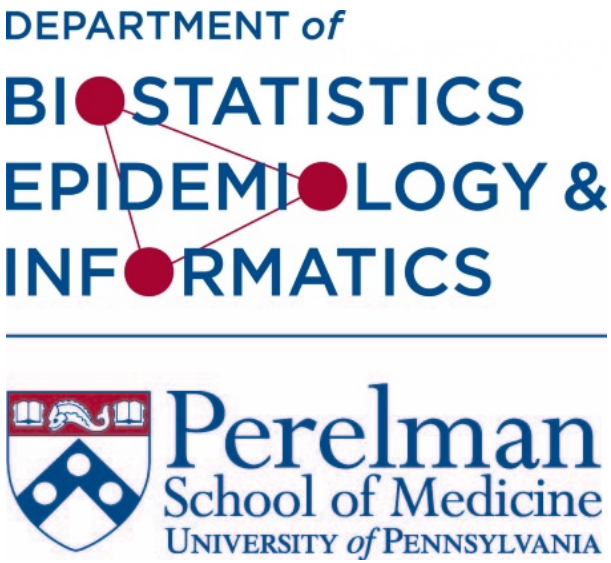
The role of data provenance in the estimation and analysis of EHR-derived phenotypes

Rebecca Hubbard, PhD

rhubb@upenn.edu
https://www.med.upenn.edu/ehr-stats/

April 26, 2019
Great Plains IDeA-CTR



Overview

I: EHR-based Phenotyping

II: Effect of Phenotyping Error on Bias and Type I Error

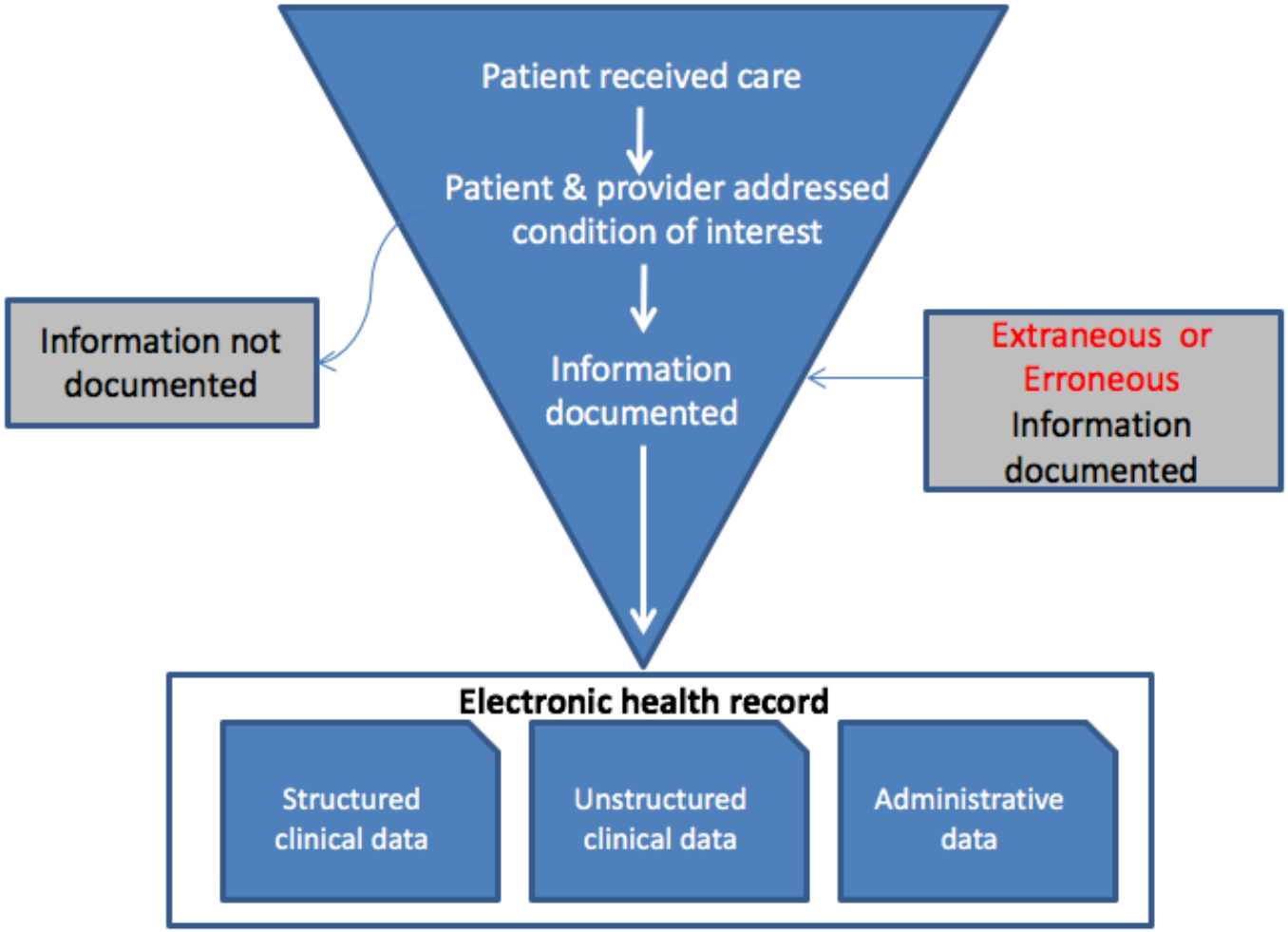
III: Accounting for Phenotyping Error in Analyses

Conclusions

EHR data provenance

- Data provenance refers to the process by which data come to be captured in the EHR
- Unlike data from a designed study, the data capture process in EHR-based studies is entirely outside the control (and often awareness) of the researcher
- Challenging aspects of data provenance for research include
  - Availability, type, and amount of data varies across patients
  - Clinical practices including frequency of visits, data that are recorded, tests that are ordered, etc may vary across clinics

EHR data provenance



Phenotype estimation using EHR data

- Phenotype = collection of characteristics describing a patient
- Motivated by lack of gold-standard for many patient characteristics of interest
- Need ways to deduce characteristics that are not explicitly recorded
- The complexities of data provenance create challenges for phenotyping

