COST-EFFECTIVE CHART REVIEW SAMPLING DESIGN TO ACCOUNT FOR PHENOTYPING ERROR IN ELECTRONIC HEALTH RECORDS (EHR) DATA

Xinyang Feng, Hyung Eun Shin

Dec 06, 2022

UNIVERSITY OF TORONTO

CONTENTS

1. INTRO
2. METHODS
3. RESULTS
4. CONCLUSION

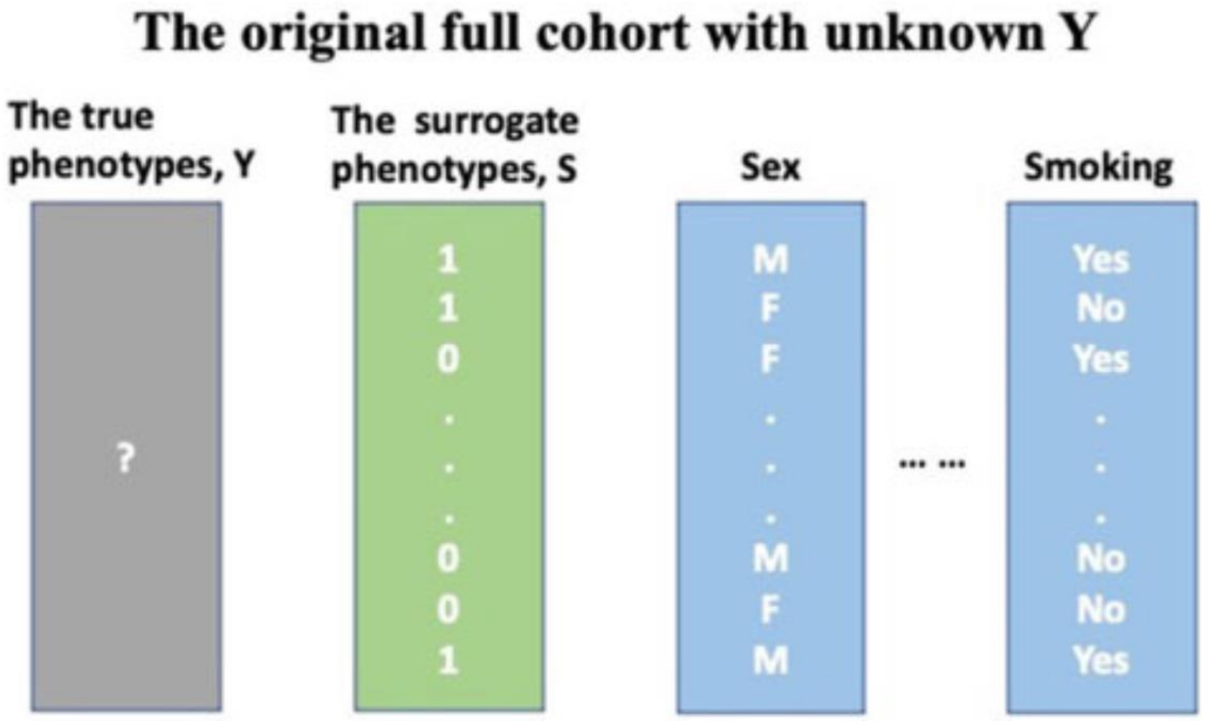
INTRO

Introduction and Motivation

- Electronic health record (EHR) data
  - A digital data system that contains medical history of patients
  - Investigates the association between risk factors and EHR-derived phenotype (binary)
- Problems:
  - EHR-derived phenotypes are likely to be misclassified; a manual chart review is required
  - Chart review can only be conducted for a validation subcohort due to limited resources
  - For rare diseases, positive phenotype patients in full cohort is even fewer
    – chart review may loss efficiency
- Research goal:

  investigate an efficient sampling procedure for association study when there is **a small number of positive phenotype subjects** in the full cohort

EHR Data



The original full cohort with unknown Y

The true phenotypes, Y	The surrogate phenotypes, S	Sex		Smoking
	1	M		Yes
	1	F		No
	0	F		Yes
	.	.		.
?	.	.	… …	.
	.	.		.
	0	M		No
	0	F		No
	1	M		Yes

