Statistical Inference with Adaptively Collected Data

Yash Deshpande

Massachusetts Institute of Technology

Youth in High-dimensions – ICTP 2020

An experiment on predicting warfarin dosage

Covariates	X_1	X_2	X_3	X_4	X_5	X_6
Dosage	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6

Fit $Y_i \approx f(X_i) = \langle X_i, \theta_0 \rangle$

The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 FEBRUARY 19, 2009 VOL. 360 NO. 8

Estimation of the Warfarin Dose with Clinical and Pharmacogenetic Data

The International Warfarin Pharmacogenetics Consortium*

ABSTRACT

BACKGROUND

Genetic variability among patients plays an important role in determining the dose of warfarin that should be used when oral anticoagulation is initiated, but practical methods of using genetic information have not been evaluated in a diverse and large population. We developed and used an algorithm for estimating the appropriate warfarin dose that is based on both clinical and genetic data from a broad population base.

METHODS

Clinical and genetic data from 4043 patients were used to create a dose algorithm that was based on clinical variables only and an algorithm in which genetic information was added to the clinical variables. In a validation cohort of 1009 subjects, we evaluated the potential clinical value of each algorithm by calculating the percentage of patients whose predicted dose of warfarin was within 20% of the actual stable therapeutic dose; we also evaluated other clinically relevant indicators.

Address reprint requests to the International Warfarin Pharmacogenetics Consortium at 300 Pasteur Dr., Ln. 301, Mail-stop 5120, Stanford, CA 94305, or at iwpc@pharmgkb.org.

*The members of the International Warfarin Pharmacogenetics Consortium are listed in the Appendix. The members of the writing committee take responsibility for the content of this article.

This article (10.1056/NEJMoa0809329) was updated on October 14, 2009, at NEJM.org.

N Engl J Med 2009;360:753-64.

What if data is collected adaptively?

Batch 1

X_1 X_2 X_3

Y_1 Y_2 Y_3

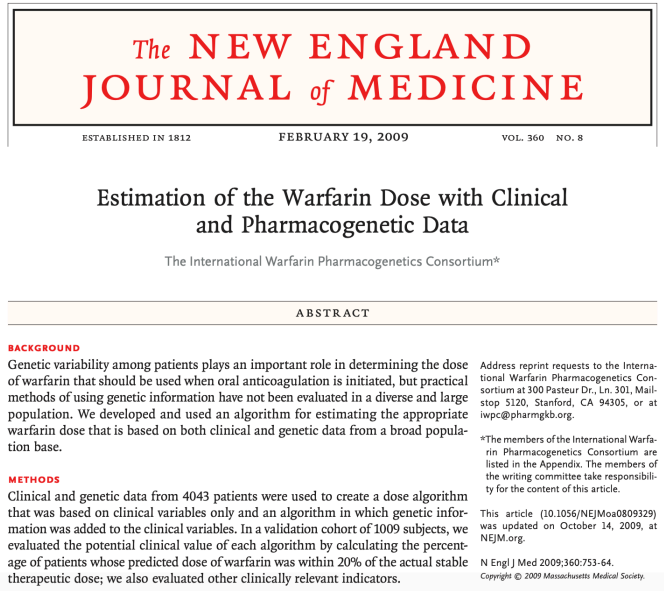
Batch 2

X_4 X_5 X_6

Y_4 Y_5 Y_6

…

Fit $Y_i \approx f(X_i) = \langle X_i, \theta_0 \rangle$



The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 FEBRUARY 19, 2009 VOL. 360 NO. 8

Estimation of the Warfarin Dose with Clinical and Pharmacogenetic Data

The International Warfarin Pharmacogenetics Consortium*

ABSTRACT

BACKGROUND

Genetic variability among patients plays an important role in determining the dose of warfarin that should be used when oral anticoagulation is initiated, but practical methods of using genetic information have not been evaluated in a diverse and large population. We developed and used an algorithm for estimating the appropriate warfarin dose that is based on both clinical and genetic data from a broad population base.

METHODS

Clinical and genetic data from 4043 patients were used to create a dose algorithm that was based on clinical variables only and an algorithm in which genetic information was added to the clinical variables. In a validation cohort of 1009 subjects, we evaluated the potential clinical value of each algorithm by calculating the percentage of patients whose predicted dose of warfarin was within 20% of the actual stable therapeutic dose; we also evaluated other clinically relevant indicators.