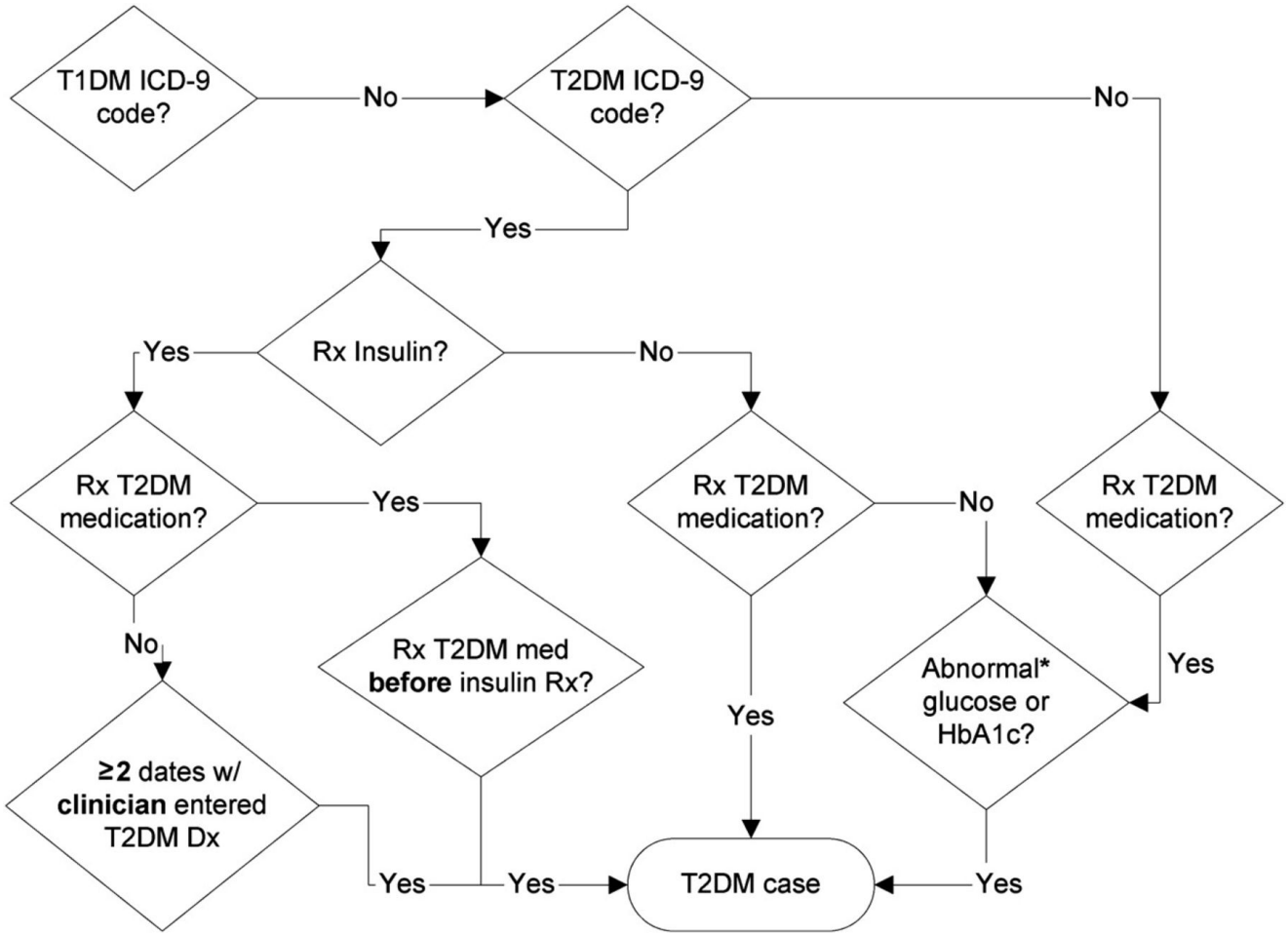
Rule-based Phenotyping

- Most of the existing literature on EHR-derived phenotyping relies on "clinical decision rules"
- Algorithm based on clinical knowledge of the phenotype and coding practices
  - Simple or complex
  - Including one data element or many
  - May include a time component
- May incorporate structured data as well as unstructured data, often via NLP

Example: Rule-based Phenotyping for T2DM

Variable type	Examples	Format
Diabetes diagnosis	• T2DM • T1DM • DM NOS	ICD-9/10 codes
Medications	• Insulin • Metformin	Prescribing data
Co-morbidities	• PCOS • Obesity	ICD-9/10 codes
Biomarkers	• Glucose • HbA1c	Procedure codes for test administration; numerical results