Relationship between Y and X: $logit\{P(Y_i = 1|\boldsymbol{X_i})\} = \boldsymbol{X_i\beta}$

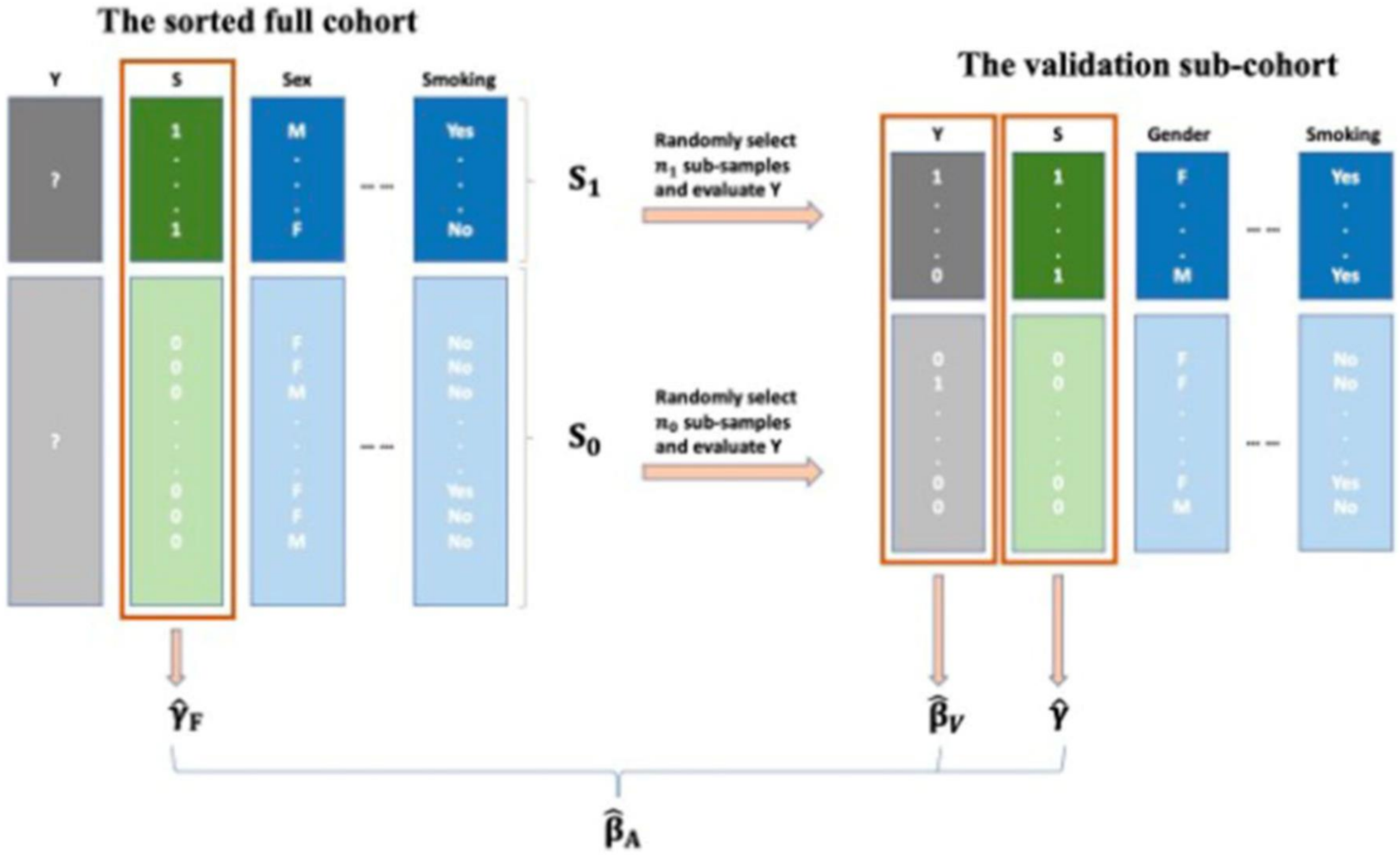
Previous Works

- 1. Original uniform (Ori-Unif) sampling
  - Cannot guarantee a balanced number of positive and negative phenotype patients in the subcohort
- 2. Original biased (Ori-Bias) sampling
  - Only explores the relationship between true phenotype and covariates but fails to take information in surrogate phenotype into account.
- 3. Augmented uniform (Aug-Unif) sampling
  - Introduces information in surrogate phenotype into the estimator
  - But uses a uniform sampling procedure