Address reprint requests to the International Warfarin Pharmacogenetics Consortium at 300 Pasteur Dr., Ln. 301, Mail-stop 5120, Stanford, CA 94305, or at iwpc@pharmgkb.org.

*The members of the International Warfarin Pharmacogenetics Consortium are listed in the Appendix. The members of the writing committee take responsibility for the content of this article.

This article (10.1056/NEJMoa0809329) was updated on October 14, 2009, at NEJM.org.

N Engl J Med 2009;360:753-64.

- Does this affect statistical estimation, i.e. consistency, error rates, …?
- … statistical inference? i.e. confidence intervals, p-values, …?

Compare two scenarios

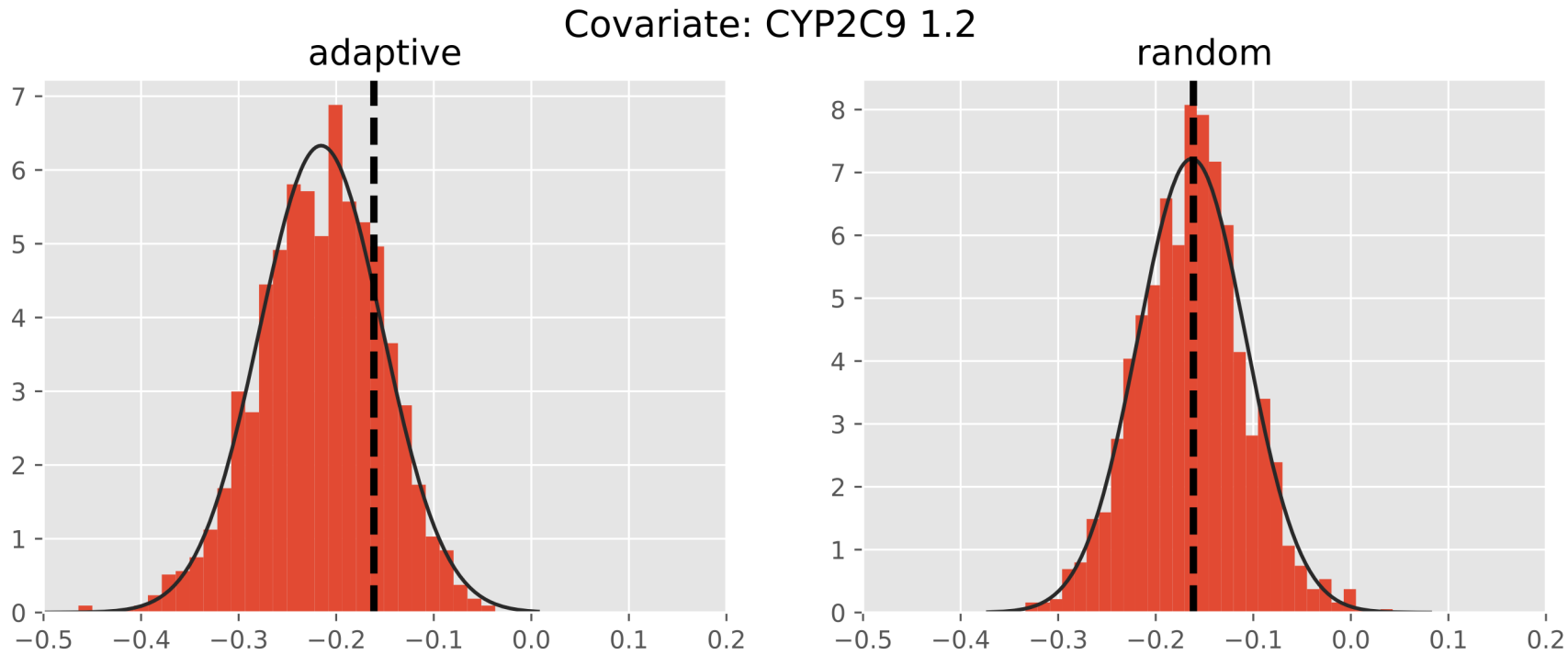
Adaptive collection

- Sample $n_1 = 250$ randomly and compute intermediate estimate on first batch $\widehat{\theta}^{int}$
- Sample $n_2 = 250$ from top 15 percentile of predicted dose $\langle X, \widehat{\theta}^{int} \rangle$

