Penn Medicine

Department of Biostatistics, Epidemiology and Informatics

Practical solutions for working with electronic health records data

Yong Chen, PhD
Rebecca Hubbard, PhD
TA: Jessie Tong (UPenn)

August 08, 2022

DEPARTMENT of BIOSTATISTICS EPIDEMIOLOGY & INFORMATICS

Schedule

- 2:00 PM – 2:45 PM **Correcting for bias due to EHR data errors**
- 2:45 PM – 3:00 PM **Break**
- 3:00 PM – 4:15 PM **Distributed analyses with case studies**
- 4:15 PM – 4:45 PM **Tutorial (with demo using synthetic data)**
- 4:45 PM – 5:00 PM **Conclusions and Wrap Up**

Correcting for bias due to EHR data errors

Reproducibility of RWD based findings

Circulation Research

MY ALERTS | SIGN IN | JOIN

Submit your article

AHA Journals | Journal Information | All Issues | Subjects | Features | Resources & Education | For Authors & Reviewers

This site uses cookies. By continuing to browse this site you are agreeing to our use of cookies.

Click here for more information.

Home > Circulation Research > Vol. 116, No. 1 > Reproducibility in Science

FREE ACCESS

REVIEW ARTICLE

PDF/EPUB

Tools Share

Jump to

Abstract

Problem

Why Is This Important?

Details | Related | References | Figures

Reproducibility in Science

Improving the Standard for Basic and Preclinical Research

C. Glenn Begley and John P.A. Ioannidis

Originally published 2 Jan 2015 | https://doi.org/10.1161/CIRCRESAHA.114.303819 | Circulation Research. 2015;116:116–126

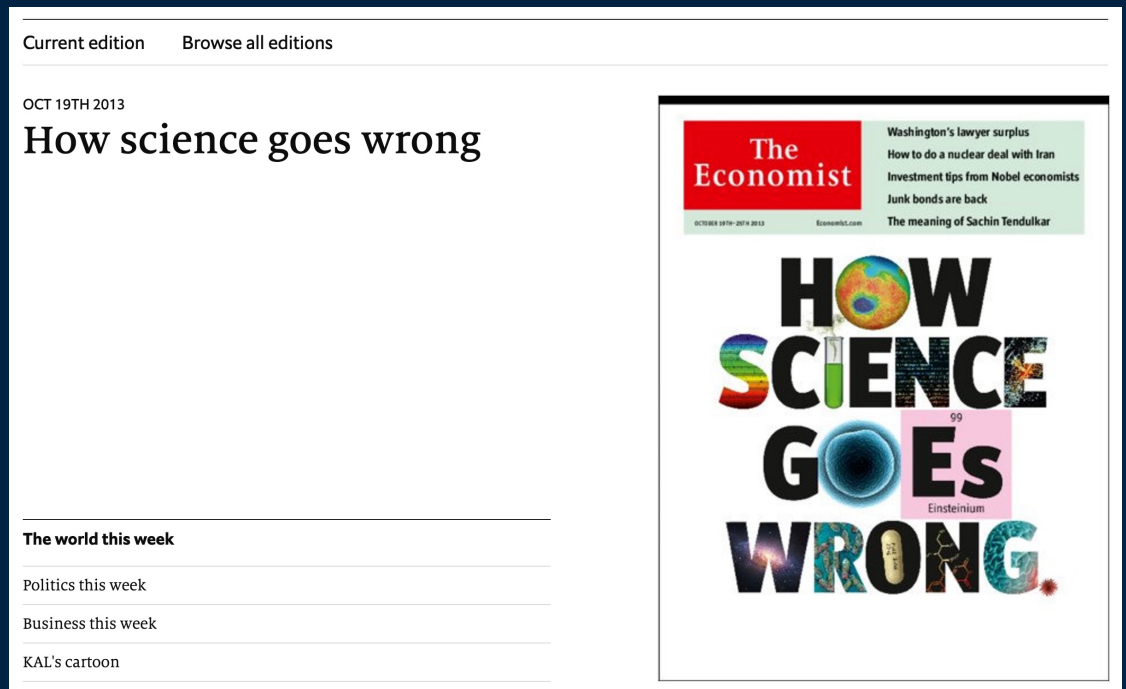
Abstract

Medical and scientific advances are predicated on new knowledge that is robust and reliable and that serves as a solid foundation on which further advances can be built. In biomedical research, we are in the midst of a revolution with the generation of new data and scientific publications at a previously

Circulation Research

January 2, 2015
Vol 116, Issue 1

Article Information



Current edition | Browse all editions

OCT 19TH 2013

How science goes wrong

The world this week

Politics this week

Business this week

KAL's cartoon

The Economist

Washington's lawyer surplus
How to do a nuclear deal with Iran
Investment tips from Nobel economists
Junk bonds are back
The meaning of Sachin Tendulkar

HOW SCIENCE GOES WRONG

Bias

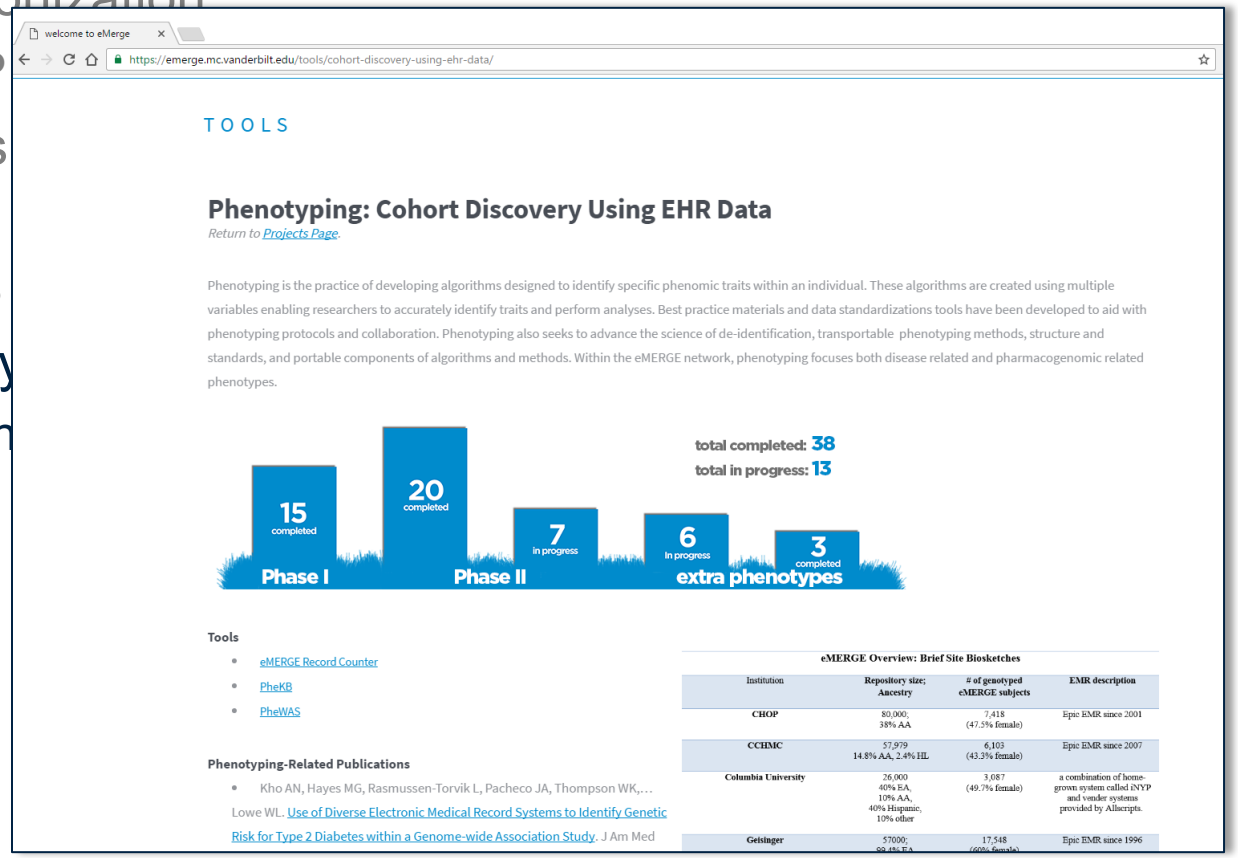
- Bias due to data quality and lack of data harmonization
  - Existing efforts: PCORnet CDM, OHDSI OMOP

Bias

- Bias due to data quality and lack of data harmonization
  - Existing efforts: PCORnet CDM, OHDSI OMOP
- Bias due to heterogeneous populations across different studies
  - Comparability of cohorts across studies
  - Criteria of inclusion/exclusion of patient cohorts

Bias

- Bias due to data quality and lack of data harmonization
  - Existing efforts: PCORnet CDM, OHDSI OMOP
- Bias due to heterogeneous populations across
  - Comparability of cohorts across studies
  - Criteria of inclusion/exclusion of patient cohorts
- Bias due to “imperfect” performance of phenoty
  - We need more high-quality phenotyping algorith
    - Existing effort: PheKB



Bias

- Bias due to data quality and lack of data harmonization
  - Existing efforts: PCORnet CDM, OHDSI OMOP
- Bias due to heterogeneous populations across different studies
  - Comparability of cohorts across studies
  - Criteria of inclusion/exclusion of patient cohorts
- Bias due to “imperfect” performance of phenotyping algorithms
  - We need more high-quality phenotyping algorithms
    - Existing effort: PheKB
  - We need to be cautious about the generalizability and portability of these algorithms when taken to a different study population

Bias

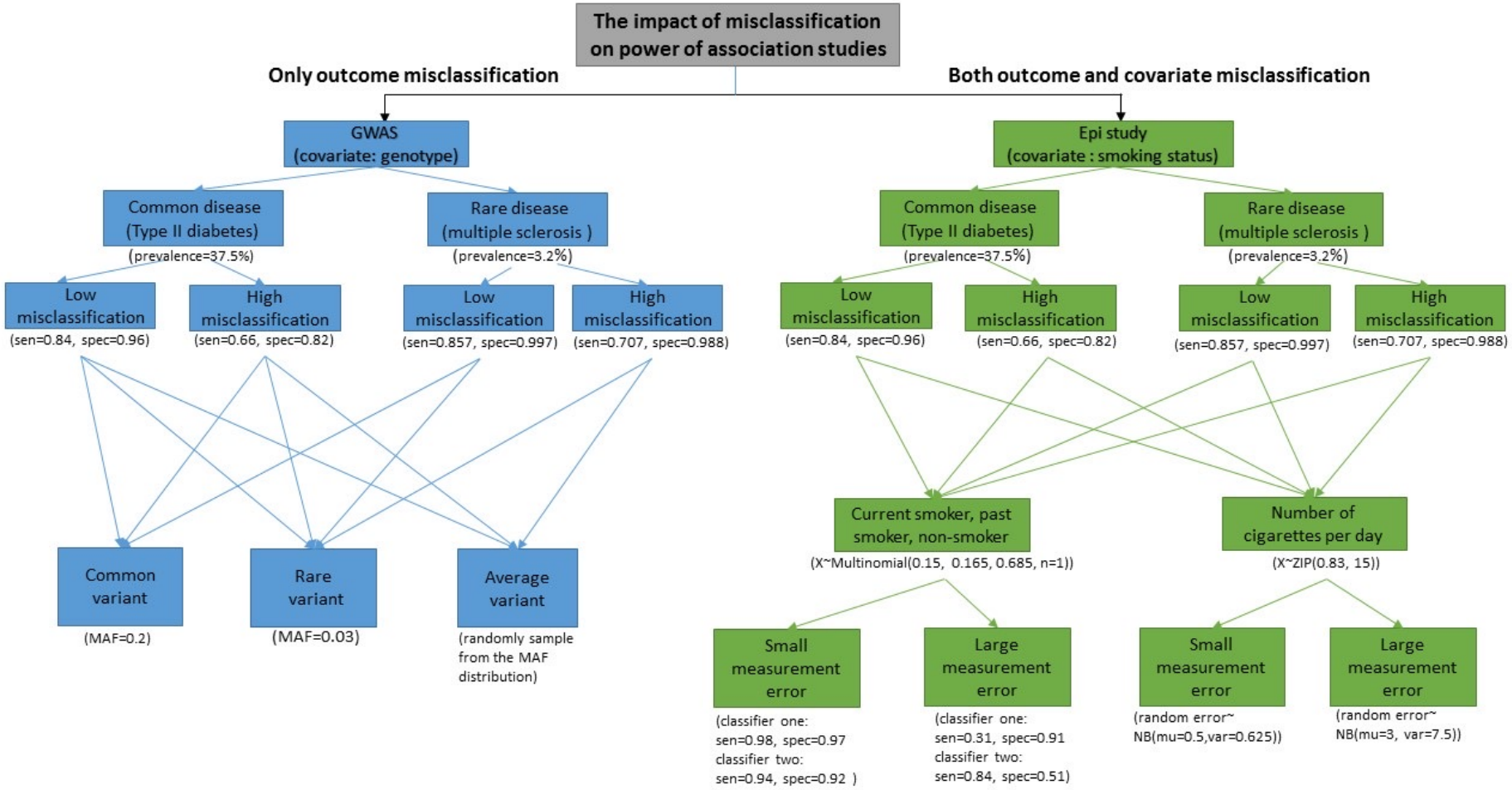
- Bias due to data quality and lack of data harmonization
  - Existing efforts: PCORnet CDM, OHDSI OMOP
- Bias due to heterogeneous populations across different studies
  - Comparability of cohorts across studies
  - Criteria of inclusion/exclusion of patient cohorts
- Bias due to “imperfect” performance of phenotyping algorithms
  - We need more high-quality phenotyping algorithms
    - Existing effort: PheKB
  - We need to be cautious about the generalizability and portability of these algorithms when taken to a different study population
  - **What can we do?**

Bias due to “imperfect” performance of phenotyping algorithms

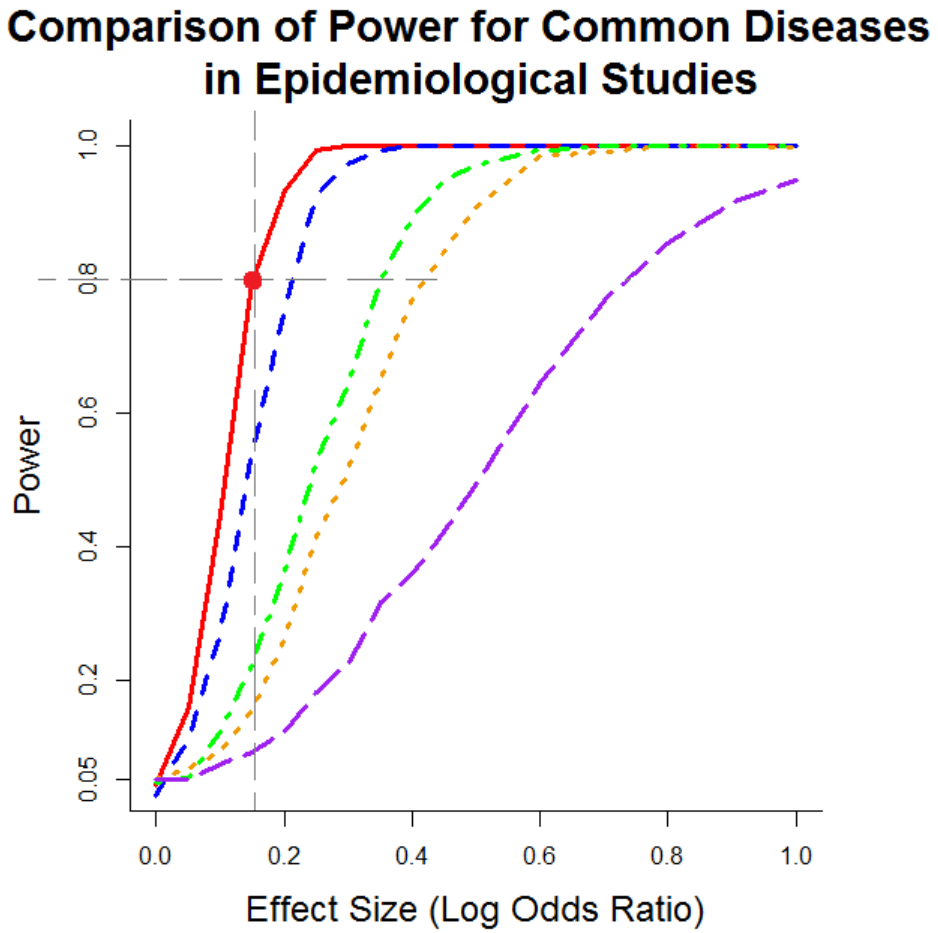
- **What can we do?**

1. **Quantify the impacts of phenotyping error in subsequent analyses, including**
   - Association analysis (e.g., understand the impacts to loss of statistical power and inflation of Type I errors). Both contributed negatively to the reproducibility of RWD-based findings.

Statistical Issues for EHR Based Association Studies

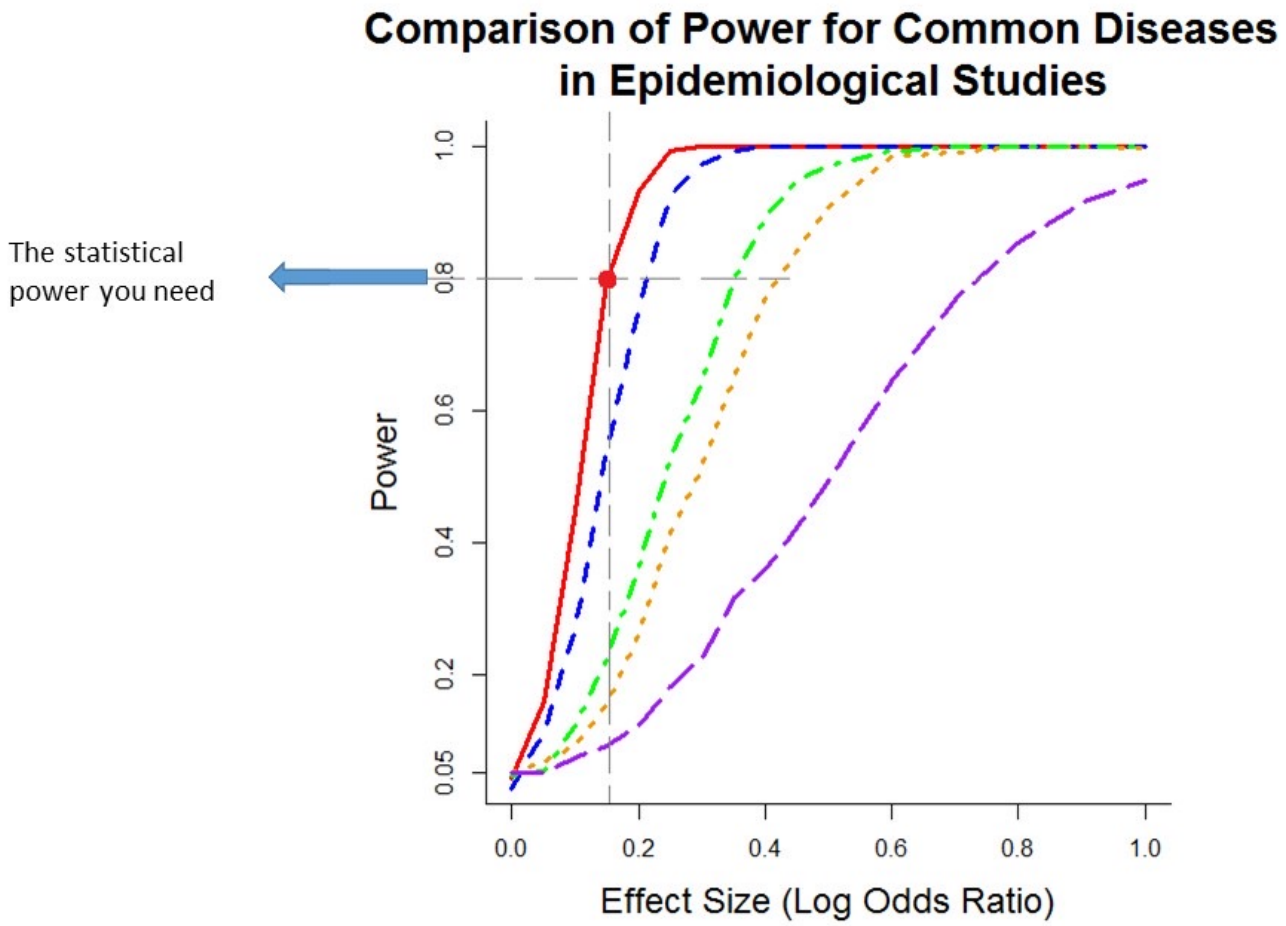


Results --- Loss of power



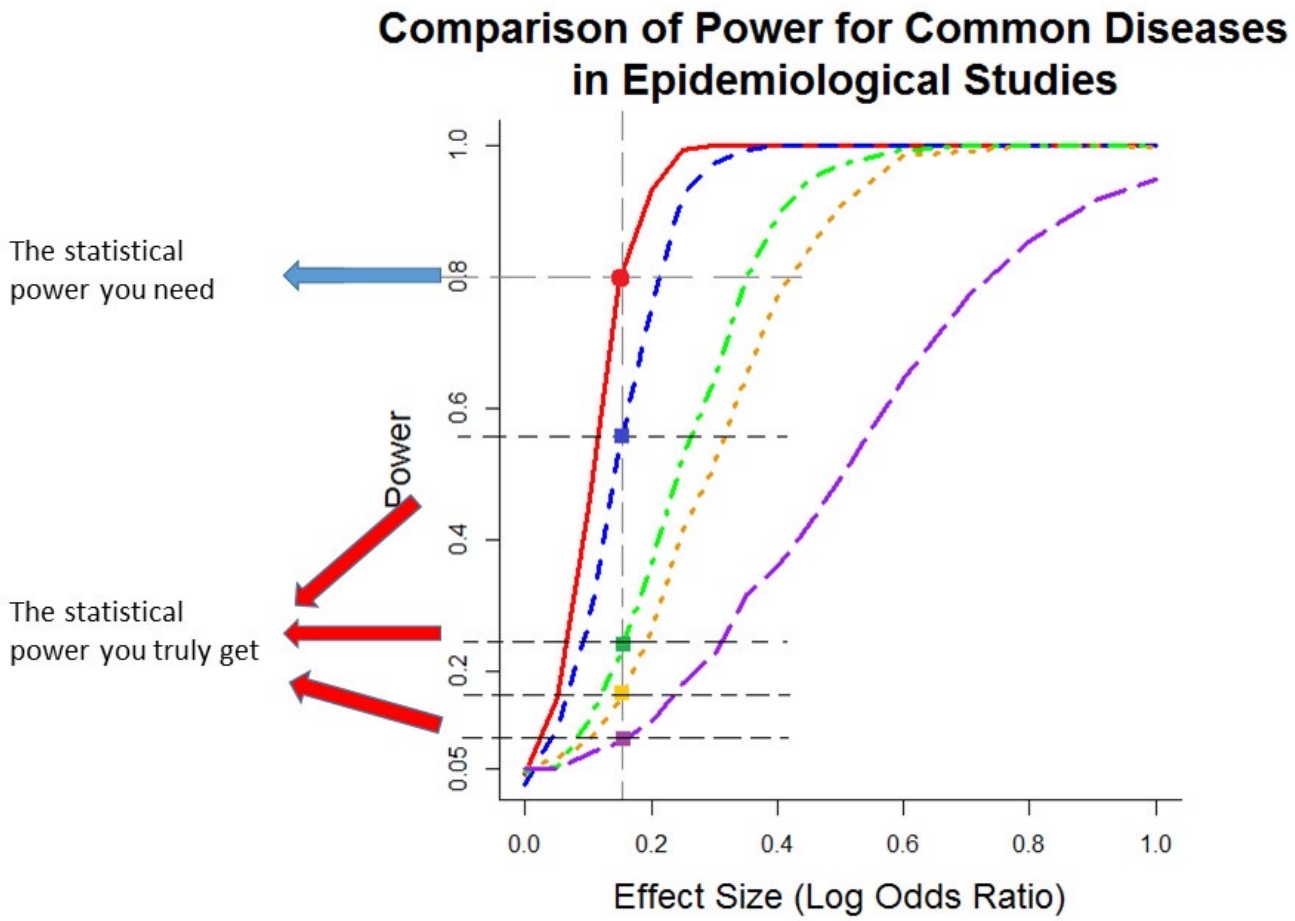
Duan, R., Cao, M., Wu, Y., Huang, J., Denny, J.C., Xu, H. and Chen, Y., 2016. An empirical study for impacts of measurement errors on EHR based association studies. In *AMIA Annual Symposium Proceedings* (Vol. 2016, p. 1764). American Medical Informatics Association

Results --- Loss of power



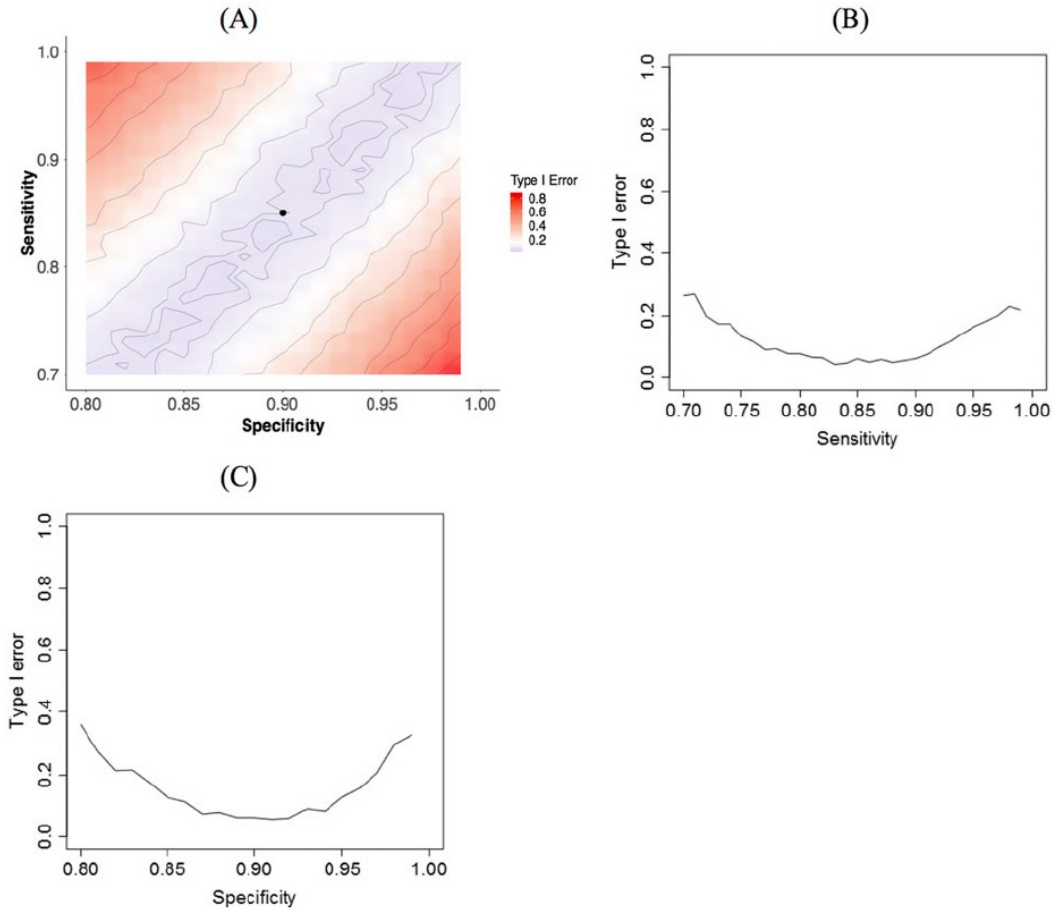
Duan, R., Cao, M., Wu, Y., Huang, J., Denny, J.C., Xu, H. and Chen, Y., 2016. An empirical study for impacts of measurement errors on EHR based association studies. In *AMIA Annual Symposium Proceedings* (Vol. 2016, p. 1764). American Medical Informatics Association

Results --- Loss of power



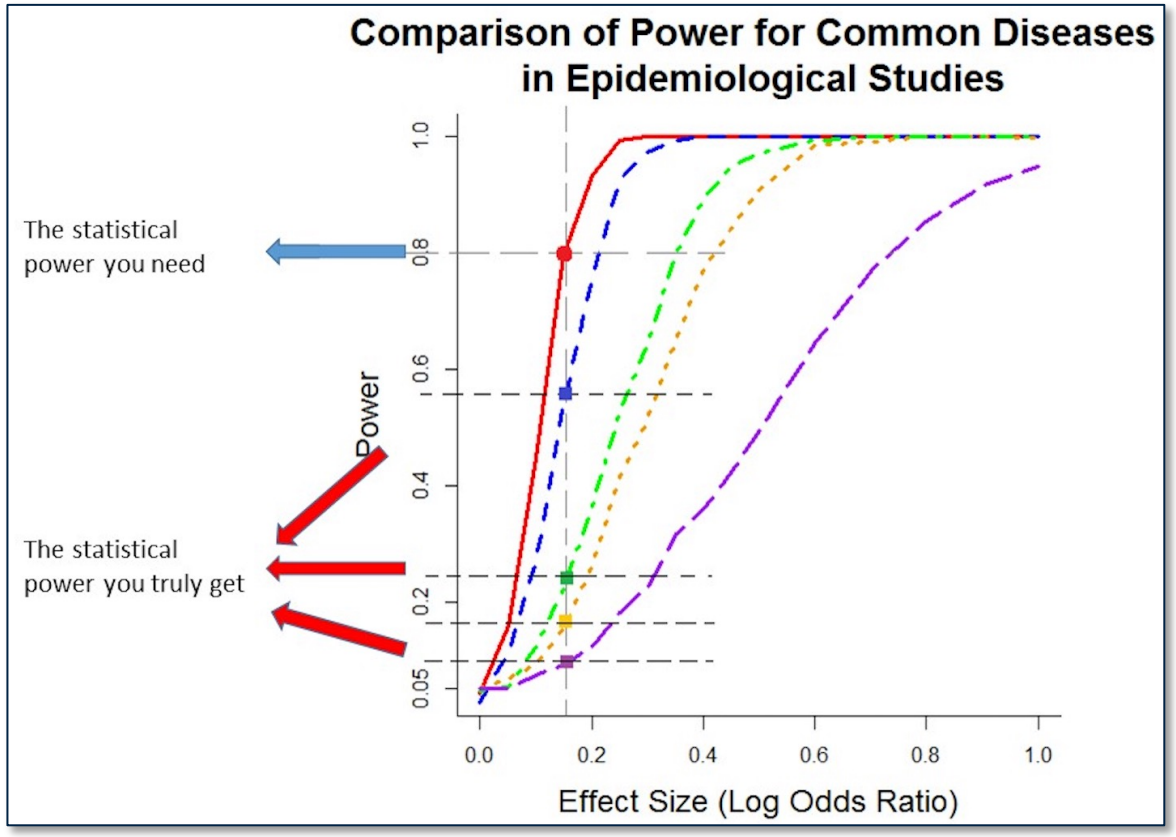
Duan, R., Cao, M., Wu, Y., Huang, J., Denny, J.C., Xu, H. and Chen, Y., 2016. An empirical study for impacts of measurement errors on EHR based association studies. In *AMIA Annual Symposium Proceedings* (Vol. 2016, p. 1764). American Medical Informatics Association

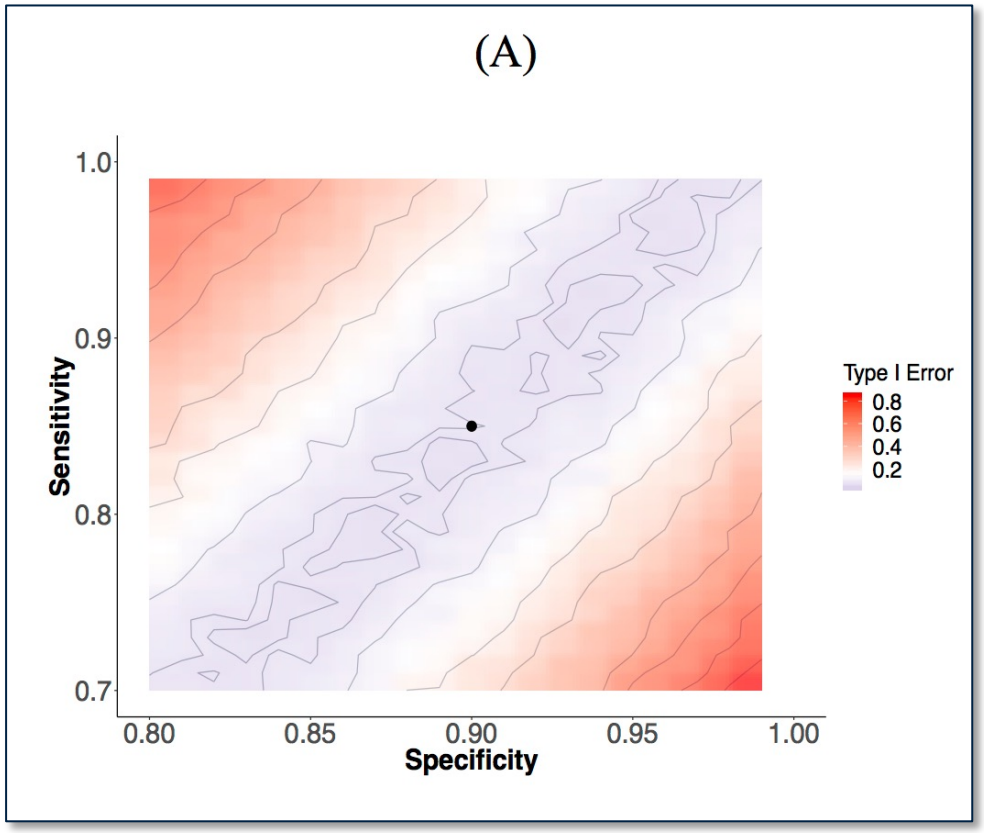
Results --- Inflated Type I Error



Chen, Y., Wang, J., Chubak, J. and Hubbard, R.A., 2019. Inflation of type I error rates due to differential misclassification in EHR-derived outcomes: Empirical illustration using breast cancer recurrence. *Pharmacoepidemiology and drug safety*, *28*(2), pp.264-268

Impact in Type I error and Power

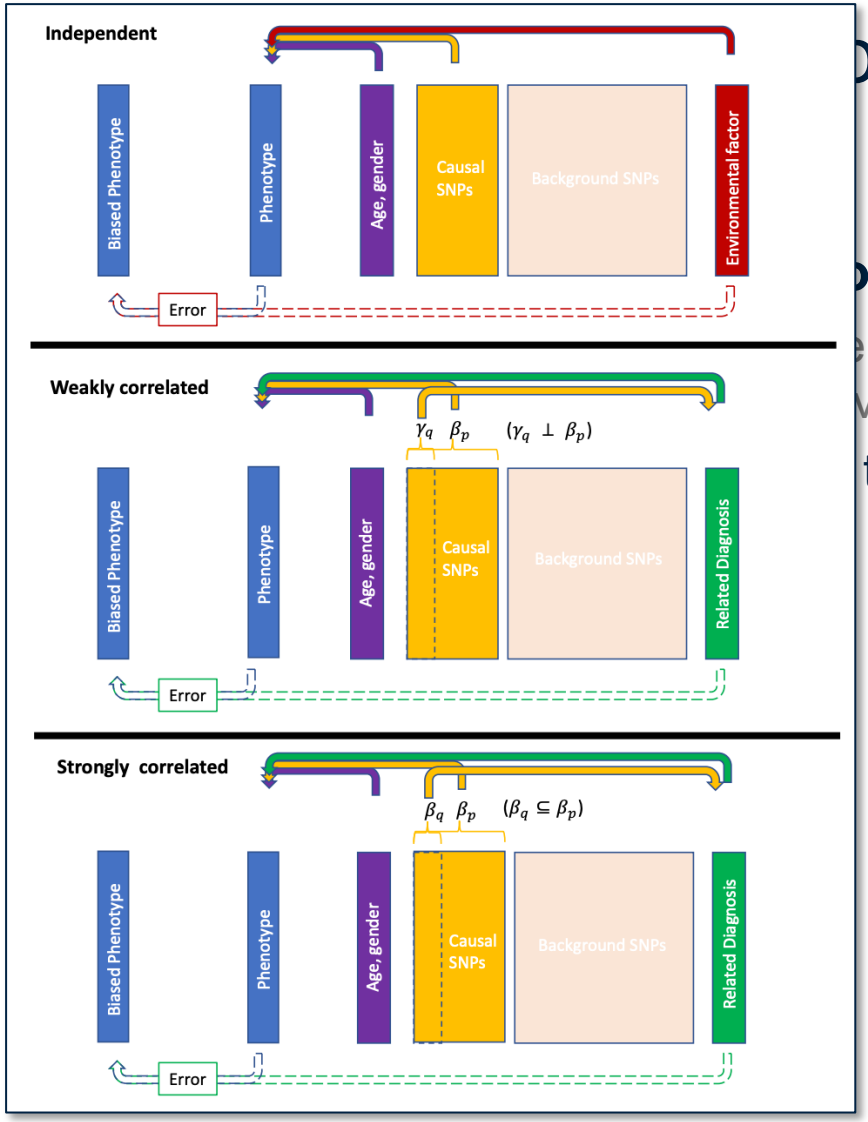


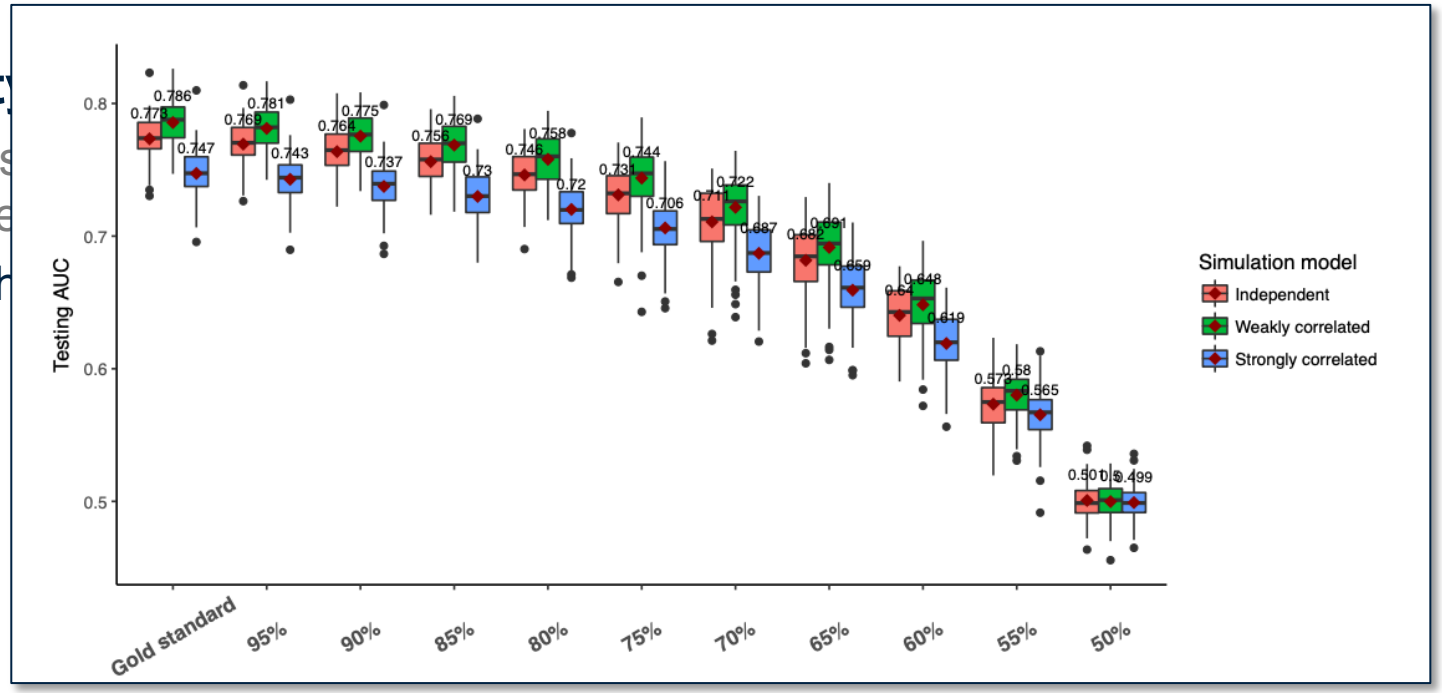


Duan, R., Cao, M., Wu, Y., Huang, J., Denny, J.C., Xu, H. and Chen, Y., 2016. An empirical study for impacts of measurement errors on EHR based association studies. In *AMIA Annual Symposium Proceedings* (Vol. 2016, p. 1764). American Medical Informatics Association

Chen, Y., Wang, J., Chubak, J. and Hubbard, R.A., 2019. Inflation of type I error rates due to differential misclassification in EHR-derived outcomes: Empirical illustration using breast cancer recurrence. *Pharmacoepidemiology and drug safety*, 28(2), pp.264-268.

performance of phenotyping algorithms





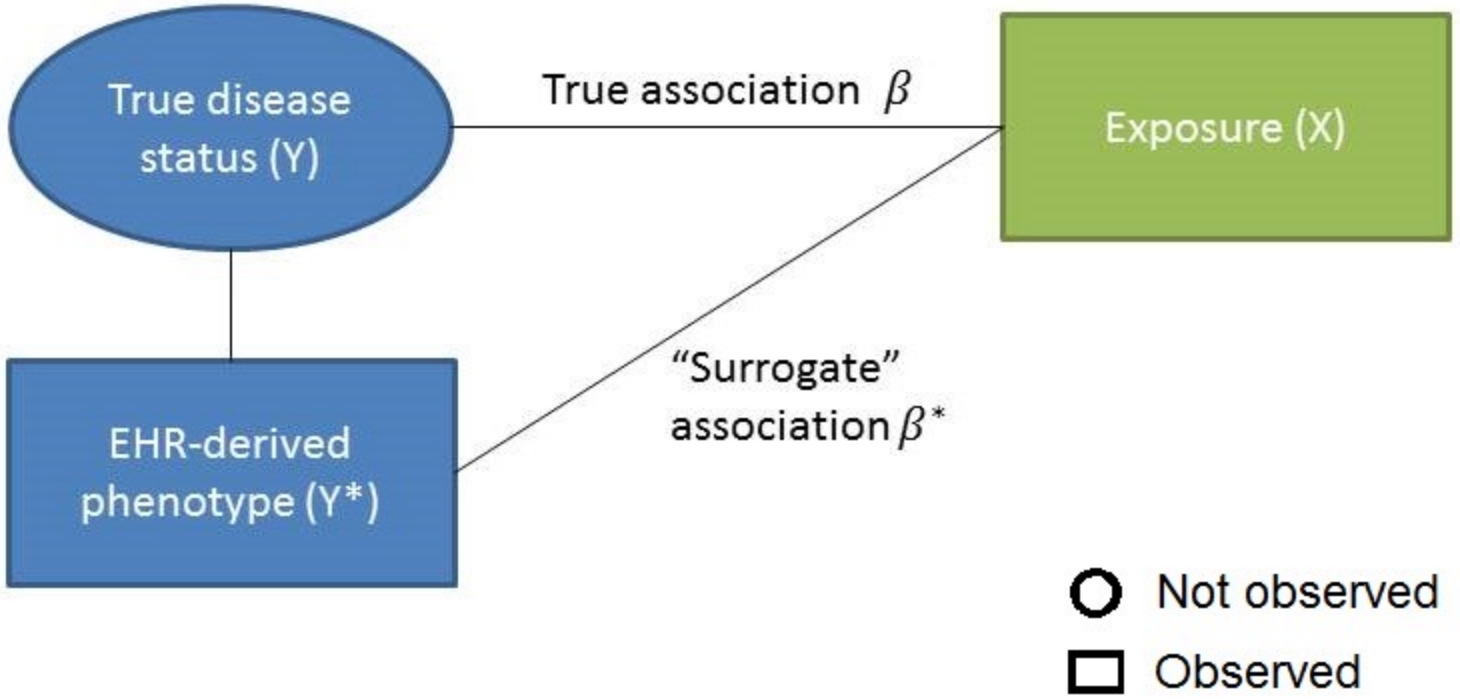
Li, R, Tong, J, Duan, R, Chen, Y and Moore, J (June, 2020), Evaluation of phenotyping errors on polygenic risk score predictions, International Conference on Bioinformatics Models, Methods and Algorithms

Bias due to “imperfect” performance of phenotyping algorithms

- **What can we do?**

1. **Quantify the impacts of phenotyping error in subsequent analyses**
2. **Bias reduction/correction**
    - Chart reviews for the disease status are crucial. One should always budget for chart reviews when possible.

