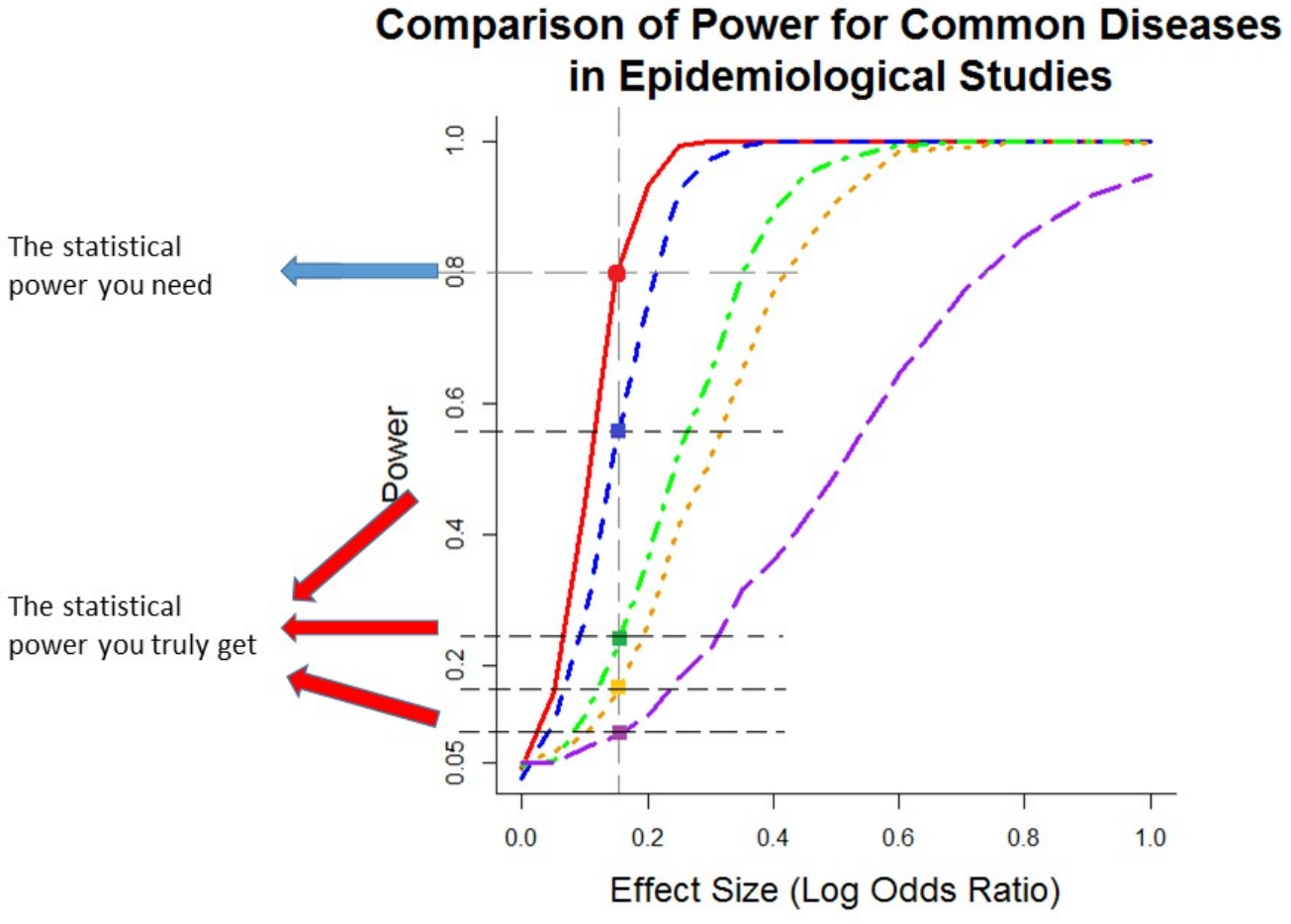
Simulation Results

Epidemiological Association for Smoking Status:



Simulation Results

Epidemiological Association for Smoking Status:



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.