METHODS

Proposed Method

- Outcome-dependent sampling design for cost-effective chart review with augmented estimation procedure (OSCA)



OSCA

- Assumption: surrogate phenotype and the true phenotype are non-differentially associated
  - Fixed probabilities of making a correct chart review conditional on the observed phenotype

$$p_1 = P(S_i = 1 | Y_i = 1) \text{ and } p_0 = P(S_i = 0 | Y_i = 0).$$

- Improvement of OSCA
  - Fewer subjects are required for chart review to achieve the same statistical power
  - Smaller estimated standard error; higher statistical efficiency

RESULTS

RESULTS

- 3 Simulations

  →Own synthetic data sets

- 1 real data set
  - Real data: Colon cancer recurrence EHR data set in the KPW healthcare system
  - Proved closer estimates to the golden standard

SIMULATION1: DATA GENERATION

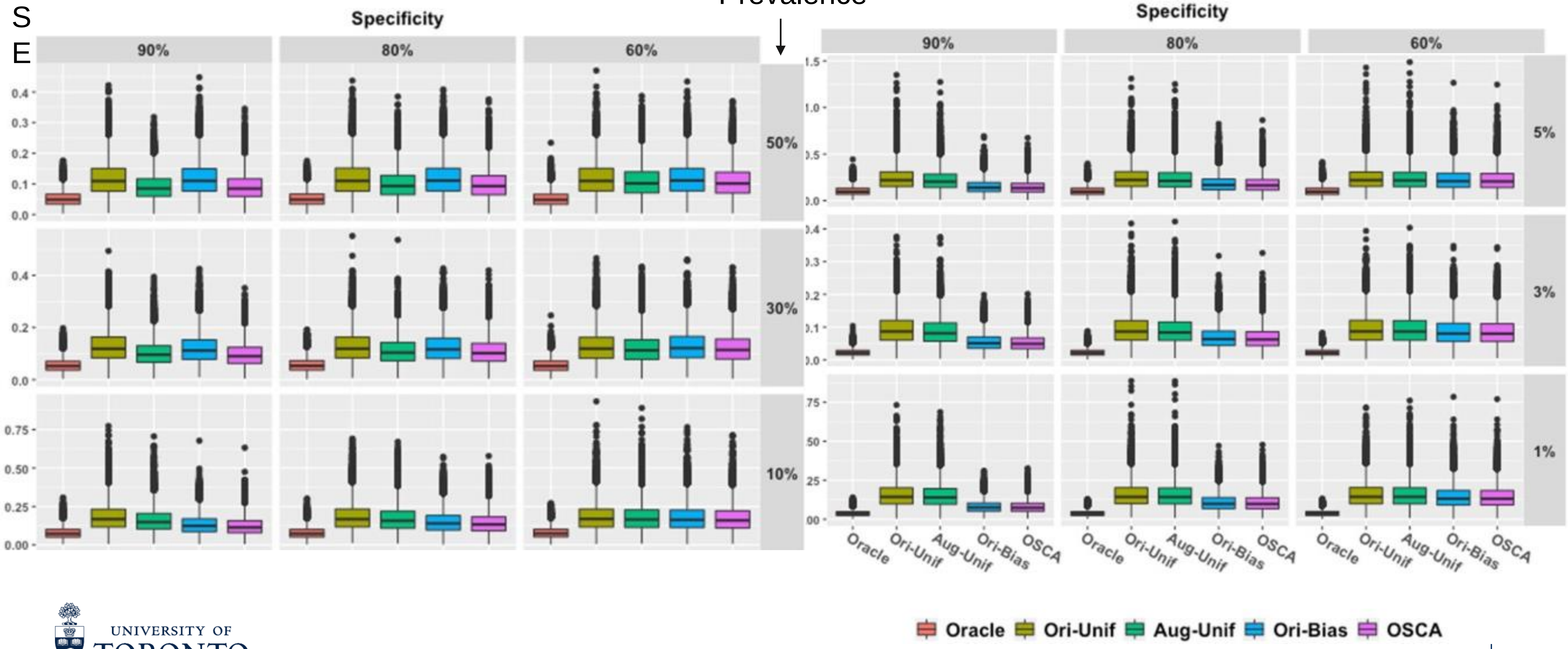
EMPIRICAL COVERAGE PROBABILITIES AND CONFIDENCE INTERVALS

- $\mathrm{logit}\{P(Y_i = 1|X_1, X_2, X_3)\} = b_0 + X_1 + X_2 + X_3$
- Prevalence: {~1%, ~3%, ~5%, ~10%, ~30%, ~50%}
    - For ~5%, ~10%, ~30%, ~50%: N=2000 , n=600
    - For ~1%, ~3%: N=8000 , n=2000

SIMULATION 1: RESULTS

M
S
E

Prevalence



UNIVERSITY OF
TORONTO

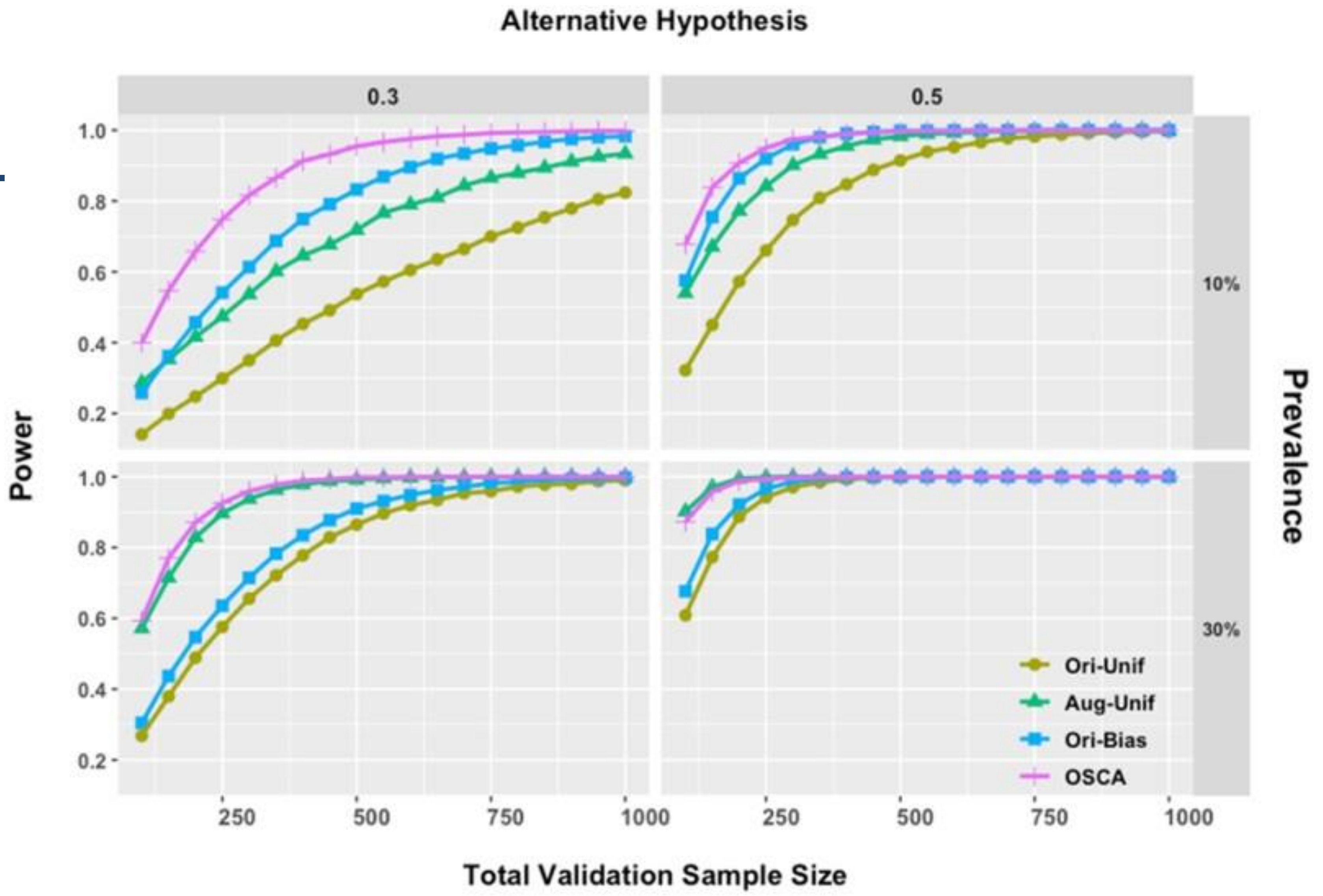
SIMULATION 2:

