Penn Medicine

An Augmented Estimation Procedure for EHR-based Association Studies with Multiple Surrogate Outcomes

Yiwen Lu[1,2,*], Jiayi Tong[1,*], Rebecca A Hubbard[1], Yong Chen[1]

[1]Perelman School of Medicine, The University of Pennsylvania, Philadelphia, PA, USA [2]Applied Mathematics and Computational Science, The University of Pennsylvania, Philadelphia, PA, USA *Contributed equally

Background

Objective

- Use multiple surrogates to conduct augmented procedure

Why do we study it?

- Manual chart review are often constrained by time and cost limitations[1-3]
- Cannot get true estimation directly

Existing methods

Method 1	use validation set only	Inefficient
Method 2	use full surrogate only	Biased
Method 3 (Tong2019)	Augmented estimation with single surrogate[4]	Unbiased but we want more efficiency

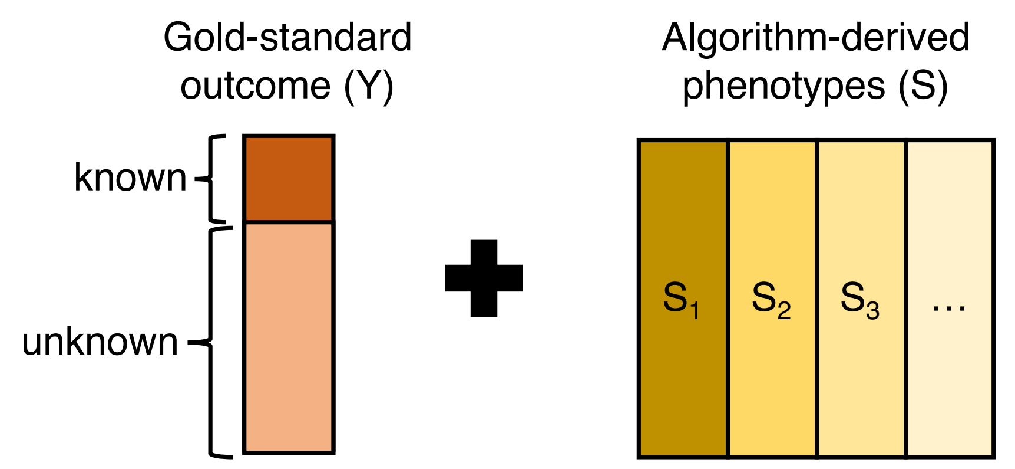
Proposed Method

Algorithm

1. Obtain
   (a) $\widehat{\boldsymbol{\beta}}_V$ using validation set
   (b) $\widehat{\gamma}_F^k, \widehat{\gamma}_V^k$ using full set and validation part of k-th surrogate respectively
2. Compute covariance matrices $\Omega, \boldsymbol{\Sigma}, \boldsymbol{\Sigma}^*$ of $\hat{\boldsymbol{\beta}}_V - \boldsymbol{\beta}_1$ joint with $\widehat{\gamma}_V - \widehat{\gamma}_F$.
3. Obtain the proposed augmented estimator $\widehat{\boldsymbol{\beta}}_{AM}$ by

$$\widehat{\boldsymbol{\beta}}_{AM} = \hat{\boldsymbol{\beta}}_V - \widehat{\Omega}^T \hat{\Sigma}^{*-1} (\widehat{\gamma}_V - \widehat{\boldsymbol{\gamma}}_F)$$