Demonstrate the bias in association analysis



Ignore the Misclassification?

	True Classification				Observed Classification		
	X=1	X=0			X=1	X=0	
T=1	40	20	60	T=1	38	24	62
T=0	60	80	140	T=0	62	76	138
	100	100	200		100	100	200

*Sensitivity=0.8, Specificity=0.9.

No misclassification $\hat{\beta} = 0.98,\ SE(\hat{\beta}) = 0.323$

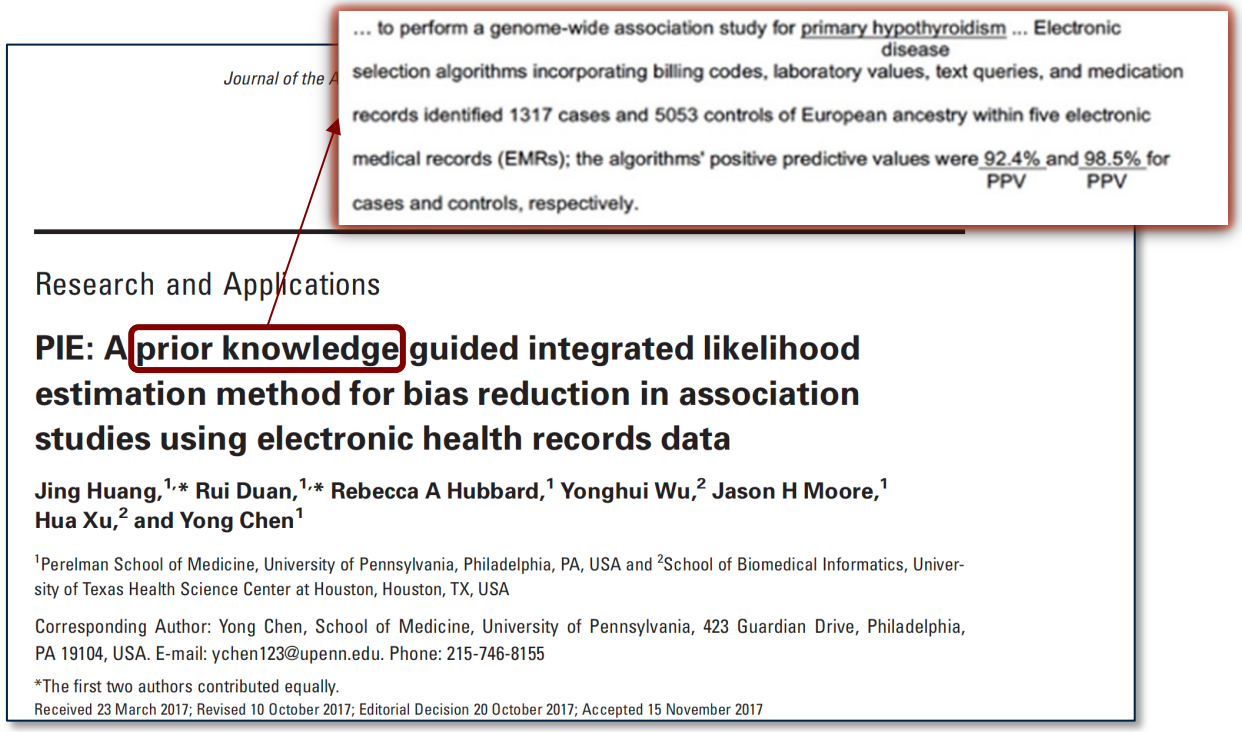
Misclassification $\tilde{\beta} = 0.667,\ \mathrm{SE}(\tilde{\beta}) = 0.478$

Bias due to “imperfect” performance of phenotyping algorithms

- **What can we do?**

1. **Quantify the impacts of phenotyping error in subsequent analyses**
2. **Bias reduction/correction**
   - Chart reviews for the disease status are crucial. One should always budget for chart reviews when possible.
   - In some situations, chart reviews are not available (e.g., budget constraints, time constraints), bias reduction (not correction) could still be conducted.

Bias reduction using **P**rior-knowledge-guided **I**ntegrated-likelihood **E**stimation (PIE) idea



... to perform a genome-wide association study for primary hypothyroidism (disease) ... Electronic selection algorithms incorporating billing codes, laboratory values, text queries, and medication records identified 1317 cases and 5053 controls of European ancestry within five electronic medical records (EMRs); the algorithms' positive predictive values were 92.4% (PPV) and 98.5% (PPV) for cases and controls, respectively.

Journal of the A

Research and Applications

PIE: A prior knowledge guided integrated likelihood estimation method for bias reduction in association studies using electronic health records data

Jing Huang,[1,*] Rui Duan,[1,*] Rebecca A Hubbard,[1] Yonghui Wu,[2] Jason H Moore,[1] Hua Xu,[2] and Yong Chen[1]

[1]Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA and [2]School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX, USA

Corresponding Author: Yong Chen, School of Medicine, University of Pennsylvania, 423 Guardian Drive, Philadelphia, PA 19104, USA. E-mail: ychen123@upenn.edu. Phone: 215-746-8155

*The first two authors contributed equally.

Received 23 March 2017; Revised 10 October 2017; Editorial Decision 20 October 2017; Accepted 15 November 2017

Huang, J., Duan, R., Hubbard, R.A., Wu, Y., Moore, J.H., Xu, H. and Chen, Y., 2018. PIE: A prior knowledge guided integrated likelihood estimation method for bias reduction in association studies using electronic health records data. *Journal of the American Medical Informatics Association*, 25(3), pp.345-352.

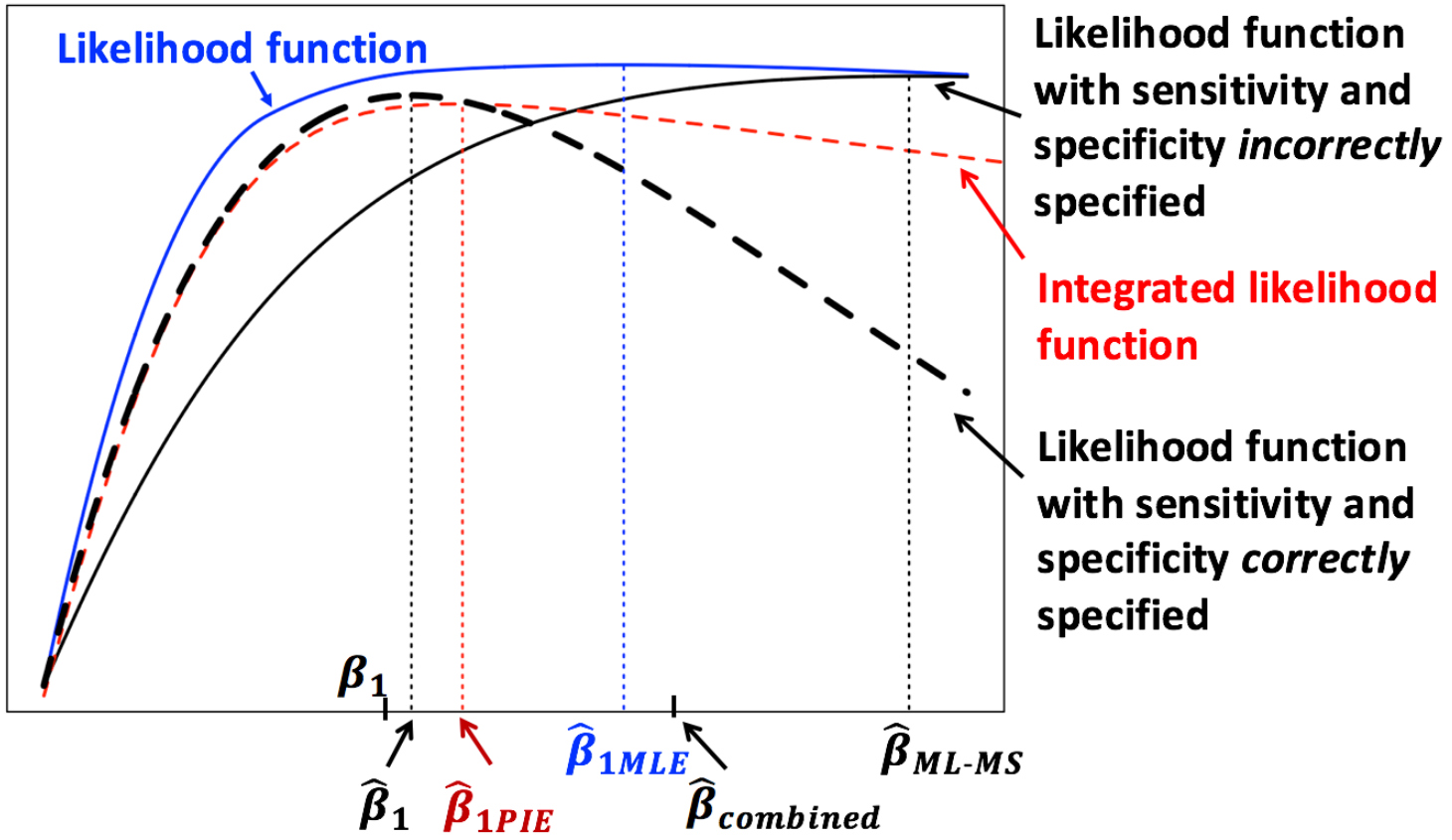
An Integrated Likelihood Method to account for phenotyping error

- What is an Integrated Likelihood? (Severini 1998, 2010 Biometrika; Berger et al. 1999 Statistical Science)

$$L_l(\beta) = \iint L(\beta, \alpha_1, \alpha_0)\pi(\alpha_1, \alpha_0)d\alpha_1 d\alpha_0$$

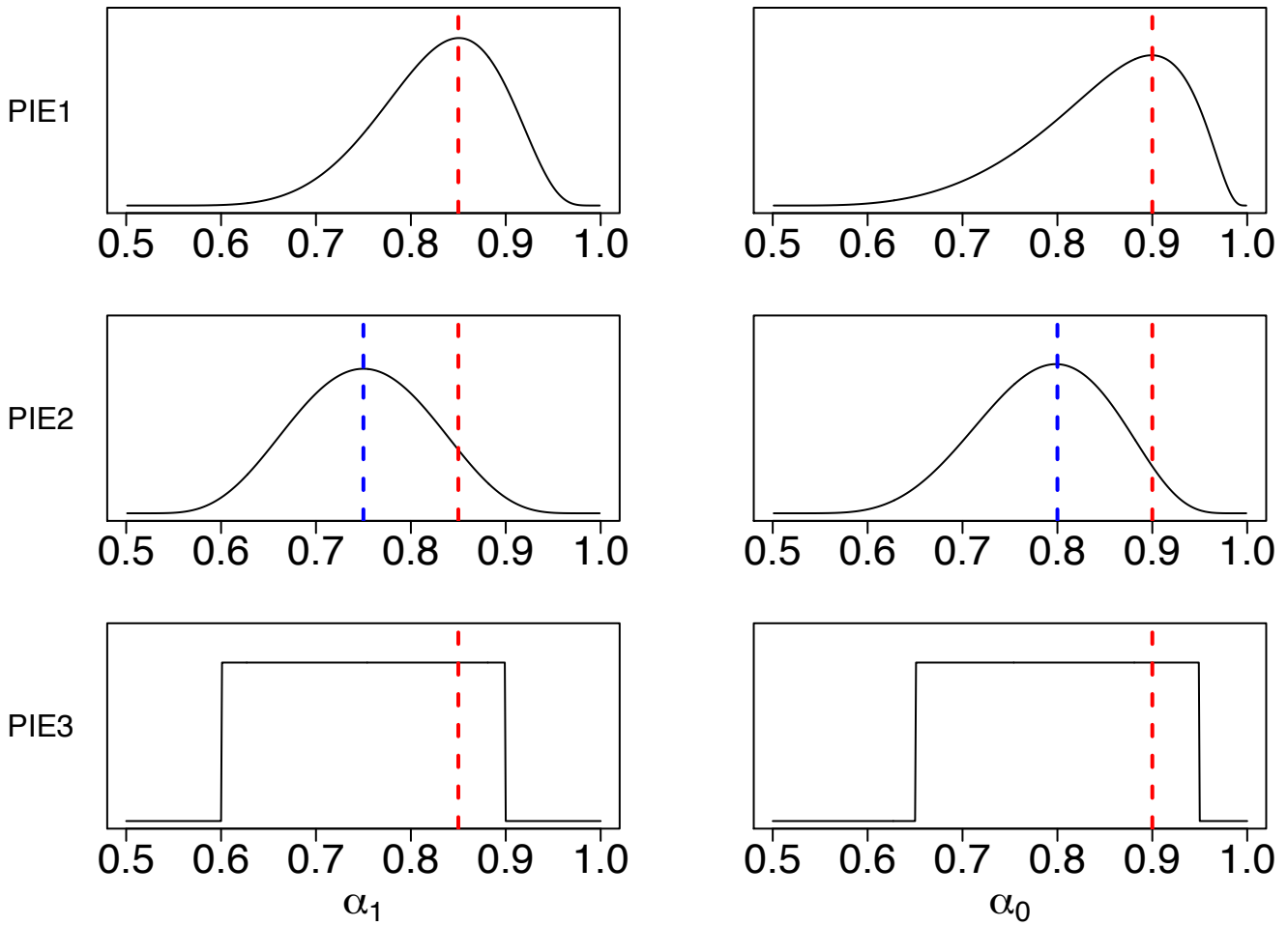
- $L(\beta, \alpha_1, \alpha_0)$ is the standard likelihood function.
- $\pi(\alpha_1, \alpha_0)$ is a prior function of sensitivity and specificity.
- Fully account for the feature of EHR data: prior knowledge of phenotyping error being available.
- Prior Knowledge Guided Integrated Likelihood Estimation Method (**PIE**)

How Integrated Likelihood Works --- Shape of Likelihood Functions



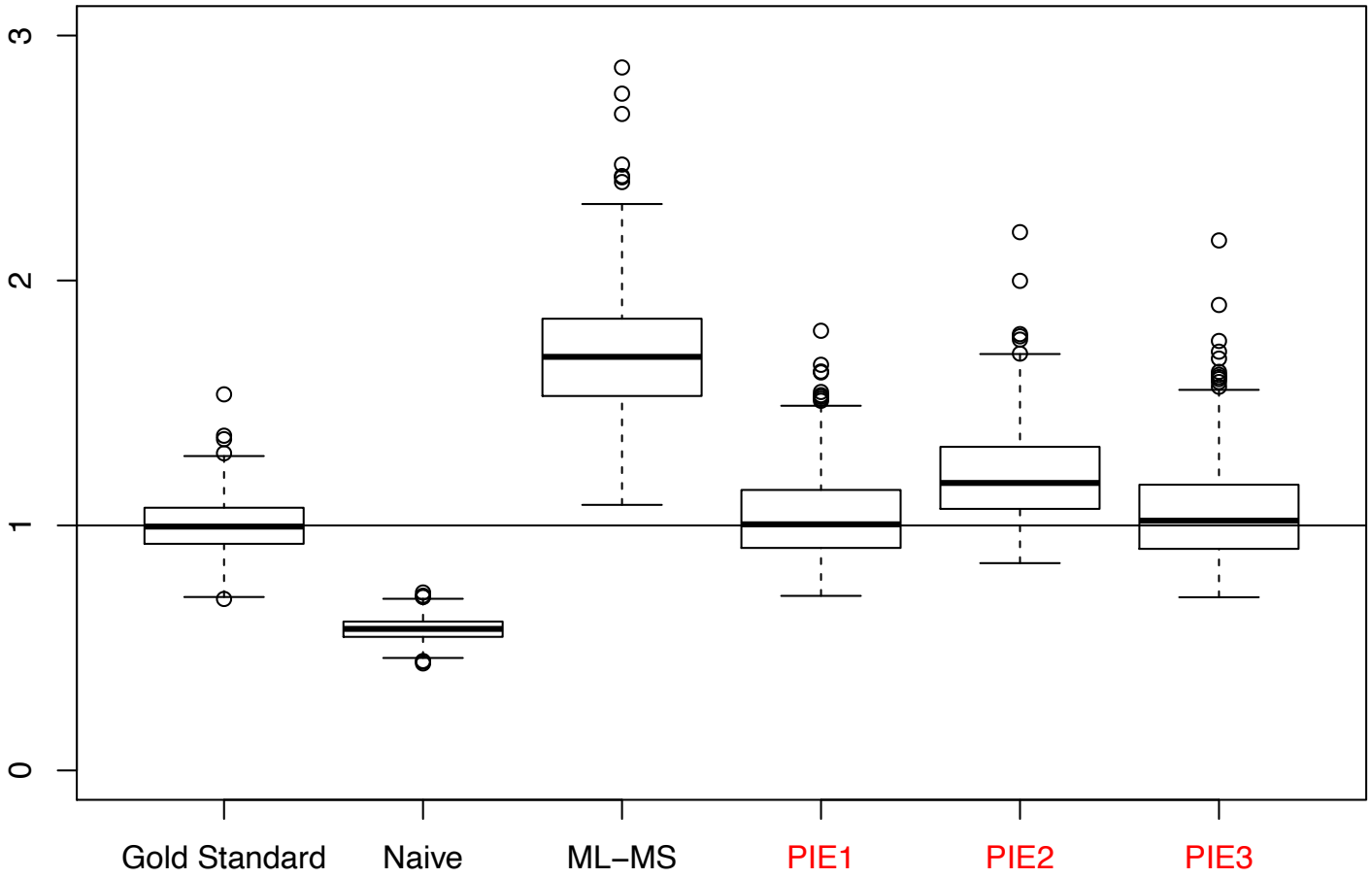
Huang, J., Duan, R., Hubbard, R.A., Wu, Y., Moore, J.H., Xu, H. and Chen, Y., 2018. PIE: A prior knowledge guided integrated likelihood estimation method for bias reduction in association studies using electronic health records data. *Journal of the American Medical Informatics Association*, *25*(3), pp.345-352.

Distribution of Priors



Huang, J., Duan, R., Hubbard, R.A., Wu, Y., Moore, J.H., Xu, H. and Chen, Y., 2018. PIE: A prior knowledge guided integrated likelihood estimation method for bias reduction in association studies using electronic health records data. *Journal of the American Medical Informatics Association*, *25*(3), pp.345-352.

Comparison of Bias



Huang, J., Duan, R., Hubbard, R.A., Wu, Y., Moore, J.H., Xu, H. and Chen, Y., 2018. PIE: A prior knowledge guided integrated likelihood estimation method for bias reduction in association studies using electronic health records data. *Journal of the American Medical Informatics Association*, *25*(3), pp.345-352.

Evaluation of PIE using Negative Controls

Idea: Analyze multiple real datasets where the true association is known and see if the estimator is consistent with the truth (Schuemie et al., 2020).

- **Negative controls**: covariate – outcome pairs where no association is believed to exist. (E.g., diclofenac – ingrowing nail)
- Evaluate PIE on 200 datasets corresponding to the 200 negative controls.

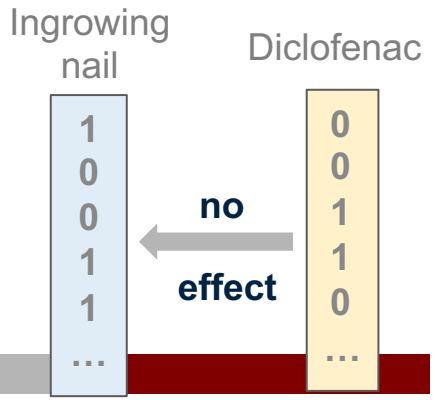
Evaluation of PIE using Negative Controls

Idea: Analyze multiple real datasets where the true association is known and see if the estimator is consistent with the truth (Schuemie et al., 2020).

200 data of 'negative controls'

Use PIE to analyze the 200 data

Evaluate whether PIE produces the expected results.



Obtain 200 results

How many 95% confidence intervals, out of 200, covering the truth?
---- Expected to be around 190

Bias due to “imperfect” performance of phenotyping algorithms

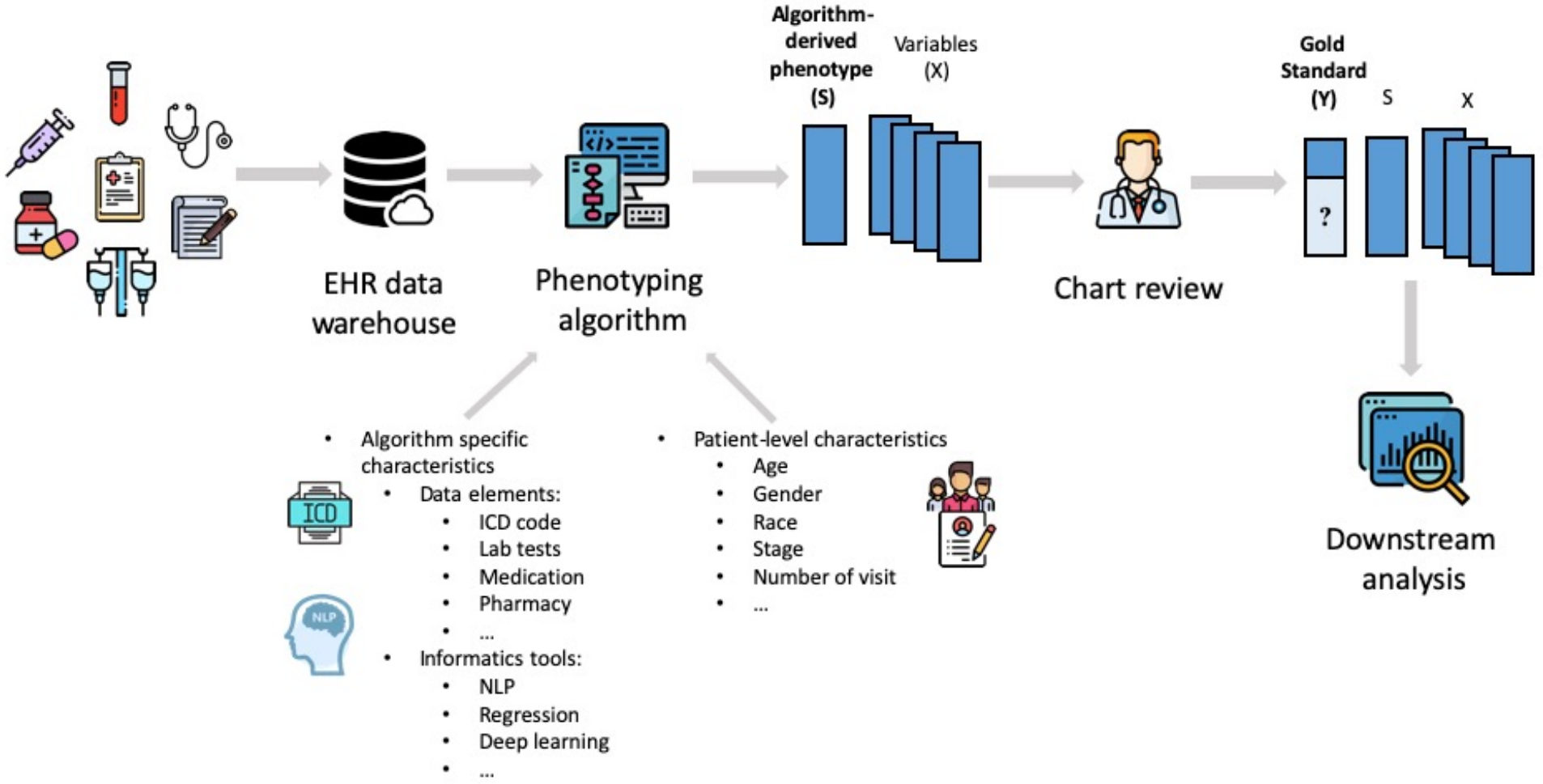
- **What can we do?**

1. **Quantify the impacts of phenotyping error in subsequent analyses**
2. **Bias reduction/correction**
   - Chart reviews for the disease status are crucial. One should always budget for chart reviews when possible.
   - In some situations, chart reviews are not available (e.g., budget constraints, time constraints), bias reduction (not correction) could still be conducted.
   - Now we have a budget for chart review. The questions are:
     - **How to best use the chart review results?**

Bias correction

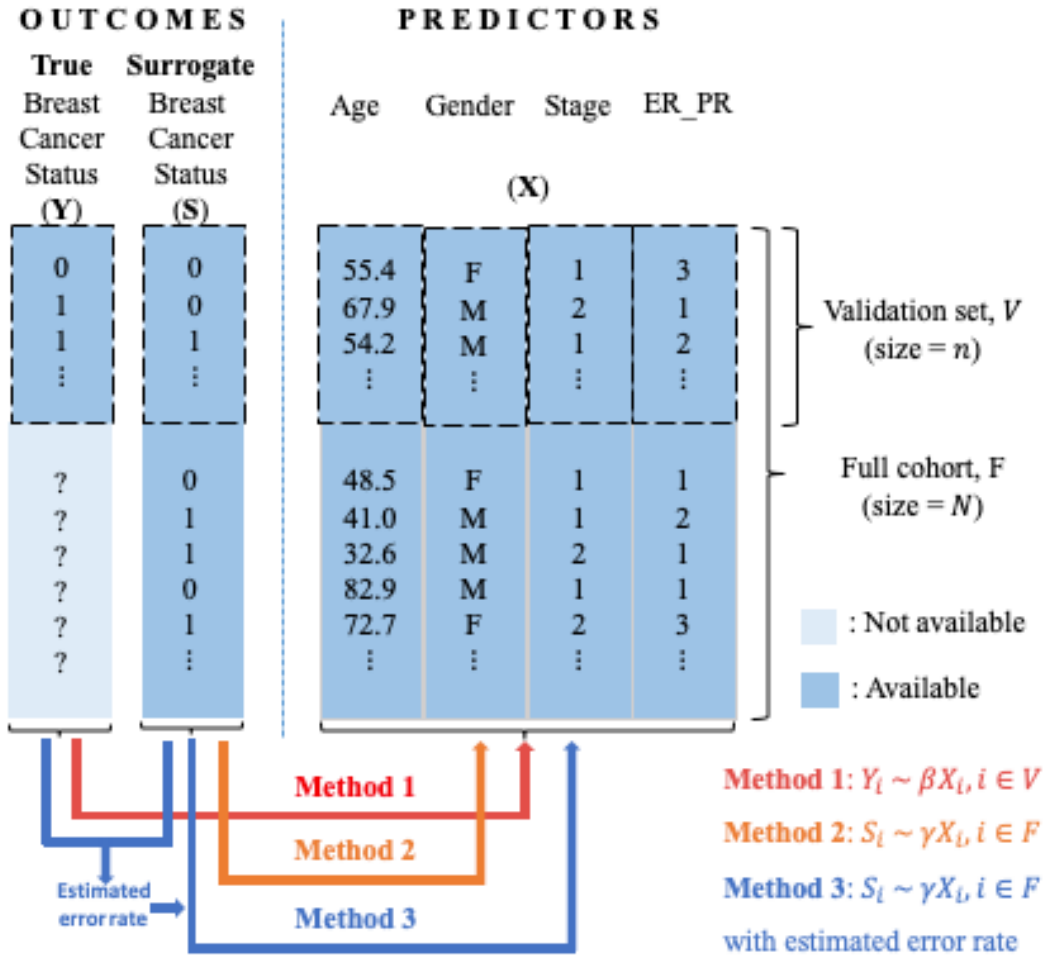
- “Despite we have some chart reviews, let’s not waste the results from phenotyping algorithms”
- Maximize the use of surrogate outcome derived from EHR-based phenotyping algorithms

Availability of Validation Data



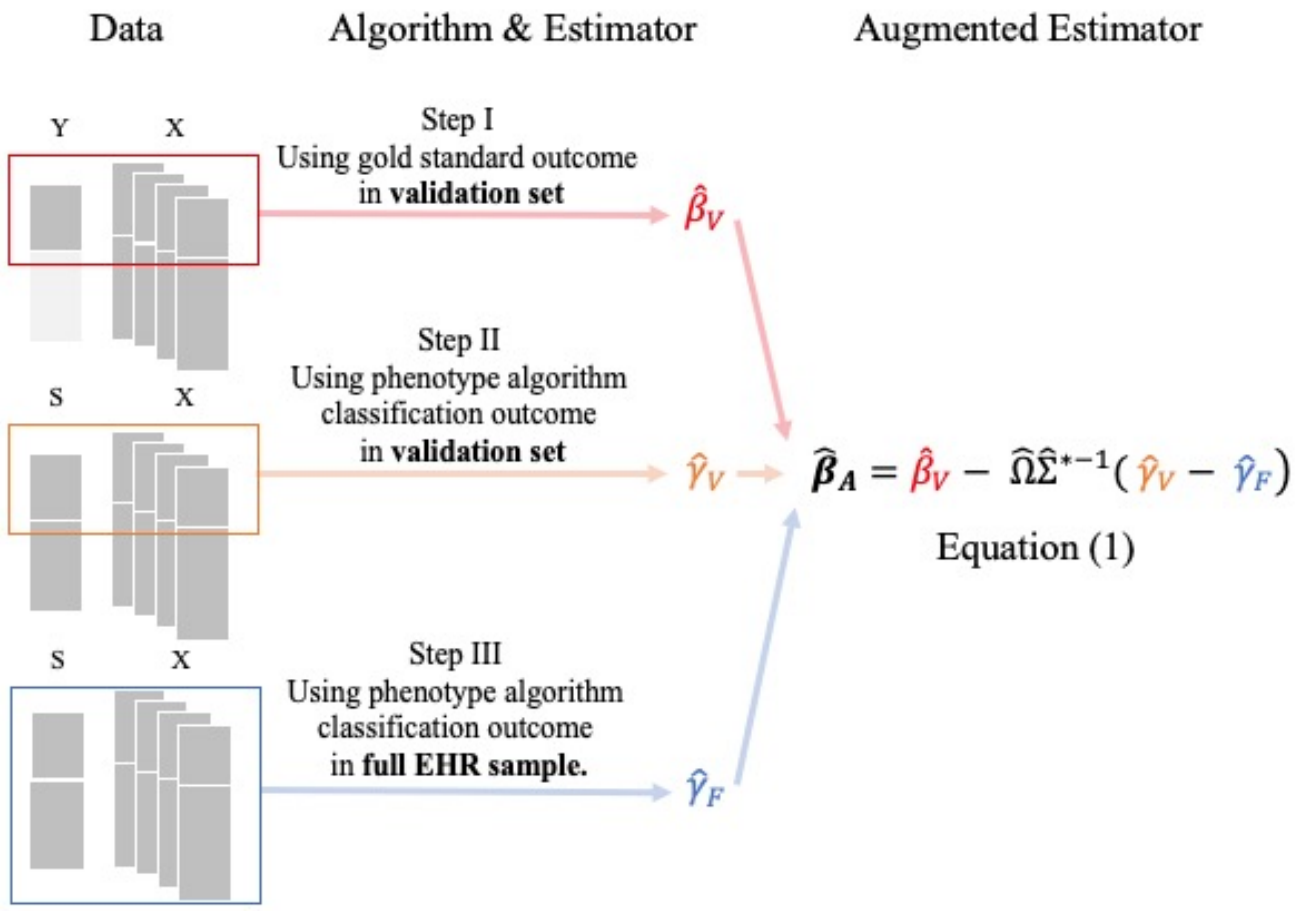
Tong, J., Huang, J., Chubak, J., Wang, X., Moore, J.H., Hubbard, R.A. and Chen, Y., 2020. An augmented estimation procedure for EHR-based association studies accounting for differential misclassification. *Journal of the American Medical Informatics Association*, *27*(2), pp.244-253.

Existing methods



Validation data are randomly sampled

Proposed method – seemingly unrelated regression (Chen 2002 JRSS-b)



Tong, J., Huang, J., Chubak, J., Wang, X., Moore, J.H., Hubbard, R.A. and Chen, Y., 2020. An augmented estimation procedure for EHR-based association studies accounting for differential misclassification. *Journal of the American Medical Informatics Association*, *27*(2), pp.244-253.