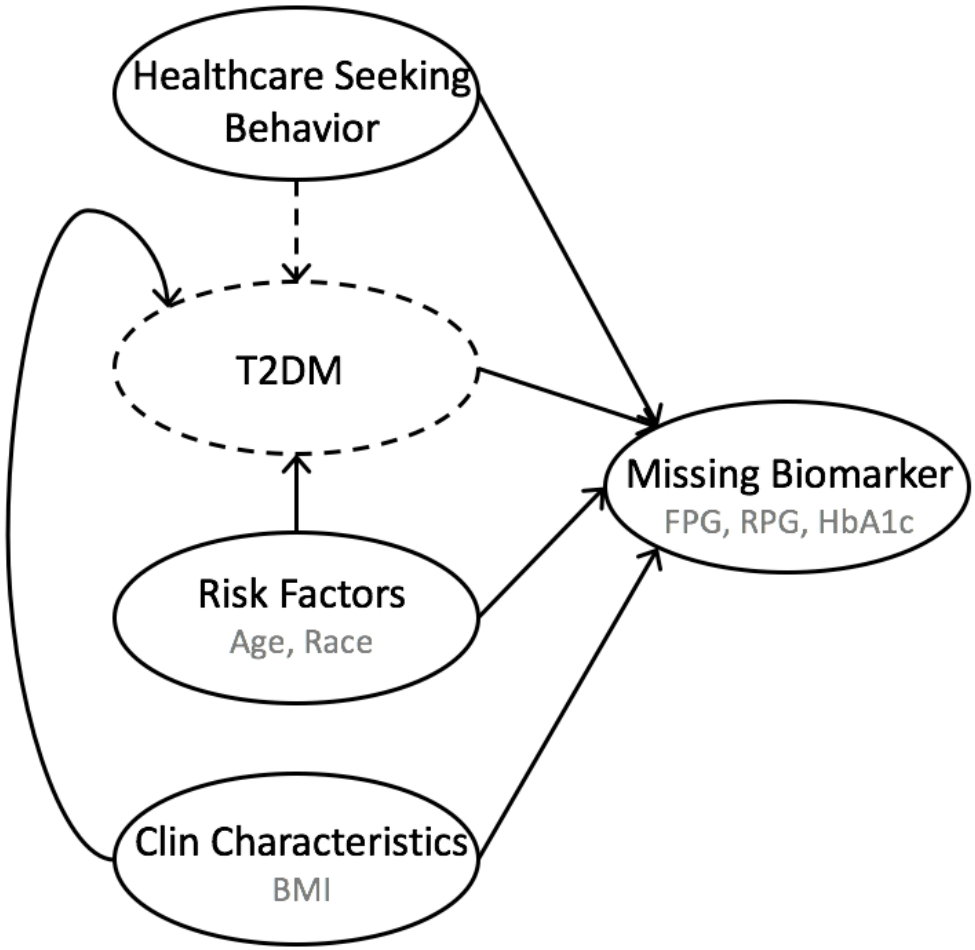
Example: eMERGE T2DM Rule



Kho et al. *J Am Med Inform Assoc* 2012;19:212-218

MNAR missingness mechanism

- Missingness likely depends on underlying T2DM status directly
- Risk factors may influence missingness through T2DM (symptoms) or directly (screening)
- Patients' interaction with the healthcare system also affects observation process
- Example of patient-driven observation



Missingness and rule-based phenotypes

- Typically, rule-based phenotypes have used a naive approach to missingness
- Absence of evidence = evidence of absence
- For conditions where all high risk individuals are evaluated for disease this may be reasonable
- However, it ignores the fact that EHR represent a combination of biological information and information about interaction with health care system

A latent phenotype model

Unobserved true phenotype	$Y_i \sim \text{Bernoulli}(\theta_i)$
Observable features (e.g., codes, medications, biomarkers)	$\boldsymbol{X}_i \sim D(\boldsymbol{\mu}^X_{ik} \mid Y_i = k)$
Missingness in features	$\boldsymbol{R}_i \sim D(\boldsymbol{\mu}^{\boldsymbol{R}}_{ik} \mid Y_i = k)$
Priors for model parameters	$\pi(\theta_i)$, $\pi(\mu^X_{ik})$, etc

$$L(\theta_i, \boldsymbol{\mu}^{\boldsymbol{X}}_i, \boldsymbol{\mu}^{\boldsymbol{R}}_i) = \sum_{k=0,1} P(Y_i = k|\theta_i) \prod_{j=1}^{J} f(R_{ij}|Y_i = k, \boldsymbol{\mu}^{\boldsymbol{R}}_{ik}) f(X_{ij}|Y_i = k, \boldsymbol{\mu}^{\boldsymbol{X}}_{ik})^{R_{ij}}$$

Posterior distribution for $\theta_i|\boldsymbol{X}_i, \boldsymbol{R}_i$ can be used as a measure of the phenotype

Hubbard et al. 2019. A Bayesian latent class approach for EHR-based phenotyping. *Statistics in Medicine*. doi:10.1002/sim.7953.

Why Bayesian estimation?

- Bayesian framework combines strengths of formal statistical prediction and clinical knowledge-base
  - Data can be used to identify patterns of data elements indicative of disease
  - Likelihood incorporates all data elements available for an individual
- Expert opinion on predictive performance of biomarkers incorporated into prior distributions

Application to PEDSnet data

- We applied this approach to an EHR-derived data set from two PEDSnet sites
- Children age 10-18 years, at least two clinical encounters between 2001-2017 separated by at least 3 years
- On at least one occasion BMI z-score in excess of the 95th percentile for age and sex
- Cohort consisted of 32,553 children from site A and 24,342 children from site B

T2DM Predictors in PEDSnet cohort

	Site A	**Site B**
	N = 32,553	N = 24,342
	Mean (SD)	**Mean (SD)**
Random Glucose	95.0 (35.0)	101.8 (44.5)
Hemoglobin A1c	5.8 (1.2)	6.0 (1.4)
	N (%)	**N (%)**
Endocrinologist	2,411 (7.4)	4,617 (19.0)
Metformin	357 (1.1)	1,460 (6.0)
Insulin	360 (1.1)	691 (2.8)
T1D Codes	408 (1.3)	787 (3.2)
T2D Codes	164 (0.5)	365 (1.5)
Missing glucose	6,382 (19.6)	8,204 (33.7)
Missing HbA1c	29,057 (89.3)	18,630 (76.5)
eMERGE T2DM	111 (0.3)	207 (0.9)

Posterior means and CIs for model parameters

	Site A		Site B	
	Posterior Mean	95% CI	Posterior Mean	95% CI
Mean shift in glucose	135.24	(131.21, 139.25)	141.24	(138.87, 143.59)
T2DM code sensitivity	0.20	(0.16, 0.24)	0.26	(0.23, 0.29)
T2DM code specificity	1.00	(1.00, 1.00)	0.99	(0.99, 0.99)
Endocrinologist code sensitivity	0.95	(0.93, 0.97)	0.98	(0.97, 0.99)
Endocrinologist code specificity	**0.94**	**(0.94, 0.94)**	**0.84**	**(0.83, 0.84)**
Metformin code sensitivity	0.29	(0.25, 0.33)	0.33	(0.30, 0.36)
Metformin code specificity	**0.99**	**(0.99, 0.99)**	**0.95**	**(0.95, 0.95)**
OR missing glucose	**0.38**	**(0.31, 0.46)**	**0.20**	**(0.17, 0.23)**