POWER ANALYSIS

- $\text{logit}\{P(Y_i = 1|X)\} = b_0 + \beta_1 X$

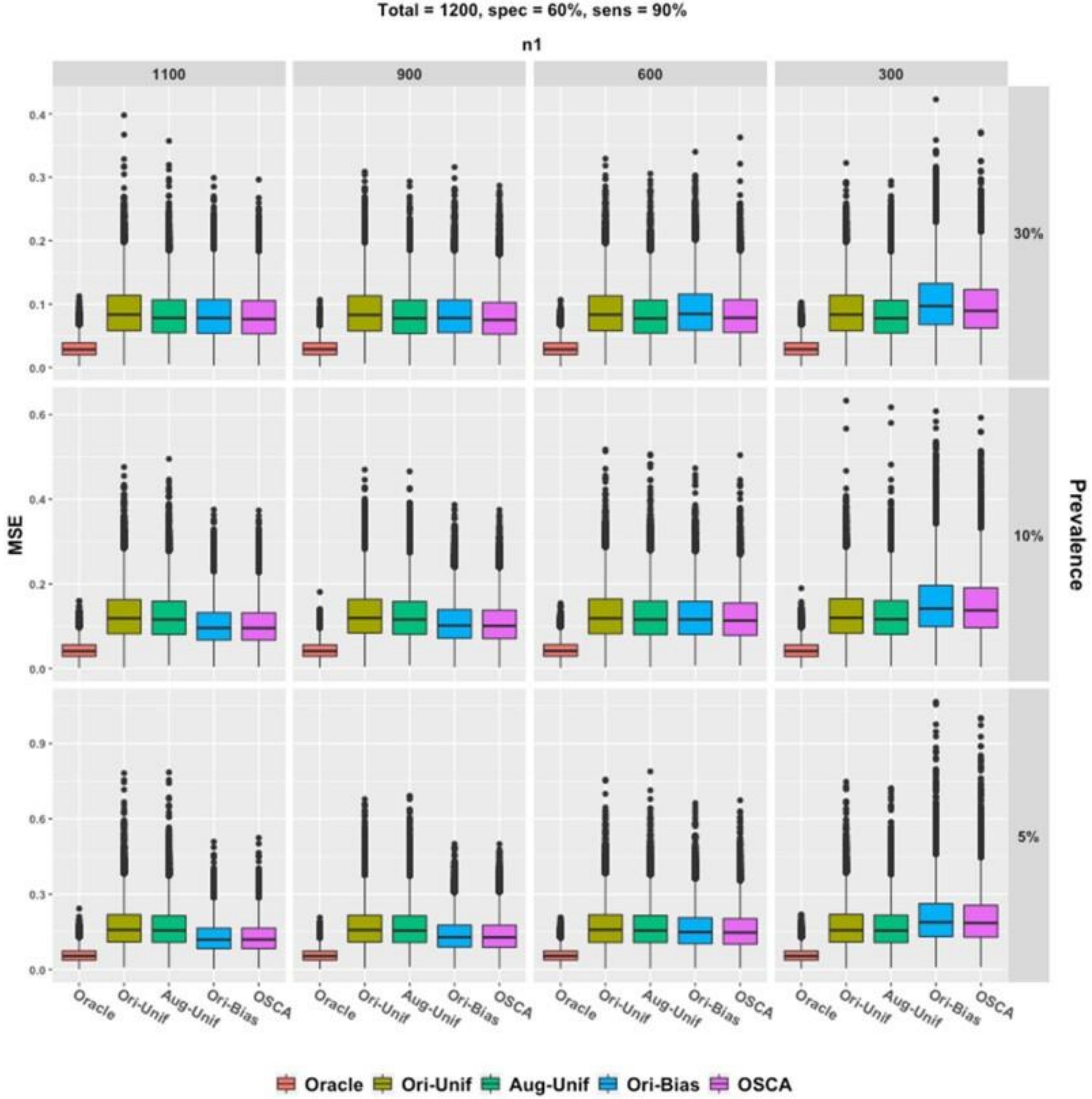
- $H_0\colon\ \beta_1 = 0$

- $H_1\colon \beta_1$ =0.3 or 0.5



SIMULATION 3:

EFFECT OF IMBALANCED SAMPLING



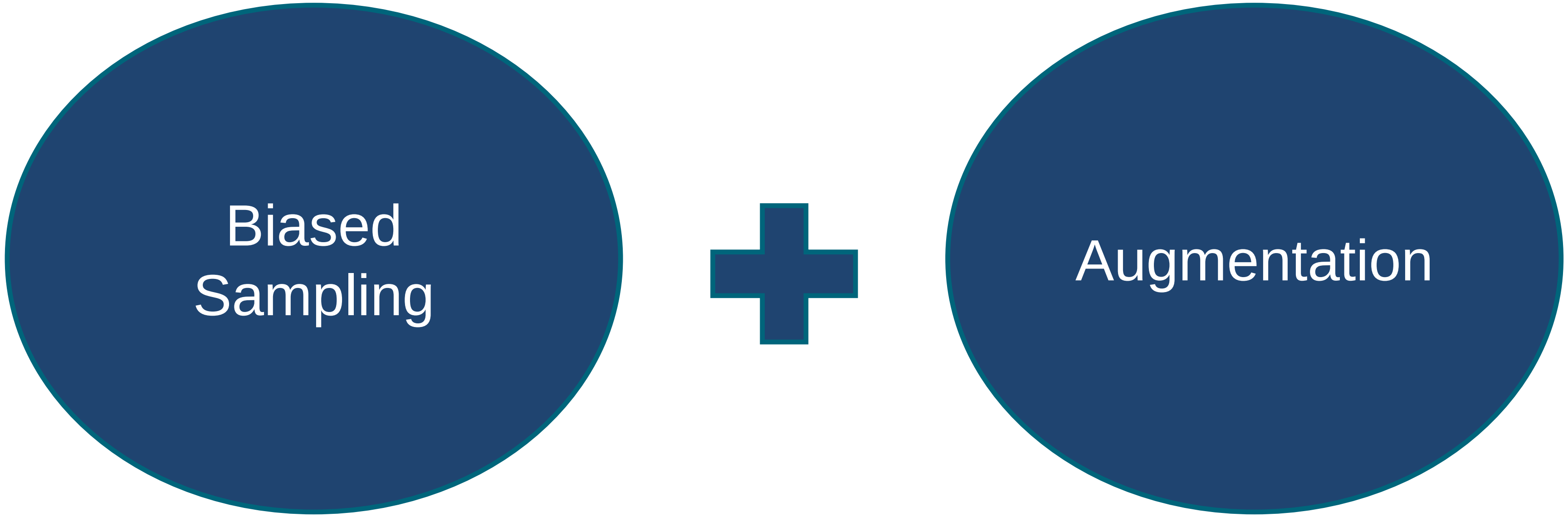
CONCLUSION

PROBLEMS?

- EHR data: an electronic medical records of patients
- phenotyping error in EHR-derived outcomes
  - Systematic bias
- a manual chart review is required
  - Time-consuming

OSCA

the outcome-dependent sampling design for cost-effective chart review with augmented estimation procedure



Biased Sampling

Augmentation

CONCLUSION

- Simulation 1: concentrated MSE box especially when disease is rare
- Simulation 2: Need smaller sub-cohort to reach the same level power
- Simulation 3: validated effectiveness of biased sampling
- In conclusion
    - reduces estimation bias while maintaining low variance
    - Is cost-effective in chart review.

Thank You!