Existing three methods to be compared with Tong et al. (2019)

- Model (1) Validation data only: association between the covariates and true phenotype can be estimated using only the validation data through a logistic regression model:

$$logit(Pr(Y = 1)) = \beta_0 + \beta_1 X$$

  - Pros and Cons: Low bias, but low efficiency

- Model (2) Naive method: association can be estimated using the full EHR sample through a logistic regression model:

$$logit(Pr(S = 1)) = \gamma_0 + \gamma_1 X$$

  - Pros and Cons: High efficiency, but large bias

Model (3) Magder & Hughes (1997)

- Sensitivity and specificity are estimated using information on the true disease status and algorithm-derived phenotype in the validation set:

$$\text{Estimated sensitivity} = \hat{S} = \frac{\sum_{i \in V, S_i = 1} Y_i}{\sum_{i \in V} Y_i}$$

$$\text{Estimated specificity} = \hat{P} = \frac{\sum_{i \in V, S_i = 0} Y_i}{\sum_{i \in V} Y_i}$$

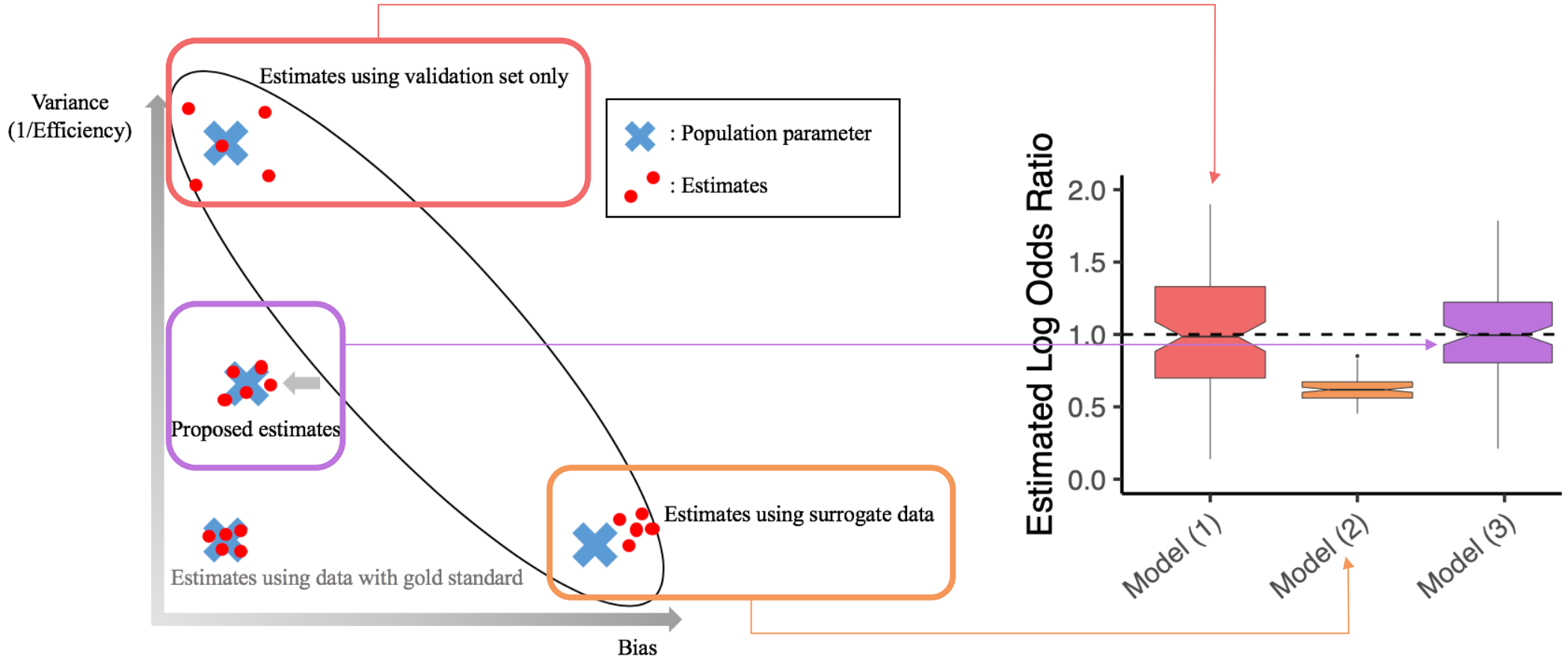
- Then, the estimates are plugged into the following likelihood function to estimate parameter of interests:

$$L(\beta_0, \beta_1) = \prod_{i=1}^{N} p_i^{S_i} (1 - p_i)^{1 - S_i}$$

$$\text{where } p_i = Pr(S_i = 1) = (1 - \hat{P}) + (\hat{P} + \hat{S} - 1) expit(\beta_0 + \beta_1 X_i)$$

- Pros and Cons: Small bias and variance if misclassification is correctly specified, but potential for high bias if misclassification is incorrectly specified (e.g., cannot account for differential misclassification).

Intuition


Variance
(1/Efficiency)
Estimates using validation set only
: Population parameter
: Estimates
Proposed estimates
Estimates using data with gold standard
Estimates using surrogate data
Bias
Estimated Log Odds Ratio
2.0
1.5
1.0
0.5
0.0
Model (1)
Model (2)
Model (3)

Proposed method

For estimators in Model (1) and (2), we obtain the joint distribution of $\hat{\beta}_V - \beta_1$ and $\hat{\gamma}_V - \hat{\gamma}_F$ as follows:

$$n^{1/2}\begin{pmatrix}\hat{\beta}_V - \beta_1 \\ \hat{\gamma}_V - \hat{\gamma}_F\end{pmatrix} \to N\left(0, \begin{pmatrix}\Sigma & \Omega \\ \Omega & \Sigma^*\end{pmatrix}\right)$$

as $n \to \infty$.

Note that $\hat{\gamma}_V - \hat{\gamma}_F$ is observed, $\hat{\beta}_V - \beta_1$ conditioning on $\hat{\gamma}_V - \hat{\gamma}_F$ is asymptotically normal with

- mean: $\Omega\Sigma^{*-1}(\hat{\gamma}_V - \hat{\gamma}_F)$
  * This mean is approximately 0 for a moderate size validation set.
- variance: $\frac{\Sigma - \Omega\Sigma^{*-1}\Omega'}{n}$.

The augmented estimator can be written as:

$$\hat{\beta}_A = \hat{\beta}_V - \Omega\Sigma^{*-1}(\hat{\gamma}_V - \hat{\gamma}_F)$$

where

$$\hat{\Omega} = \frac{1-\rho}{n}\sum_{i\in V}(Y_i - \hat{Y}_i)X_i\hat{I}^{-1}(\beta_1)[(S_i - \hat{S}_i)X_i\hat{I}^{-1}(\gamma)]'$$

$$\hat{\Sigma} = \frac{1-\rho}{n}\sum_{i\in V}(S_i - \hat{S}_i)X_i\hat{I}^{-1}(\gamma_1)[(S_i - \hat{S}_i)X_i\hat{I}^{-1}(\gamma)]'$$

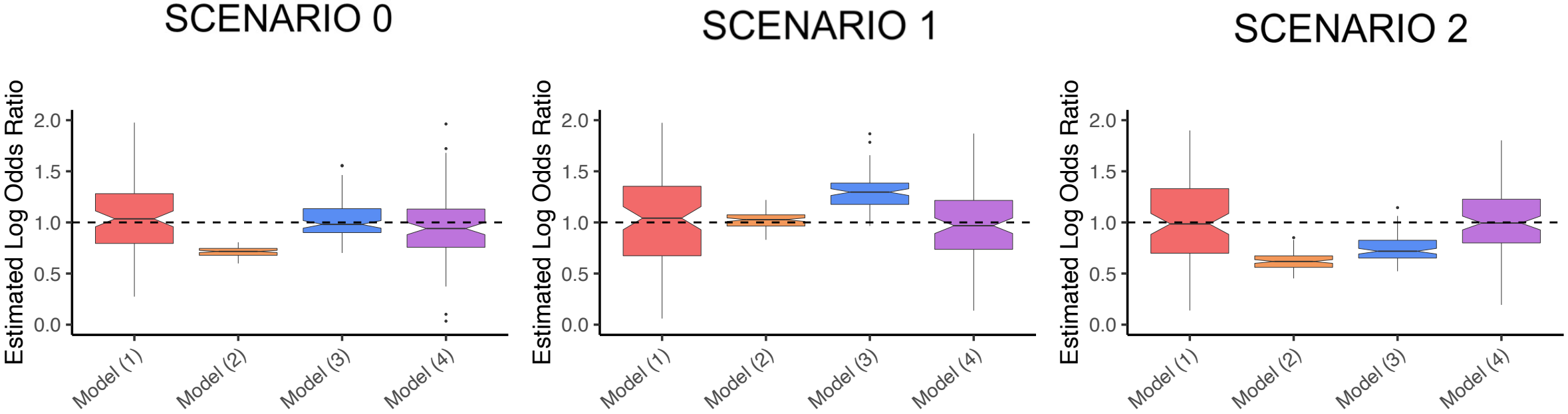
$$\hat{I}(\beta) = \frac{1}{n}\hat{Y}(1-\hat{Y})XX'$$

$$\hat{I}(\gamma) = \frac{1}{n}\hat{S}(1-\hat{S})XX'$$

Simulation results

	Non-differential Misclassification		Differential Misclassification			
	Scenario 0		Scenario 1		Scenario 2	
	P	S	P	S	P	S
Non-exposure group	0.90	0.85	0.90	0.85	0.90	0.85
Exposure group	0.90	0.85	0.80	0.95	0.95	0.80

Non-differential

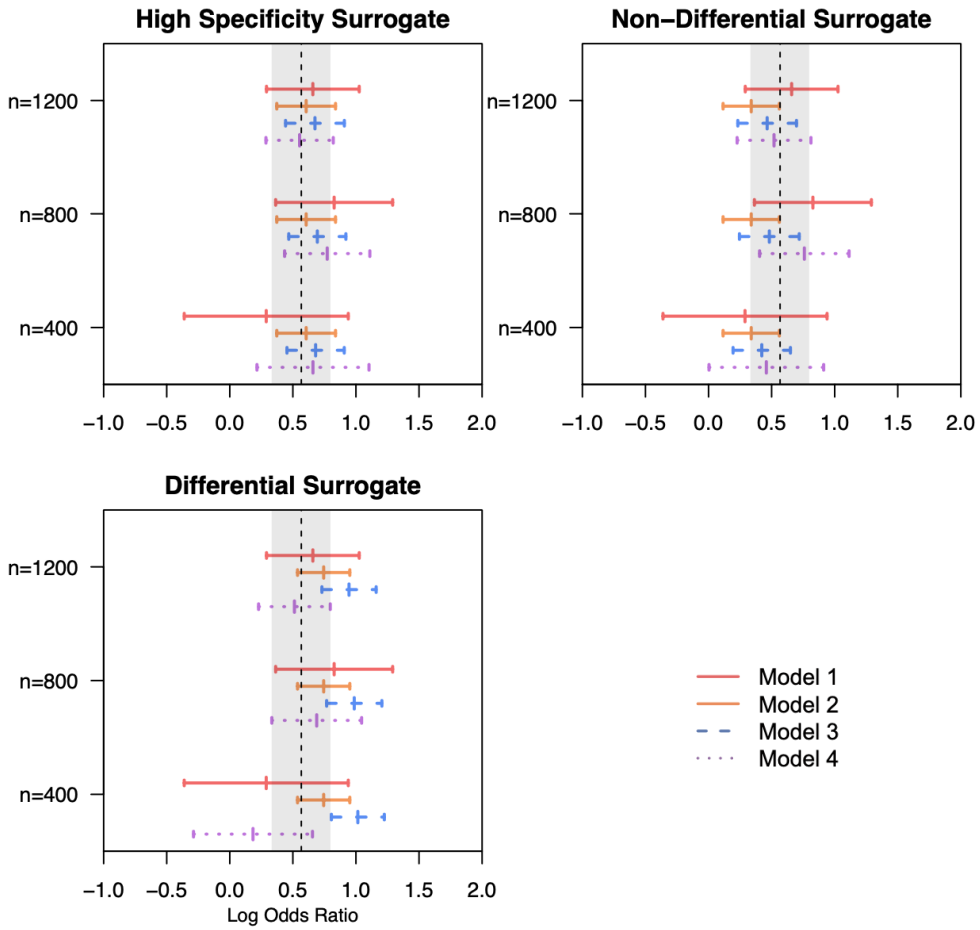


Tong, J., Huang, J., Chubak, J., Wang, X., Moore, J.H., Hubbard, R.A. and Chen, Y., 2020. An augmented estimation procedure for EHR-based association studies accounting for differential misclassification. *Journal of the American Medical Informatics Association*, *27*(2), pp.244-253.

Data analysis results

- Data from BRAVA study
  - An investigation of risk factors for second primary or recurrent
  - breast cancers (jointly termed second breast cancer events [SBCE]) in women with a personal history of breast cancer
  - 3,152 women enrolled in Kaiser Permanente Washington (KPWA)
  - The gold standard and phenotype algorithm classified SBCE phenotype are provided for all women.

Variable: Stage



Bias correction

- “Under rare disease setting, random sampling for chart review is obviously inefficient”
- We need procedure for case-enrichment

Cost-effective Sampling Design to Account for Phenotyping Error

Journal of the American Medical Informatics Association, 29(1), 2022, 52–61
doi: 10.1093/jamia/ocab222
Advance Access Publication Date: 26 October 2021
Research and Applications

Research and Applications

A cost-effective chart review sampling design to account for phenotyping error in electronic health records (EHR) data

Ziyan Yin[1], Jiayi Tong[2], Yong Chen [2], Rebecca A. Hubbard[2], and Cheng Yong Tang[1]

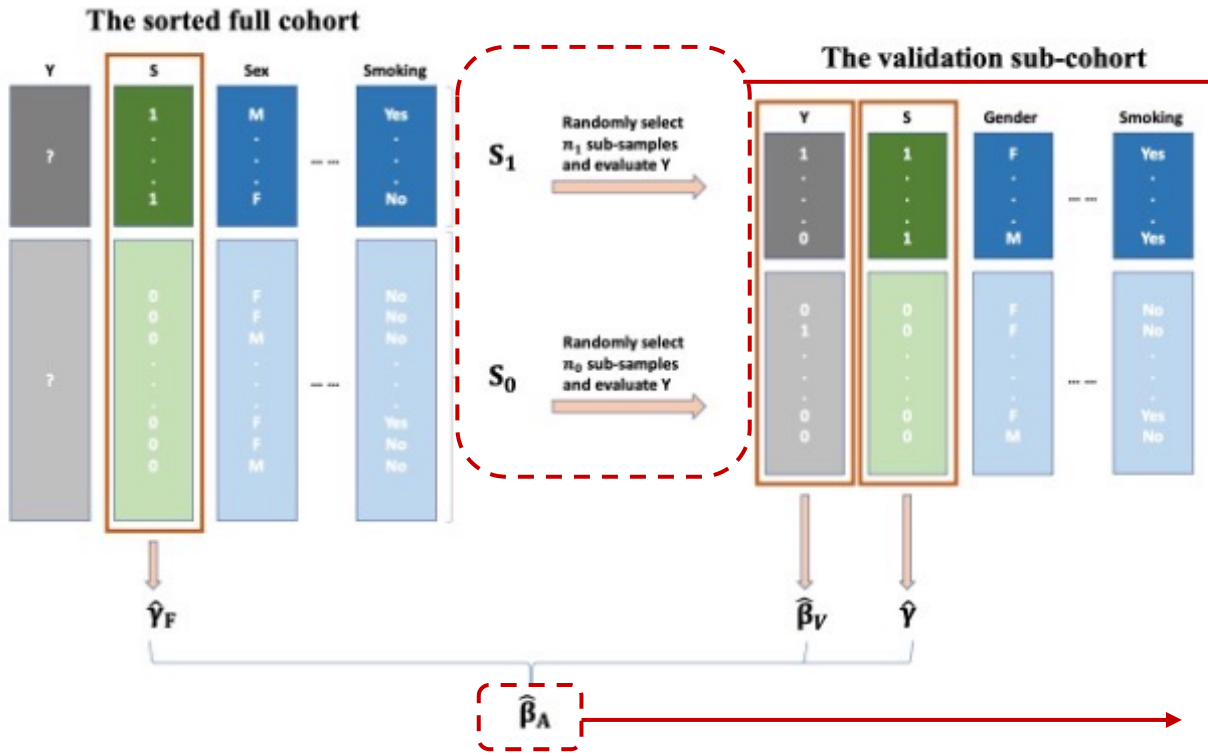
Algorithm

1. Split the original full cohort into 2 sub-groups: "S-positive" (S_1) and "S − negative (S_0)."
2. Uniformly select n_0 samples from S_0 and n_1 samples from S_1 to construct a new subcohort V.
   Let h_1 and h_0 be the sampling ratios in S_1 and S_0, respectively.
   Perform the manual chart review in V and obtain the true phenotype Y.
3. In the full cohort, fit weighted logistic regression for S and obtain the MLE estimator
   $\hat{\gamma}_F$; for the $i-$th subject, the weight is h_1 if $S=1$ or h_0 if $S=0$;
4. Within V, fit unweighted logistic regression for S *and* Y separately and obtain the "working" MLE $\hat{\gamma}_V$ and $\hat{\beta}_V$.
5. Construct the final estimator $\hat{\beta}_A = \hat{\beta}_V - \hat{H}_Y^{-1}\hat{G}_{SY}\hat{G}_S^{-1}\hat{H}_S(\hat{\gamma}_V - \hat{\gamma}_F)$ and obtain the MLE estimator
   $\hat{V} = n^{-1}\{\hat{H}_Y^{-1} - (1-nN^{-1})\hat{H}_Y^{-1}\hat{G}_{SY}\hat{G}_S^{-1}\hat{G}_{SY}^{\top}\hat{H}_Y^{-1}\}$,
   where $n = n_0 + n_1$ and the definition of $\hat{H}_Y$, $\hat{G}_{SY}$, $\hat{G}_S$, and $\hat{H}_S$ are given in Supplementary Appendix. Under mild conditions, our estimator $\hat{\beta}_A$ is approximately distributed $N\{\beta_0 + (c, 0^{\top})^{\top}, \hat{V}\}$.

Rare disease

- Outcome-dependent sampling
  - Sampling based on the surrogate outcome (i.e., S)

Key changes compared with Tong et al. 2019 JAMIA



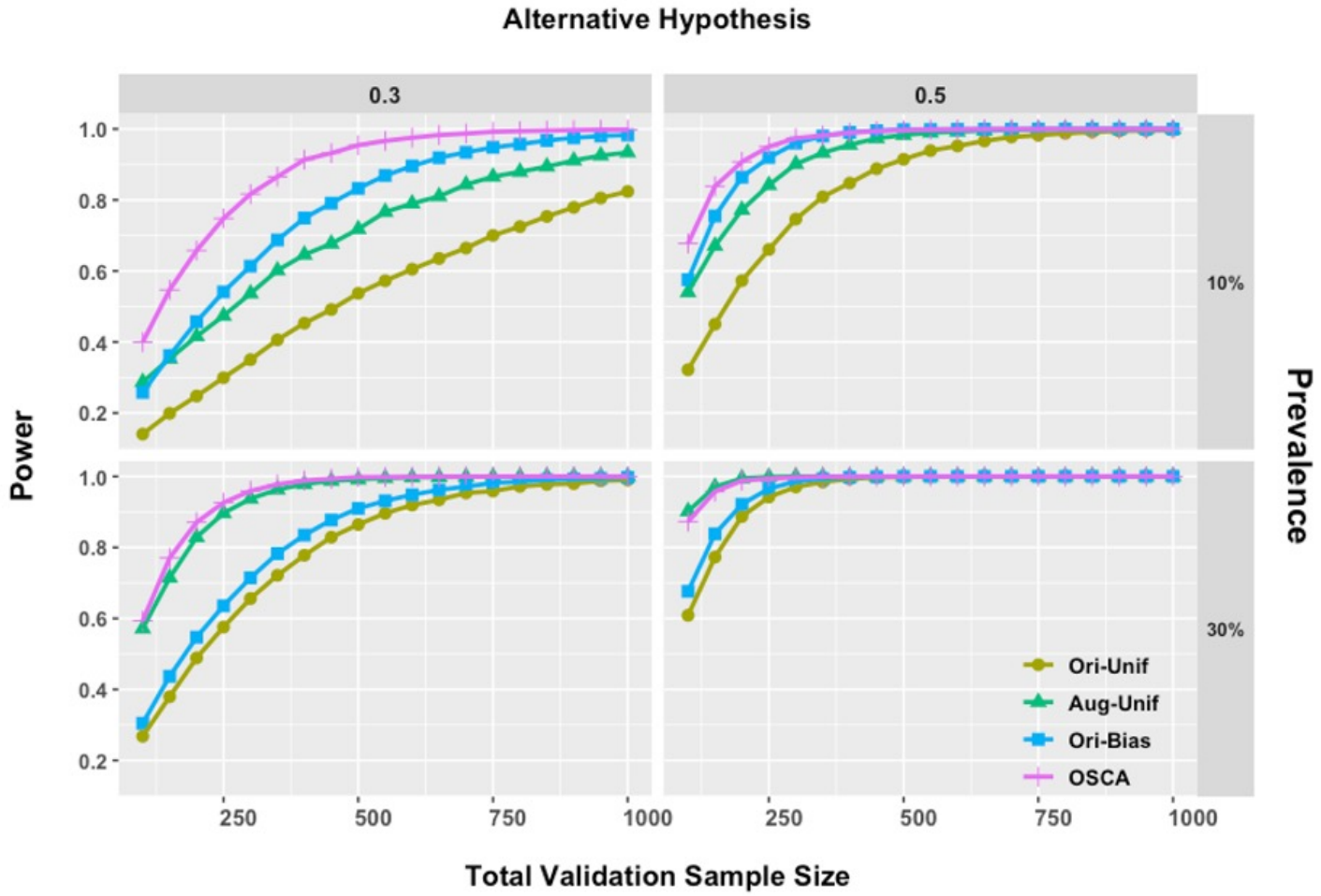
Change 1: split the data into sub-groups based on surrogate outcomes

Change 2: $\hat{\beta}_A = \hat{\beta}_V - \hat{H}_{SY}\hat{H}_S^{-1}(\hat{\gamma}_V - \hat{\gamma}_F)$,

where $\hat{H}_{SY}$ is the estimated covariance matrix between $n^{1/2}(\hat{\beta}_V - \beta_0)$ and $n^{1/2}(\hat{\gamma}_V - \gamma_0)$, and $\hat{H}_S$ is the estimated covariance matrix of $N^{1/2}(\hat{\gamma}_F - \gamma_0)$

Yin, Z., Tong, J., Chen, Y., Hubbard, R.A. and Tang, C.Y., 2021. A cost-effective chart review sampling design to account for phenotyping error in electronic health records (EHR) data. ***Journal of the American Medical Informatics Association***.

Simulation results -- Power comparison

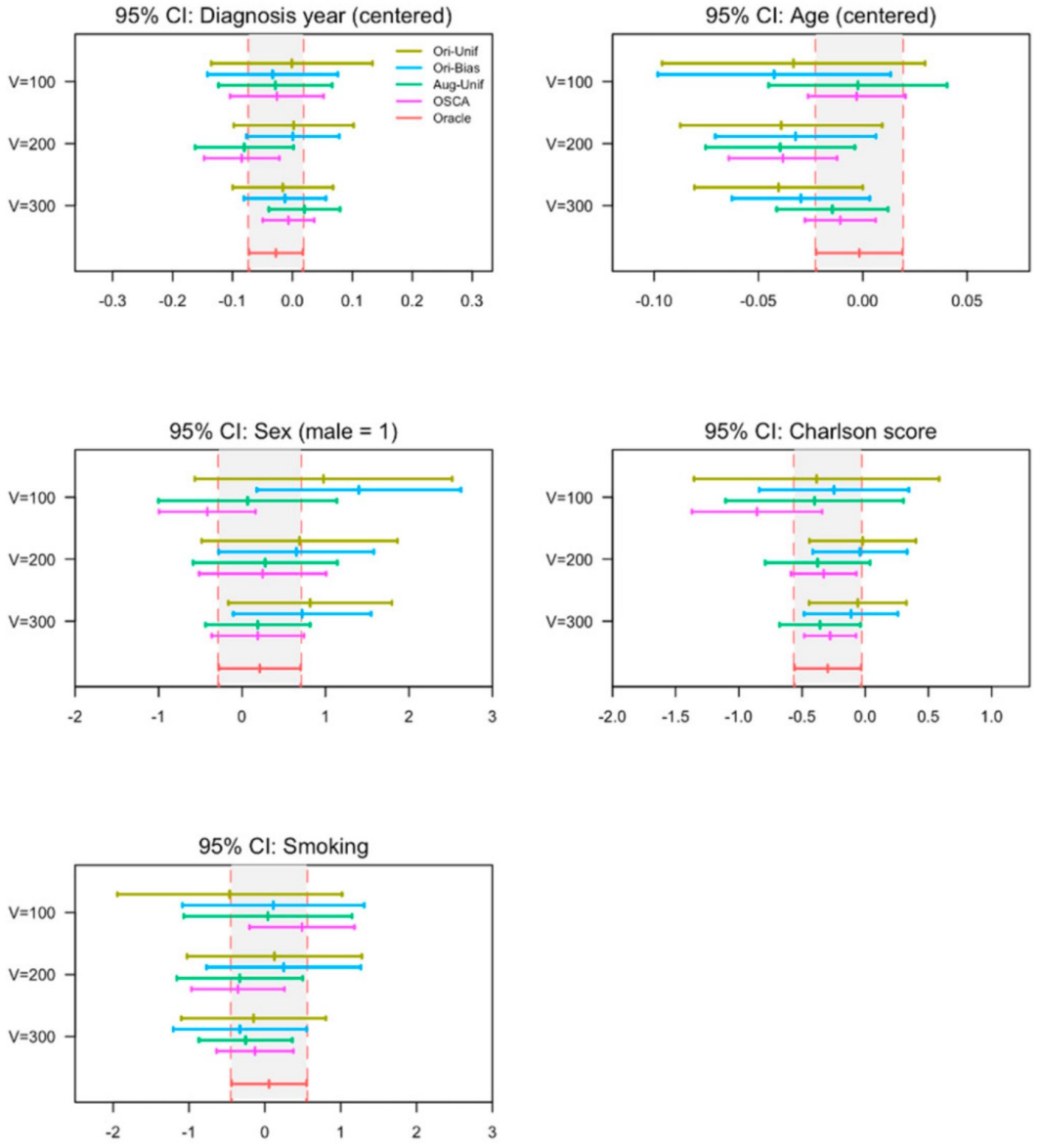


- Ori-Unif: random sampling and use validation data only
- Ori-Bias: outcome-dependent sampling and use validation data only
- Aug-Unif: random sampling with data augmentation
- OSCA: outcome-dependent sampling with data augmentation

Yin, Z., Tong, J., Chen, Y., Hubbard, R.A. and Tang, C.Y., 2021. A cost-effective chart review sampling design to account for phenotyping error in electronic health records (EHR) data. *Journal of the American Medical Informatics Association*.

Data analysis results

- EHR data on colon cancer recurrence in a cohort of patients with a primary colon cancer diagnosed and treated in the KPW healthcare system
- 1063 patients
  - age 18 years or older at the time of diagnosis of a stage I–IIIA colon cancer between 1995 and 2014.



Bias due to “imperfect” performance of phenotyping algorithms

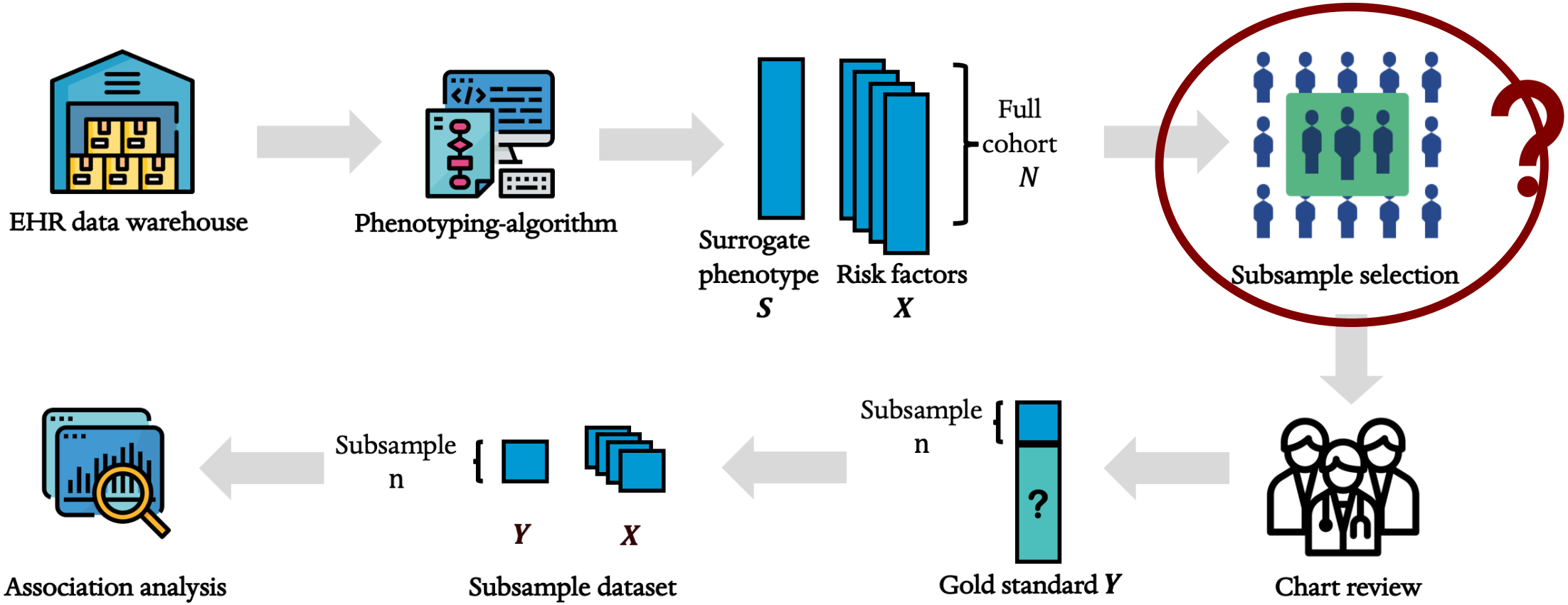
- **What can we do?**

1. **Quantify the impacts of phenotyping error in subsequent analyses**
2. **Bias reduction/correction**
   - Chart reviews for the disease status are crucial. One should always budget for chart reviews when possible.
   - In some situations, chart reviews are not available (e.g., budget constraints, time constraints), bias reduction (not correction) could still be conducted.
   - Now we have a budget for chart review. The questions are:
     - How to best use the chart review results?
     - **If investigators are involved in the design stage for chart review, we can maximize the statistical efficiency via “data-adaptive” design for multi-wave sampling in order to conduct chart reviews on the “most informative” subjects**

Bias correction

- data-adaptive two-wave sampling
- “not only case enrichment, but also maximizing the estimation efficiency for regression coefficients”

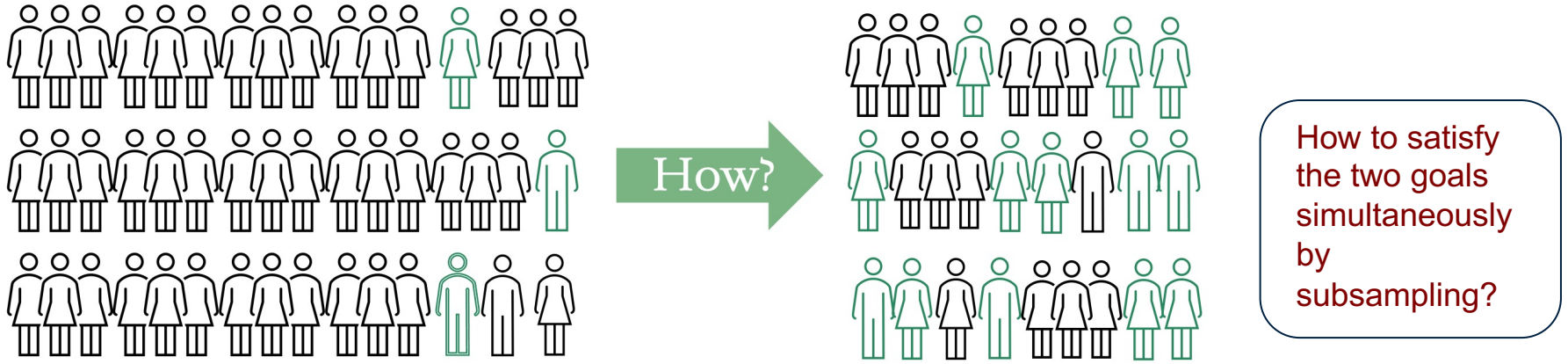
From EHRs to association analysis



Liu, X, Chubak, J, Hubbard, R and Chen, Y (2021) SAT: a Surrogate Assisted Two-wave case boosting sampling method, with application to EHR-based association studies. *Journal of the American Medical Informatics Association*, *27*(2), pp.244-253.

Existing subsample selection methods

- To handle rare disease:
  - Surrogate-guided sampling (SGS): enrich the cases in a subsample (e.g., Tan and Heagerty, 2019)
  - Local case-control (LCC) sampling: select influential subsamples (e.g., Fithian and Hastie, 2014)
- To handle true phenotype absence
  - Optimal Sampling Under Measurement Constraints: nearly response-free MSE minimization (OSUMC, Zhang et al. 2021)



How to satisfy the two goals simultaneously by subsampling?

Two goals of subsample selection:
1. Boost cases
2. Minimize the MSE of the estimates

Our solution:
1. Surrogate-assisted case-boosting sampling
2. MSE minimization targeted sampling

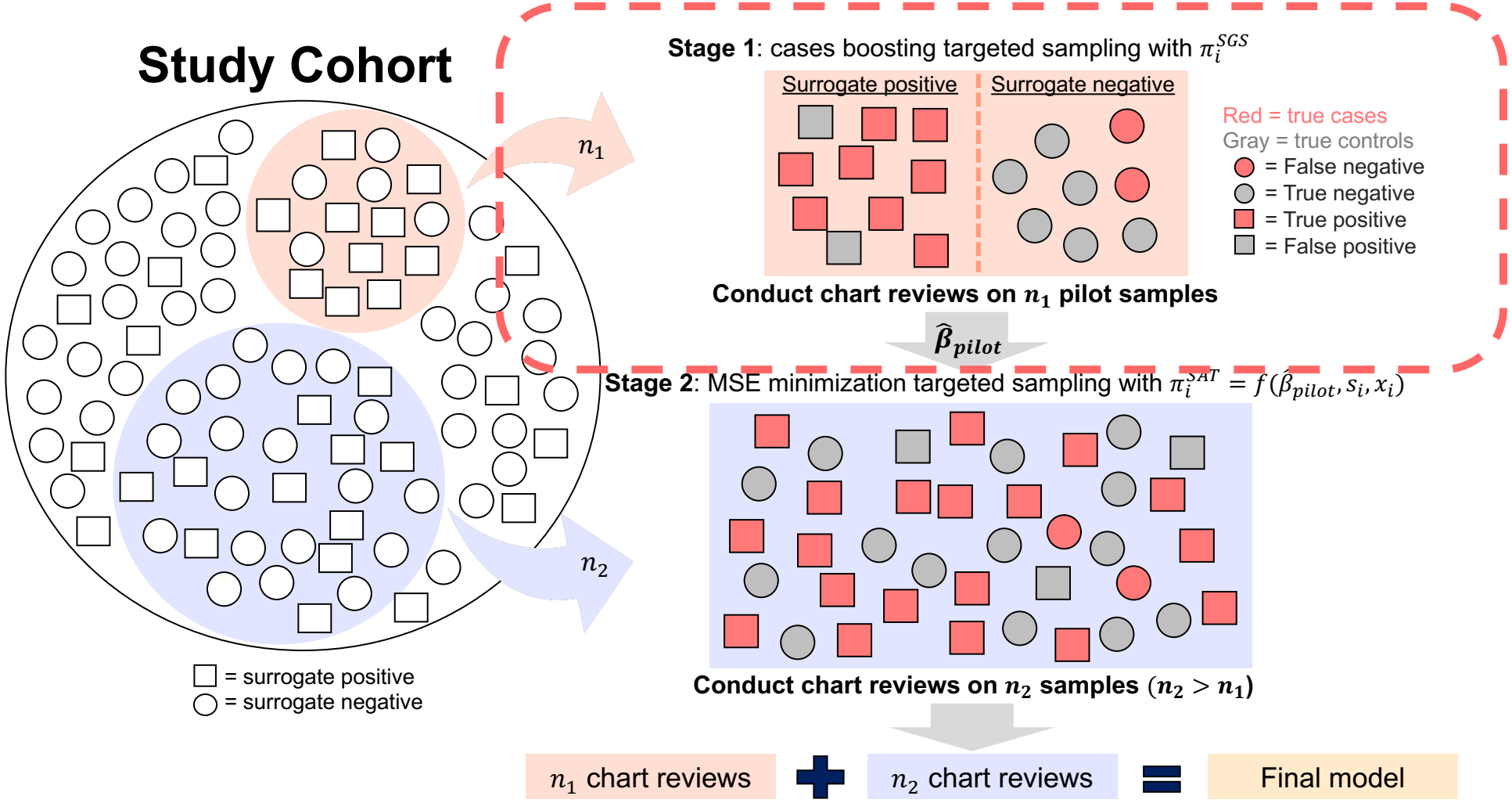
SAT: surrogate assisted two-wave case boosting sampling

- Step 1: **case boosting pilot subsample selection** -- applying **SGS** to enrich cases
  - Divide the whole dataset into two strata by $s_i = 1$ or $s_i = 0$ and apply different sampling probability π_i^{SGS}
    - $\boldsymbol{\pi_i^{SGS} = \frac{R}{n_{s1}}}$ **for** $\boldsymbol{s_i = 1}$
    - $\boldsymbol{\pi_i^{SGS} = \frac{1-R}{n_{s0}}}$ **for** $\boldsymbol{s_i = 0}$

    Notations:
      - n_{s1}: the number of positive surrogates in the whole dataset
      - n_{s0}: the number of negative surrogates in the whole dataset
      - $\boldsymbol{R = P(s_i = 1 | z_i = 1)}$: the user-specified case proportion parameter to adjust for the **case proportion** in the subsample
      - $z_i = 1$ indicate that the i-th subject is in the subsample and 0 otherwise
  - Keep sampling with replacement to obtain a pilot subsample of size n_1
  - Collect true phenotypes for the selected samples
  - Fit a **weighted** logistic regression on the pilot subsample to get the pilot estimator

$$\hat{\beta}_{pilot} = \arg\min_{\beta} \frac{1}{n_1} \sum_{i=1}^{n_1} \frac{1}{\pi_i^{SGS}} \left[y_i \log p_i(\beta) + (1 - y_i) \log\left(1 - p_i(\beta)\right) \right]$$