Overview

I: EHR-based Phenotyping

II: Effect of Phenotyping Error on Bias and Type I Error

III: Accounting for Phenotyping Error in Analyses

Conclusions

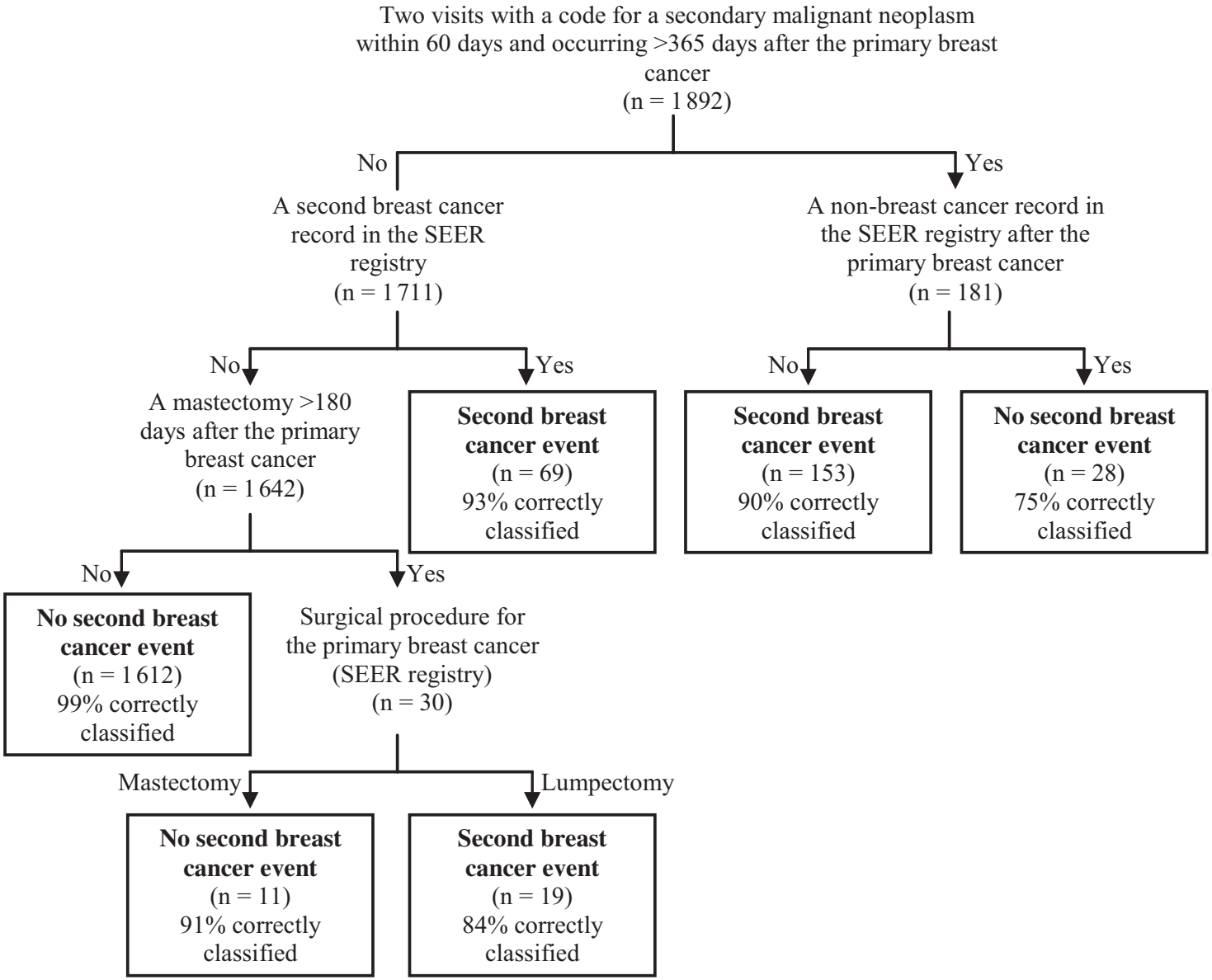
Error in EHR derived phenotypes

- EHRs provide the opportunity to identify novel risk factors
- However, EHR-derived phenotypes may exhibit exposure-dependent differences in data quality
  - More data available for patients with high intensity of contact with healthcare system (higher sensitivity among exposed)
  - High intensity patient also have more opportunity for erroneous codes to appear in charts (lower specificity)
- Example: Second breast cancer event (SBCE) in women with a history of breast cancer
  - Algorithm identifies SBCE with Se = 88%, Sp = 99%
  - Can algorithm be used to identify date of SBCE?
  - What are implications for estimation and hypothesis testing if imperfectly ascertained outcomes are used?