Random collection

- Sample $n = 500$ randomly without replacement

Least squares on the warfarin dataset



- Adaptive data collection may not affect estimator errors in size

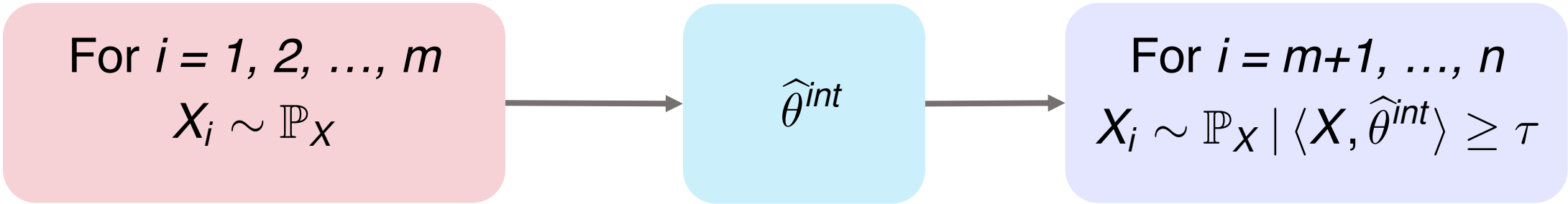
- However, estimators *can be biased*, i.e. it can affect error *shape*

My goal for today

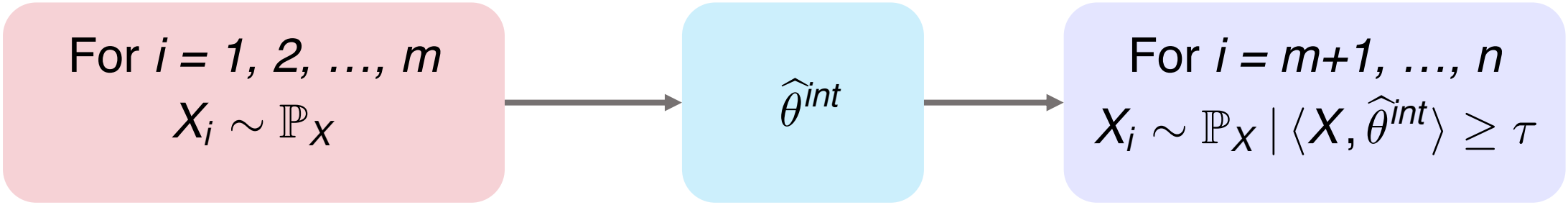
- A simple model for adaptive data collection
- Prior theory: what fails?
- A method: online debiasing, and a theorem
- Open problems and connections

Model

- Parameter $\theta_0 \in \mathbb{R}^p$
- Sample of size n: $Y_i = \langle X_i, \theta_0 \rangle + \varepsilon_i, \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma^2)$
- Data collected in two batches



Two reasons to focus on this model



- Approximates bandits, adaptive designs in medicine
- Captures key issue of adaptively collected data:
  future data (or X's) depend on past outcomes or Y, ε's

What does prior theory say?

Theorem (Lai, Wei 1985, Abbasi-Yadkori et al 2011)

Under mild conditions on $\mathbb{P}_X$, $\mathbb{E}\|\widehat{\theta}^{\mathrm{LS}} - \theta_0\|_2^2 \lesssim \sigma^2 \frac{p \log n}{n}$

$$\widehat{\theta}^{\mathrm{LS}} = \frac{1}{n}\widehat{\Sigma}_n^{-1} X^n Y^n$$

$$\widehat{\Sigma}_n = \frac{1}{n}\sum_i X_i X_i^T$$

Theorem (Lai, Wei 1985)

If, in addition, data collection is **stable** i.e. $\widehat{\Sigma}_n \mathbb{E}\{\widehat{\Sigma}_n^{-1}\} \to I_p$

$$\sqrt{n}\widehat{\Sigma}_n^{1/2}(\widehat{\theta}^{LS} - \theta_0) \Rightarrow \mathrm{N}(0, \sigma^2 I_p)$$

Consistency robust to adaptive collection, inferential theory is not!