SAT: surrogate assisted two-wave case boosting sampling



SAT: surrogate assisted two-wave case boosting sampling

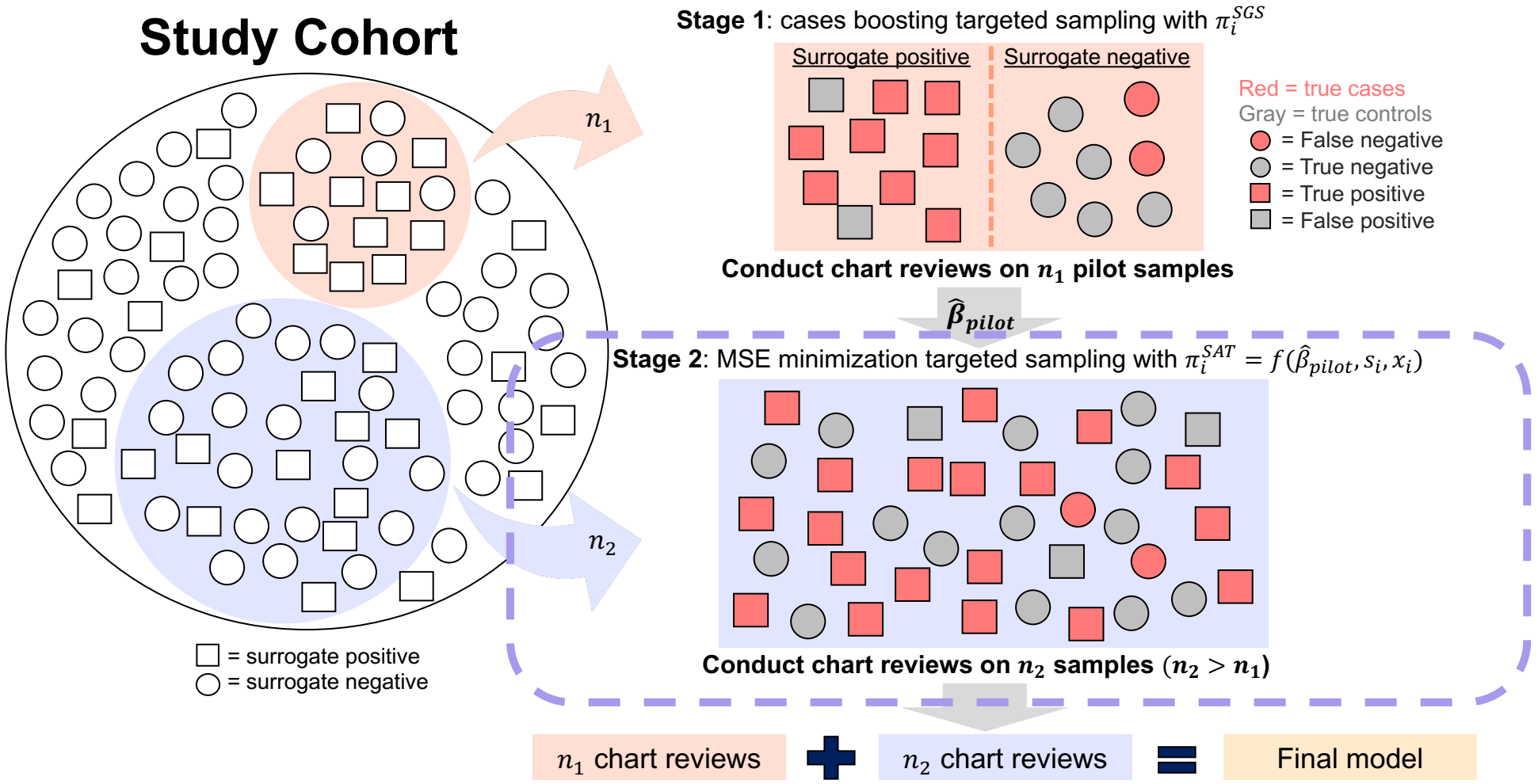
- Step 2: **MSE minimization targeted subsample selection -- minimizes the asymptotic MSE** of the subsample estimator by selecting sampling probabilities
  - Optimal subsampling procedure motivated by the A-optimality criterion (OSMAC, Wang et al. 2018):

$$\pi_i^{OSMAC} = \frac{|\boldsymbol{y_i} - p_i(\hat{\beta}_{pilot})|\,||\,M_X^{-1}x_i||_2}{\sum_{j=1}^{N}|f(s_i) - p_j(\hat{\beta}_{pilot})|\,||M_X^{-1}\,x_j||_2},$$

$$M_X = \sum_{i=1}^{N} p_i(\hat{\beta}_{pilot})(1 - p_i(\hat{\beta}_{pilot}))x_i x_i^T,\ p_i(\hat{\beta}_{pilot}) = \left(1 + e^{-x_i^T\hat{\beta}_{pilot}}\right)^{-1}$$

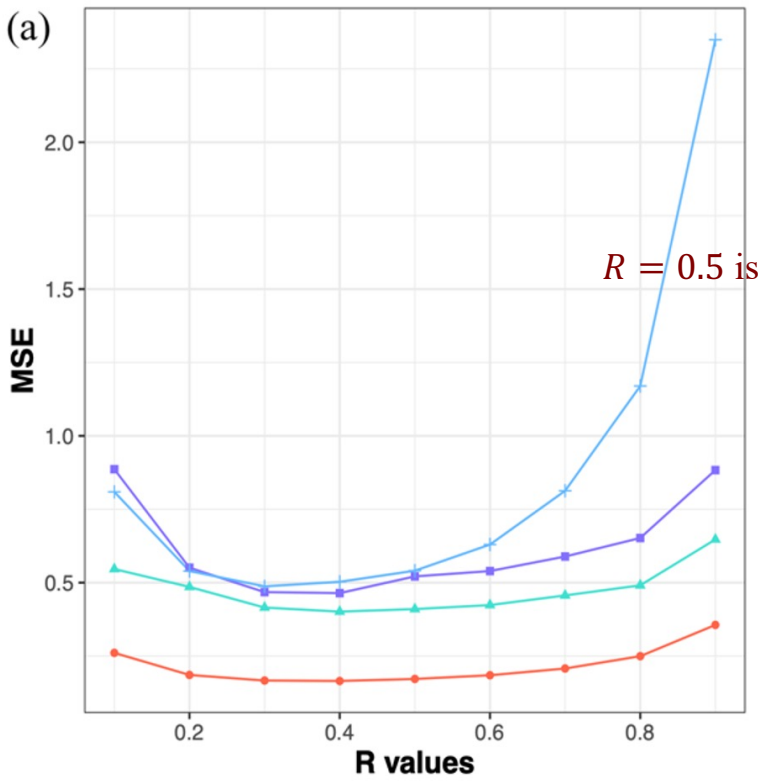
    - SAT-S: get π_i^{SAT-S} by **replacing $\boldsymbol{y_i}$ with $\boldsymbol{s_i}$**
    - SAT-cY: get π_i^{SAT-cY} by **replacing $\boldsymbol{y_i}$ with $\boldsymbol{E(y_i = 1|s_i = s)}$**
  - Conduct the second-stage sampling by sampling with replacement to get a subsample of size n_2 $(n_2 > n_1)$
  - Collect true phenotypes for the selected samples
  - Pool all the subsamples $n = n_1 + n_2$ and conduct weighted regression
    - $\hat{\beta}_{SAT} = \arg\min_{\beta} \frac{1}{n_1+n_2}(\sum_{i\in pilot\ set}\frac{1}{\pi_i^{SGS}}\left[y_i \log p_i(\beta) + (1-y_i)\log(1 - p_i(\beta))\right] + \sum_{j\in second\ stage\ set}\frac{1}{\pi_j^{SAT}}\left[y_j \log p_j(\beta) + (1-y_j)\log\left(1 - p_j(\beta)\right)\right])$.

SAT: surrogate assisted two-wave case boosting sampling



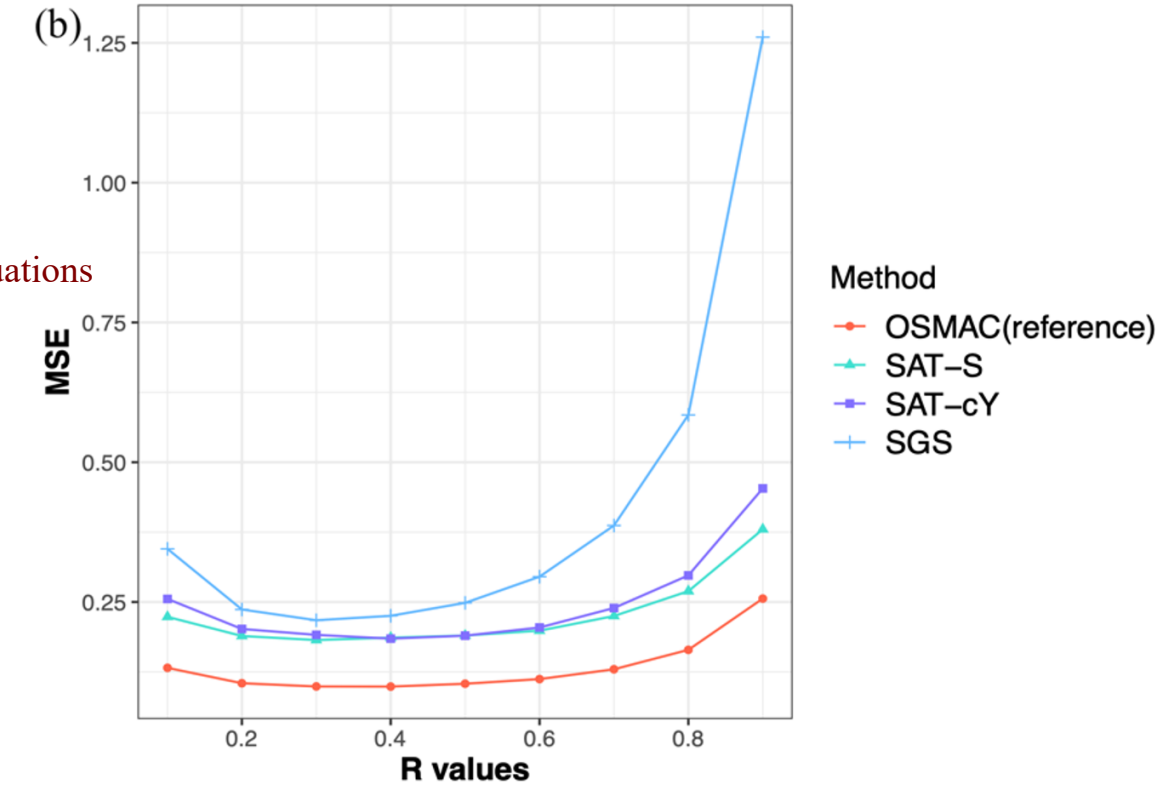
Simulation results – R's effects

- Setting 4A: case prevalence 1.56% ($n_1 = 300$ and $n_2 = 800$)



$R = 0.5$ is a safe choice for most situations

- Setting 4A: case prevalence 5.39% ($n_1 = 300$ and $n_2 = 800$)



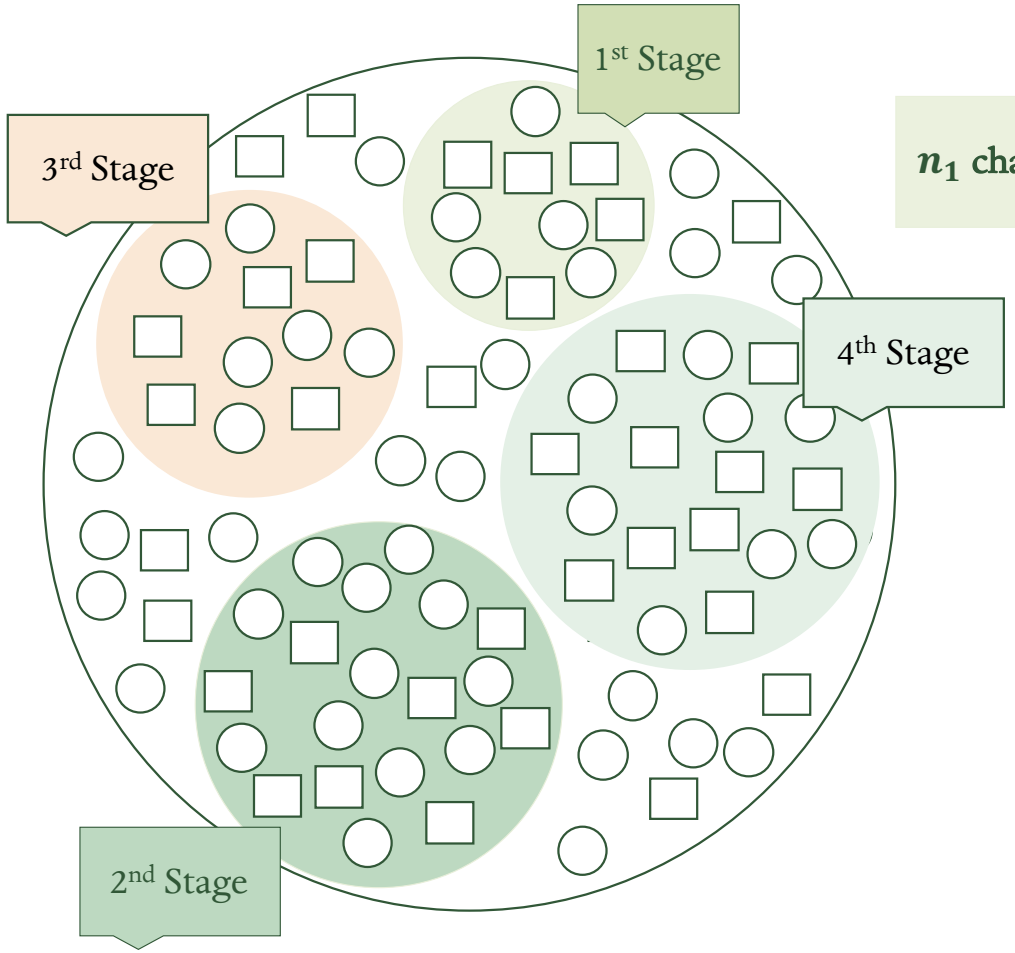
SAT's benefits

- Compared to OSUMC (outcome-free sampling):
  - SAT also handles low prevalence outcome
  - SAT further exploits the information in the surrogate to aid sampling
- Compared to OSMAC (outcome-dependent-optimal sampling):
  - SAT does not require true phenotypes for the full cohort
- Compared to SGS (Surrogate-guided sampling):
  - SAT utilizes an additional MSE minimization rule in the second-stage sampling to improve estimation accuracy

Remaining issues:

- SAT relies on the asymptotic property of the subsample estimator derived in OSMAC which does not consider the imbalanced class distribution. Asymptotic properties of the subsample estimator in the imbalanced class scenario need to be developed.
- For extremely rare outcomes (of prevalence 1% or lower), even the OSMAC estimator has a large bias. The sampling approach is not the best choice and other methods should be considered.
- In some investigations of rare diseases, a case-finding algorithm is commonly employed to capture as many cases as possible. These algorithms guarantee a high sensitivity of the surrogate, but have poor specificity. Extensions of SAT are needed.
- Extension to settings that utilizes multiple surrogates.
- Extension to data-adaptive multi-wave sampling.

Future work: multi-wave sampling



n_1 chart reviews + n_2 chart reviews → Model 1

guide

+ n_3 chart reviews → Model 2

guide

+ .....

= Final model

TIME FOR A BREAK

Distributed Analysis

Motivation: a large-scale CER research using observational data

THE LANCET
Volume 394, Issue 10211, 16–22 November 2019, Pages 1816-1826

Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis

Marc A Suchard, Martijn J Schuemie, Harlan M Krumholz, Seng Chan You, RuiJun Chen, Nicole Pratt, Christian G Reich, Jon Duke, David Madigan, George Hripcsak, Patrick B Ryan

Lancet 2019; 394: 1816–26
Published Online October 24, 2019 https://doi.org/10.1016/S0140-6736(19)32317-7
See Comment page 1782

Department of Biostatistics, Fielding School of Public Health (Prof M A Suchard MD, M J Schuemie PhD), and Department of Biomathematics, David Geffen School of Medicine at UCLA (Prof M A Suchard), University of California, Los Angeles, CA, USA; Epidemiology Analytics, Janssen Research &

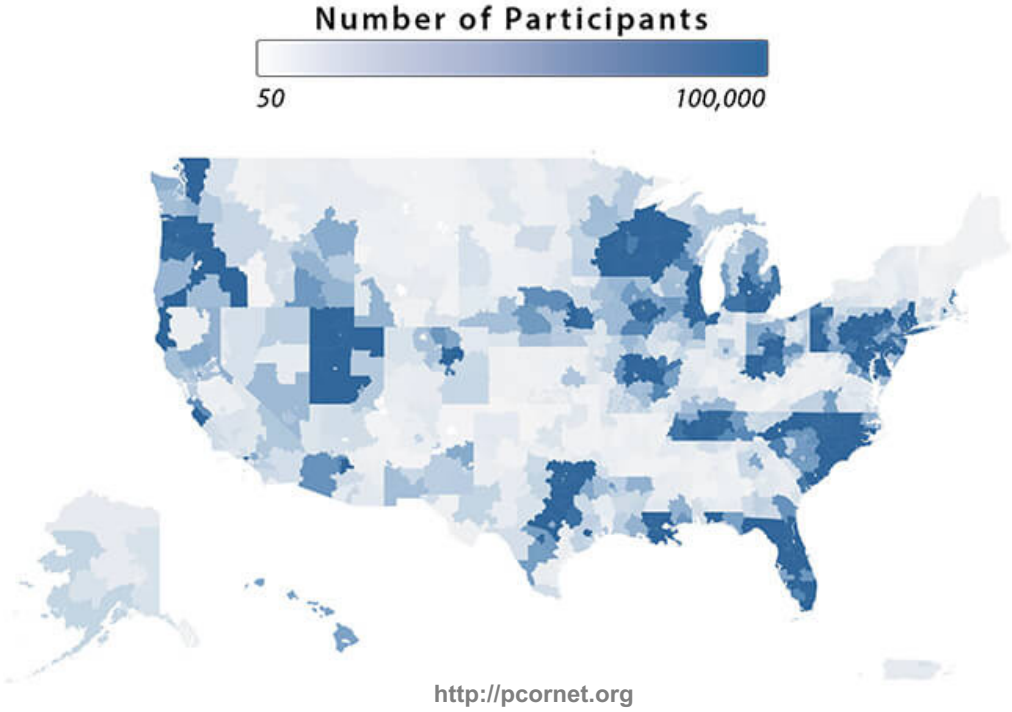
Summary

Background Uncertainty remains about the optimal monotherapy for hypertension, with current guidelines recommending any primary agent among the first-line drug classes thiazide or thiazide-like diuretics, angiotensin-converting enzyme inhibitors, angiotensin receptor blockers, dihydropyridine calcium channel blockers, and non-dihydropyridine calcium channel blockers, in the absence of comorbid indications. Randomised trials have not further refined this choice.

Methods We developed a comprehensive framework for real-world evidence that enables comparative effectiveness and safety evaluation across many drugs and outcomes from observational data encompassing millions of patients, while minimising inherent bias. Using this framework, we did a systematic, large-scale study under a new-user cohort design to estimate the relative risks of three primary (acute myocardial infarction, hospitalisation for heart failure, and stroke) and six secondary effectiveness and 46 safety outcomes comparing all first-line classes across a global network of six administrative claims and three electronic health record databases. The framework addressed residual confounding, publication bias, and p-hacking using large-scale propensity adjustment, a large set of control outcomes, and full disclosure of hypotheses tested.

Motivation: Distributed Health Data Networks

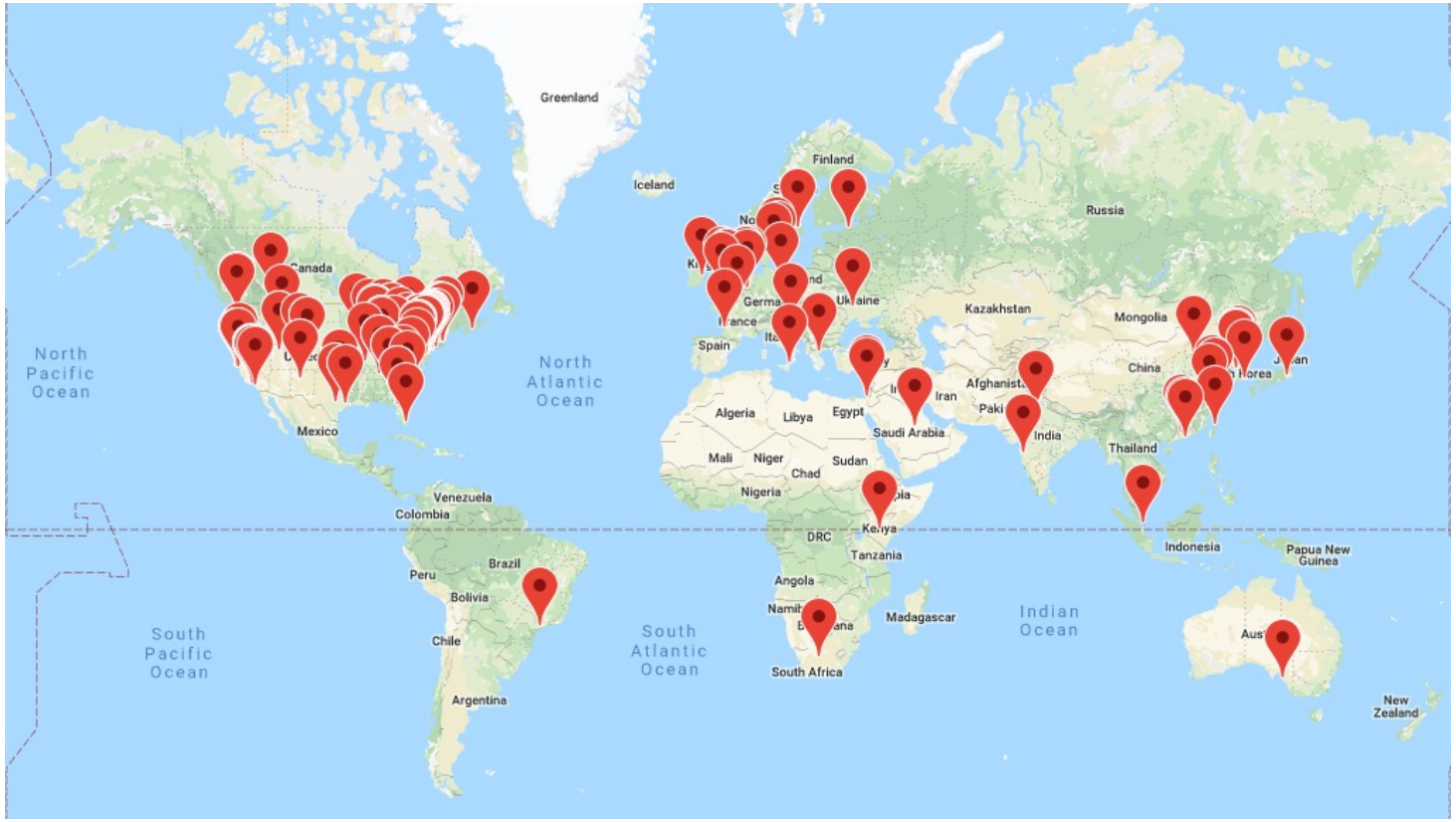
- No data centralization: data holders maintain control over data
  - Participants adopt common data model (CDM)
  - Analyses performed distributively (email, central server, etc.)
  - No patient-level data transfer
- Sentinel Initiative
  - FDA: Post-market safety surveillance
  - 16 data partners contribute billing data, EHRs
- PCORnet
  - National patient-centered clinical research network for comparative effectiveness research
  - 68 million patients



Motivation: OHDSI

OHDSI's global research community

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS



Promise and challenges of multi-site studies

- Covers broader population
  - Results are more generalizable
  - Better statistical power
  - Opportunities of studying rare diseases
- Ecosystem of bringing together expertise from different investigators

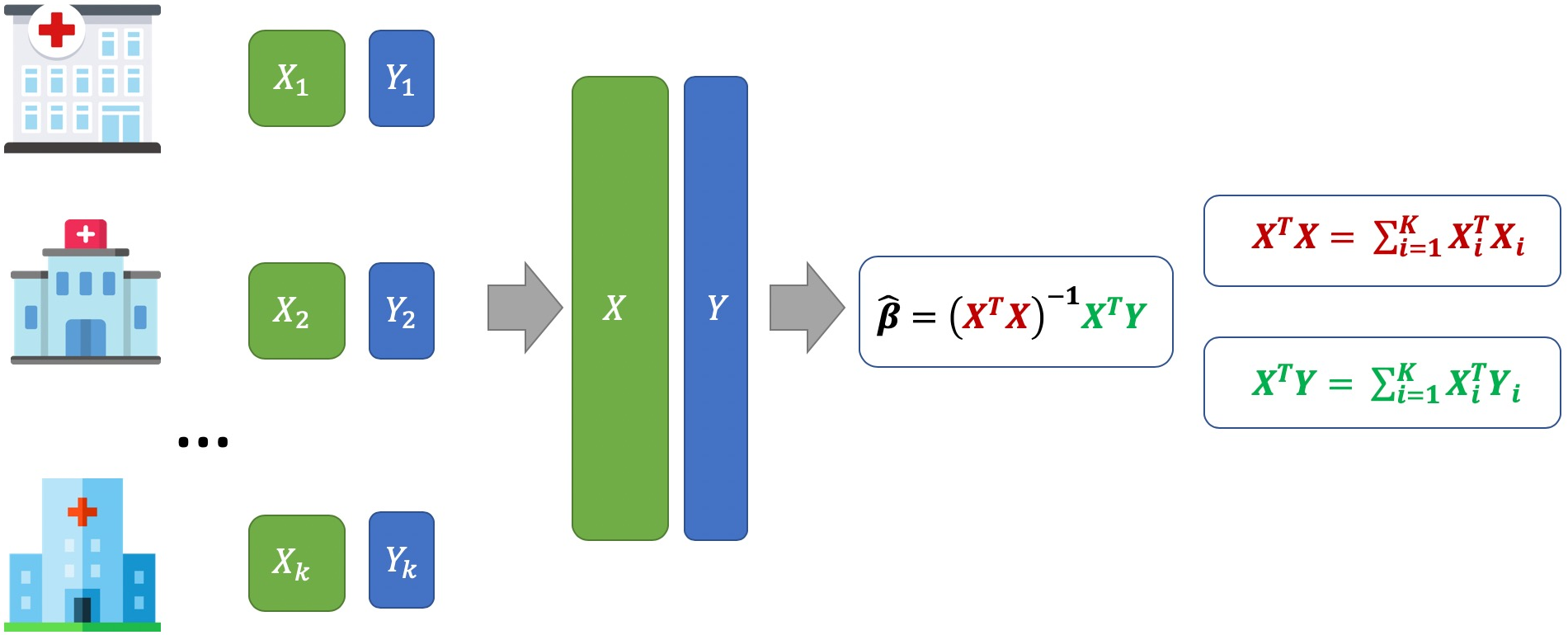
- Sharing individual-patient level data is challenging

- Need iterative collaboration/coordination among investigators from different institutes

  (consuming lots of time, effort, funding)

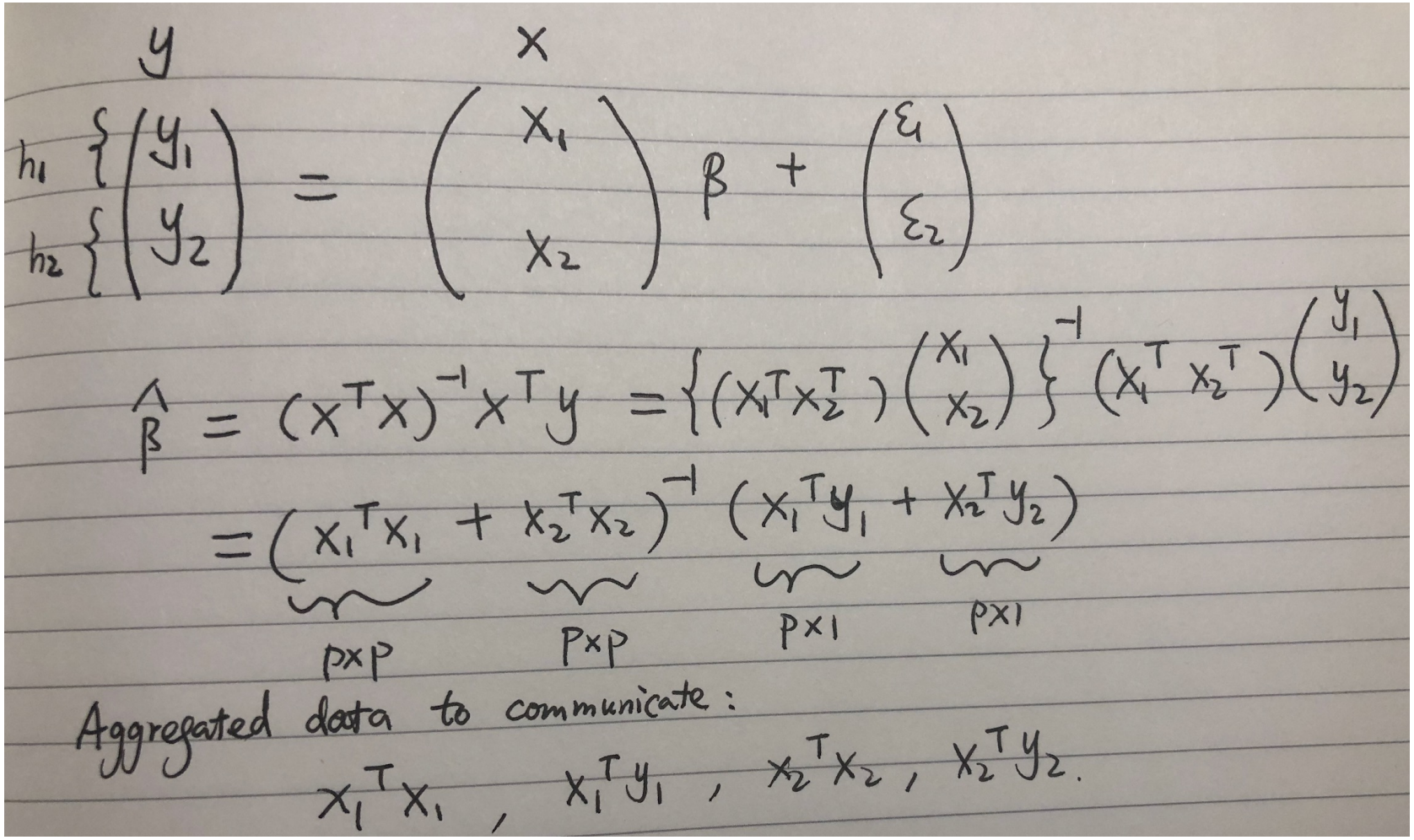
Distributed Linear Regression - a useful result

- Do we really need to share patient-level data?



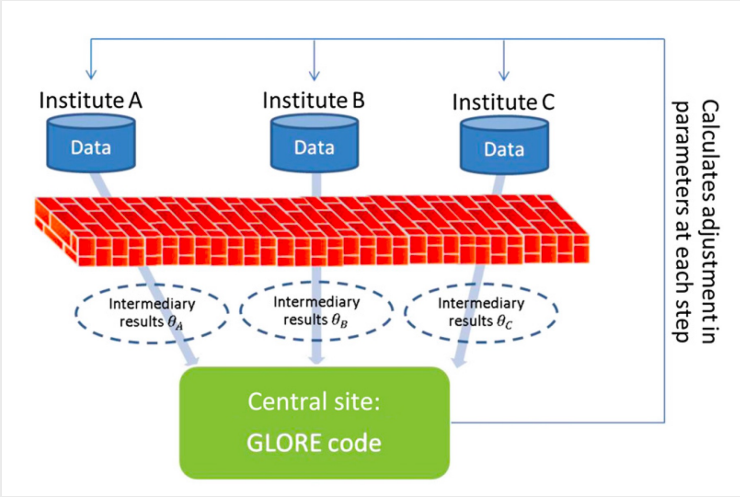
Chen et al.(2006) Regression cubes with lossless compression and aggregation.

A bit more detail



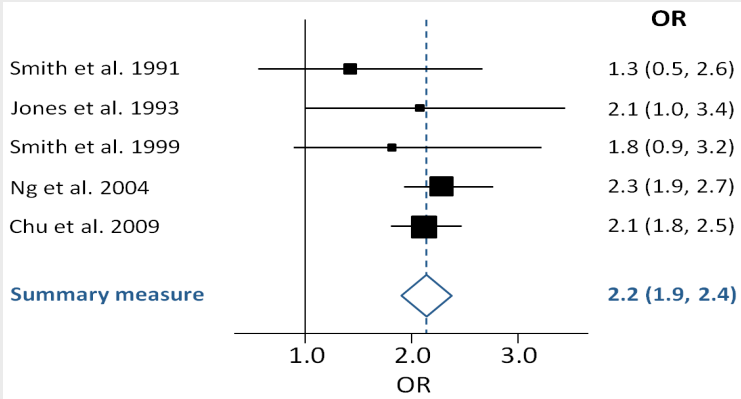
Existing algorithms of distributed analysis

- **Iterative ---** iteratively updates the estimates using aggregated data. Often each step is a distributed Newton-Raphson update. Implemented in pSCANNER.
  - **Binary outcome**: Grid Binary LOgistic REgression (GLORE) (Wu et al. 2012)
  - **Time-to-event outcome**: WebDISCO: a Web service for distributed Cox model (Lu et al. 2015)



Wu et al. 2012, *JAMIA*

- **One-shot (non-iterative) ---** only requires the collaborative sites to exchange aggregated data ***once***
  - Averaging local estimates
    - Meta-analysis, distributed PCA (Fan et al. 2018), distributed LDA (Lu et al. 2019), ...
  - Surrogate likelihood (Jordan et al. 2018)



Our Standards

Privacy-preserving --- only aggregated data are communicated

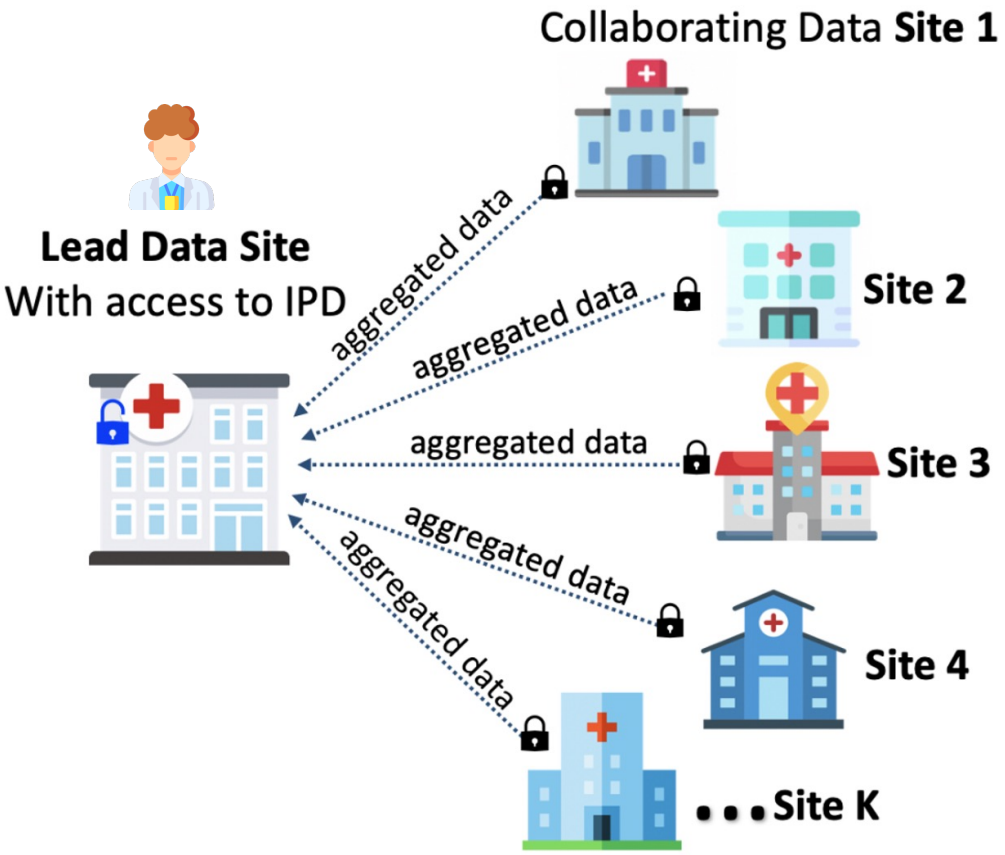
Formal privacy-protection techniques (e.g., differential privacy, homomorphic encryption) can be added, out of the scope of this talk

Accurate --- estimates from algorithms are close to the pooled analysis results

Communication-efficient --- avoid iterative communications across collaborating sites

Inspiration: an interesting observation – unique architecture of distributed research networks

- An investigator at a hospital do have access to the patient-level data of that hospital
- Collaborating hospitals can share aggregated data



JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2019, VOL. 114, NO. 526, 668–681: Theory and Methods
https://doi.org/10.1080/01621459.2018.1429274

Taylor & Francis
Taylor & Francis Group

Communication-Efficient Distributed Statistical Inference

Michael I. Jordan[a], Jason D. Lee[b], and Yun Yang[c]

[a]Department of Statistics, University of California Berkeley, Berkeley, CA; [b]Institute of Computational and Mathematical Engineering, Stanford University, Cupertino, CA; [c]Statistical Science, Duke University, Durham, NC

ABSTRACT

We present a *communication-efficient surrogate likelihood* (CSL) framework for solving distributed statistical inference problems. CSL provides a communication-efficient surrogate to the global likelihood that can be used for low-dimensional estimation, high-dimensional regularized estimation, and Bayesian inference. For low-dimensional estimation, CSL provably improves upon naive averaging schemes and facilitates the construction of confidence intervals. For high-dimensional regularized estimation, CSL leads to a minimax-optimal estimator with controlled communication cost. For Bayesian inference, CSL can be used to form a communication-efficient quasi-posterior distribution that converges to the true posterior. This quasi-posterior procedure significantly improves the computational efficiency of Markov chain Monte Carlo (MCMC) algorithms even in a nondistributed setting. We present both theoretical analysis and experiments to explore the properties of the CSL approximation. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received December 2016
Revised December 2017

KEYWORDS
Communication efficiency;
Distributed inference;
Likelihood approximation

Can we use local hospital's patient-level data and collaborating hospitals' aggregated data to do multi-site data analysis?

ODAL – One-Shot Distributed Algorithm for Logistic Regression

Journal of the American Medical Informatics Association, 27(3), 2020, 376–385
doi: 10.1093/jamia/ocz199
Advance Access Publication Date: 9 December 2019
Research and Applications

Research and Applications

Learning from electronic health records across multiple sites: A communication-efficient and privacy-preserving distributed algorithm

Rui Duan [1], Mary Regina Boland [1], Zixuan Liu[2], Yue Liu,[3] Howard H. Chang,[4] Hua Xu,[5] Haitao Chu,[6] Christopher H. Schmid,[7] Christopher B. Forrest,[8] John H. Holmes,[1] Martijn J. Schuemie [9], Jesse A. Berlin,[9] Jason H. Moore,[1] and Yong Chen [1]

Algorithm 1 ODAL1

1. Initial value: obtain $\bar{\beta} = \arg\max_{\beta} L_1(\beta)$ using data in the local site 1.
2. Initial communication: transfer $\bar{\beta}$ to the other sites.
3. For site j = 2 to K,
4. **do** compute $\nabla L_j(\bar{\beta})$ using equation 5
5. transfer $\nabla L_j(\bar{\beta})$ to site 1
6. **end**
7. Compute the surrogate likelihood $\tilde{L}^1(\beta)$ using equation 2
8. Obtain $\tilde{\beta}^1 = \arg\max_{\beta} \tilde{L}^1(\beta)$
9. Obtain $V(\tilde{\beta}^1)$ using Supplementary Material equation S1
10. return $\tilde{\beta}^1$ and $V(\tilde{\beta}^1)$.