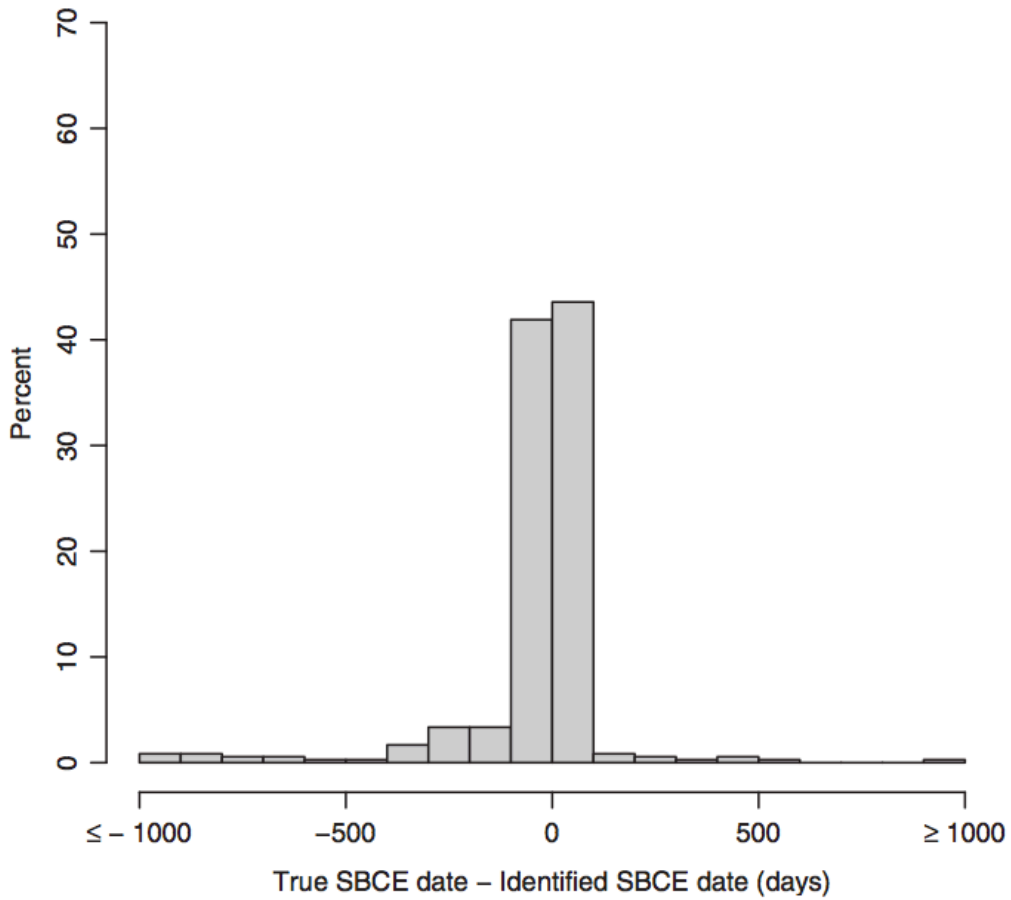
Second breast cancer events

- COMBO study developed algorithm to identify SBCEs using a combination of cancer registry and EHR data
- Validated against manual chart review
- We explored how well dates assigned based on this algorithm agreed with gold-standard
- 407 chart-reviewed SBCEs, 358 (88%) identified by algorithm

High specificity algorithm



Error in date assignment for SBCE



- 82% of events were within 60 days of algorithm-based date
- Is this good enough?

Simulation study for imperfect time to event outcomes

- Conducted a simulation study with event and error rates for dates motivated by SBCE study
- Estimated HRs using imperfectly assigned SBCE dates and compared to true HRs used to simulate data

Sensitivity/specificity	Error in date
Non-differential	Non-differential
Non-differential	Later event detection in exposed
Non-differential	Earlier event detection and less variability
Non-differential	Later event detection and more variability
Higher sensitivity/lower specificity	Non-differential
Higher sensitivity/lower specificity	Earlier event detection and less variability

Simulation study for imperfect time to event outcomes

Sensitivity/specificity	Error in date	% Bias in HR
Non-differential	Non-differential	-2.2
Non-differential	Later event detection in exposed	0.4
Non-differential	Earlier event detection and less variability	-0.9
Non-differential	Later event detection and more variability	-3.8
Higher sensitivity/lower specificity	Non-differential	6.5
Higher sensitivity/lower specificity	Earlier event detection and less variability	8.1

Chubak J et al. 2017. An electronic health record-based algorithm to ascertain the date of second breast cancer events using automated data. *Med Care*. 55(12):e81-e87.

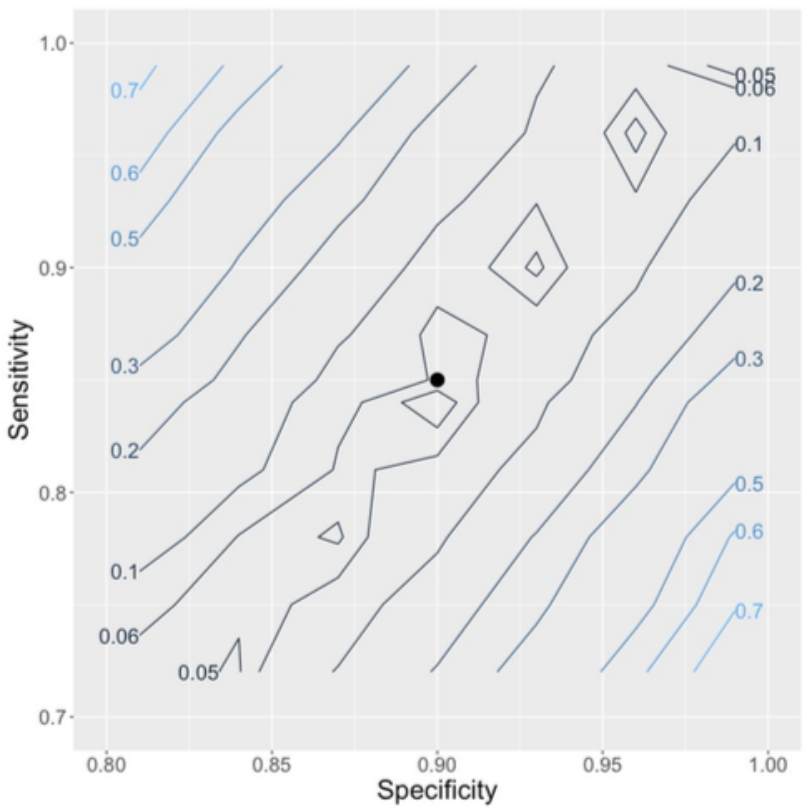
Type I error due to phenotyping error

- In addition to bias, inflated type I error rates are of high importance as they indicate the frequency of spuriously identified risk factors
- Using COMBO data on EHR-derived SBCE and patient and cancer characteristics, we simulated an exposure variable (E) that was independent of the outcome
- However, the sensitivity and specificity of the surrogate outcome (Y^*) varied according to exposure status.

Type I error due to phenotyping error

- We then analyzed the association between Y^* and E using logistic regression
- We varied the difference in sensitivity and specificity between exposed and unexposed across a range of values, with sensitivity in the unexposed fixed at 0.85 and specificity fixed at 0.9.
- Each scenario was repeated 1,000 times
- Type I error was computed as the proportion of hypothesis tests rejected at the $\alpha = 0.05$ level across the 1,000 simulations

Type I error results



- Holding specificity equal in exposed and unexposed individuals, when sensitivity was 10% higher in exposed individuals compared to unexposed (i.e., 0.95 vs 0.85) the type I error rate increased to 14%.
- Similarly, holding sensitivity equal between the two groups, a 10% decrease in specificity between exposed and unexposed individuals (i.e., 0.80 vs 0.90) resulted in a type I error rate of 33%.

Chen Y et al. 2018. Inflation of type I error rates due to differential misclassification in EHR-derived outcomes: Empirical illustration using breast cancer recurrence. *Pharmacoepidemiol Drug Safety*. doi:10.1002/pds.4680.

Overview

I: EHR-based Phenotyping

II: Effect of Phenotyping Error on Bias and Type I Error

III: Accounting for Phenotyping Error in Analyses

Conclusions

What can we do about phenotyping error?

