Enhancing Precision and Validity: Leveraging Multiple Error-Prone Phenotypes in EHR-Based Association Studies

Yiwen Lu[1,2,*], Jiayi Tong[1,*], Rebecca A Hubbard[1], Yong Chen[1,3,4,5]

[1] Center for Health Analytics and Synthesis of Evidence (CHASE), Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[2] The Graduate Group in Applied Mathematics and Computational Science, School of Arts and Sciences, University of Pennsylvania, Philadelphia, PA, USA
[3] Leonard Davis Institute of Health Economics, Philadelphia, PA, USA
[4] Penn Medicine Center for Evidence-based Practice (CEP), Philadelphia, PA, USA
[5] Penn Institute for Biomedical Informatics (IBI), Philadelphia, PA, USA

[*]: Equally contributed

Background

With the increasing availability of electronic health record (EHR) data, association studies leveraging large-scale EHR data have gained prominence in epidemiological research. A key aspect of utilizing EHR data is the development of computable phenotypes, which are measurable or observable characteristics or conditions derived from EHRs. Over the past decade, significant efforts have been devoted to deriving computable phenotyping algorithms for various conditions and characteristics[1,2]. However, the performance of these algorithms can vary across different study populations. Neglecting the presence of phenotyping error in EHR-derived phenotypes can lead to inflated type I error rates and reduced statistical power, ultimately impacting the reproducibility of EHR-based findings. This issue has been observed in our earlier studies[3,4], highlighting the need to mitigate phenotyping errors in EHR-based research.

One effective approach to mitigate phenotyping errors is manual chart review to verify the true disease status. As demonstrated in studies by Williamson et al.[5], Inacio et al.[6], and Tian et al.[7], manual chart review involves thoroughly examining medical records to ascertain disease status. However, manual chart review is often constrained by time and cost limitations, making it feasible to review only a limited subset of patients' EHRs. Despite this limitation, it is crucial to leverage the validated phenotype obtained from the small subset of patients to address potential bias and improve the accuracy of phenotyping in the broader population.

To tackle this issue, Tong et al.[8] proposed a simple yet effective bias-correction procedure that optimally combines the manual chart reviews (available for a subset of patients) and computable phenotypes derived from a phenotyping algorithm (available for all patients). This procedure yields an augmented estimator with asymptotically zero bias and enhanced statistical efficiency compared to the validation set alone. However, with the increasing availability of multiple phenotyping algorithms, such as those based on ICD codes alone, ICD codes + specialty care, or ICD codes + medications[9-11], relying solely on a single phenotyping algorithm is limiting as it fails to fully leverage the diverse phenotypes derived from all available algorithms.

This study presents a novel bias-correction approach that harnesses the collective power of all available phenotypes derived from multiple algorithms. By simultaneously incorporating these diverse phenotypes,

our proposed procedure maintains asymptotic unbiasedness and enhances the estimator's statistical efficiency while saving the time and cost consumption of the manual chart review. It represents a significant improvement over the previous state-of-the-art method introduced by Tong et al.[8], which solely relies on a single phenotyping algorithm. We rigorously evaluate the performance of our approach through extensive simulation studies and a colon cancer data analysis, considering both non-differential and differential misclassification scenarios. Our findings demonstrate the effectiveness and robustness of our method in addressing the challenges posed by phenotyping error. With the potential to yield more reliable and precise associations, our proposed approach offers valuable insights into population-based research using EHR data.

Methods

The proposed estimator (method 4, denoted as $\widehat{\boldsymbol{\beta}}_{AM}$) incorporates the information from an increasing number of multiple surrogates compared with the work of Tong et al. (method 3). and guarantees higher efficiency. In practice, we first obtain $\widehat{\boldsymbol{\beta}}_{V}$ using the validation set (method 1). Then, for each surrogate outcome, an estimator $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{F}}^{\boldsymbol{k}}$, can be estimated (method 2). We further obtain an estimator $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{V}}^{\boldsymbol{k}}$ using the k-th surrogate in the validation data set as the bridging estimator to obtain the proposed estimator $\widehat{\boldsymbol{\beta}}_{\boldsymbol{AM}} = \hat{\boldsymbol{\beta}}_{\boldsymbol{V}} - \widehat{\Omega}^{T}\hat{\Sigma}^{*-1}(\widehat{\boldsymbol{\gamma}}_{\boldsymbol{V}} - \widehat{\boldsymbol{\gamma}}_{\boldsymbol{F}})$.

We conducted simulation studies to demonstrate the performance of the proposed method by comparing it with the existing methods (i.e., Methods 1-3). The simulation studies had two major settings: nondifferential (Setting A) and differential (Setting B) misclassifications. Under each setting, we simulated the cases where two to five are independent (Case 1) or correlated (Case 2).