Algorithm 2 ODAL2

1. Initial value: obtain $\bar{\beta} = \arg\max_{\beta} L_1(\beta)$ using data in the local site 1.
2. Initial communication: transfer $\bar{\beta}$ to the other sites.
3. For site j = 2 to K,
4. do compute $\nabla L_j(\bar{\beta}), \nabla^2 L_j(\bar{\beta})$
5. transfer $\nabla L_j(\bar{\beta}), \nabla^2 L_j(\bar{\beta})$ to site 1
6. end
7. Compute the surrogate likelihood $\tilde{L}^2(\beta)$ using equation (6)
8. Obtain $\tilde{\beta}^2 = \arg\max_{\beta} \tilde{L}^2(\beta)$
9. Obtain $V(\tilde{\beta}^2)$ using Supplementary Material equation S2
10. return $\tilde{\beta}^2$ and $V(\tilde{\beta}^2)$

The likelihood functions

- **Combined likelihood function** (if data could be shared)

$$L(\beta) = \frac{1}{nK}\sum_{j=1}^{K}\sum_{i=1}^{n}\{y_{ij}x_{ij}^{T}\beta - \log\{1+\exp(x_{ij}^{T}\beta)\}\}$$

- **Local likelihood function** (assume local site to be the first site, j=1)

$$L_1(\beta) = \frac{1}{n}\sum_{i=1}^{n}\{y_{i1}x_{i1}^{T}\beta - \log\{1+\exp(x_{i1}^{T}\beta)\}\}$$

How to borrow aggregated information from other sites to make $L_1(\beta)$ more like $L(\beta)$?

The surrogate likelihood (SL) approach

- For an initial value $\bar{\beta}$,

$$L(\beta) = L(\bar{\beta}) + \nabla L(\bar{\beta})^T(\beta - \bar{\beta}) + \sum_{t=2}^{\infty} \frac{1}{t!} \nabla^t L(\bar{\beta})(\beta - \bar{\beta})^{\otimes t}$$

$$L_1(\beta) = L_1(\bar{\beta}) + \nabla L_1(\bar{\beta})^T(\beta - \bar{\beta}) + \sum_{t=2}^{\infty} \frac{1}{t!} \nabla^t L_1(\bar{\beta})(\beta - \bar{\beta})^{\otimes t}$$

$$\sum_{t=2}^{\infty} \frac{1}{t!} \nabla^t L_1(\bar{\beta})(\beta - \bar{\beta})^{\otimes t} = L_1(\beta) - L_1(\bar{\beta}) - \nabla L_1(\bar{\beta})^T(\beta - \bar{\beta})$$

First-order SL function

$$\tilde{L}^1(\beta) = L_1(\beta) + \{\nabla L(\bar{\beta}) - \nabla L_1(\bar{\beta})\}^T \beta$$

$$\nabla L(\bar{\beta}) = \frac{1}{K} \sum \nabla L_j(\bar{\beta}); \quad \nabla L_j(\bar{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \{y_{ij} - \text{expit}(x_{ij}^T \bar{\beta})\} x_{ij}$$

Increase the approximation accuracy

$$L(\beta) = L(\bar{\beta}) + \nabla L(\bar{\beta})^T(\beta - \bar{\beta}) + \frac{1}{2}\nabla^2 L(\bar{\beta})(\beta - \bar{\beta})^{\otimes 2} + \sum_{t=3}^{\infty} \frac{1}{t!}\nabla^t L(\bar{\beta})(\beta - \bar{\beta})^{\otimes t}$$

$$L_1(\beta) = L_1(\bar{\beta}) + \nabla L_1(\bar{\beta})^T(\beta - \bar{\beta}) + \frac{1}{2}\nabla^2 L_1(\bar{\beta})(\beta - \bar{\beta})^{\otimes 2} + \sum_{t=3}^{\infty} \frac{1}{t!}\nabla^t L_1(\bar{\beta})(\beta - \bar{\beta})^{\otimes t}$$

Second-order SL function

$$\tilde{L}^2(\beta) = L_1(\beta) + \{\nabla L(\bar{\beta}) - \nabla L_1(\bar{\beta})\}^T\beta + \frac{1}{2}(\beta - \bar{\beta})^T\{\nabla^2 L(\bar{\beta}) - \nabla^2 L_1(\bar{\beta})\}^T(\beta - \bar{\beta})$$

$$\nabla^t L(\bar{\beta}) = \frac{1}{K}\sum \nabla^t L_j(\bar{\beta}), \text{for } t = 1, 2.$$

$$\nabla L_j(\bar{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\{y_{ij} - \text{expit}(x_{ij}^T\bar{\beta})\}x_{ij};\ \nabla^2 L_j(\bar{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\text{expit}(x_{ij}^T\bar{\beta})\{1 - \text{expit}(x_{ij}^T\bar{\beta})\}x_{ij}\ x_{ij}^T.$$

Surrogate likelihood estimates

- First-order algorithm (ODAL1)

$$\tilde{\beta}^1 = \text{argmax}_\beta \, \tilde{L}^1(\beta)$$

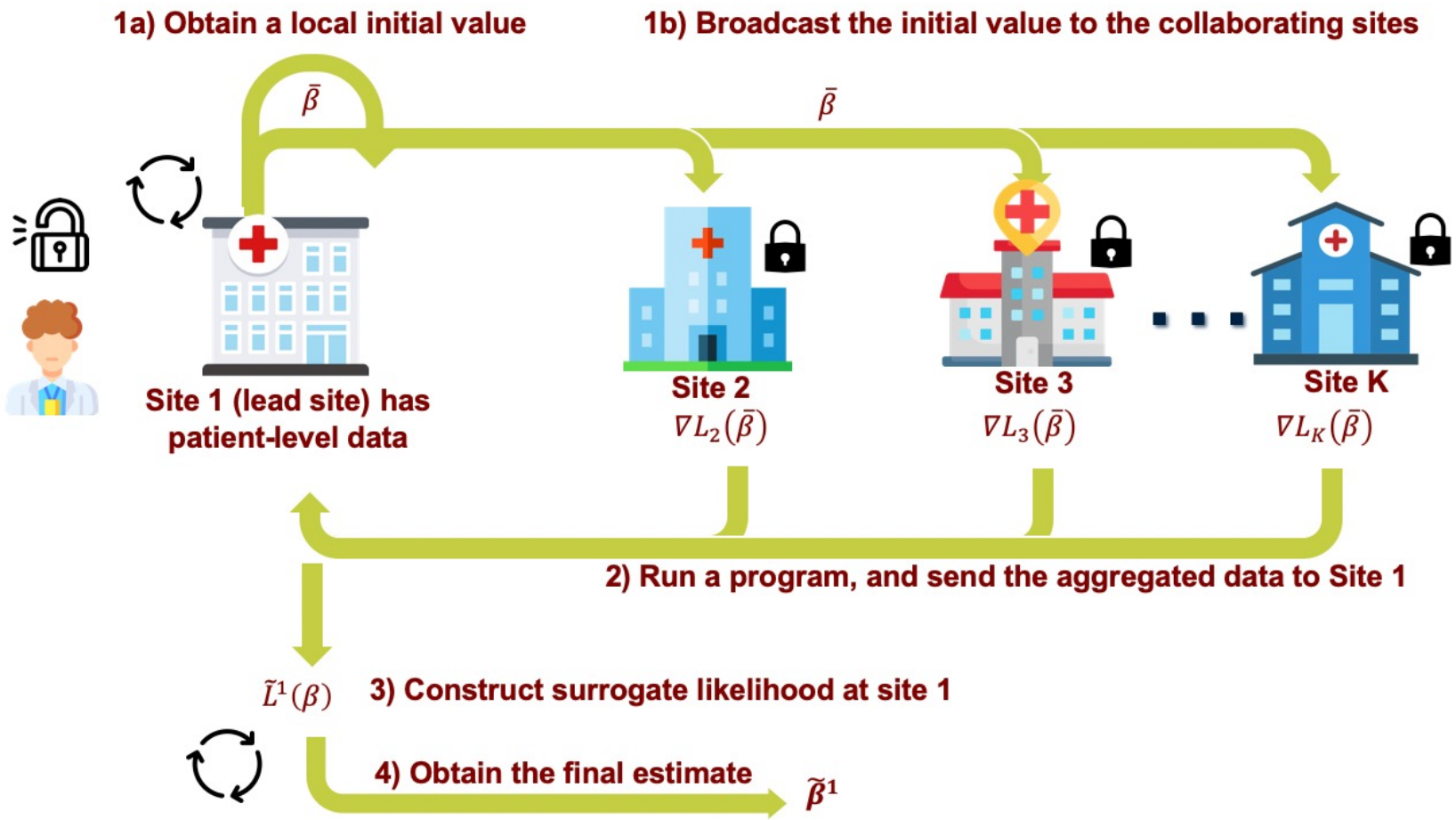
- Second-order algorithm (ODAL2)

$$\tilde{\beta}^2 = \text{argmax}_\beta \, \tilde{L}^2(\beta)$$

- Initial estimator:

$$\bar{\beta} = \text{argmax}_\beta \, L_1(\beta)$$

ODAL step-by-step illustration:



Statistical Inference

Theorem

Under mild regularity conditions, the proposed estimator $\tilde{\beta}$ *satisfies,*

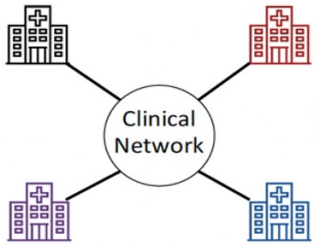
$$\sqrt{Kn}(\tilde{\beta} - \beta^*) \to N(0, \{\mathbb{E}\nabla^2 f(y, x; \beta^*)\}^{-1})$$

as $n \to \infty$, *and* $K \ll n$.

- Inference at local site (no extra communication)

$$\hat{V} = \frac{1}{Kn}\{\nabla^2 L_1(\tilde{\beta})\}^{-1}$$

ODAL advantages:

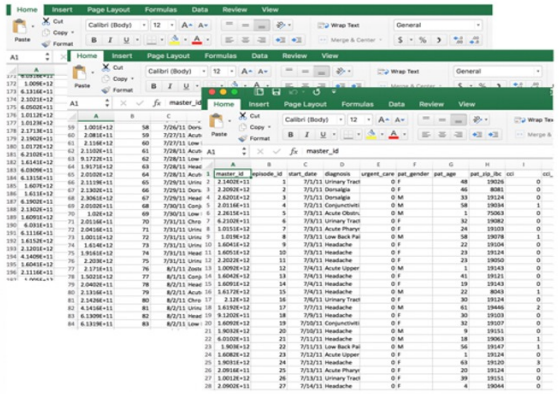


Research question:
What are the risk factors of acute myocardial infarction?

Fit a logistic model with 4 risk factors.

How to perform multicenter analysis?

Strategy 1: pooled analysis



Strategy 2: ODAL

Two results are nearly the same. (Duan el al. 2019)

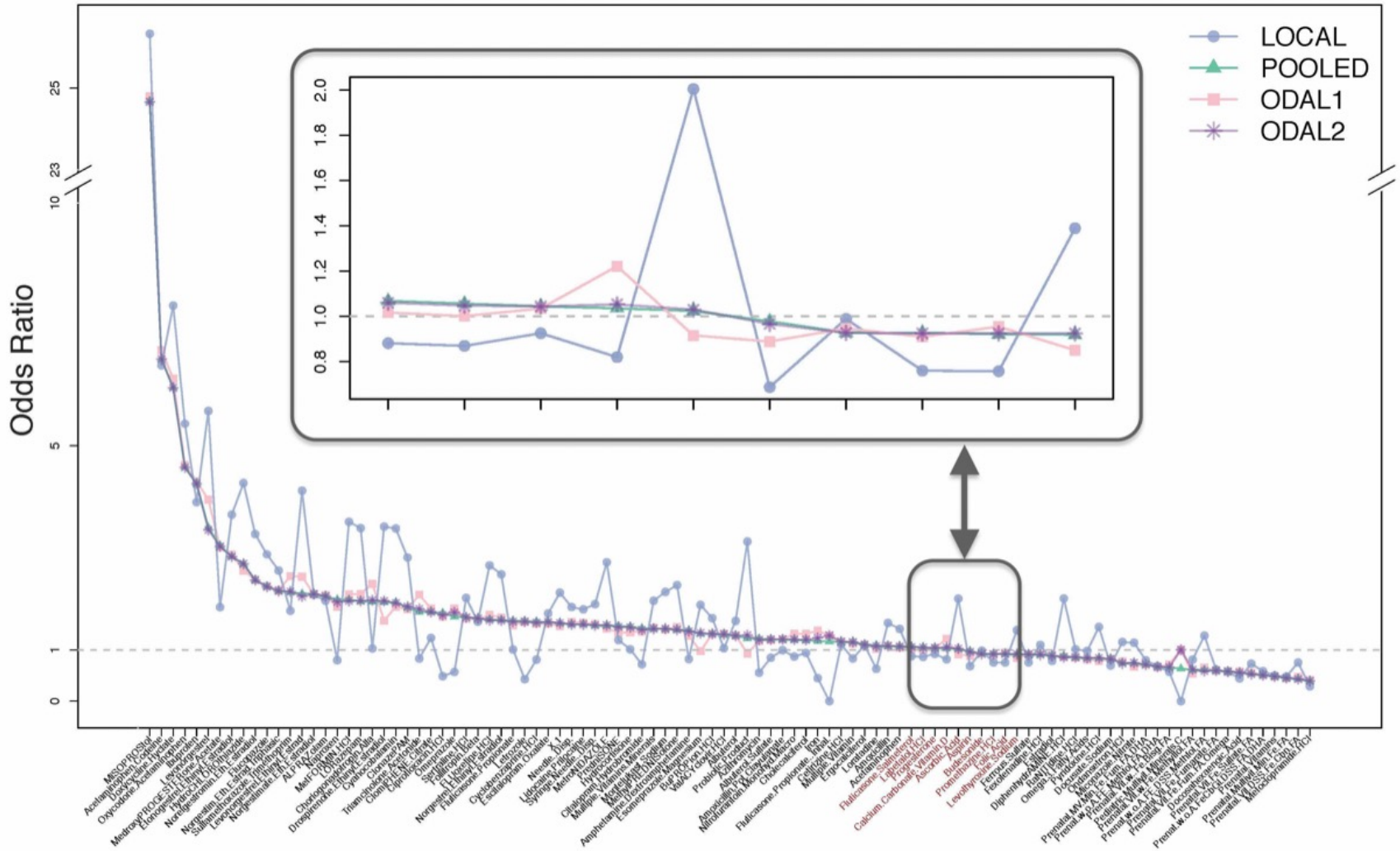
Evaluation through EHR data from Penn Medicine

- Outcome: normal pregnancy (Z34 ICD-10 codes or a V22 ICD-9 code) or a fetal loss (ICD-9 code 630-639 or ICD-10 code O00-O08)
- Exposure: 100 most common medications prescribed during pregnancy, prevalence ranging from 0.05% - 20%.

Table 1. Demographics of Pregnancies Treated at UPenn Health System (UPHS)

Demographics	Normal Pregnancy (N=30,810)	Fetal Loss (M=4,763)	P-value
Race			
White *	13911 (45.2%)	2291 (48.1%)	
African American	12918 (41.9%)	1871 (39.3%)	
Other	1916 (6.2%)	274 (5.8%)	
Asian	2065 (6.7%)	327 (6.9%)	
Age	29.40	32.15	<0.001
Weight (pounds)	123.45	115.43	<0.001
Body Mass Index	16.95	16.61	0.043

Results



PDA MENU

ODAL

ODAC

ODAP

ODAH

Hetero-aware

DLMM

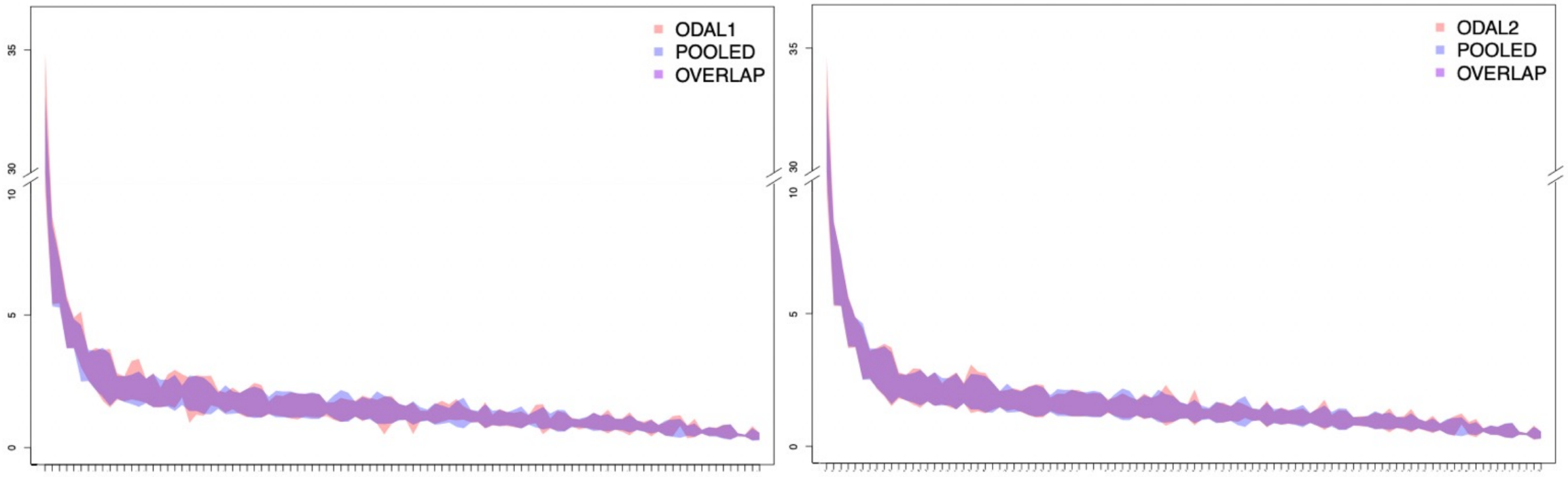
dPQL

ODACH

dCLR

dist-EM

Results - inference



- ODAL2 needs to transfer an extra hessian matrix.
- ODAL2 provides more accurate estimation than ODAL1.

What's next?

ODAC – One-shot Distributed Algorithm for Cox Proportional Hazards model

Journal of the American Medical Informatics Association, 27(7), 2020, 1028–1036
doi: 10.1093/jamia/ocaa044
Advance Access Publication Date: 6 July 2020
Research and Applications

Research and Applications

Learning from local to global: An efficient distributed algorithm for modeling time-to-event data

Rui Duan,[1,†] Chongliang Luo,[1,†] Martijn J. Schuemie,[2,†] Jiayi Tong,[1] C. Jason Liang,[3] Howard H. Chang,[4] Mary Regina Boland,[1] Jiang Bian,[5,6] Hua Xu,[7] John H. Holmes,[1] Christopher B. Forrest,[8] Sally C. Morton,[9] Jesse A. Berlin,[10] Jason H. Moore,[1] Kevin B. Mahoney,[11] and Yong Chen[1]

ODAC algorithm

(1) Initialization

In site $k = 1$ to K,

do

Fit a Cox model and obtain the local estimate $\hat{\beta}_k$ and the variance estimate $\hat{V}_k$; ***broadcast*** $\hat{\beta}_k$, $\hat{V}_k$, and the set of unique event time points in site k.

end

(2) Local surrogate estimator

In Site $k = 1$ to K,

do

obtain $\bar{\beta}$ using (4), and all the unique event time points across all sites $t_1 \ldots t_d$; calculate and broadcast the intermediate summary-level statistics $U_j(\mathcal{T})$, $W_j(\mathcal{T})$ and $Z_j(\mathcal{T})$; construct the surrogate likelihood $\tilde{L}_k(\beta)$ by (3) treating the k-th site as the local site; obtain and broadcast $\tilde{\beta}_k$ and the variance $\tilde{V}_k$;

end

(3) Evidence synthesis

Obtain $\tilde{\beta}$ using (5) by plugging in $\tilde{\beta}_k$ and $\tilde{V}_k$.

Return $\tilde{\beta}$.

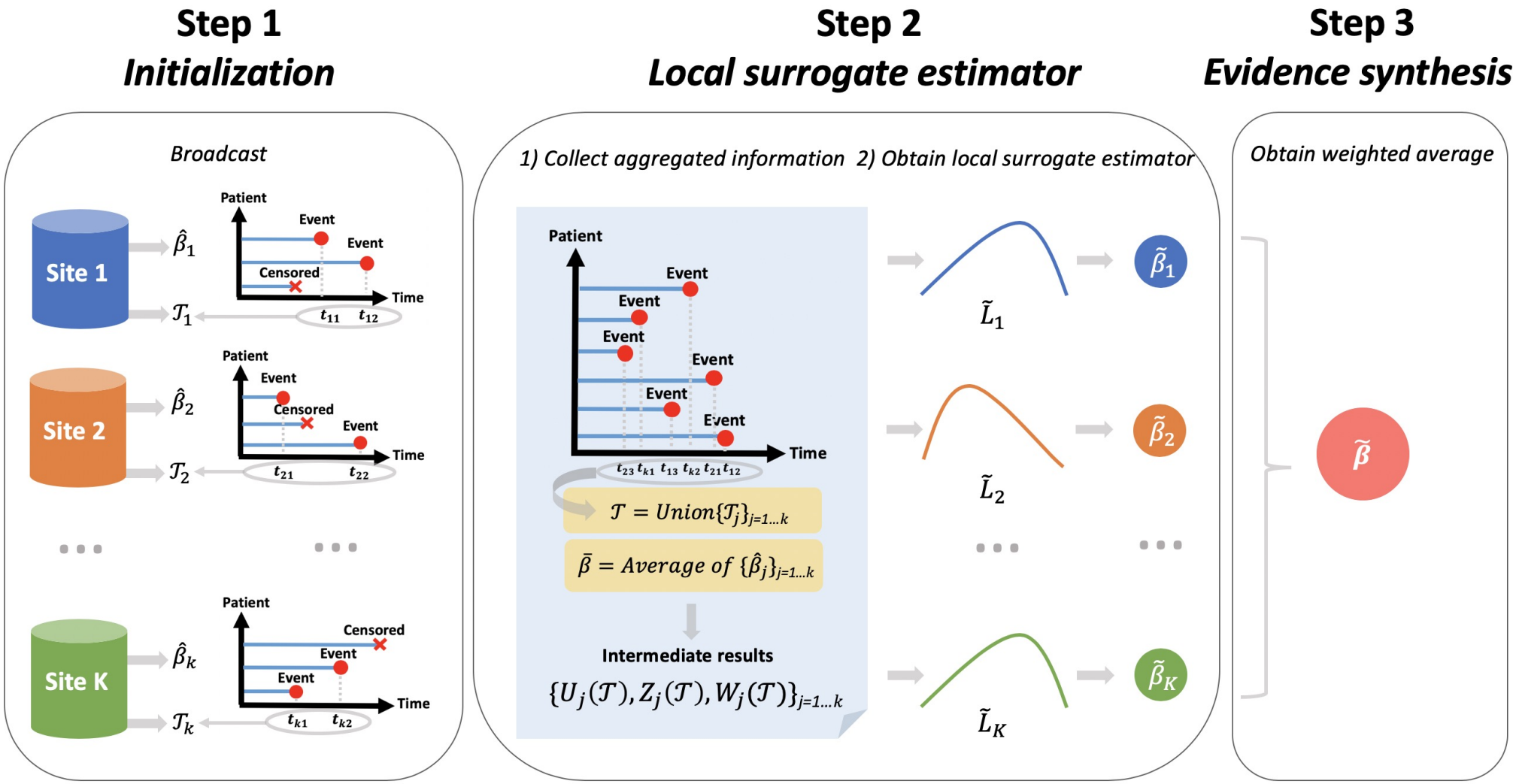
New Challenges

- For the i-th observation in the j-th site
  - t_{ij} --- time to event, δ_{ij} ---censoring indicator , x_{ij} --- observed risk factors
  - Hazard function: $\lambda(t) = \lim_{dt \to 0} \frac{P(t \le T < T+dt | T \ge t)}{dt}$
- Cox PH Model assumes
  - $\lambda(t|X) = \lambda_0(t)\exp(X\beta)$
- Combined log partial likelihood function

$$L(\beta) = \frac{1}{nK} \sum_{j=1}^{K} \sum_{i=1}^{n} \delta_{ij} \log \frac{\exp(x_{ij}^T \beta)}{\sum_{(l,m) \in \boldsymbol{R(t_{ij})}} \exp(x_{lm}^T \beta)}$$

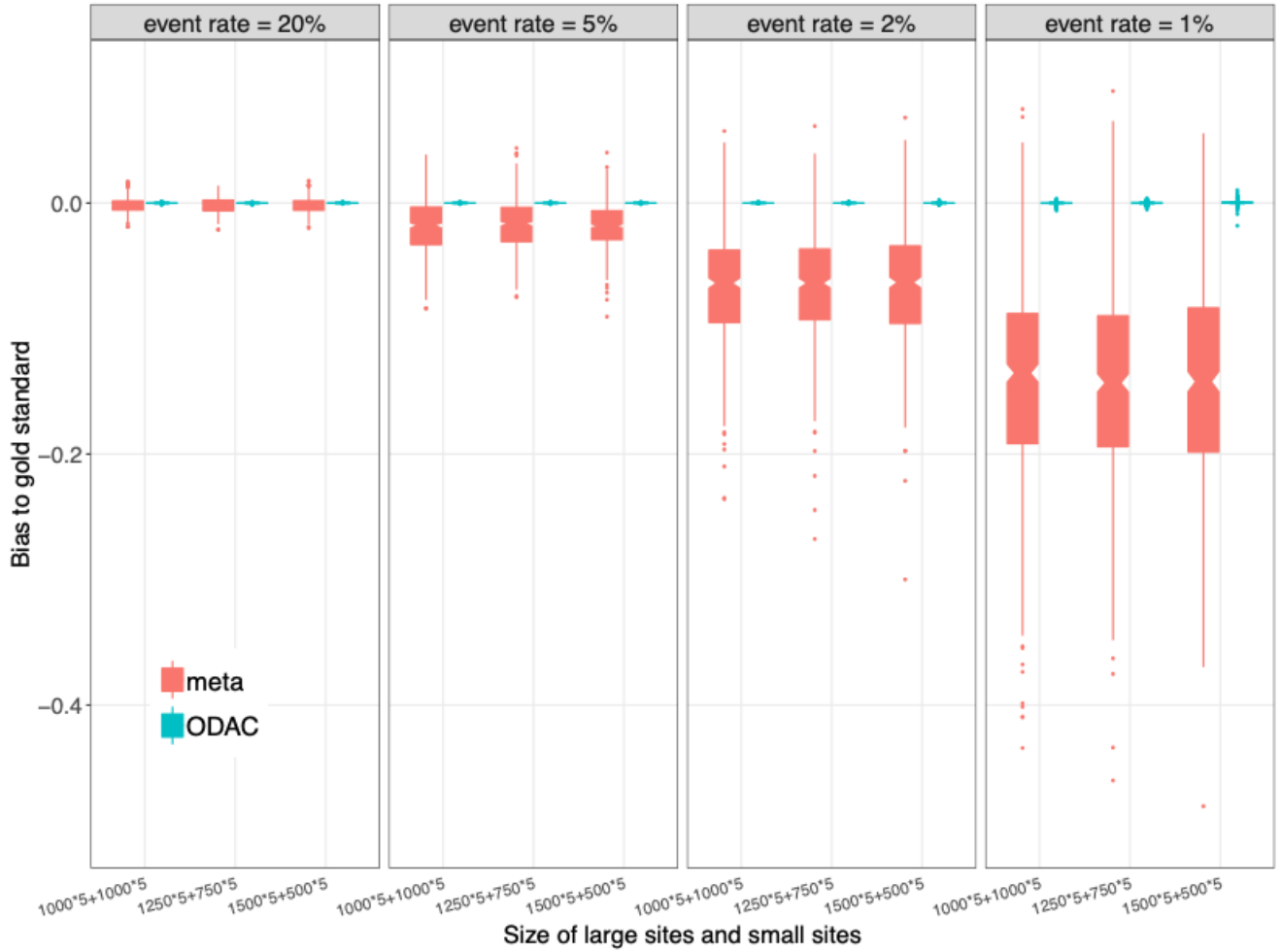
  - Combined likelihood function cannot be written as sum of individual terms
  - Distributions of covariates vary from site to site

ODAC : One-shot Distributed Algorithm for Cox Proportional Hazards Model



Duan R, et al. Learning from local to global: An efficient distributed algorithm for modeling time-to-event data. Journal of the American Medical Informatics Association. 2020 Jul;27(7):1028-1036.

Benefit in studying rare events



- Meta-analysis has increasing bias when event is rarer.
- ODAC provides estimates close to the pooled analysis.

Application to OHDSI

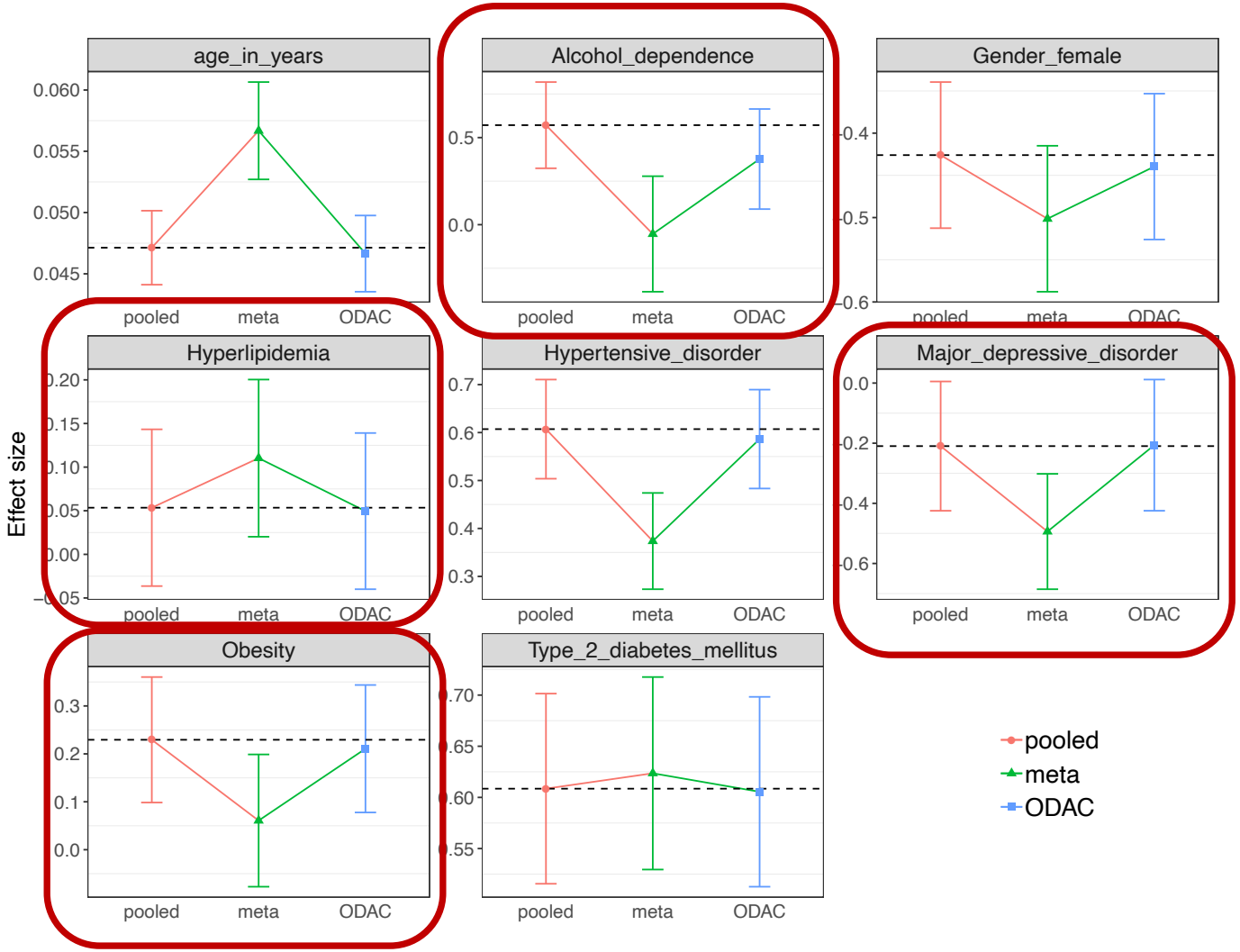
- Four claims datasets.
- Population: pharmacologically-treated major depressive disorder
- Outcome: acute myocardial infarction (AMI)
- Cox regression model with 8 risk factors.

Table 1. Characteristics of the 4 claims datasets at the Observational Health Data Sciences and Informatics

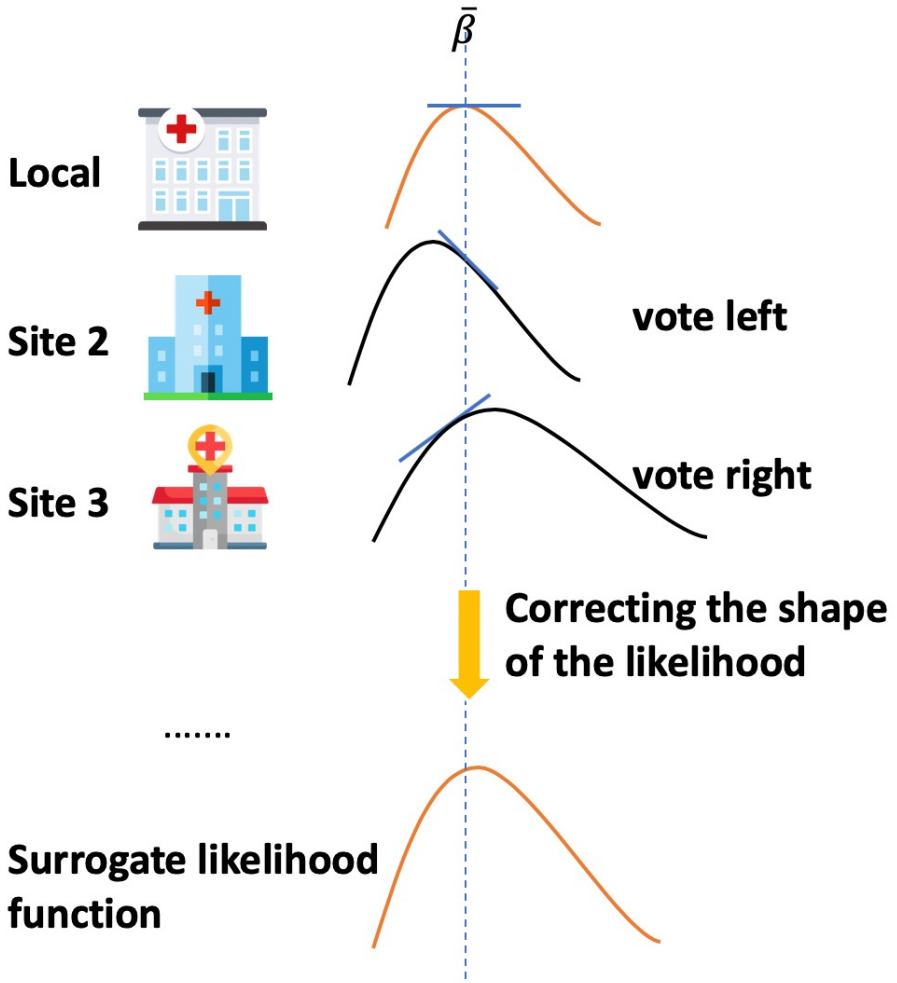
Dataset	CCAE	MDCD	MDCR	Optum
Subjects	64 222	59 861	69 164	62 348
Median age, y	43	35	71	47
Female, %	69.21	73.82	68.08	69.68
Congestive heart failure, %	0.70	3.06	7.58	2.61
Hypertensive disorder, %	20.81	31.80	57.70	32.96
Ischemic heart disease, %	1.70	3.82	10.27	4.10
Type 2 diabetes mellitus, %	7.49	14.63	21.83	12.71
Coronary arteriosclerosis, %	2.39	4.92	18.43	5.75
Renal failure syndrome, %	0.69	2.67	2.31	2.49
Transient cerebral ischemia, %	0.41	0.64	2.32	0.71
Hyperlipidemia, %	20.96	22.00	43.21	33.85
Obesity, %	7.15	16.54	6.71	9.62
Alcohol dependence, %	1.79	2.94	1.01	2.29
Major depressive disorder, %	4.17	3.55	3.16	3.34
Acute myocardial infarction, %	0.26	0.75	2.03	0.51
Stroke, %	0.24	0.73	1.75	0.58

Results (AMI)

- **Risk factors for acute myocardial infarction (AMI):**
  gender
  age
  obesity
  alcohol dependence
  hypertensive
  major depressive disorder
  type 2 diabetes mellitus
  hyperlipidemia



Intuitions: a nice interpretation of surrogate likelihood



- Meta-analysis requires point estimate and standard error.
- In ODAL/ODAC, slopes help to correct the shape of local likelihood function.

Recap: definition of surrogate likelihood

The surrogate likelihood (SL) approach

- For an initial value $\bar{\beta}$,

$$L(\beta) = L(\bar{\beta}) + \nabla L(\bar{\beta})^T(\beta - \bar{\beta}) + \sum_{t=2}^{\infty} \frac{1}{t!} \nabla^t L(\bar{\beta})(\beta - \bar{\beta})^{\otimes t}$$

$$L_1(\beta) = L_1(\bar{\beta}) + \nabla L_1(\bar{\beta})^T(\beta - \bar{\beta}) + \sum_{t=2}^{\infty} \frac{1}{t!} \nabla^t L_1(\bar{\beta})(\beta - \bar{\beta})^{\otimes t}$$

$$\sum_{t=2}^{\infty} \frac{1}{t!} \nabla^t L_1(\bar{\beta})(\beta - \bar{\beta})^{\otimes t} = L_1(\beta) - L_1(\bar{\beta}) - \nabla L_1(\bar{\beta})^T(\beta - \bar{\beta})$$

First-order SL function

$$\tilde{L}^1(\beta) = L_1(\beta) + \{\nabla L(\bar{\beta}) - \nabla L_1(\bar{\beta})\}^T \beta$$

$$\nabla L(\bar{\beta}) = \frac{1}{K} \sum \nabla L_j(\bar{\beta}); \quad \nabla L_j(\bar{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \{y_{ij} - \text{expit}(x_{ij}^T \bar{\beta})\} x_{ij}$$

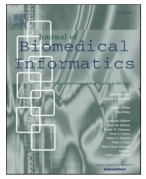
ODAP – One-shot Distributed Algorithm for Quasi-Poisson regression

Journal of Biomedical Informatics 131 (2022) 104097

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin



Distributed Quasi-Poisson regression algorithm for modeling multi-site count outcomes in distributed data networks

Mackenzie J. Edmondson [a], Chongliang Luo [a], Md. Nazmul Islam [b], Natalie E. Sheils [b], John Buresh [b], Zhaoyi Chen [c,d], Jiang Bian [c,d], Yong Chen [a,*]

[a] Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA
[b] Optum Labs at UnitedHealth Group, Minnetonka, MN, USA
[c] Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Gainesville, FL, USA
[d] Cancer Informatics Shared Resource, University of Florida Health Cancer Center, Gainesville, FL, USA

Algorithm:

Input: Patient-level data $X = \{X_{i1}\}, y = \{y_{i1}\}$ from lead institution, as well as parameter estimates $\widehat{\beta}_j$ and $\widehat{\sigma}_j^2$, first order-gradients $\nabla QL_j(\widehat{\beta})$, and second-order gradients $\nabla^2 QL_j(\widehat{\beta})$ from non-lead institutions, where i, j denote the patient and institution indices, respectively.