- We have seen that phenotyping error can lead to substantial bias and inflated type I error
- Numerous statistical methods have been developed to account for misclassified outcomes
- Despite this, the vast majority of EHR-based analyses in the applied literature use standard methods with no correction for misclassification

An approach for predicted probabilities

- Increasingly, phenotyping is using statistical or machine learning approaches that provide predicted probabilities of phenotype, $\hat{p}$
- More sophisticated phenotyping allows for covariate-specific phenotypes
- Sinnott et al. 2014 developed a bias correction approach for analyses using these predicted probabilities as outcomes
- Suppose we wish to estimate the association between a phenotype, Y, and exposure, Z adjusting for confounders W

$$g(P(Y = 1|Z, W)) = \alpha + \beta Z + \gamma W.$$

- Let $f(\hat{p}) = (\hat{p} - \mu_0)/(\mu_1 - \mu_0)$, where $\mu_k = E(\hat{p}|Y = k)$
- Sinnott et al. showed that regressing $f(\hat{p})$ on Z and W provides unbiased estimates for regression coefficients.

Sinnott et al. 2014. Improving the power of genetic association tests with imperfect phenotype derived from electronic medical records. *Human Genetics*. 133:1369-82.

A simple bias correction for risk differences

- In the context of logistic regression, this approach requires specialized software.
- In the context of risk difference regression, however, this approach gives rise to a very simple bias correcion

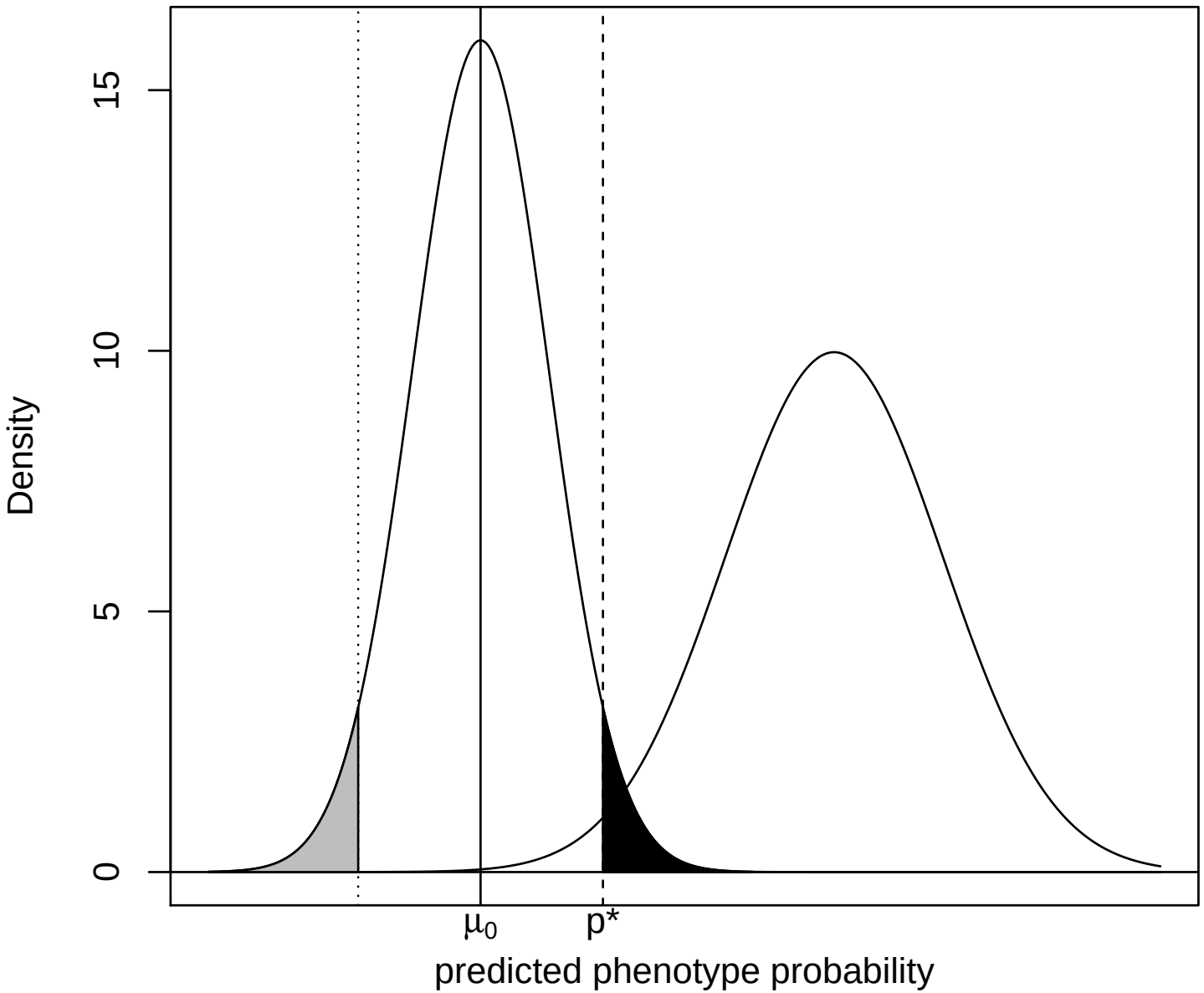
$$\begin{aligned} E(f(\hat{p})|Z,W) &= \alpha + \beta Z + \gamma W \\ E[(\hat{p}-\mu_0)/(\mu_1-\mu_0)|Z,W] &= \alpha + \beta Z + \gamma W \\ E[\hat{p}|Z,W] &= \alpha^* + (\mu_1-\mu_0)(\beta Z + \gamma W) \\ E[\hat{p}|Z,W] &= \alpha^* + \beta^* Z + \gamma^* W \end{aligned}$$

- Therefore, $\hat{\beta} = \frac{\hat{\beta}^*}{\mu_1 - \mu_0}$ is unbiased for β

One additional complication

- Unfortunately, in the EHR context μ_0 and μ_1 will only be available in data sets with validation data
- In the data set initially used to develop the phenotype this will be straightforward to calculate by taking the mean of $\hat{p}$ among cases and controls
- In data sets without validation data we typically have access to published validation results, typically including a proposed cutpoint, p^*, along with sensitivity and specificity for the dichotomized phenotype
- Using this information we can obtain estimates $\hat{\mu}_0$ and $\hat{\mu}_1$