Why is this important

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.

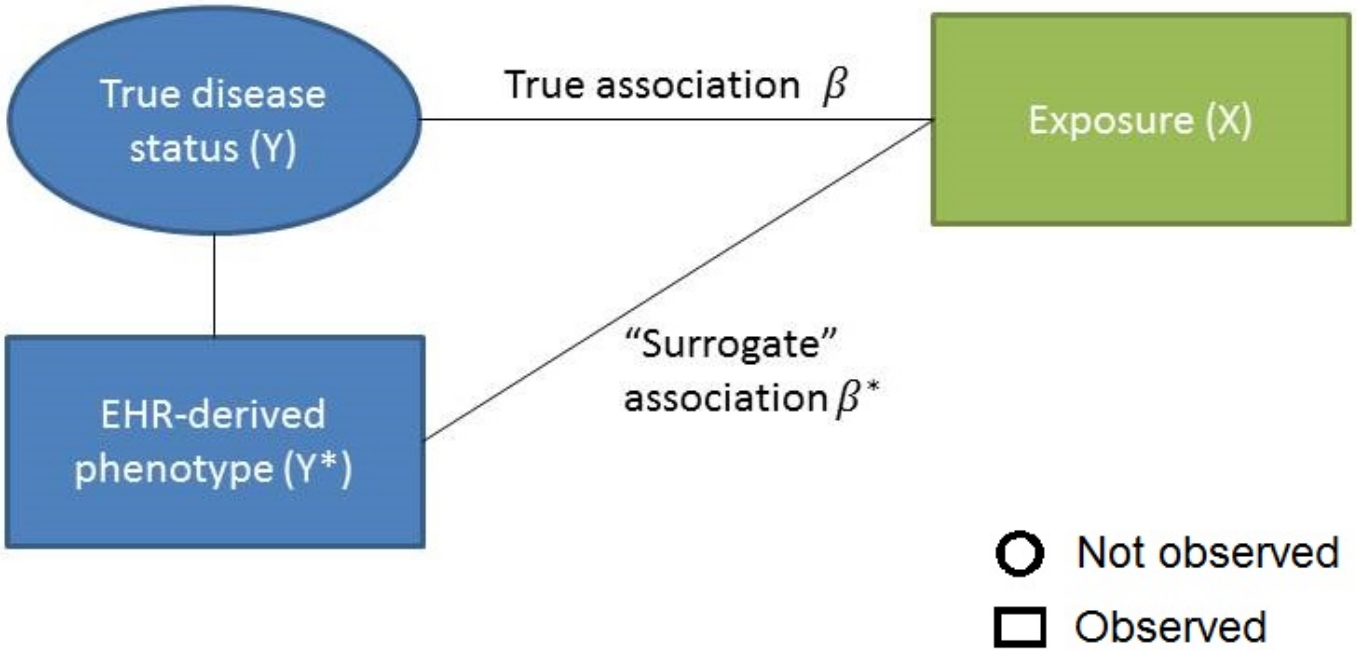
Reference

Duan, R, Cao, M, Wu, Y, Huang, J, Denny, J, Xu, H and Chen, Y, (2016), An Empirical Study for Impacts of Measurement Errors on EHR based Association Studies, AMIA annual symposium proceedings (in press).

(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



$$\beta \neq \beta^*$$

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?