Output: ODAP estimates $\widetilde{\beta}$ and variances $V(\widetilde{\beta})$.

Initialization:

At institution $j = 1, \ldots, K$, **do**

Fit quasi-Poisson regression model and obtain $\widehat{\beta}_j$ and $\widehat{\sigma}_j^2$. Send to lead institution.

end

At lead institution, compute initial estimates $\bar{\beta}$ via meta-analysis. Send to institutions $j = 2, \cdots, K$.

At institutions $j = 2, \cdots, K$, **do**

Calculate $\nabla QL_j(\bar{\beta})$ and $\nabla^2 QL_j(\bar{\beta})$ using (7) and (8), respectively. Send to lead institution.

end

Surrogate Likelihood Estimation:

At lead institution, compute $\widetilde{QL}(\beta)$ using (5).

At lead institution, obtain $\widetilde{\beta} = \text{argmax}_\beta \widetilde{QL}(\beta)$. Send $\widetilde{\beta}$ to institutions $j = 2, \cdots, K$.

Variance Calculation:

At institutions $j = 2, \cdots, K$, **do**

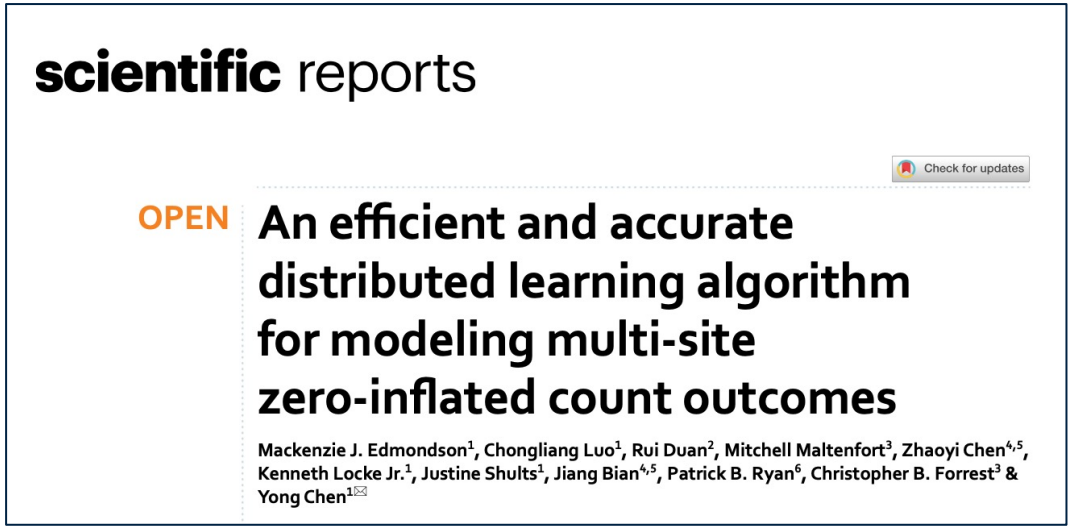
Calculate $\nabla QL_j\left(\widetilde{\beta}\right)$ and $\nabla^2 QL_j\left(\widetilde{\beta}\right)$ using (7) and (8), respectively. Send to lead institution.

end

Calculate $V(\widetilde{\beta})$ using (9).

Return $\widetilde{\beta}$ and $V(\widetilde{\beta})$.

ODAH – One-shot Distributed Algorithm for Hurdle Regression



scientific reports

Check for updates

OPEN

An efficient and accurate distributed learning algorithm for modeling multi-site zero-inflated count outcomes

Mackenzie J. Edmondson[1], Chongliang Luo[1], Rui Duan[2], Mitchell Maltenfort[3], Zhaoyi Chen[4,5], Kenneth Locke Jr.[1], Justine Shults[1], Jiang Bian[4,5], Patrick B. Ryan[6], Christopher B. Forrest[3] & Yong Chen[1]✉

Algorithm:

Input: Patient-level data $X = \{X_{i1}\}, Y = \{Y_{i1}\}$ from the lead site, as well as parameter estimates $\widehat{\boldsymbol{\beta}_j}$, $\widehat{\boldsymbol{\gamma}_j}$, $\widehat{\sigma_j^2}$, and $\widehat{\tau_j^2}$, first order-gradients ($\frac{1}{n_j}\nabla L_{1j}(\bar{\boldsymbol{\beta}})$ and $\frac{1}{n_j}\nabla L_{2j}(\bar{\boldsymbol{\gamma}})$) and second-order gradients ($\frac{1}{n_j}\nabla^2 L_{1j}(\bar{\boldsymbol{\beta}})$ and $\frac{1}{n_j}\nabla^2 L_{2j}(\bar{\boldsymbol{\gamma}})$) from coordinating sites, where i, j denote the observation and clinical site indices, respectively.
Output: Surrogate maximum likelihood estimators $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\gamma}}$.

Initialization:
1: At site $j = 1, \ldots, K$, **do**
Fit hurdle model and obtain point estimates $\widehat{\boldsymbol{\beta}_j}$ and $\widehat{\boldsymbol{\gamma}_j}$, as well as variance estimates $\widehat{\sigma_j^2}$ and $\widehat{\tau_j^2}$ of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\gamma}}$, respectively. Send $\widehat{\boldsymbol{\beta}_j}$, $\widehat{\boldsymbol{\gamma}_j}$, $\widehat{\sigma_j^2}$, and $\widehat{\tau_j^2}$ to the lead site.
end

2: At lead site, compute initial estimates $\bar{\boldsymbol{\beta}}$ and $\bar{\boldsymbol{\gamma}}$ using meta-analysis. Send to sites $j = 2, \ldots, K$.
3: At site sites $j = 2, \ldots, K$, **do**
Calculate first order-gradients ($\frac{1}{n_j}\nabla L_{1j}(\bar{\boldsymbol{\beta}})$ (S.1) and $\frac{1}{n_j}\nabla L_{2j}(\bar{\boldsymbol{\gamma}})$ (S.3)) and second-order gradients ($\frac{1}{n_j}\nabla^2 L_{1j}(\bar{\boldsymbol{\beta}})$ (S.2) and $\frac{1}{n_j}\nabla^2 L_{2j}(\bar{\boldsymbol{\gamma}})$ (S.4)). Send to lead site.
end

Surrogate Likelihood Construction/Maximization:
1: At lead site, compute surrogate log likelihoods $\widetilde{L_1}(\boldsymbol{\beta})$ (12) and $\widetilde{L_2}(\boldsymbol{\gamma})$ (13).
2: At lead site, obtain $\tilde{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \tilde{L}_1(\boldsymbol{\beta})$ and $\tilde{\boldsymbol{\gamma}} = \arg\max_{\boldsymbol{\gamma}} \tilde{L}_2(\boldsymbol{\gamma})$.
3: **Return** $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\gamma}}$.

PDA MENU
ODAL
ODAC
ODAP
ODAH
Hetero-aware
DLMM
dPQL
ODACH
dCLR
dist-EM

Heterogeneity in Clinical Settings

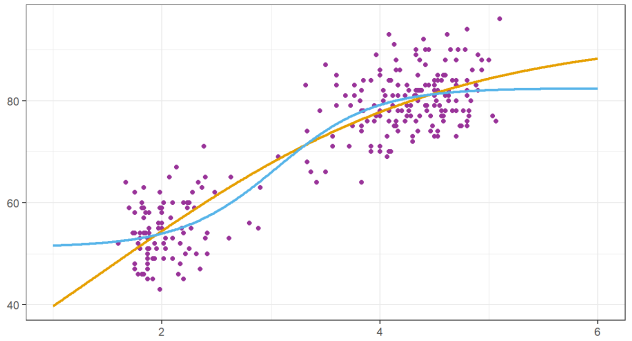
Hospital A

Hospital B

Hospital C

- Heterogeneity in **relationship between outcome and covariates, i.e., *f(y|x)***
- Heterogeneity in **distributions in covariates, *f(x). "distribution-shift"***
- Heterogeneity in **data structure across sites**
- Other…

Our current algorithms for heterogeneous multi-site data:

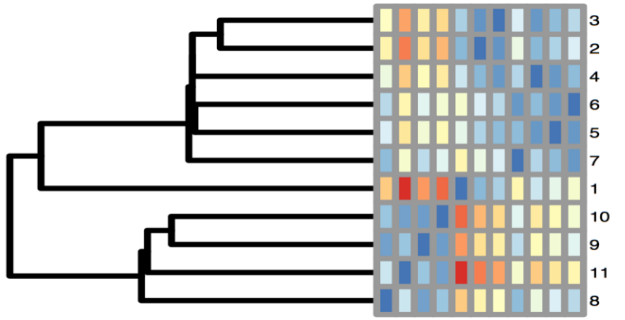


- Supervised learning/regression analysis under heterogeneity
  - GLM with incidental parameters (for site-specific effects)
    - Conditional inference based on pseudolikelihood

      Luo, Chen et al. (2022) Annals of Applied Statistics

      Tong, Chen et al (2022) NPJ Digital Medicine
    - Density-ratio tilted efficient score function

      Duan, Ning and Chen, 2021 Biometrika
  - GLM with random effects
    - ***DLMM (Luo, Chen, et al. 2022 Nature Communications)***
    - ***dPQL (Luo, Chen, et al. 2022 JAMIA)***



- Unsupervised learning/clustering algorithms under heterogeneity
  - Mixture models
    - ***Ning and Chen, 2015 Scandinavian Journal of Statistics***
    - ***Hong, Chen et al, 2017 JASA***
  - High-dimensional Latent class regressions
    - ***Chen et al., 2020 SMMR***
  - Key: **distributed EM algorithm**, latent transfer learning

A novel framework for heterogeneity-aware distributed inference

Biometrika (2022), **109**, 1, *pp.* 67–83
Printed in Great Britain
doi: 10.1093/biomet/asab007
Advance Access publication 27 May 2021

Heterogeneity-aware and communication-efficient distributed statistical inference

BY RUI DUAN

Department of Biostatistics, Harvard University,
677 Huntington Avenue, Boston, Massachusetts 02115, U.S.A.
rduan@hsph.harvard.edu

YANG NING

Department of Statistics and Data Science, Cornell University,
Comstock Hall 1188, Ithaca, New York, 14853, U.S.A.
yn265@cornell.edu

AND YONG CHEN

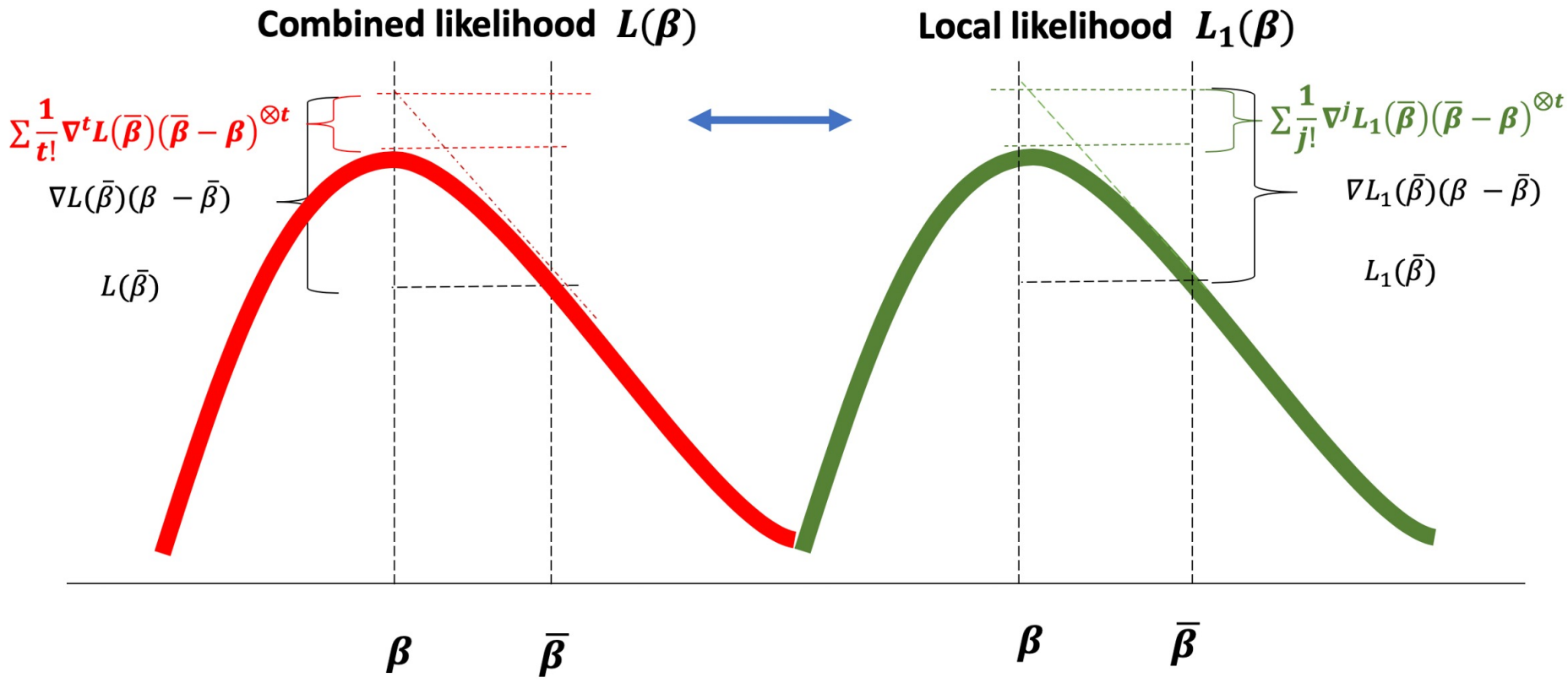
Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania,
423 Guardian Drive, Philadelphia, Pennsylvania 19104, U.S.A.
ychen123@upenn.edu

- Data in the j-th site follows

$$Y_{ij} \sim f(y; \boldsymbol{\beta}, \boldsymbol{\gamma_j})$$

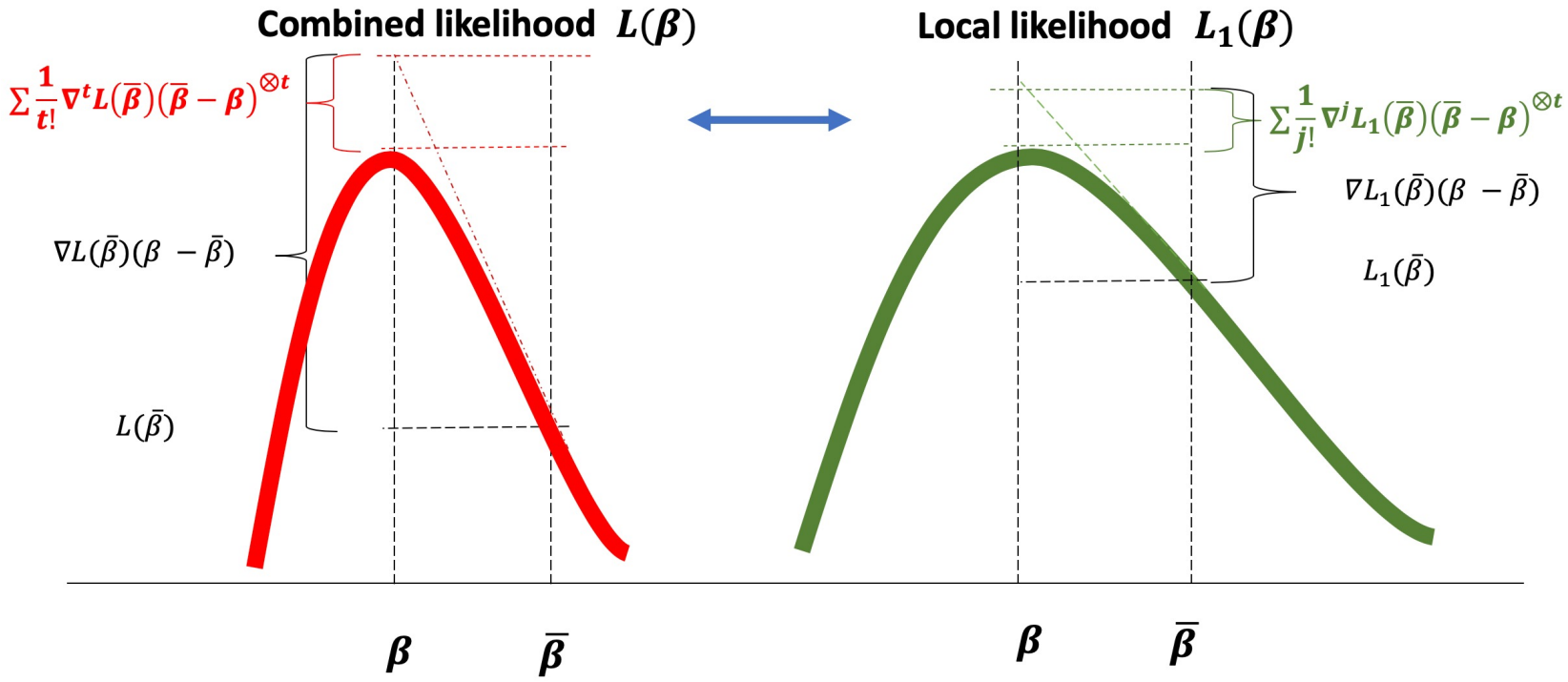
- $\boldsymbol{\beta}$ is the ***parameter of interest***
- $\boldsymbol{\gamma_j}$ is the ***site-specific nuisance/incidental parameter***--- *allow site to be a covariate variable, allow interaction terms between site and other covariates.*

Surrogate likelihood approach assumes homogeneous data



Same distribution ⇒ $\nabla^t L(\bar{\beta}) \approx \nabla^t L_1(\bar{\beta})$, for any $t \geq 2$, and any $\bar{\beta}$

Challenges



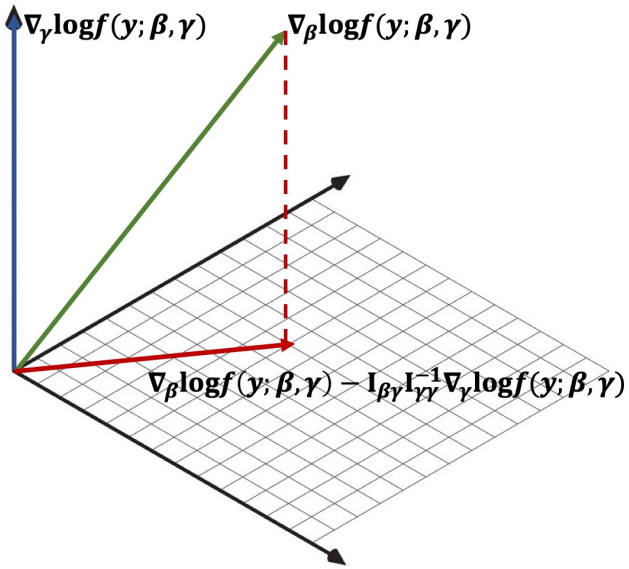
$$\nabla^t L(\bar{\beta}) \neq \nabla^t L_1(\bar{\beta})$$

Consideration 1: account for the incidental parameters

Change the target function to the efficient score function

$$L(\beta;\bar{\Gamma}) = \frac{1}{Kn}\sum_{j=1}^{K}\sum_{i=1}^{n}\log f(y_{ij};\beta,\bar{\gamma}_j)$$

$$S(\beta;\bar{\Gamma}) = \frac{1}{Kn}\sum_{j=1}^{K}\sum_{i=1}^{n}\left\{\nabla_\beta \log f(y_{ij};\beta,\bar{\gamma}_j) - I^{(j)}_{\beta\gamma}I^{(j)\,-1}_{\gamma\gamma}\nabla_\gamma \log f(y_{ij};\beta,\bar{\gamma}_j)\right\}$$



where $\bar{\gamma}_j$ is some initial estimator for γ_j obtained from Site j.

Consideration 2: account for the "different shapes" of estimating functions

Goal: Find a local estimating function and construct higher-order match

- We want to find an estimating function $g(y;\beta)$, such that, for t = 1, 2, …

$$E_{f_1}\left\{\nabla_\beta^t g(Y_{i1};\beta)\right\} = \mathrm{E}\left\{\nabla_\beta^t S(\beta;\bar{\Gamma})\right\} := \frac{1}{K}\sum_{j=1}^{K} \mathrm{E}_{f_j}\nabla_\beta^{\mathrm{t}}\left\{\nabla_\beta \log f(Y_{ij};\beta,\bar{\gamma}_j) - I_{\beta\gamma}^{(j)} I_{\gamma\gamma}^{(j)^{-1}} \nabla_\gamma \log f(Y_{ij};\beta,\bar{\gamma}_j)\right\}$$

$$\mathrm{E}_{f_1}\left\{\frac{f_j}{f_1}s(y)\right\} = \mathrm{E}_{f_j}\left\{s(y)\right\}$$

Local Function	Density ratio tilted	Function at Site j
$\mathrm{E}_{f_1}\{s(y)\}$	$\mathrm{E}_{f_1}\{\frac{f_j}{f_1}s(y)\}$	$\mathrm{E}_{f_j}\{s(y)\}$

Proof: $\mathrm{E}_{f_1}\left\{\frac{f_j}{f_1}s(y;\beta)\right\} = \int \frac{f_j}{f_1}s(y)\, f_1 dy = \int \frac{f_j}{f_1}s(y)\, f_1 dy = \int s(y)\, f_j dy = \mathrm{E}_{f_j}\{s(y)\}$

Consideration 2: account for the "different shapes" of estimating functions

Goal: Find a local estimating function and construct higher-order match

- We want to find an estimating function $g(y;\beta)$, such that, for t=1, 2, …

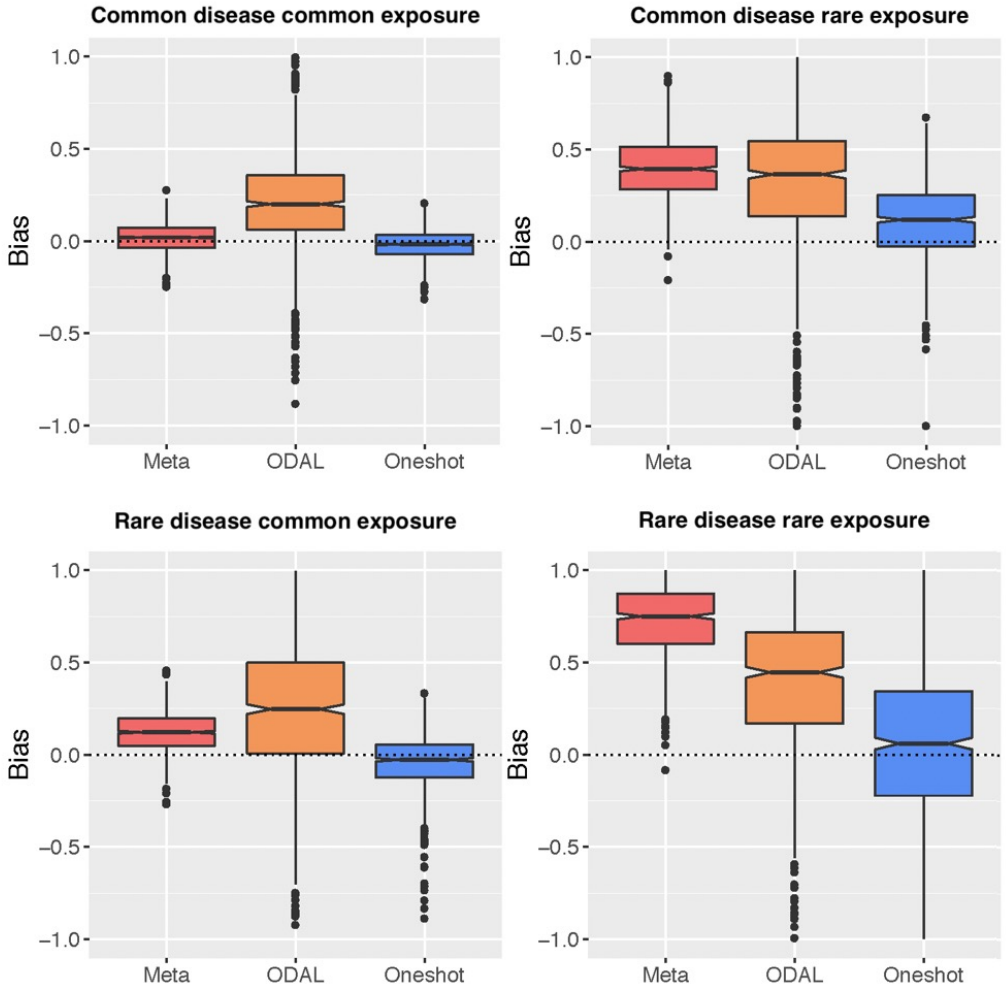
$$\boldsymbol{E_{f_1}}\left\{\boldsymbol{\nabla_\beta^t g(Y_{i1};\beta)}\right\} = \mathbf{E}\left\{\boldsymbol{\nabla_\beta^t S(\beta;\bar{\Gamma})}\right\} := \frac{1}{K}\sum_{j=1}^{K} \mathrm{E}_{f_j} \nabla_\beta^{\mathrm{t}} \left\{\nabla_\beta \log f(Y_{ij};\beta,\bar{\gamma}_j) - I_{\beta\gamma}^{(j)} I_{\gamma\gamma}^{(j)^{-1}} \nabla_\gamma \log f(Y_{ij};\beta,\bar{\gamma}_j)\right\}$$

- For initial estimator $\bar{\beta},\bar{\gamma}_j$, define

$$g(y;\beta) = \frac{1}{K}\sum_{j=1}^{K} \frac{\boldsymbol{f(y;\bar{\beta},\bar{\gamma}_j)}}{\boldsymbol{f(y;\bar{\beta},\bar{\gamma}_1)}} \left\{\nabla_\beta \log f(y;\beta,\bar{\gamma}_j) - \tilde{H}_{\beta\gamma}^{(j)} \tilde{H}_{\gamma\gamma}^{(j)^{-1}} \nabla_\gamma \log f(y;\beta,\bar{\gamma}_j)\right\}$$

Conclusion

Desired Properties	Meta-analysis	Proposed estimator (T=1)	Proposed estimator (T=2)
Consistency	consistent	consistent	consistent
Distance to the Gold Standard Estimator	$\frac{C}{\sqrt{Kn}}$	$\lesssim \frac{C}{n}$	$\lesssim \frac{C}{n\sqrt{K}} + \frac{C}{n\sqrt{n}}$
Asymptotic Normality	asymptotic normal	asymptotic normal	asymptotic normal
Asymptotic Efficient	not efficient	efficient	efficient



Duan, R., Ning, Y. and Chen, Y., 2022. Heterogeneity-aware and communication-efficient distributed statistical inference. Biometrika, 109(1), pp.67-83.

Heterogeneity in Clinical Settings

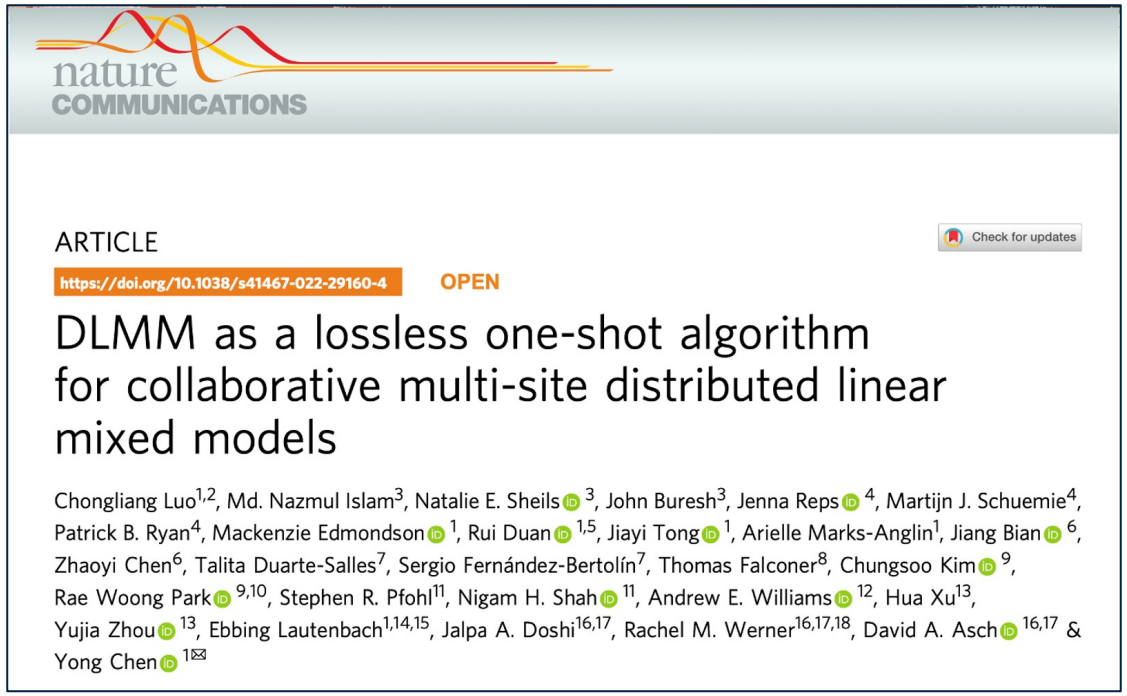
Hospital A

Hospital B

Hospital C

- Heterogeneity in **relationship between outcome and covariates, i.e., *f(y|x)***
- Heterogeneity in **distributions in covariates, *f(x). "distribution-shift"***
- Heterogeneity in **data structure across sites**
- Other…

DLMM algorithm: distributed algorithm for linear mixed models



nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-022-29160-4 OPEN

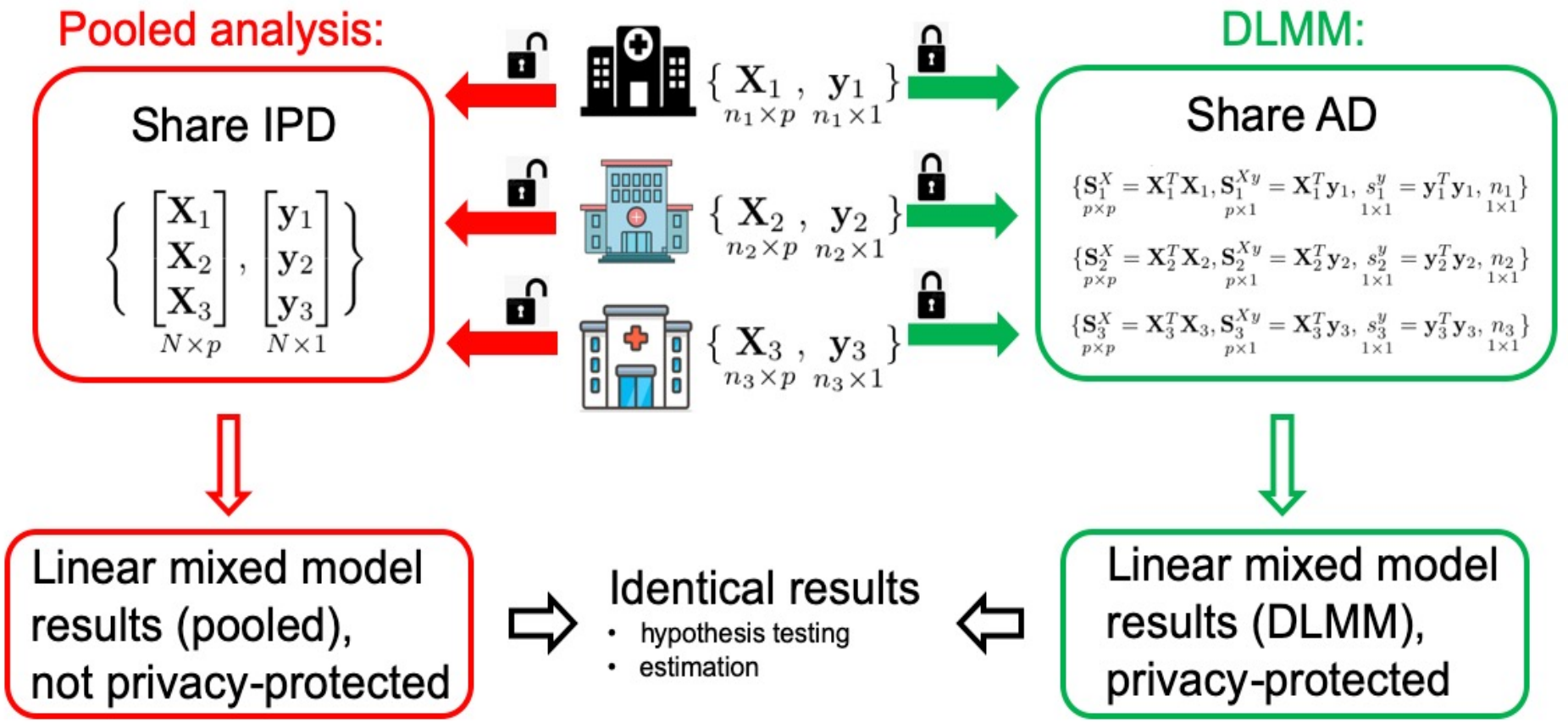
DLMM as a lossless one-shot algorithm for collaborative multi-site distributed linear mixed models

Chongliang Luo[1,2], Md. Nazmul Islam[3], Natalie E. Sheils[3], John Buresh[3], Jenna Reps[4], Martijn J. Schuemie[4], Patrick B. Ryan[4], Mackenzie Edmondson[1], Rui Duan[1,5], Jiayi Tong[1], Arielle Marks-Anglin[1], Jiang Bian[6], Zhaoyi Chen[6], Talita Duarte-Salles[7], Sergio Fernández-Bertolín[7], Thomas Falconer[8], Chungsoo Kim[9], Rae Woong Park[9,10], Stephen R. Pfohl[11], Nigam H. Shah[11], Andrew E. Williams[12], Hua Xu[13], Yujia Zhou[13], Ebbing Lautenbach[1,14,15], Jalpa A. Doshi[16,17], Rachel M. Werner[16,17,18], David A. Asch[16,17] & Yong Chen[1✉]

Box 1 ▌Pseudo-code of the distributed linear mixed model algorithm

1. In site $i = 1, \ldots, K$, calculate and share $S_i^X = X_i^T X_i$, $S_i^{Xy} = X_i^T y_i$, $s_i^y = y_i^T y_i$ and sample size n_i.
2. Perform the likelihood ratio test for the significance of random effects of each covariate by Eq. (10).
3. With the significant random effects identified by the above step, reconstruct the profile log-likelihood Equation (5) or the restricted profile log-likelihood Eq. (6), obtain the estimate $\hat{\Theta}$.
4. Obtain $\hat{\beta} = \tilde{\beta}(\hat{\Theta})$ and $\hat{\sigma}^2 = \tilde{\sigma}^2(\hat{\Theta})$ by Eqs. (3) and (4).
5. Calculate the variance of the estimated fixed effects $\hat{\beta}$ by Eq. (7).
6. Calculates the BLUPs of the random effects in each site by Eq. (11).

DLMM algorithm: illustration



DLMM algorithm: empirical validation

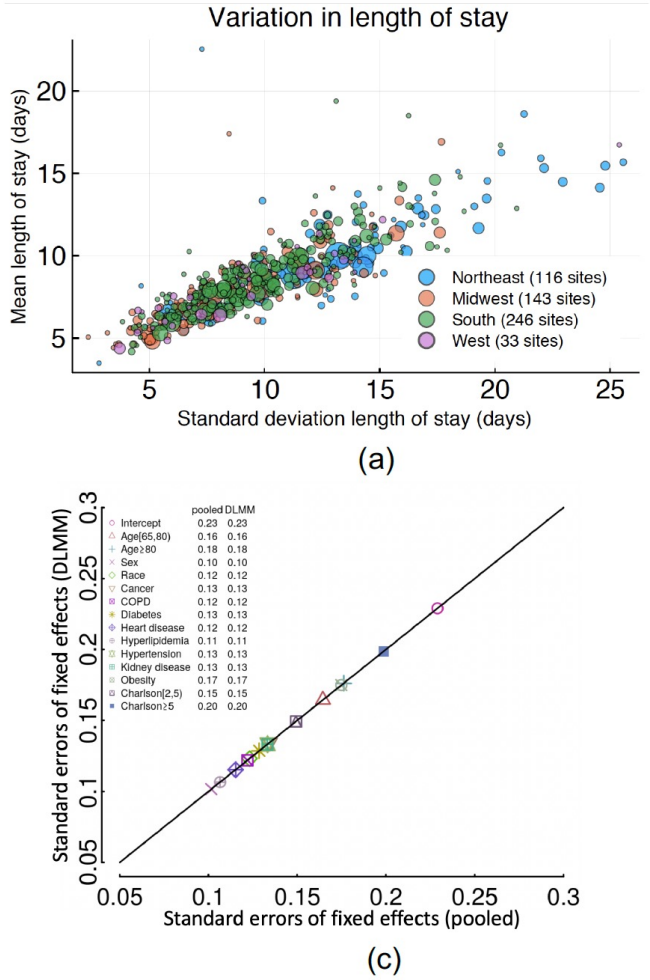
Data source: UnitedHealthGroup

Patient cohort:
patients who were admitted as inpatients to a hospital with a primary diagnosis of COVID-19 between 01/01/2020 – 09/30/2020.