Stability fails in high dimensions

Proposition (D. 2019)

If $\mathbb{P}_X$ is standard normal and batches are equal sized:

$$\text{Bias} \quad \mathbb{E}\{\widehat{\theta}^{LS} - \theta_0\} = c\theta_0 \frac{p}{n} + O\left(\frac{p}{n^{3/2}}\right)$$

- Compare bias of order $\frac{p}{n}$ to (per entry) standard deviation of order $\frac{1}{\sqrt{n}}$
- Bias negligible only if $n \gg p^2$
- However, consistency needs $n \gg p \log p$

Towards a solution: predictable estimators

- Let us try a linear estimator for $p = 1$

$$\widehat{\theta}^d = \sum_i W_i Y_i$$

As $Y_i = X_i \theta_0 + \varepsilon_i$

$$= \theta_0 + \left(\sum_i W_i X_i - 1 \right) \theta_0 + \sum_i W_i \varepsilon_i$$

- Suppose $\sum_i W_i \varepsilon_i$ stable martingale: $W_i = \text{func}(Y_1, \ldots, Y_{i-1}, X_1, \ldots, X_i)$

$$\text{bias} = \theta_0 \cdot \mathbb{E}\left(\sum_i W_i X_i - 1 \right) \qquad \text{variance} = \sigma^2 \mathbb{E}\left(\sum_i W_i^2 \right)$$

Predictable estimators to debiased estimators

- Let us try a linear estimator for $p = 1$

$$\widehat{\theta}^{d} = \widehat{\theta}^{\mathrm{LS}} + \sum_i W_i (Y_i - X_i \widehat{\theta}^{\mathrm{LS}})$$

$$= \theta_0 + \Big(\sum_i W_i X_i - 1 \Big) (\theta_0 - \widehat{\theta}^{\mathrm{LS}}) + \sum_i W_i \varepsilon_i$$

- Suppose $\sum_i W_i \varepsilon_i$ stable martingale: $W_i = \mathrm{func}(Y_1, \dots, Y_{i-1}, X_1, \dots, X_i)$

$$\text{bias} = \mathbb{E}\Big[(\widehat{\theta}^{\mathrm{LS}} - \theta_0) \Big(\sum_i W_i X_i - 1 \Big) \Big] \qquad \text{variance} = \sigma^2 \mathbb{E}\Big(\sum_i W_i^2 \Big)$$

Constructing weights: bias-variance tradeoff

- Optimize (or trade off) the bias and variance

$$W_i \in \arg\min_w \left(1 - \sum_{j<i} X_j W_j - X_i w\right)^2 + \nu w^2$$

- In general

$$W_i \in \arg\min_w \left\| 1 - \sum_{j<i} X_j W_j^T - X_i w^T \right\|_F^2 + \nu \| w \|_2^2$$

A theorem for online debiasing

Theorem (D, Mackey, Syrgkanis, Taddy 2018)

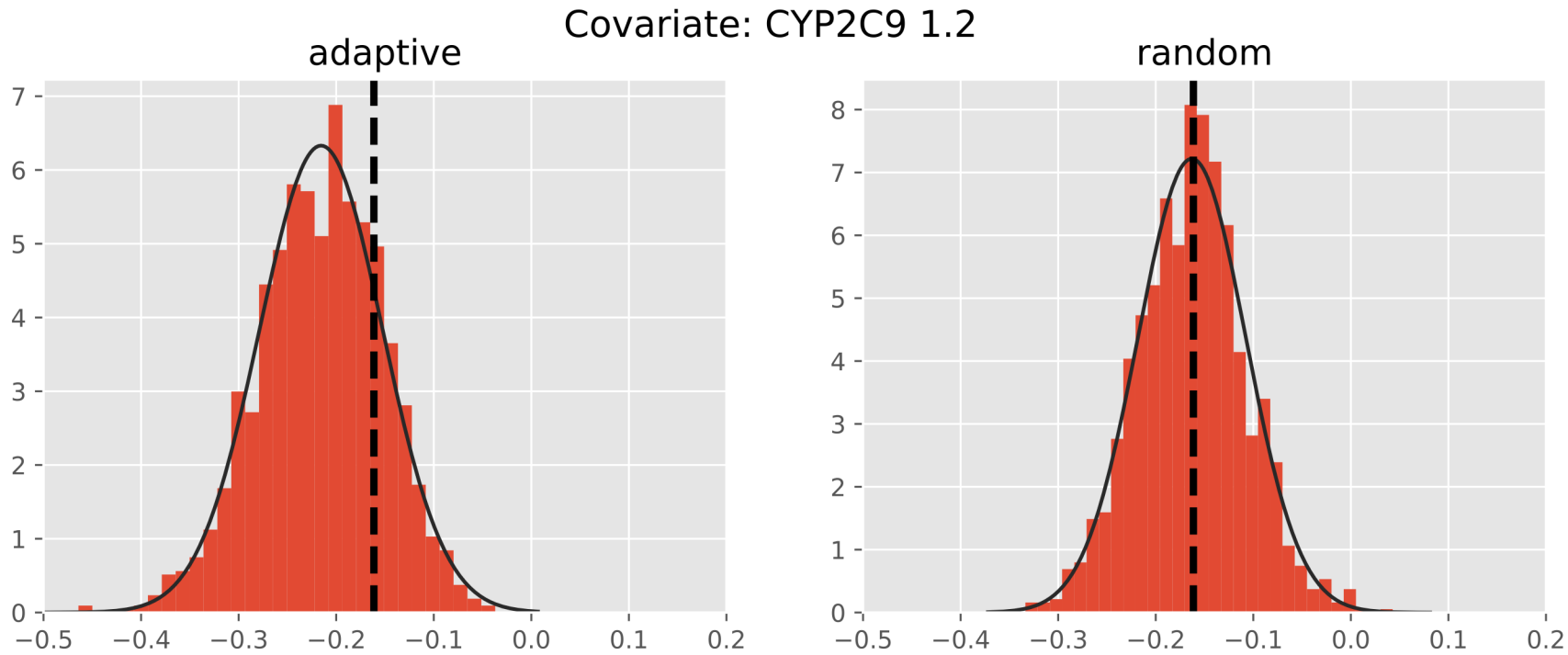
Under mild conditions on $\mathbb{P}_X$ and $n = \Omega(p \log p)$