Current Methods in Practice

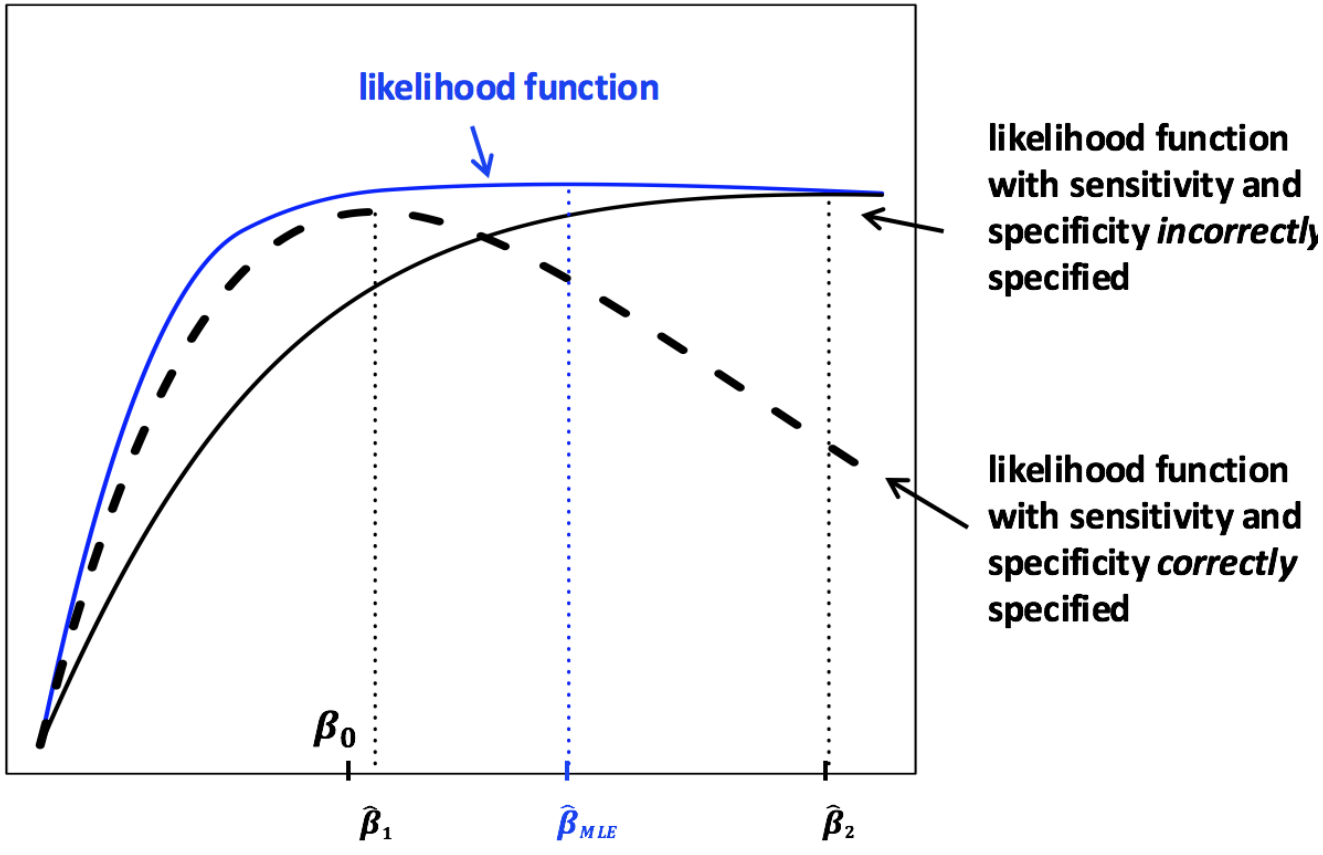
The standard likelihood function: $L(\alpha_1, \alpha_2, \beta)$

- α_1 and α_2 are sensitivity and specificity.
- β 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 $L(\alpha_1 =$ a fixed value, $\alpha_2 =$ a fixed value, $\beta)$.

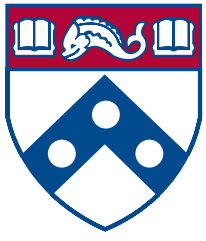
Performance of Existing Methods



Alternative: Bayesian approach

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

More to come !



Perelman
School of Medicine
University of Pennsylvania