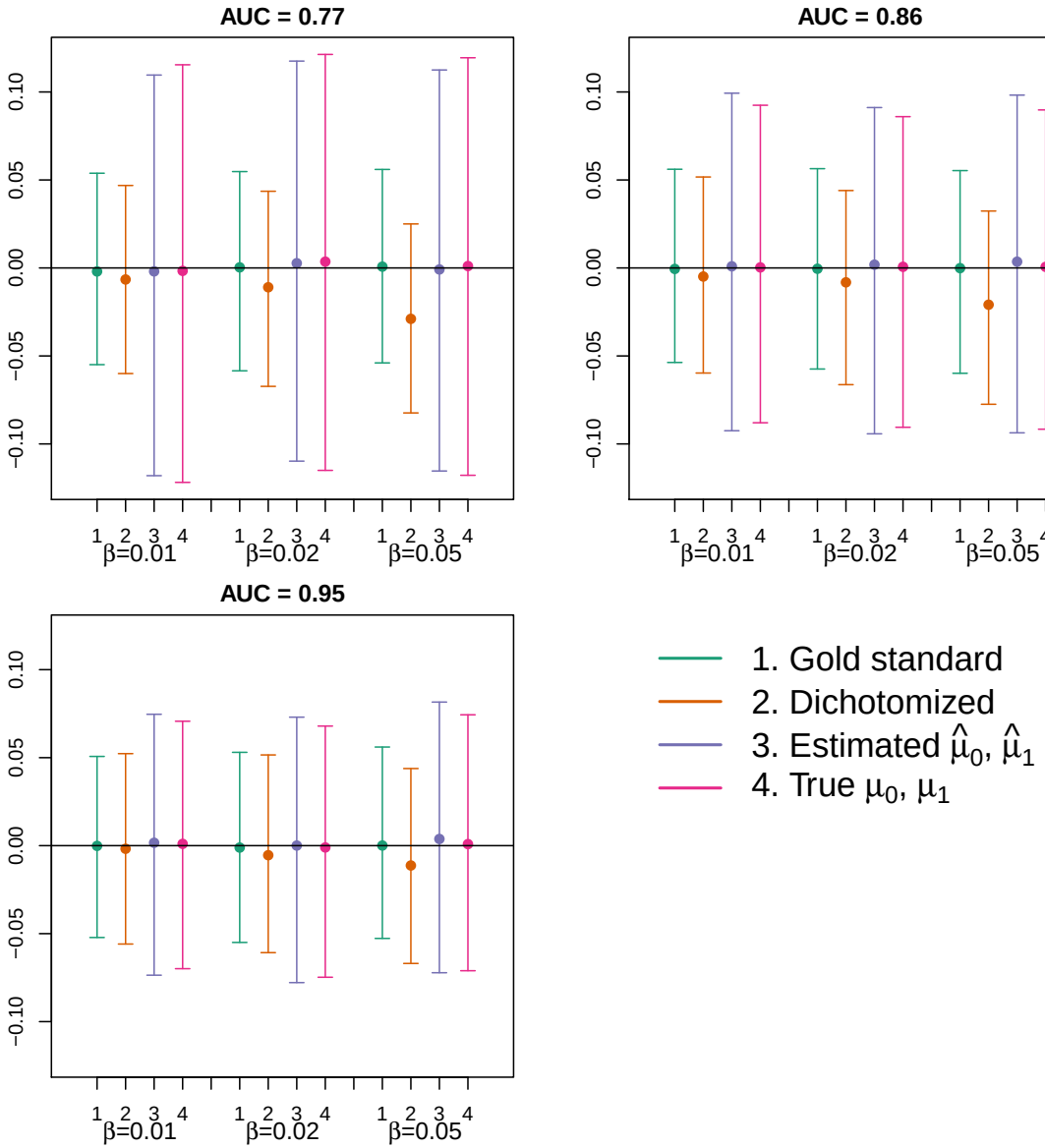
Estimating μ_0 without validation data



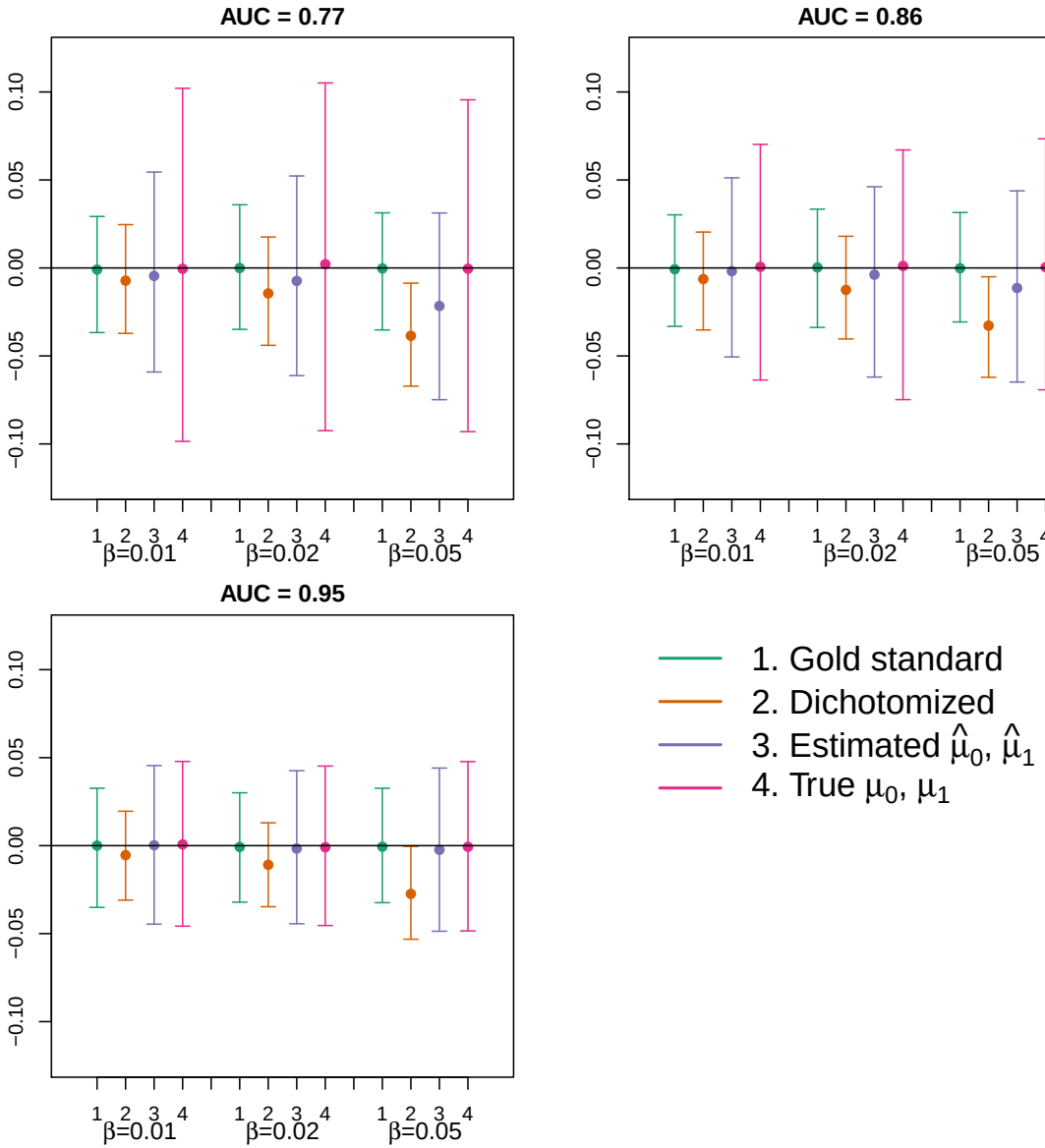
Simulation study design

- Compared
  1. Gold standard true phenotype
  2. Dichotomized phenotype based on predicted probability
  3. Bias correction using estimated $\hat{\mu}_0$ and $\hat{\mu}_1$
  4. Bias correction using true μ_0 and μ_1
- Varying: AUC of $\hat{p}$, strength of effect (β), prevalence of Y

Bias: Prevalence = 0.5



Bias: Prevalence = 0.1



Overview

I: EHR-based Phenotyping

II: Effect of Phenotyping Error on Bias and Type I Error

III: Accounting for Phenotyping Error in Analyses

Conclusions

Conclusions

- Consideration of data provenance is critical to appropriate development and analysis of phenotypes
- Efforts should be made to improve phenotypes
  - Consider routine practice for how patients are treated and how frequently
  - Don't assume phenotypes are transportable across clinical sites
  - Incorporate information on intensity of interaction with healthcare system
- Phenotyping error can result in substantial bias and type I error
- A variety of approaches exist to account for phenotyping error or conduct sensitivity analyses to determine if results are robust

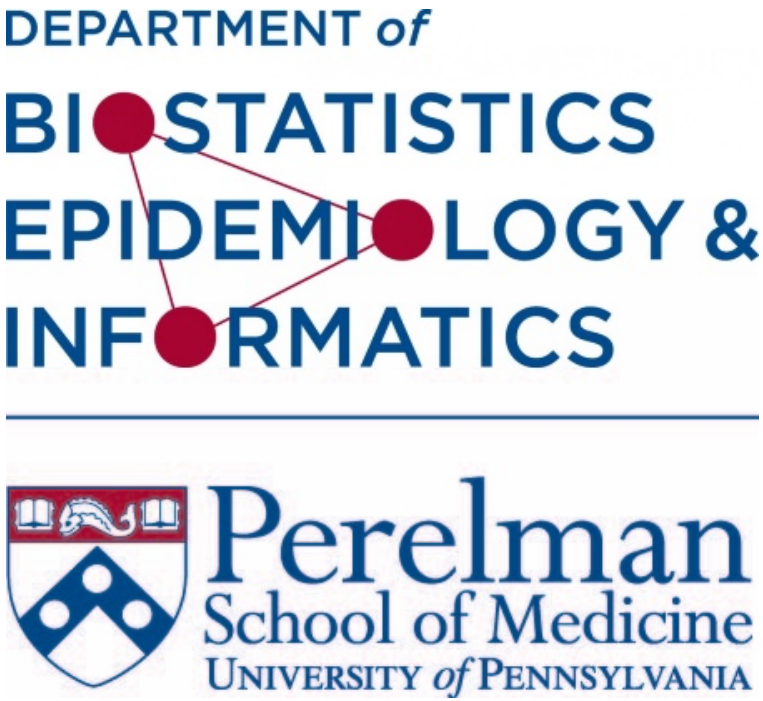
References

1. Hubbard RA, Huang J, Harton J, Oganisian A, Choi G, Utidjian L, Eneli I, Bailey LC, Chen Y. 2019. A Bayesian latent class approach for EHR-based phenotyping. *Statistics in Medicine*. doi:10.1002/sim.7953.