Outcome: Length of stay

Data dimension: K=538 sites; total number of patients N = 47,756

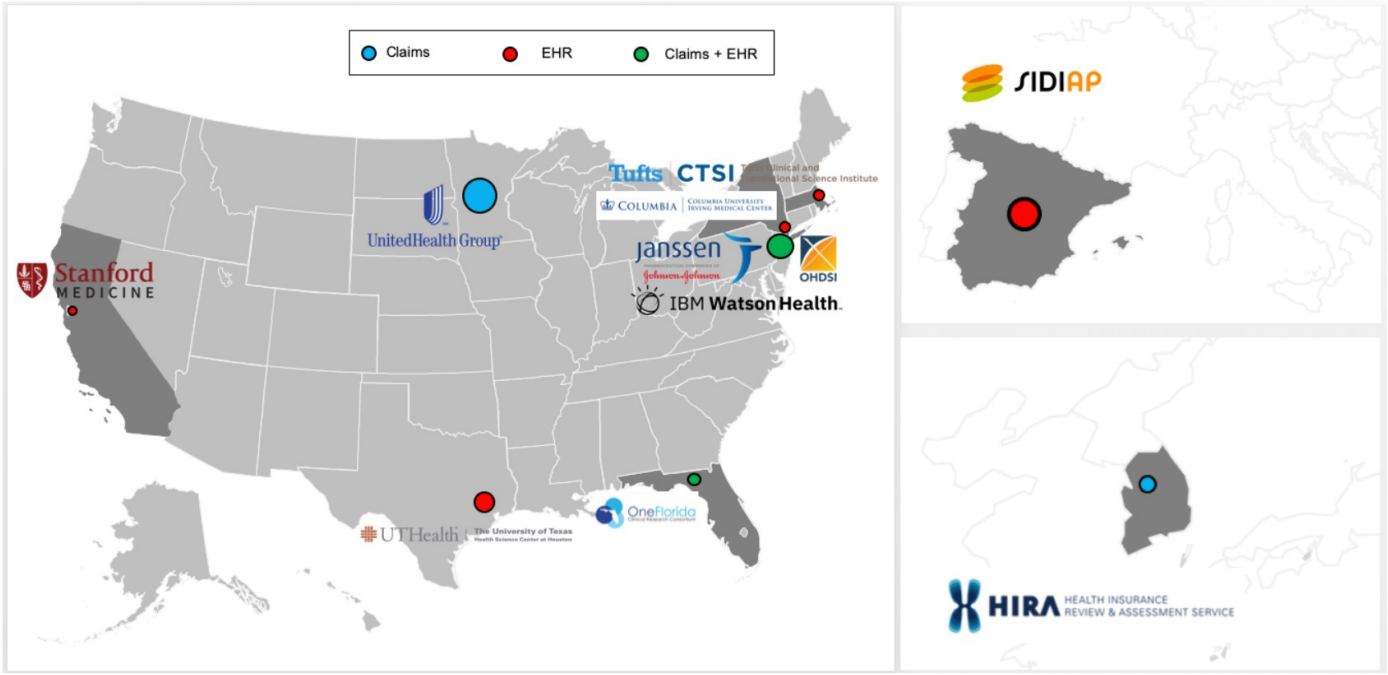
Covariates: 15 variables including age, gender, race, cancer status, heart disease, etc



DLMM algorithm: an OHDSI multi-country study

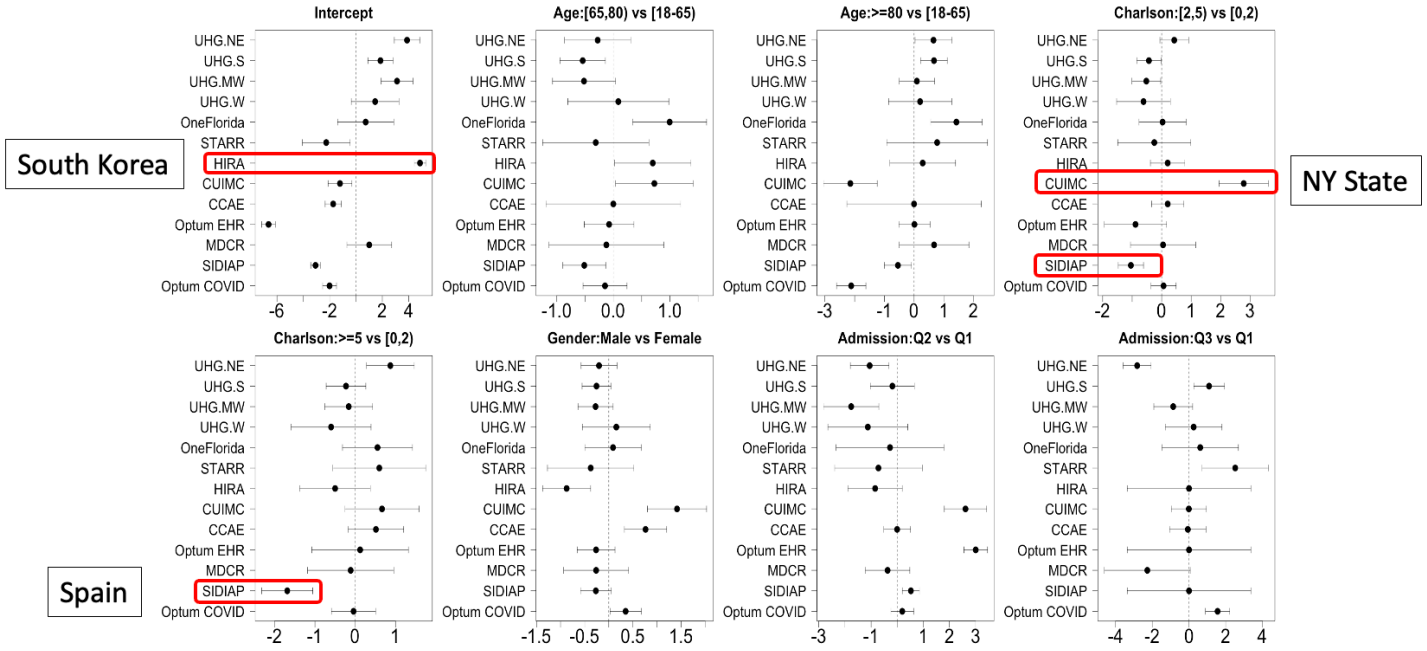
Data source:

- 14 sites
- total of N=120,609 patients across 3 countries



DLMM algorithm: an OHDSI multi-country study - results

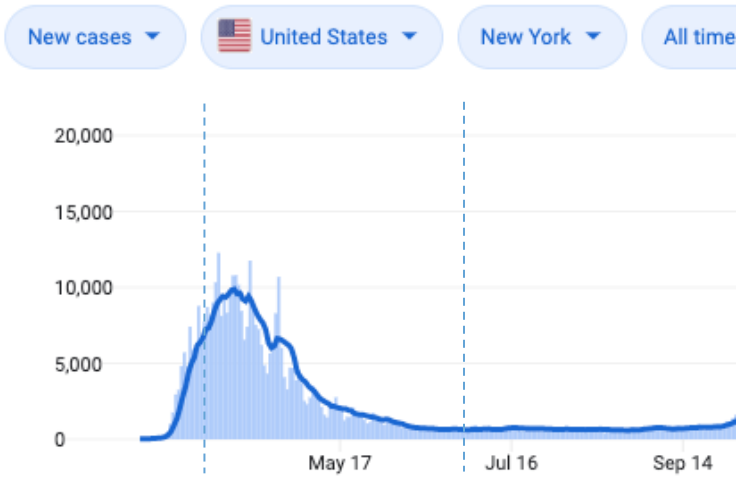
estimated random effects



- The random intercept (i.e., baseline LoS at a site, compared to the overall mean) of South Korea is much greater than 0, possibly due to different discharge criteria in South Korea for COVID patients.
- The effect of Charlson Comorbidity Index (CCI), in the New York state, is much greater than the average, suggesting higher CCI score is strongly associated with longer LoS at New York state.
- The effect of CCI in Spain data it is much smaller than the average, suggesting heterogeneous associations across sites.

COVID-19 LOS: random effects

- CUIMC (NY state) has a larger CCI effect, and SIDIAP (Spain) has a smaller CCI effect.
  - CUIMC patients were admitted from all three quarters (Q1=24.0%, Q2=52.0%, Q3=24.0%), while SIDIAP patients were admitted earlier (Q1=58.6%, Q2=41.4%),
  - patients from SIDIAP may suffer more from the hospital equipment shortage due to the early stage of the pandemic, thus sicker patients may have to be discharged sooner.



Generalized Linear Mixed Model

- What if the outcome is considered as non-Gaussian (binary, count…)?
- Example: COVID-19 mortality (binary) of multiple hospitals
- GLMM with hospital-specific random effects

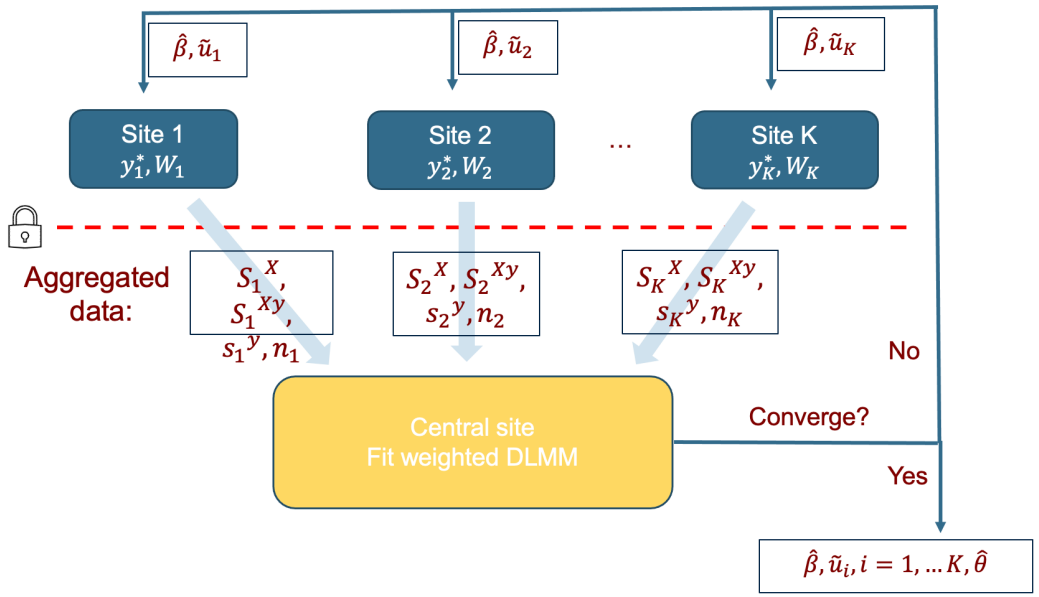
dPQL: Distributed Penalized Quasi-Likelihood Algorithm

Journal of the American Medical Informatics Association, 00(0), 2022, 1–6
https://doi.org/10.1093/jamia/ocac067
Research and Applications

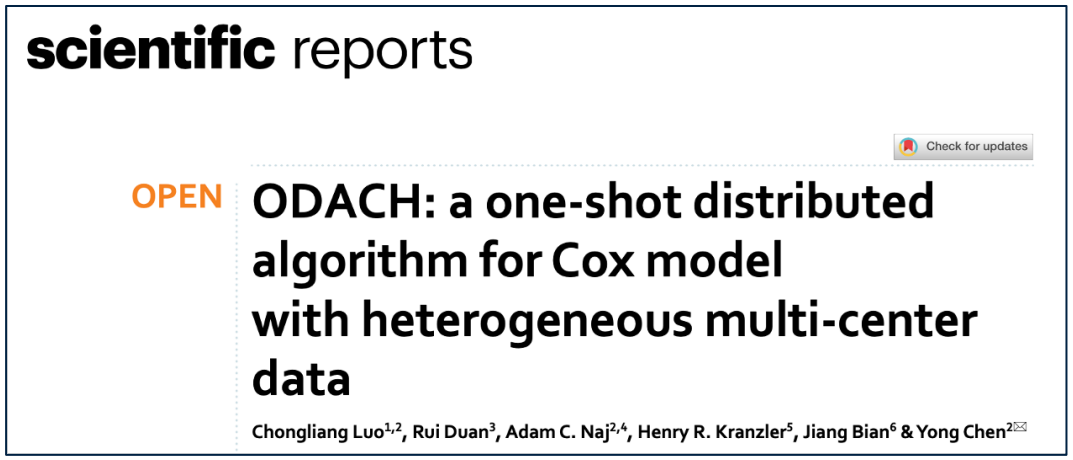
Research and Applications

dPQL: a lossless distributed algorithm for generalized linear mixed model with application to privacy-preserving hospital profiling

Chongliang Luo[1,2], Md. Nazmul Islam[3], Natalie E. Sheils[3], John Buresh[3], Martijn J. Schuemie[4], Jalpa A. Doshi[5,6], Rachel M. Werner[5,6,7], David A. Asch[5,6], and Yong Chen[1,6]



ODACH: a one-shot distributed algorithm for Cox model with heterogeneous multi-center data



scientific reports

Check for updates

OPEN

ODACH: a one-shot distributed algorithm for Cox model with heterogeneous multi-center data

Chongliang Luo[1,2], Rui Duan[3], Adam C. Naj[2,4], Henry R. Kranzler[5], Jiang Bian[6] & Yong Chen[2]✉

Box 1. Pseudo-code of the ODACH algorithm.

Algorithm ODACH

(1) Initialization

In Site k = 1 to K,

do

Fit a Cox regression model and obtain the local estimate $\hat{\beta}_k$ and the variance estimate $\hat{V}_k$

broadcast $\hat{\beta}_k$, $\hat{V}_k$.

end

(2) Aggregated data communication

In Site k = 1 to K,

do

obtain $\bar{\beta}$ using (4)

calculate and broadcast the gradients $\nabla L_j(\bar{\beta})$, and $\nabla^2 L_j(\bar{\beta})$

end

(3) Local surrogate estimator

In the leading site k = 1

do

construct the surrogate likelihood $\tilde{L}(\beta)$ in (3)

obtain $\tilde{\beta}$ by maximizing $\tilde{L}(\beta)$.

Return $\tilde{\beta}$.

dCLR: Distributed Conditional Logistic Regression Model

npj | digital medicine

www.nature.com/npjdigitalmed

ARTICLE OPEN

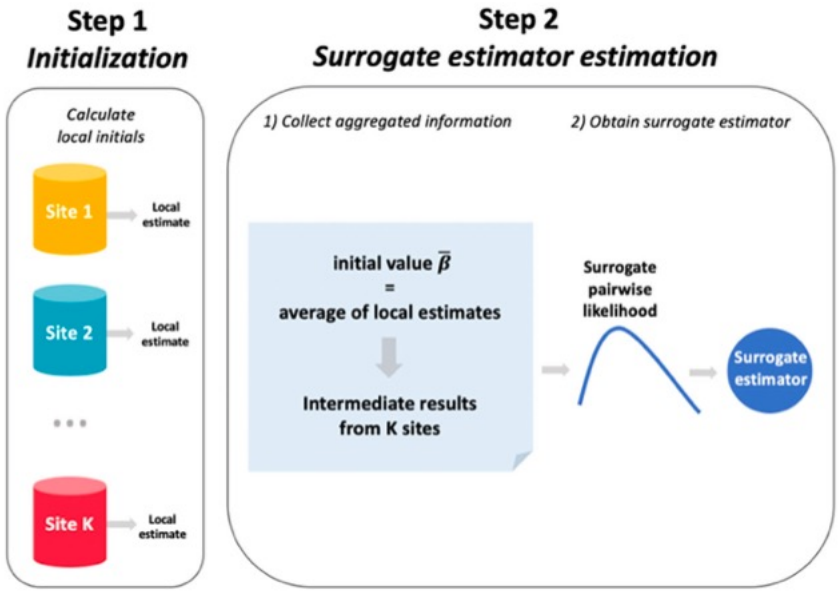
Check for updates

Distributed learning for heterogeneous clinical data with application to integrating COVID-19 data across 230 sites

Jiayi Tong[1], Chongliang Luo[2], Md Nazmul Islam[3], Natalie E. Sheils[3], John Buresh[3], Mackenzie Edmondson[1], Peter A. Merkel[1], Ebbing Lautenbach[1], Rui Duan[4] and Yong Chen[1]✉

Integrating real-world data (RWD) from several clinical sites offers great opportunities to improve estimation with a more general population compared to analyses based on a single clinical site. However, sharing patient-level data across sites is practically challenging due to concerns about maintaining patient privacy. We develop a distributed algorithm to integrate heterogeneous RWD from multiple clinical sites without sharing patient-level data. The proposed distributed conditional logistic regression (dCLR) algorithm can effectively account for between-site heterogeneity and requires only one round of communication. Our simulation study and data application with the data of 14,215 COVID-19 patients from 230 clinical sites in the UnitedHealth Group Clinical Research Database demonstrate that the proposed distributed algorithm provides an estimator that is robust to heterogeneity in event rates when efficiently integrating data from multiple clinical sites. Our algorithm is therefore a practical alternative to both meta-analysis and existing distributed algorithms for modeling heterogeneous multi-site binary outcomes.

npj Digital Medicine (2022)5:76 ; https://doi.org/10.1038/s41746-022-00615-8



dCLR: Distributed Conditional Logistic Regression Model

- Motivation
  - Extended Mantel-Haenszel regression (Liang 1987)
  - Surrogate likelihood approach (Jordan et al. 2018)
- Innovation
  - Elimination of the need of modeling heterogeneous baseline distributions
  - Robust statistical inference without parametric assumptions on the distribution of baseline patient characteristics
    - Allow any canonical link functions in the generalized linear models
    - Feasible to semiparametric extension of the genialized linear models, for example, the semiparametric proportional likelihood ratio proposed by Luo and Tsai (2012)

dCLR: Distributed Conditional Logistic Regression Model

- Stratified logistic regression model:

shared association effect across hospitals

$$logit\{Pr(y_{ij} = 1|x_{ij})\} = \alpha_i + \beta x_{ij} \quad (1)$$

site-specific prevalence of response variable

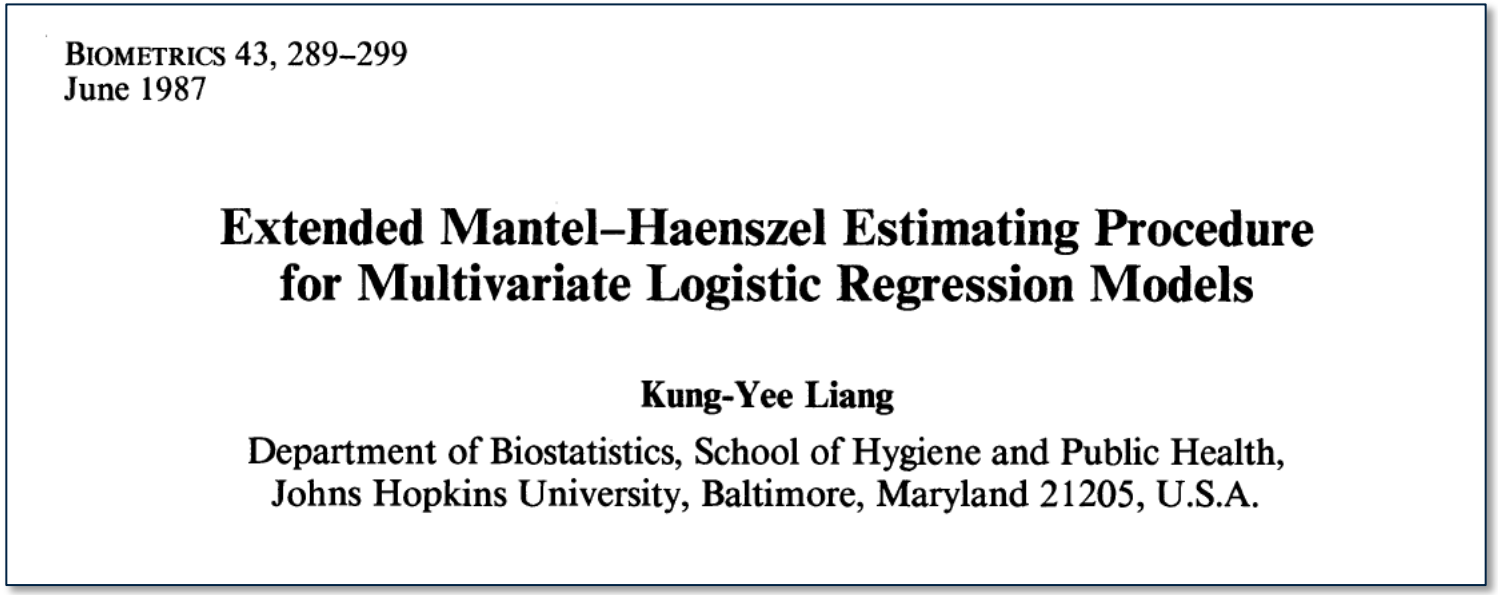
nuisance parameter

- $logit(p) = \log\left(\frac{p}{1-p}\right)$

Pairwise conditioning

- **Liang (1987)'s extended Mantel-Haenszel regression:**
  - Goal: avoid nuisance parameters (i.e., site-specific prevalence)
  - Idea: For the (j,l) pair, the density function conditioning on the order statistics $(y^{(1)}, y^{(2)})$ of (y_{ij}, y_{il}):

$$f\left(y_{ij}, y_{il} \middle| y^{(1)}, y^{(2)}, x_{ij}, x_{il}\right) = \frac{f\left(y_{ij}|x_{ij}\right) f\left(y_{il}|x_{il}\right)}{f\left(y_{ij}|x_{ij}\right) f\left(y_{il}|x_{il}\right) + f\left(y_{il}|x_{ij}\right) f\left(y_{ij}|x_{il}\right)} \quad (2)$$



BIOMETRICS 43, 289–299
June 1987

Extended Mantel–Haenszel Estimating Procedure for Multivariate Logistic Regression Models

Kung-Yee Liang

Department of Biostatistics, School of Hygiene and Public Health,
Johns Hopkins University, Baltimore, Maryland 21205, U.S.A.

Pairwise conditioning

- **Liang (1987)'s extended Mantel-Haenszel regression:**
  - Goal: avoid nuisance parameters (i.e., site-specific prevalence)
  - Idea: For the (j,l) pair, the density function conditioning on the order statistics $(y^{(1)}, y^{(2)})$ of (y_{ij}, y_{il}):

$$f\left(y_{ij}, y_{il} \middle| y^{(1)}, y^{(2)}, x_{ij}, x_{il}\right) = \frac{f\left(y_{ij}\middle|x_{ij}\right) f\left(y_{il}|x_{il}\right)}{f\left(y_{ij}\middle|x_{ij}\right) f\left(y_{il}|x_{il}\right) + f\left(y_{il}\middle|x_{ij}\right) f\left(y_{ij}|x_{il}\right)} \quad (2)$$

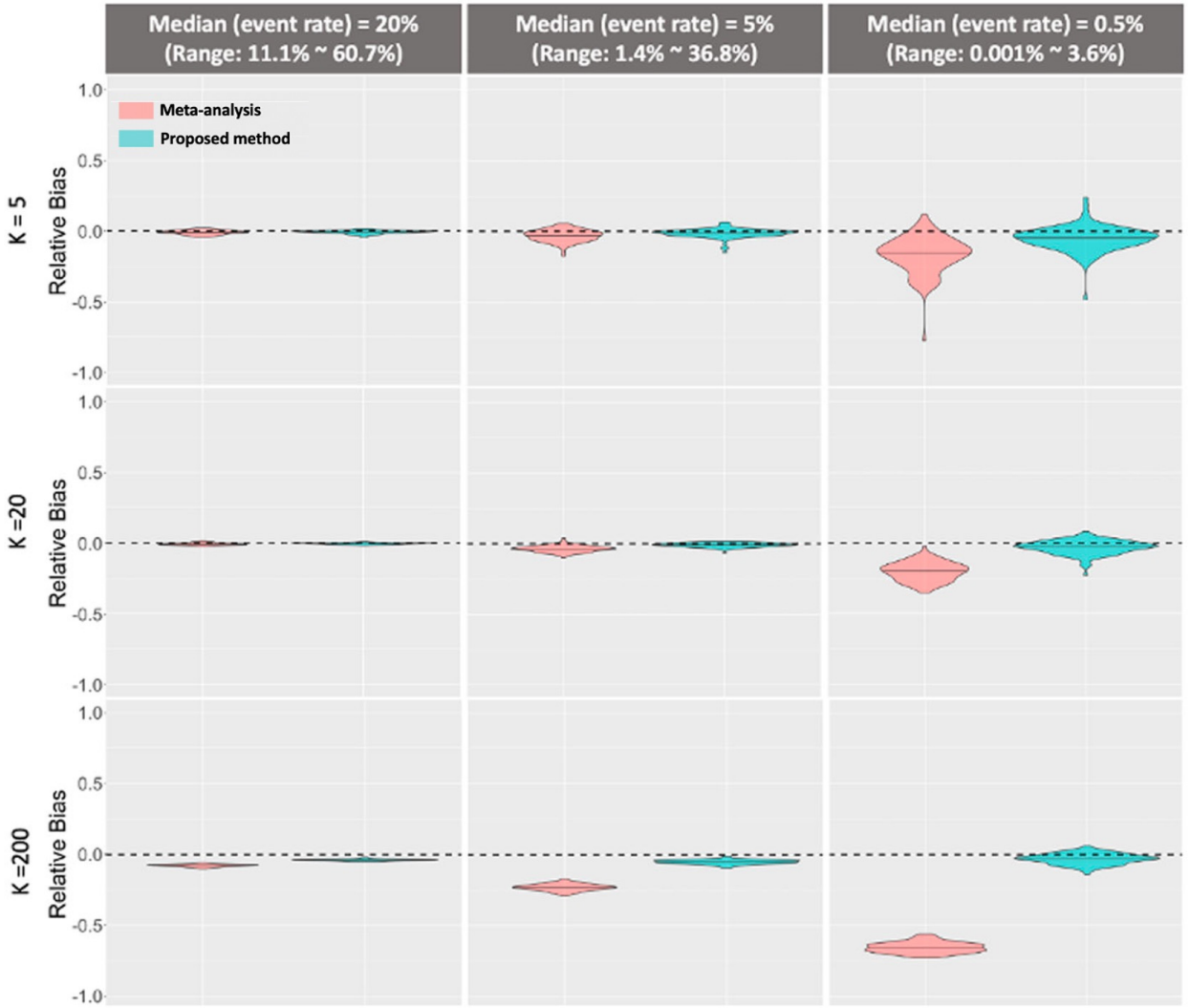
- **Pairwise likelihood for the i-th site:**

$$L_i(\beta) = \prod_{1 \le j < l \le n} \left[1 + exp\{-\left(y_{ij} - y_{il}\right)\left(x_{ij} - x_{il}\right)^T \beta\}\right]^{-1} \quad (3)$$

- **Overall pairwise likelihood function for all sites (if all patient-level data are available):**

$$L^*(\beta) = \prod_{i=1}^{K} L_i(\beta) = \prod_{i=1}^{K} \prod_{1 \le j < l \le n} \left[1 + exp\left\{-\left(y_{ij} - y_{il}\right)\left(x_{ij} - x_{il}\right)^T \beta\right\}\right]^{-1} \quad (4)$$

Simulation results



PDA MENU

ODAL

ODAC

ODAP

ODAH

Hetero-aware

DLMM

dPQL

ODACH

dCLR

dist-EM

Data application -- COVID-19 data across 230 sites

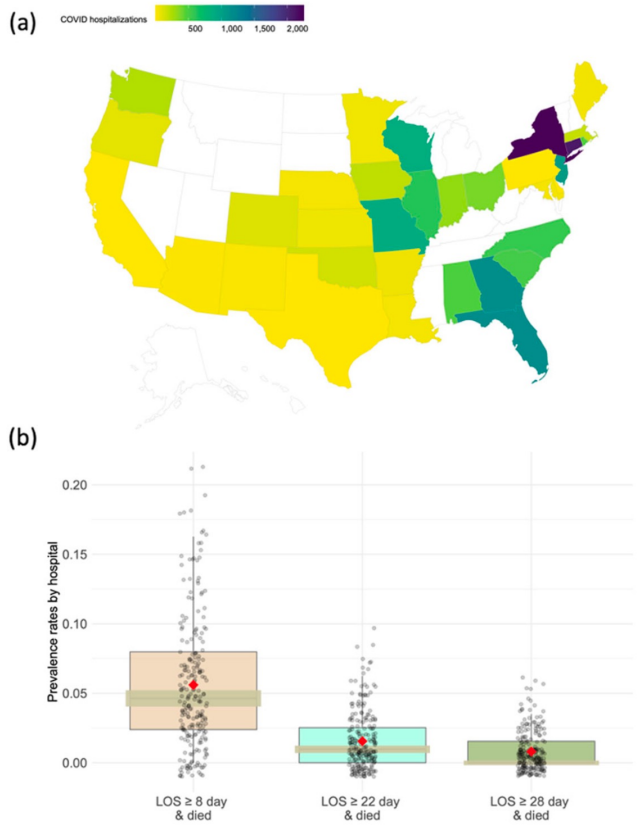


Fig. 1 Summary of real-world data from 230 sites. a COVID-19 cases distribution: number of COVID-19 hospitalizations included in the study are represented across 47 states created by open-source R package *usmap*[34] (https://cran.r-project.org/web/packages/usmap/usmap.pdf) (**b**) Box plots of the prevalence rates of composite outcomes of 230 hospitals.

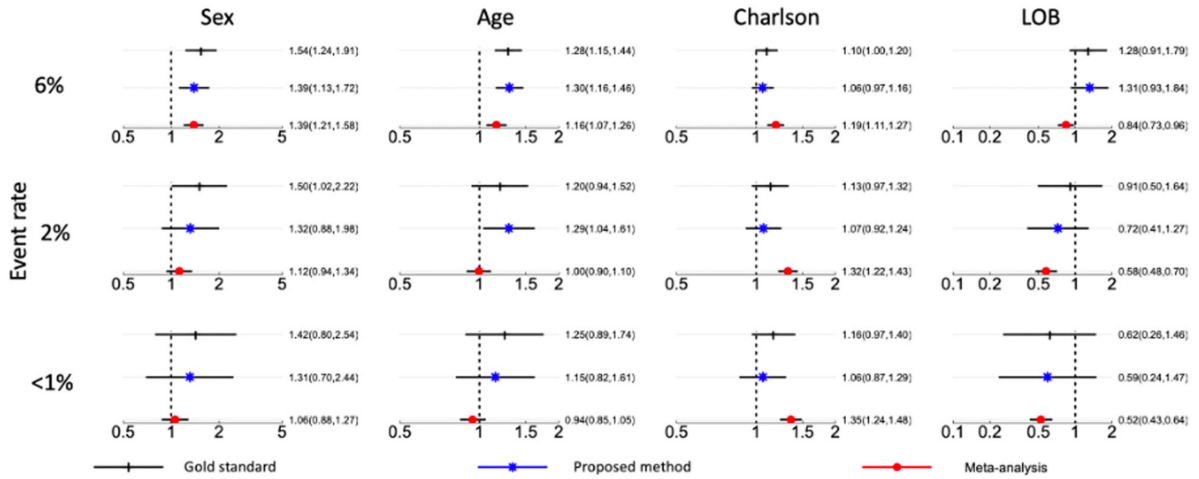
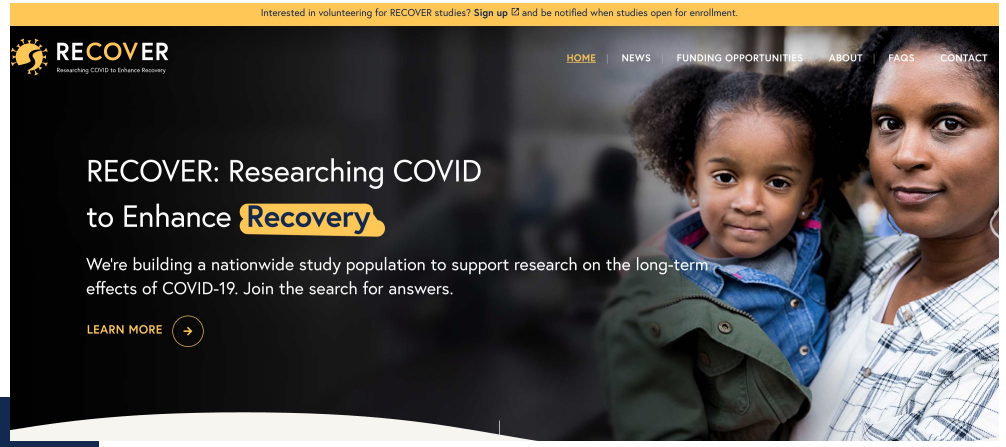


Fig. 4 Real-world data analysis results. Point estimates and 95% confidence intervals (CI) for the association (in odds ratio scale) between the LOS (i.e., length of stay) and covariates (i.e., sex, age, Charlson score, line of business, from left to right). Each row represents an event rate of the outcome: 6%, 2%, and <1% from top to bottom. Each column represents the estimation of the covariate.

RECOVER Initiative

RECOVER, a research initiative from the National Institutes of Health (NIH), seeks to understand, prevent, and treat PASC, including Long COVID. **PASC** stands for post-acute sequelae of SARS-CoV-2 and is a term scientists are using to study the potential consequences of a SARS-CoV-2 infection.



https://recoverCOVID.org/

PI for pediatric RECOVER:
Christopher Forrest (CHOP)

PI for adult RECOVER:
Rainu Kaushal (Weill Cornell)

Biostatistics Core Director:
Yong Chen
for PCORnet Pediatric RECOVER

Analytic tasks in year 1 (2021-2022)

- Team 1: Latent class analyses to identify MIS-C (Multisystem Inflammatory Syndrome in Children) subphenotypes and non-MIS-C PASC phenotypes
  - Federated learning approach to characterizing PASC
- Team 2: Disease trajectories following COVID-19 infection
  - To evaluate the impact of COVID-19 infection on disease trajectories for children with chronic medical conditions
  - To assess the risk of PASC in children with chronic medical conditions (compared to children without chronic medical conditions)
- Team 3: Epidemiological associations of PASC features; PASC computable phenotyping
- Team 4: Disparities in PASC at the individual and features using geospatial data; focus on disparities

Existing knowledge from reports at CDC (MMWR)

Morbidity and Mortality Weekly Report

COVID-19–Associated Multisystem Inflammatory Syndrome in Children — United States, March–July 2020

TABLE 1. Characteristics of patients (N = 570) reported with multisystem inflammatory syndrome in children (MIS–C) — United States, March–July 2020

Characteristic	No. (%)				
		Latent class analysis group*			
	Total (N = 570)	Class 1 (n = 203)	Class 2 (n = 169)	Class 3 (n = 198)	p value
Sex					
Female	**254 (44.6%)**	87 (42.9%)	81 (47.9%)	86 (43.4%)	0.57
Male	**316 (55.4%)**	116 (57.1%)	88 (52.1%)	112 (56.6%)	
Age (yrs), median (IQR)	**8 (4–12)**	9 (6–13)	10 (5–15)	6 (3–10)	<0.01
Race/Ethnicity					
Hispanic	**187 (40.5%)**	62 (36.9%)	62 (46.6%)	63 (39.1%)	0.03
Black, non–Hispanic	**153 (33.1%)**	66 (39.3%)	39 (29.3%)	48 (29.8%)	
White, non–Hispanic	**61 (13.2%)**	22 (13.1%)	15 (11.3%)	24 (14.9%)	
Other	**26 (5.6%)**	8 (4.8%)	6 (4.5%)	12 (7.5%)	
Multiple	**18 (3.9%)**	9 (5.4%)	5 (3.8%)	4 (2.5%)	
Asian	**13 (2.8%)**	1 (0.6%)	3 (2.3%)	9 (5.6%)	
American Indian/Alaskan Native	**3 (0.6%)**	0 (0.0%)	3 (2.3%)	0 (0.0%)	
Native Hawaiian/Pacific Islander	**1 (0.2%)**	0 (0.0%)	0 (0.0%)	1 (0.6%)	
Unknown	**108 (—)**	35 (—)	36 (—)	37 (—)	
Outcome					
Died	**10 (1.8%)**	1 (0.5%)	9 (5.3%)	0 (0.0%)	<0.01
Days in hospital, median (IQR)	**6 (4–9)**	8 (6–11)	6 (4–10)	5 (4–8)	<0.01
1	**16 (3.2%)**	3 (1.8%)	3 (2.0%)	10 (5.4%)	<0.01
2–7	**304 (60.2%)**	86 (50.3%)	87 (58.8%)	131 (70.4%)	
8–14	**149 (29.5%)**	66 (38.6%)	41 (27.7%)	42 (22.6%)	
≥15	**36 (7.1%)**	16 (9.4%)	17 (11.5%)	3 (1.6%)	
Missing	**65 (—)**	32 (—)	21 (—)	12 (—)	
ICU admission	**364 (63.9%)**	171 (84.2%)	105 (62.1%)	88 (44.4%)	<0.01
Days in ICU, median (IQR)	**5 (3–7)**	5 (4–7)	6 (3–9)	3 (2–5)	<0.01
Underlying medical conditions					<0.01
Obesity	**146 (25.6%)**	60 (29.6%)	49 (29.0%)	37 (18.7%)	0.02
Chronic lung disease	**48 (8.4%)**	18 (8.9%)	17 (10.1%)	13 (6.6%)	0.46
Clinical characteristic					

TABLE 1. (*Continued*) Characteristics of patients (N = 570) reported with multisystem inflammatory syndrome in children (MIS–C) — United States, March–July 2020

Characteristic	No. (%)				
		Latent class analysis group*			
	Total (N = 570)	Class 1 (n = 203)	Class 2 (n = 169)	Class 3 (n = 198)	p value
Respiratory**	**359 (63.0%)**	155 (76.4%)	129 (76.3%)	75 (37.9%)	<0.01
Cough	**163 (28.6%)**	51 (25.1%)	67 (39.6%)	45 (22.7%)	<0.01
Shortness of breath	**149 (26.1%)**	66 (32.5%)	59 (34.9%)	24 (12.1%)	<0.01
Chest pain or tightness	**66 (11.6%)**	33 (16.3%)	24 (14.2%)	9 (4.5%)	0.01
Pneumonia††	**110 (19.3%)**	47 (23.2%)	62 (36.7%)	1 (0.5%)	<0.01
ARDS	**34 (6.0%)**	14 (6.9%)	17 (10.1%)	3 (1.5%)	<0.01
Pleural effusion§§	**86 (15.8%)**	49 (24.7%)	29 (18.4%)	8 (4.2%)	<0.01
Neurologic	**218 (38.2%)**	107 (52.7%)	70 (41.4%)	41 (20.7%)	<0.01
Headache	**186 (32.6%)**	90 (44.3%)	63 (37.3%)	33 (16.7%)	<0.01
Renal	**105 (18.4%)**	77 (37.9%)	28 (16.6%)	0 (0.0%)	<0.01
Acute kidney injury	**105 (18.4%)**	77 (37.9%)	28 (16.6%)	0 (0.0%)	<0.01
Other					
Periorbital edema	**27 (4.7%)**	13 (6.4%)	5 (3.0%)	9 (4.5%)	0.32
Cervical lymphadenopathy >1.5 cm diameter	**76 (13.3%)**	28 (13.8%)	18 (10.7%)	30 (15.2%)	0.43
SARS COV–2 testing					
Any laboratory test done	**565 (99.1%)**	200 (98.5%)	169 (100.0%)	196 (99.0%)	0.39
Any positive laboratory test¶¶ (% among tested)	**565 (100.0%)**	200 (100.0%)	169 (100.0%)	196 (100.0%)	NA
PCR positive/Serology negative, not done, or missing***	**147 (25.8%)**	1 (0.5%)	142 (84.0%)	4 (2.0%)	<0.01
Serology positive/PCR negative†††	**263 (46.1%)**	138 (68.0%)	0 (0.0%)	125 (63.1%)	<0.01
PCR positive/Serology positive	**155 (27.2%)**	61 (30.0%)	27 (16.0%)	67 (33.8%)	<0.01
Epidemiologic link only, with no testing	**5 (0.9%)**	3 (1.5%)	0 (0.0%)	2 (1.0%)	<0.01
Treatment§§§					
IVIG¶¶¶	**424 (80.5%)**	174 (87.9%)	96 (62.7%)	154 (87.5%)	<0.01
Steroids	**331 (62.8%)**	145 (73.2%)	80 (52.3%)	106 (60.2%)	<0.01
Antiplatelet medication	**309 (58.6%)**	113 (57.1%)	69 (45.1%)	127 (72.2%)	<0.01
Anticoagulation medication	**233 (44.2%)**	92 (46.5%)	76 (49.7%)	65 (36.9%)	0.03
Vasoactive medications	**221 (41.9%)**	129 (65.2%)	64 (41.8%)	28 (15.9%)	<0.01
Respiratory support, any	**201 (38.1%)**	104 (52.5%)	79 (51.6%)	18 (10.2%)	<0.01
Intubation and mechanical ventilation	**69 (13.1%)**	37 (18.7%)	30 (19.6%)	2 (1.1%)	<0.01
Immune modulators	**119 (22.6%)**	52 (26.3%)	34 (22.2%)	33 (18.8%)	0.18
Dialysis	**2 (0.4%)**	0 (0.0%)	2 (1.3%)	0 (0.0%)	0.08

Penn Medicine

Subphenotypes for **post-acute phase** among children is unknown

Morbidity and Mortality Weekly Report

COVID-19–Associated Multisystem Inflammatory Syndrome in Children — United States, March–July 2020

Shana Godfred-Cato, DO[1]; Bobbi Bryant, MP
Katherine Roguski, MPH[1]; Bailey Wallace, MPH[1]
Maura K. Lash, MPH[3]; Kathleen H. Reilly, PhD[3]
Nottasorn Plipat, MD, PhD[8]; Gillian Richardson
Susan Hrapcak, MD[1]; Deblina Datta, MD[1]; Sap

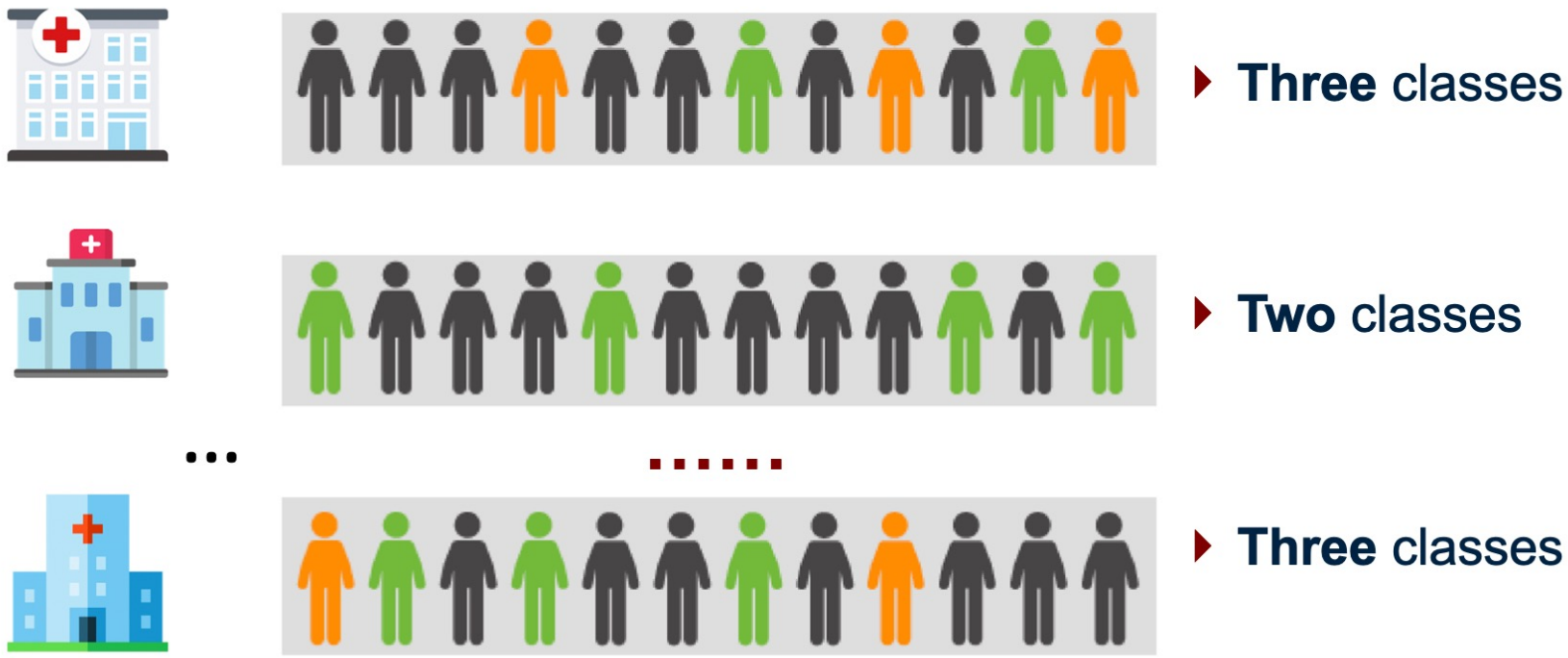
Multicenter latent class analysis for collaborative subphenotyping

How to use aggregated data across hospitals to jointly characterize subphenotypes, while allowing for between-site heterogeneity

A naïve solution:

- Fit mixture model within each hospital, and then integrate results (via some ad hoc procedures).