$$(nV_n)^{-1/2}(\widehat{\theta}_1^d - \theta_{0,1}) \Rightarrow \mathsf{N}(0, \sigma^2)$$

where variance $V_n = n^{-1}\left(\sum_i W_i W_i^T\right)_{11}$ is order one

- Stability is built-in, so consistency of $\widehat{\theta}^{\mathsf{LS}}$ is enough!
- Regularization ν depends on $\mathbb{E}\lambda_{\min}(\widehat{\Sigma}_n)$

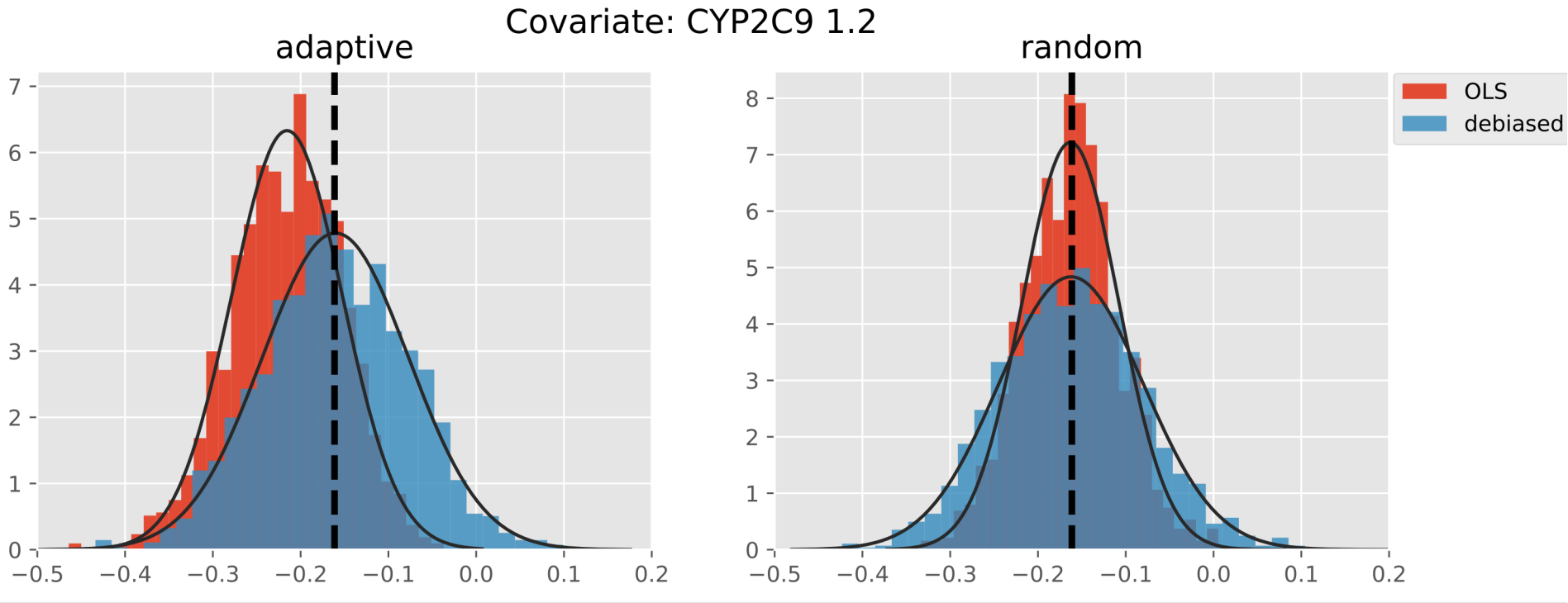
Recall our experiment: before debiasing



- Adaptive data collection may not affect estimator errors in size

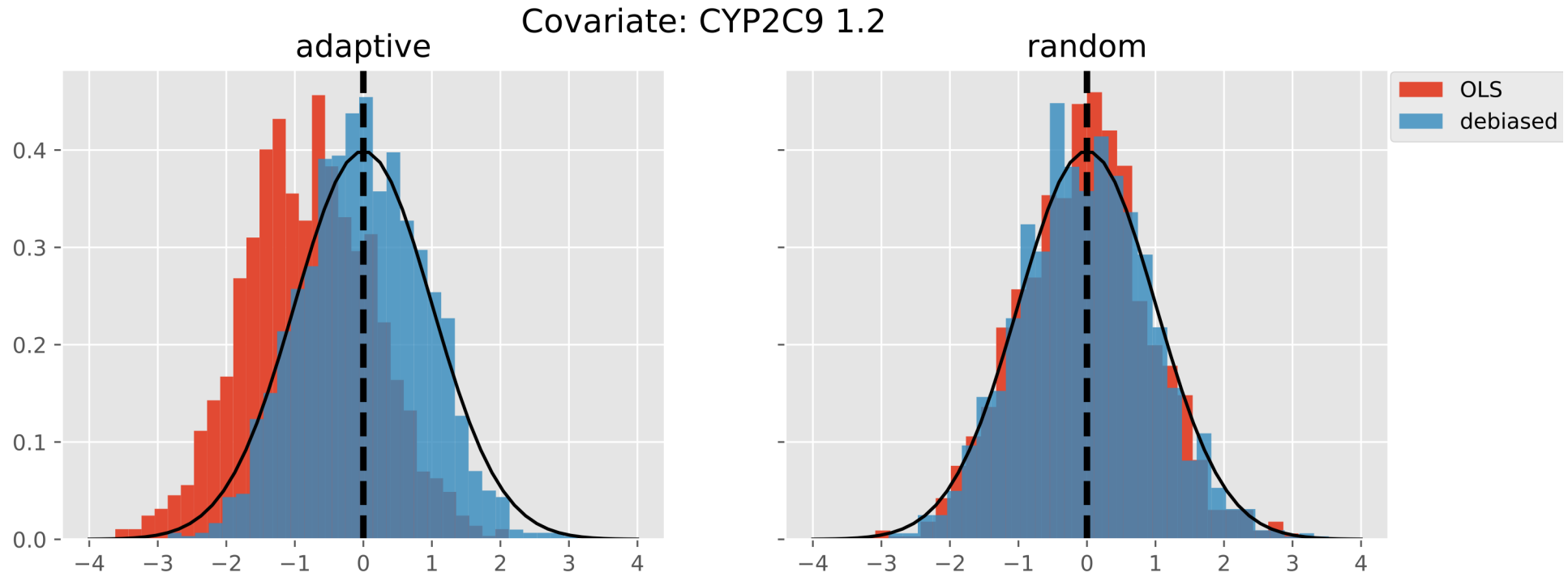
- However, estimators *can be biased*, i.e. it can affect error *shape*

After debiasing



- Online debiasing removes the bias with the *same data*
- Variance of debiased estimator is somewhat inflated

After debiasing: normalized errors



- Online debiasing removes the bias due to adaptive sampling
- In general, corrects for defects in all moments

Flexibility to base estimators: the LASSO

Theorem (D, Javanmard, Mehrabi 2019)

Suppose θ_0 is s-sparse, intermediate and final estimator is LASSO. Then, online debiasing works provided*

$$n \gg (s \log p)^2$$

- Sample size requirement scales with s, not ambient dimension p
- Online debiasing adapts to geometry of LASSO

To conclude

- Online debiasing: a robust, flexible ‘wrapper’ for inference.
  different base estimators, non-linear models, …
- Key issue: what data do we get to see?
  bandits, reinforcement learning and causal inference
- Data provenance crucial for valid inference