For all settings, we set the intercept of the logistic regression model to be -0.5 and the true value of the association to be 0.5. The binary true phenotype Y was generated using covariates $\boldsymbol{X}$ and association parameters (i.e., $\boldsymbol{Y}$ ~ Bernoulli $(-0.5 + 0.1\boldsymbol{X_1} + \boldsymbol{0.5X_2})$), where $\boldsymbol{X_1}$ is a continuous variable and a binary variable $\boldsymbol{X_2}$ is treated as the exposure of interest. For the case where the surrogates were independent, the surrogate outcomes were generated from Y and $\boldsymbol{X}$ based on specified values of sensitivity and specificity. For the case where the surrogates were correlated, they were generated sequentially. In the nondifferential misclassification setting the sensitivity was 0.9, and the specificity was 0.95 for all exposure levels. In the differential misclassification setting, the sensitivity was 0.9, and the specificity was 0.95 for the non-exposure group; the sensitivity and specificity for the exposure group are 0.95 and 0.9, respectively.

To demonstrate the applicability of our method, we applied our method to analyze a colon cancer dataset collected from the Kaiser Permanente Washington (KPW) healthcare system, containing 1063 patients aged 18 or older at the time of diagnosis of stage I-IIIA colon cancer between 1995 and 2014. We included two algorithm-derived phenotypes with different cutoff criteria: RECUR_MAX_ACCURACY (sensitivity=0.725, specificity=0.979) and RECUR_MAX_YOUDON (sensitivity=0.870, specificity=0.899). We compared the point estimates and 95% confidence intervals (CI) for the association (in log odds ratio scale) between covariates and the occurrence of colon cancer events using different methods, where the

validation ratio is approximately 0.3.

Results

The simulation results under different settings are presented in Figure 1. In each subfigure, the first plot shows the magnitude of relative bias reduction compared to using Method 2 with a single surrogate, and the second plot represents the magnitude of standard error reduction compared to using Method 3 with a single surrogate.

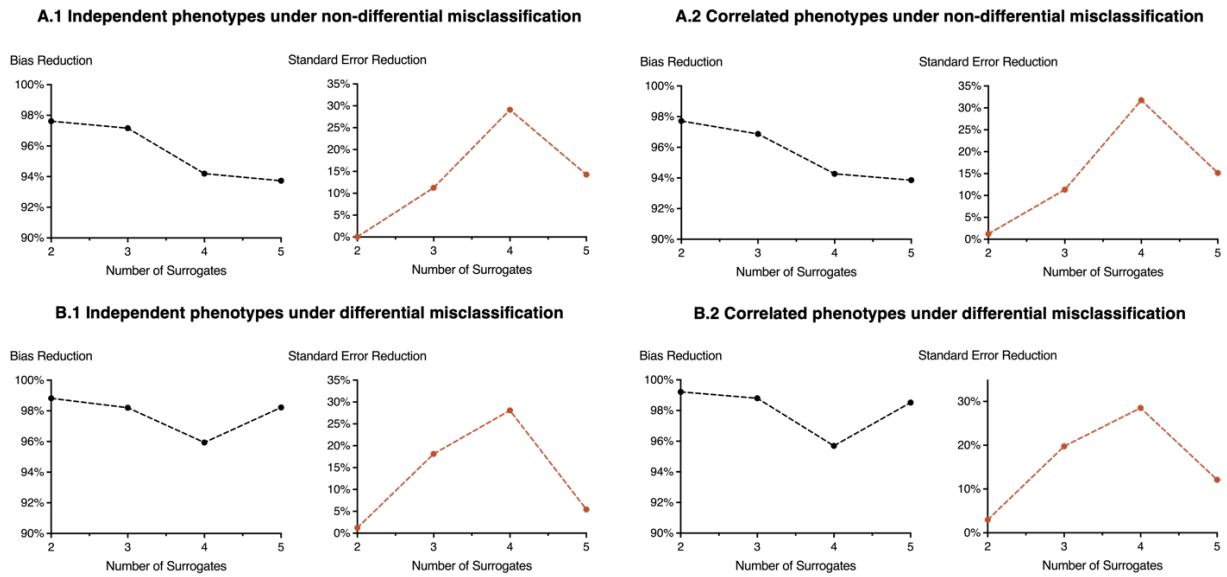

Figure 1. Simulation results of independent (Case 1) and correlated (Case 2) phenotypes with non-differential (Setting A) and differential (Setting B) misclassification. The full sample size of the presenting results is 10000 and the validation set size is 2000.

Given the time and cost constraints of chart reviews, our evaluation of the proposed method was carried out within a specific dataset from the Kaiser Permanente Washington (KPW) healthcare system. This dataset is particularly valuable as it includes a comprehensive gold standard of recurrence flag information for all individuals, enabling a robust comparison between our results and ground truth. Nevertheless, it is noteworthy that our proposed method is designed to accept generic input, hence it is versatile and suitable for application to diverse datasets, including those in the OHDSI framework. The code to implement the proposed method can be found in https://github.com/yiwenluiris/Multiple-Surrogate-Estimation-Sample-Code. Figure 2 presents the results of Method 1-4 applied to the real-world colon cancer data. We observe that the proposed method (Method 4, red solid line) provides a similar estimate to that based on the gold standard in the full sample (vertical dashed line) while reducing uncertainty compared to Method 3 (orange lines) using a single surrogate.