Proposed method
Augmented estimation with all surrogates available

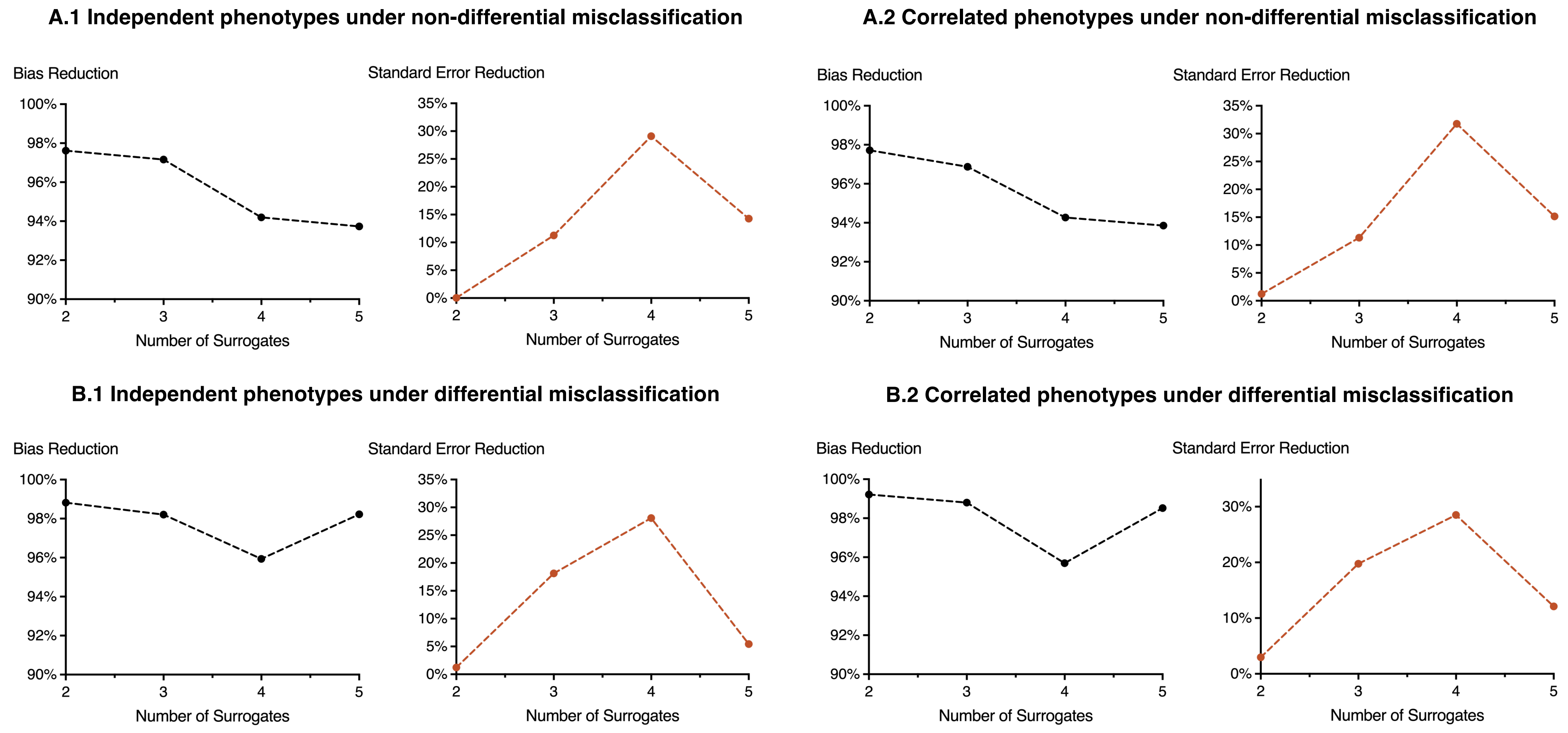


Unbiased and most efficient

Result

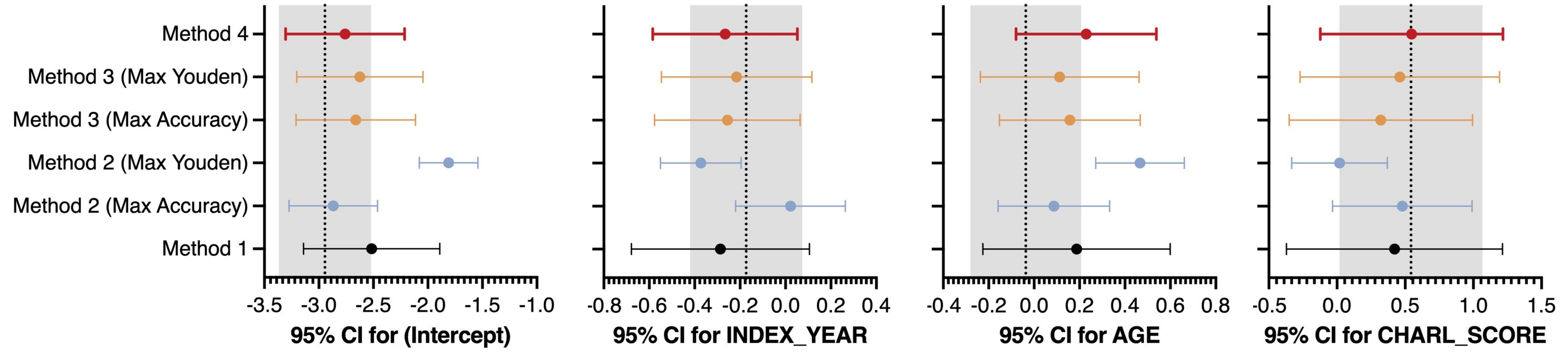
Simulation study

Two settings
Non-differential/differential misclassification
Two cases
Independent/correlated surrogates
Validation ratio range
0.04-0.4
Number of surrogates
2-5



Real data evaluation

Colon cancer in the Kaiser Permanente Washington (KPW) healthcare system



References
[1] Williamson T, Green ME, Birtwhistle R, Khan S, Garies S, Wong ST, et al. Validating the 8 CPCSSN Case Definitions for Chronic Disease Surveillance in a Primary Care Database of Electronic Health Records. The Annals of Family Medicine. 2014 Jul 1;12(4):367–72.
[2] Inacio MCS, Paxton EW, Chen Y, Harris J, Eck E, Barnes S, et al. Leveraging Electronic Medical Records for Surveillance of Surgical Site Infection in a Total Joint Replacement Population. Infect Control Hosp Epidemiol. 2011 Apr;32(4):351–9.
[3 Tian TY, Zlateva I, Anderson DR. Using electronic health records data to identify patients with chronic pain in a primary care setting. J Am Med Inform Assoc. 2013 Dec;20(e2): e275–80.
[4] Tong J, Huang J, Chubak J, Wang X, Moore JH, Hubbard RA, et al. An augmented estimation procedure for EHR-based association studies accounting for differential misclassification. Journal of the American Medical Informatics Association. 2020 Feb 1;27(2):244–53.

Contact: yiwenlu@sas.upenn.edu