Issue 1:
The number of latent classes estimated in each hospital may be different.

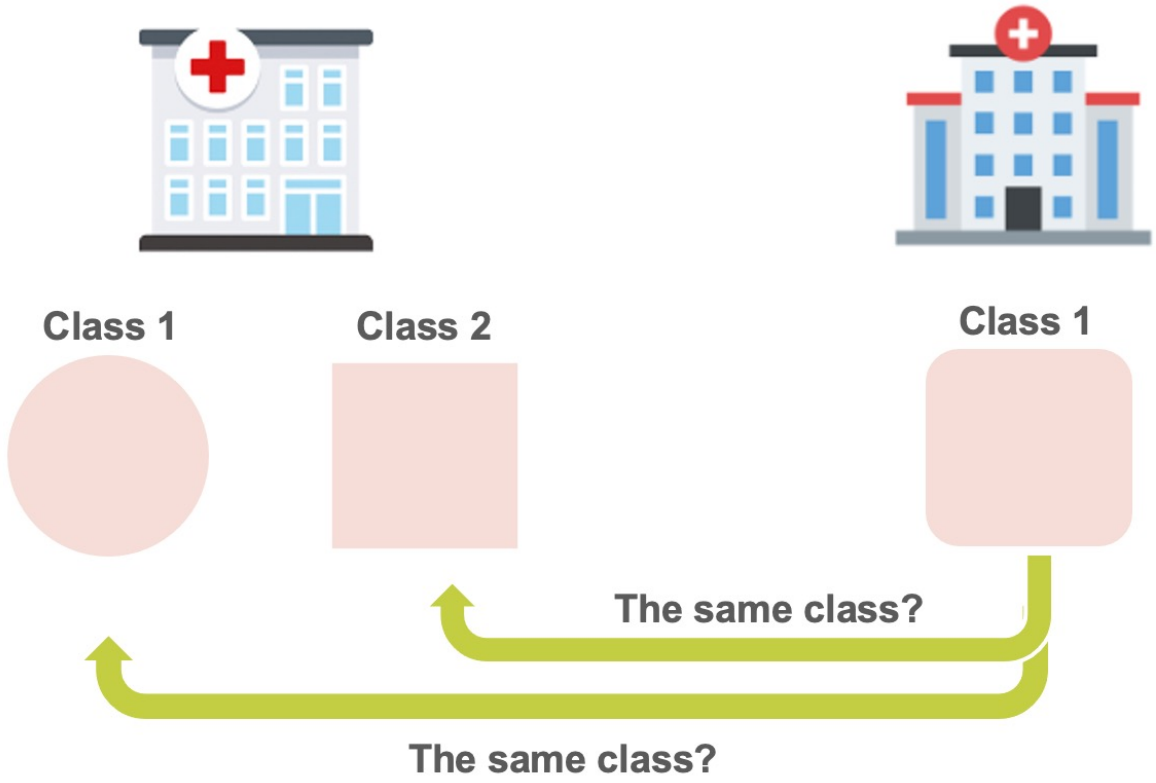


A naïve solution:

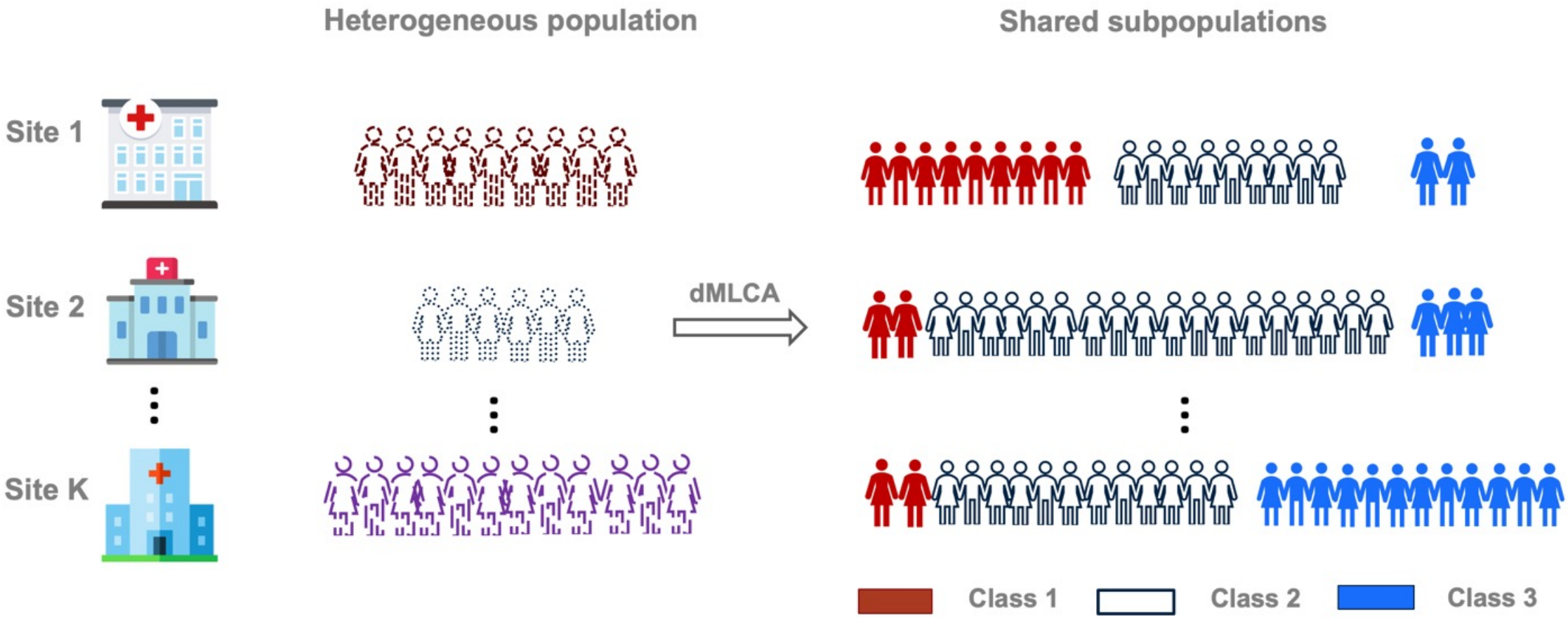
- Fit mixture model within each hospital, and then integrate results (via some ad hoc procedure).

Issue 2:

The difference between latent classes may be ambiguous



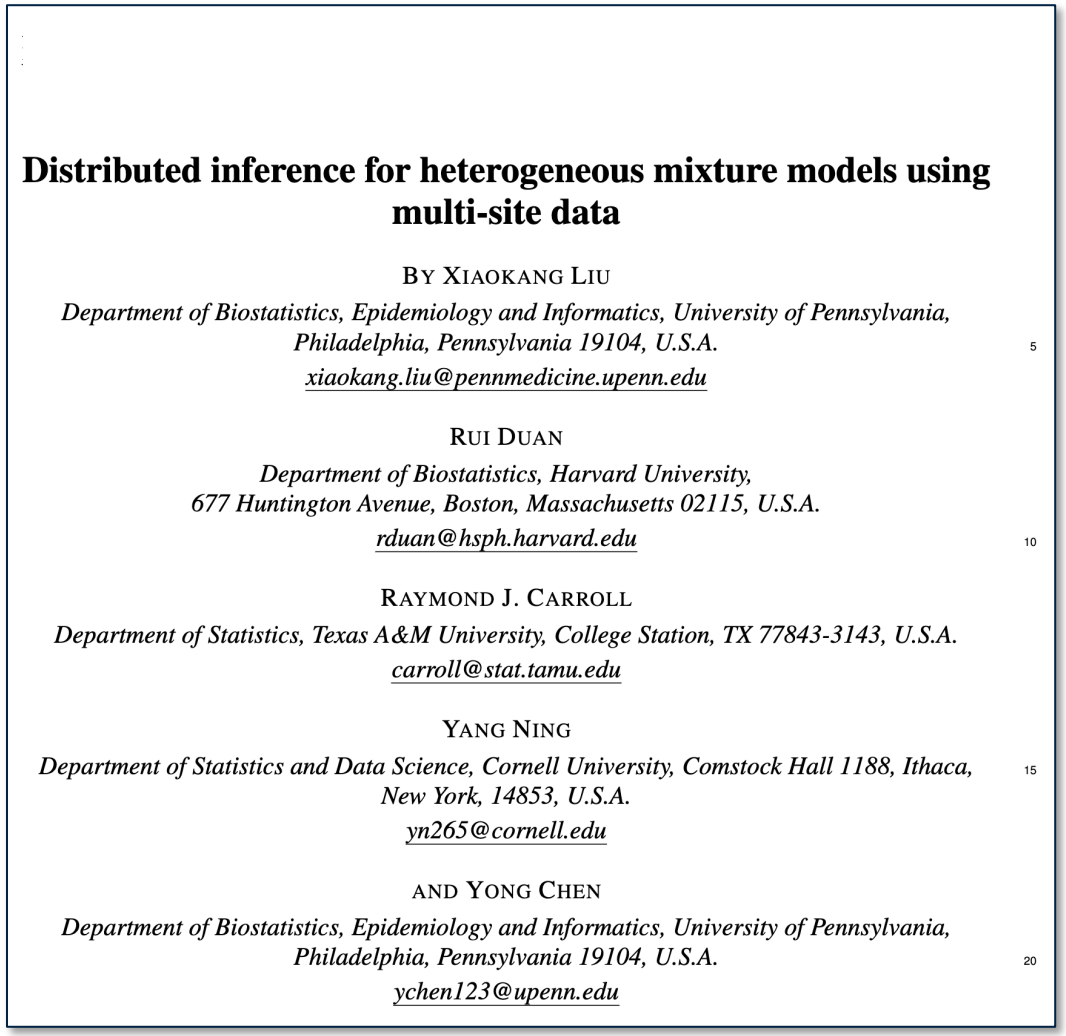
Consider a setting with 3 subpopulations



Distributed-EM algorithm:

- Mixture model formulation naturally leads to the use of EM algorithm
- Density ratio tilting technique to modify the "Q-function" in the EM algorithm to ensure the matching of high order gradients
- Construct "surrogate Q-function" in EM algorithm
- Theoretical guarantees

Liu et al., 2022 (under review)



Distributed inference for heterogeneous mixture models using multi-site data

BY XIAOKANG LIU

Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, Pennsylvania 19104, U.S.A.

xiaokang.liu@pennmedicine.upenn.edu

RUI DUAN

Department of Biostatistics, Harvard University, 677 Huntington Avenue, Boston, Massachusetts 02115, U.S.A.

rduan@hsph.harvard.edu

RAYMOND J. CARROLL

Department of Statistics, Texas A&M University, College Station, TX 77843-3143, U.S.A.

carroll@stat.tamu.edu

YANG NING

Department of Statistics and Data Science, Cornell University, Comstock Hall 1188, Ithaca, New York, 14853, U.S.A.

yn265@cornell.edu

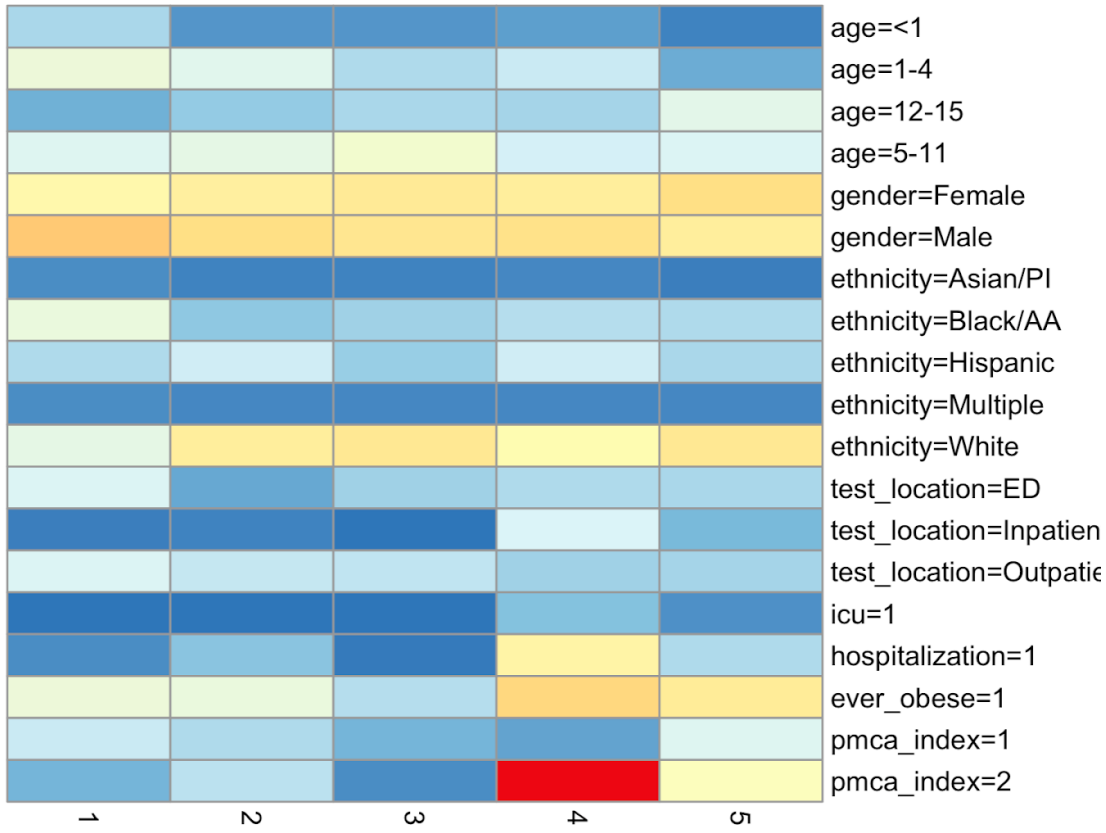
AND YONG CHEN

Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, Pennsylvania 19104, U.S.A.

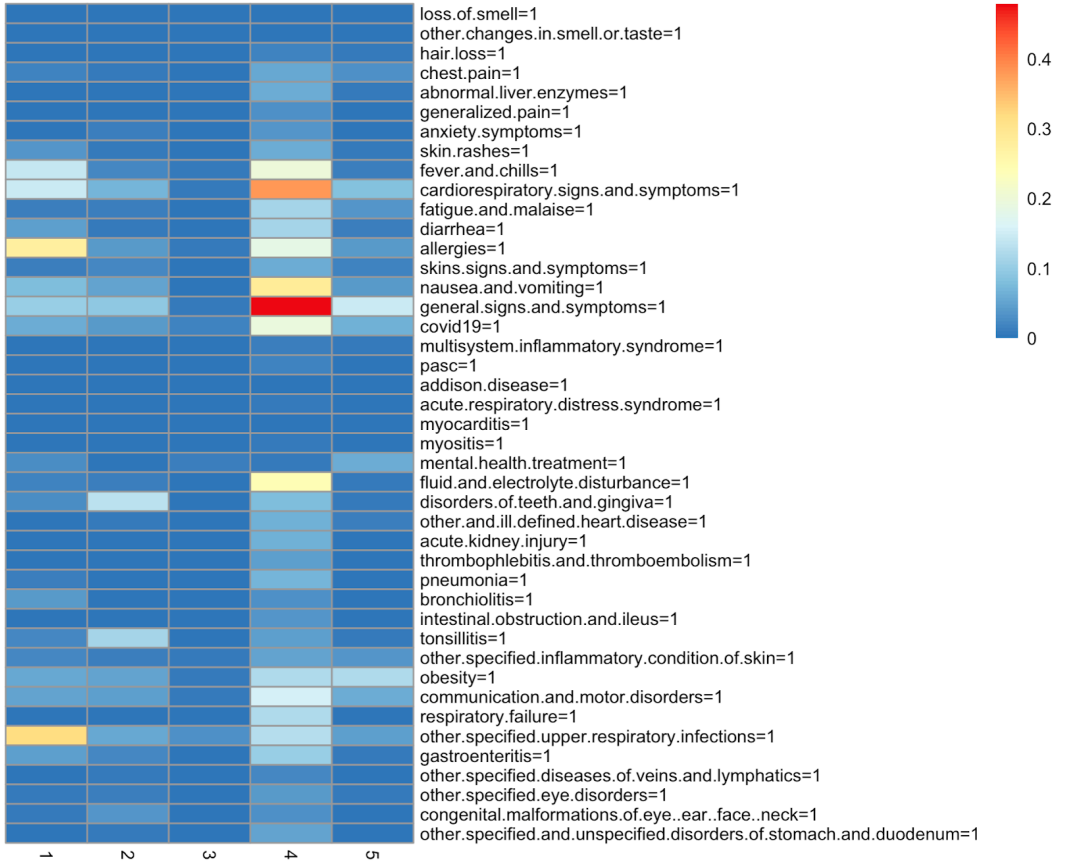
ychen123@upenn.edu

Multi-site subphenotyping (working version): 5 latent classes

Covariates



Syndromic and systemic features



PDA MENU

ODAL

ODAC

ODAP

ODAH

Hetero-aware

DLMM

dPQL

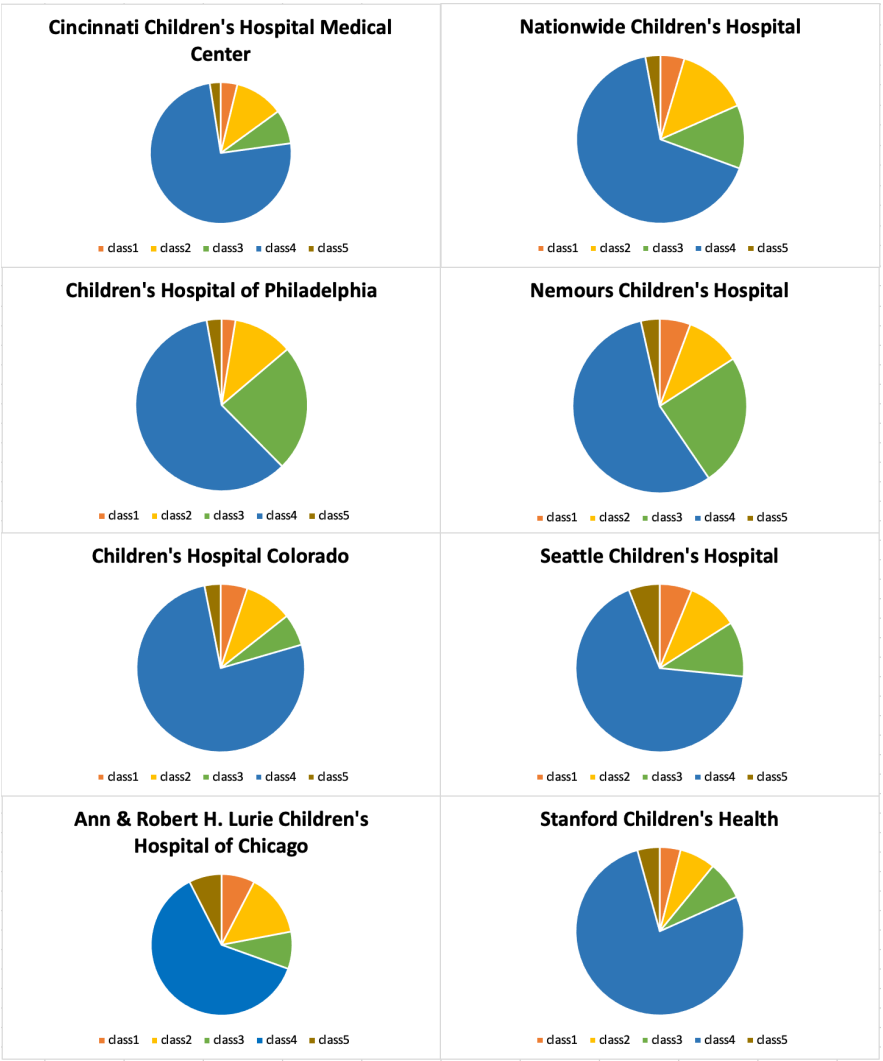
ODACH

dCLR

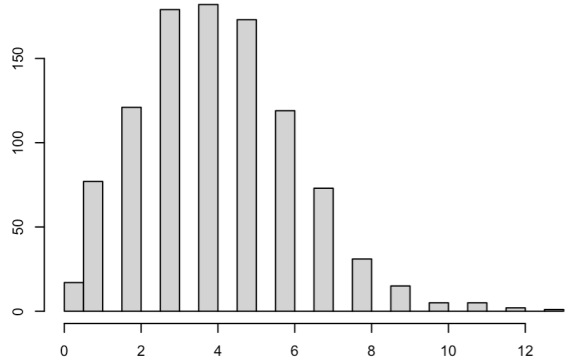
dist-EM

Interpretations of latent classes and between-site heterogeneity

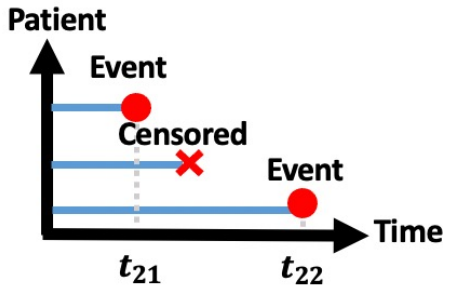
Class 1 (10.7%) (Mental health)	Class 2 (5.6%) (Respiratory PASC)	Class 3 (3.1%) (Post-hospitalization and post-icu PASC)	Class 4 (14.4%) (Younger children with persistent symptoms; syndromic PASC)	Class 5 (66.2%) (Mild PASC)
Age >= 12; Anxiety disorders; Minor depression; Attention deficit hyperactivity disorder; Autism spectrum disorder; Major depression; antidepressants; Development delay; Conduct disorder	pmca=1; Age = 5-11; Ethnicity = Black/AA Asthma; Allergies; Other specified upper respiratory infections; Cardiorespiratory signs; Obesity; Mental health treatment; Nasal congestion; cough; Other drugs for obstructive airway diseases inhalants; Adrenergics inhalants	pmca=2; General signs and symptoms; Cardiorespiratory signs and symptoms; Sleep wake disorders; Constipation; Abdominal signs and symptoms; Drugs of peptic ulcer and gastro oesophageal reflux disease; **Existence of many clusters**	Age<=4; pmca=0; Other specified upper respiratory infections; Fever and chills; Nausea and vomiting; Skin rashes; Bronchiolitis; Diarrhea; gastroenteritis; Nasal congestion; cough;	pmca=0; Age = 5-11; healthy



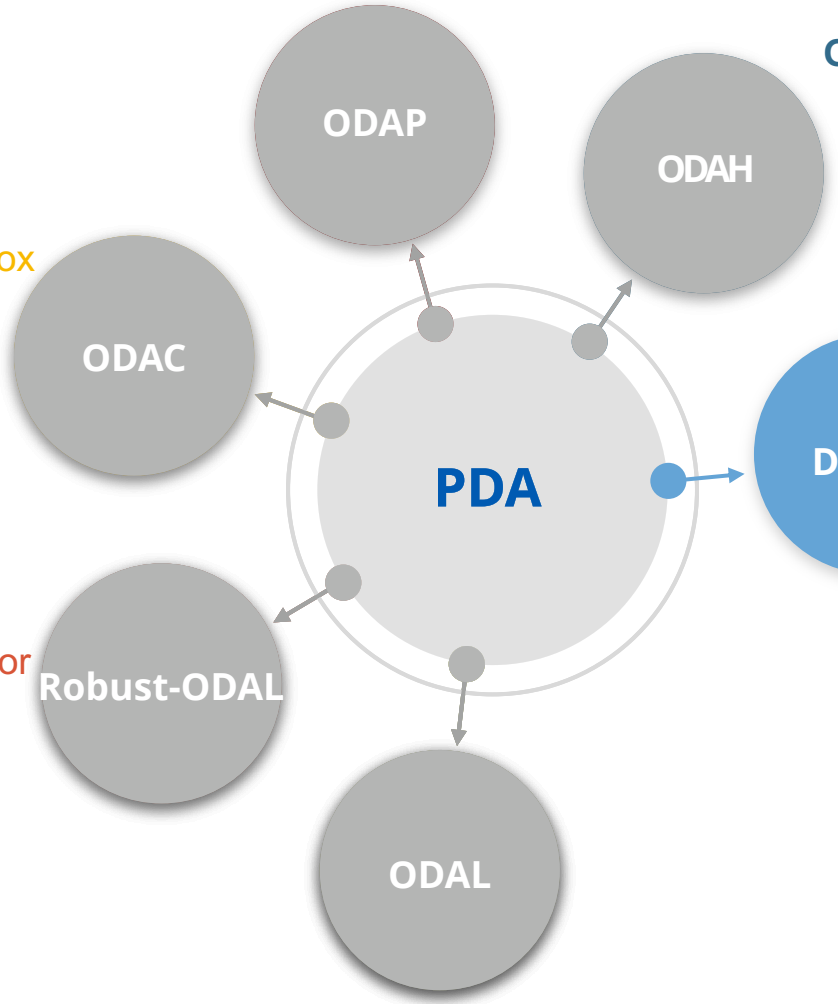
One-shot Distributed Algorithms for Poisson regression



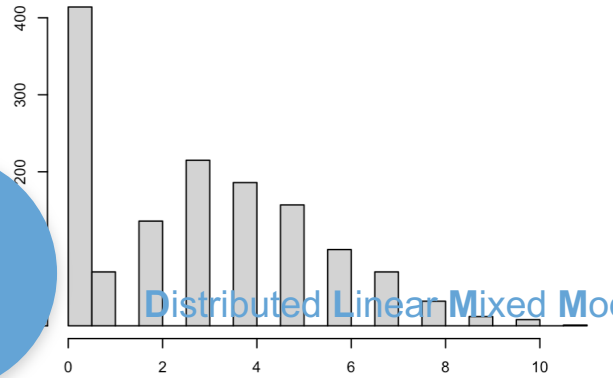
Cox



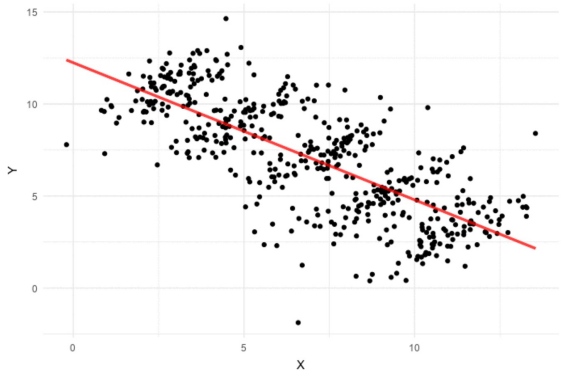
Robust ms for



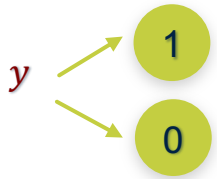
One-shot Distributed Algorithms for Hurdle regression



Distributed Linear Mixed Model



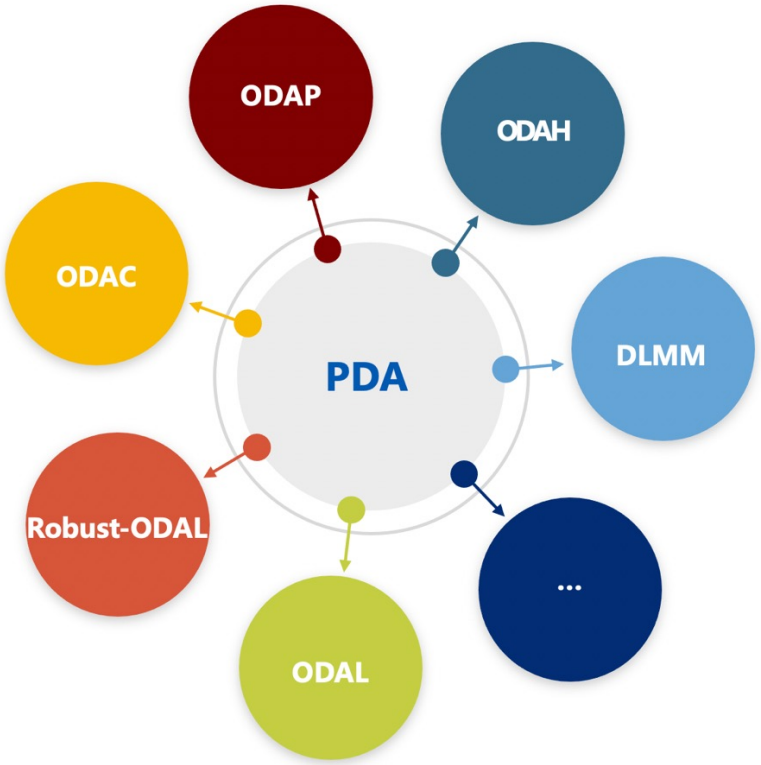
One-shot Distributed Algorithms for Logistic regression



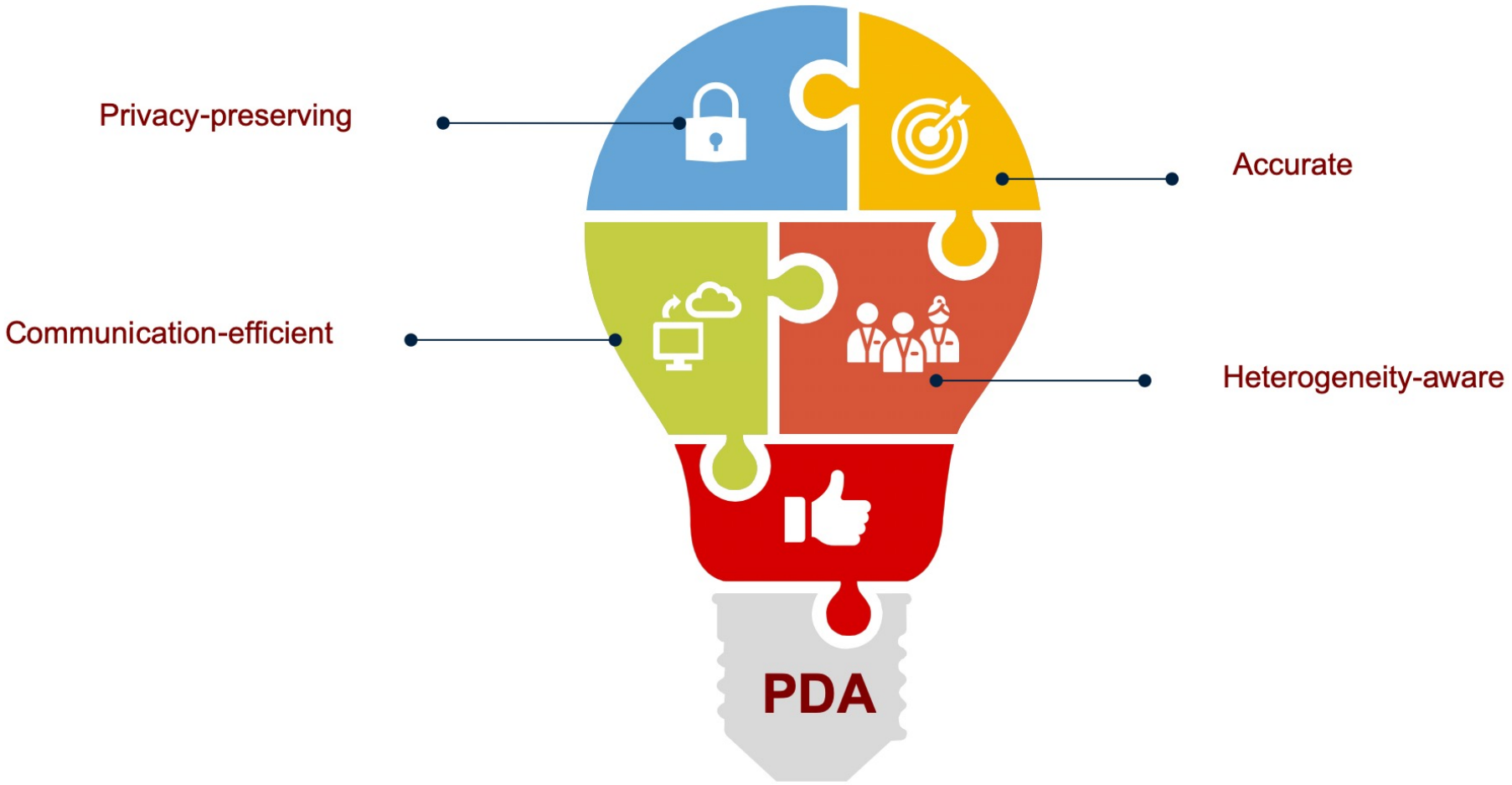
PDA as a collection of distributed algorithms

PDA : a solution for next generation data sharing

Privacy-preserving Distributed Algorithms

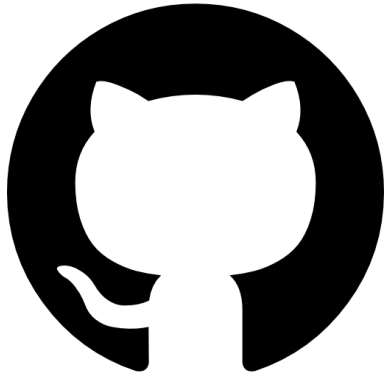


R package ‘pda’: four principles



A Solution for Next Generation Data Sharing for Collaborative Modeling

Privacy-preserving Distributed Algorithms



Tutorial