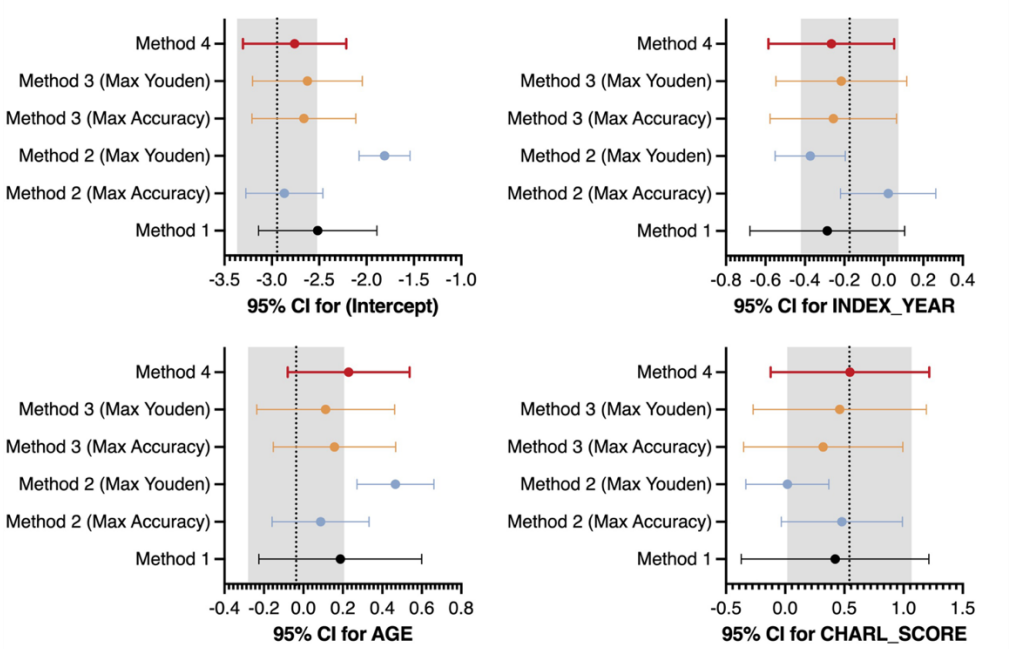

Figure 2. Point estimates and 95% confidence intervals (CI) for the association (in log odds ratio scale) between covariates and the occurrence of colon cancer events, where validation sample size was 319 (validation ratio ≈ 0.3).

Conclusion

In summary, we have presented an augmented estimation procedure that effectively addresses both non-differential and differential classifications by leveraging the full potential of multiple independent or correlated algorithm-derived phenotypes. Our comprehensive simulation studies and real-world data analysis demonstrate that the proposed method outperforms existing approaches, yielding unbiased association estimates with improved statistical efficiency. By integrating multiple error-prone algorithm-derived phenotypes with the gold-standard phenotype, our proposed estimator offers an enhanced procedure for estimating the association between risk factors and specific clinical outcomes using EHR data. Our novel bias-correction approach contributes to the reliability and reproducibility of findings, ultimately fostering progress in evidence-based healthcare and population health research.

References

1. Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, Peissig PL, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. Journal of the American Medical Informatics Association. 2016 Nov 1;23(6):1046–52.
2. Zheng NS, Feng Q, Kerchberger VE, Zhao J, Edwards TL, Cox NJ, et al. PheMap: a multi-resource knowledge base for high-throughput phenotyping within electronic health records. Journal of the American Medical Informatics Association. 2020 Nov 1;27(11):1675–87.
3. Duan R, Cao M, Wu Y, Huang J, Denny JC, Xu H, et al. An Empirical Study for Impacts of Measurement Errors on EHR based Association Studies. AMIA Annu Symp Proc. 2017 Feb 10; 2016:1764–73.
4. Chen Y, Wang J, Chubak J, Hubbard RA. Inflation of type I error rates due to differential misclassification in EHR-derived outcomes: Empirical illustration using breast cancer recurrence. Pharmacoepidemiol Drug Saf. 2019 Feb;28(2):264–8.

5. 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.
6. 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.
7. 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.
8. 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.
9. Hong C, Liao KP, Cai T. Semi-supervised validation of multiple surrogate outcomes with application to electronic medical records phenotyping. Biometrics. 2019 Mar;75(1):78–89.
10. Viana FAC, Haftka RT, Steffen V. Multiple surrogates: how cross-validation errors can help us to obtain the best predictor. Struct Multidisc Optim. 2009 Oct;39(4):439–57.
11. Liang X, Wang Z, Sha Q, Zhang S. An Adaptive Fisher's Combination Method for Joint Analysis of Multiple Phenotypes in Association Studies. Sci Rep. 2016 Oct 3;6(1):34323.