2. Chubak J, Hubbard RA, Zhu W, Buist DSM, Onega T. 2017. An electronic health record-based algorithm to ascertain the date of second breast cancer events using automated data. *Medical Care*. 55(12):e81-e87.
3. Chen Y, Wang J, Chubak J, Hubbard RA. 2018. Inflation of type I error rates due to differential misclassification in EHR-derived outcomes: Empirical illustration using breast cancer recurrence. *Pharmacoepidemiology & Drug Safety*. doi:10.1002/pds.4680.
4. Sinnott JA, Dai W, Liao KP, Shaw SY, Ananthakrishnan AN, Gainer VS, Karlson EW, Churchill S, Szolovits P, Murphy S, Kohane I. 2014. Improving the power of genetic association tests with imperfect phenotype derived from electronic medical records. *Human Genetics*. 133:1369-82.

Acknowledgments

- Jinbo Chen
- Yong Chen
- Grace Choi
- Jessica Chubak
- Joanna Harton
- Jing Huang
- Arman Oganisian
- Jianqiao Wang
- Weiwei Zhu

This work was partially supported through a Patient-Centered Outcomes Research Institute (PCORI) Award (ME-1511-32666).
All statements in this presentation, including its findings and conclusions, are solely those of the authors and do not necessarily represent the views of PCORI, its Board of Governors or Methodology Committee.


DEPARTMENT of
BIOSTATISTICS
EPIDEMIOLOGY &
INFORMATICS
Perelman
School of Medicine
University of Pennsylvania