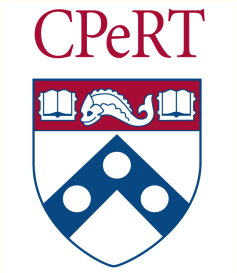
Instrumental Variable Methods for Continuous Outcomes that Accommodate Non-ignorable Missing Baseline Values

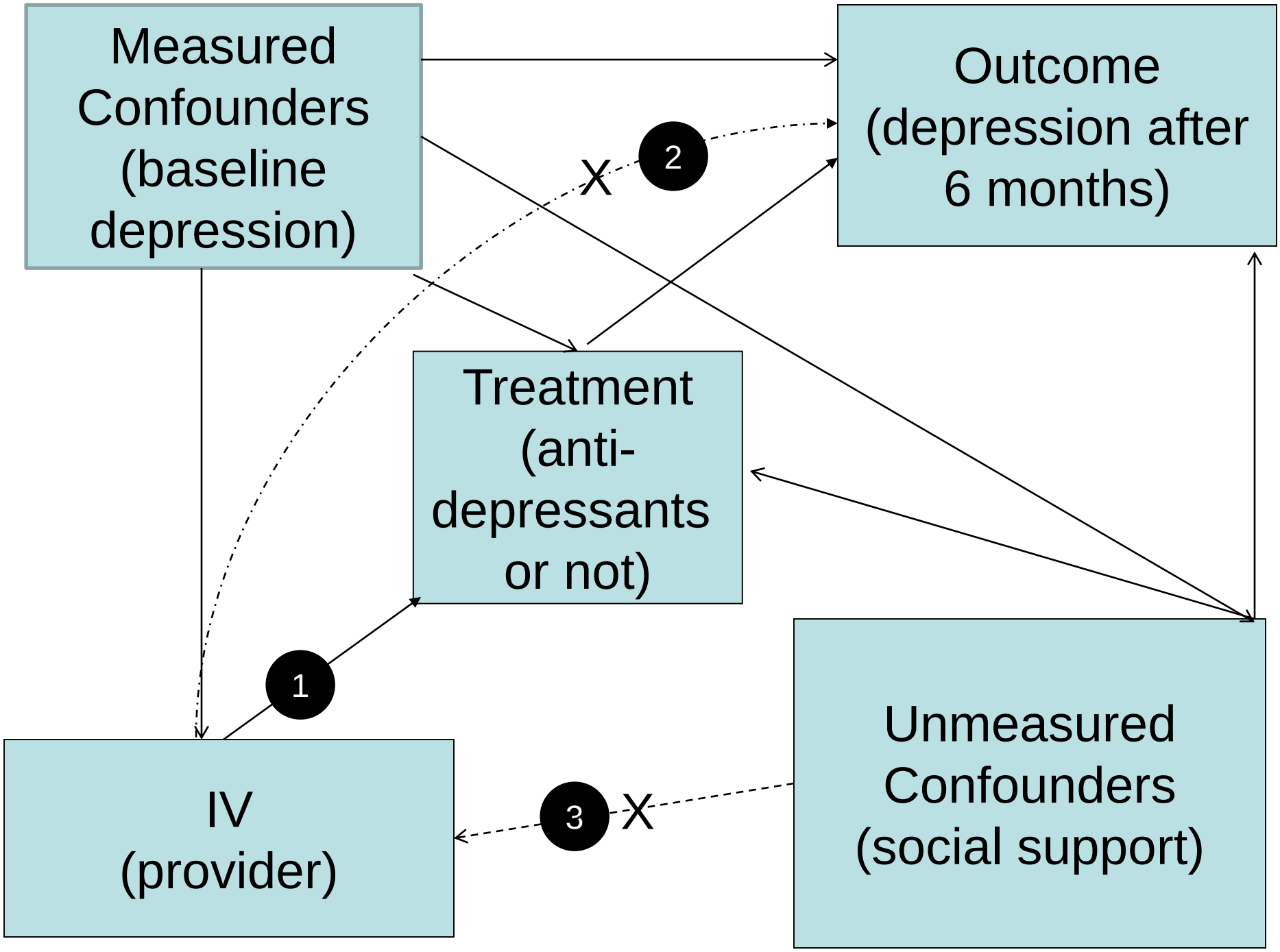
Ashkan Ertefaie Sean Hennessy James Flory Dylan Small

The Wharton School; Center for Pharmacoepidemiology Research and Training,
Perelman School of Medicine,
University of Pennsylvania



Confounding by Indication and Instrumental Variables

- In observational studies of health outcomes, the effects of a treatment may be confounded with the reasons a patient receives the treatment, "confounding by indication."
  - Patients receiving antidepressant drugs are more, not less, likely than others to be depressed, not because the drugs cause depression, but because the drugs are given to people who are depressed.
- Instrumental variables are used to break up confounding by indication.
- The key assumptions for a pretreatment variable to be a valid instrumental variable (IV) are:
  - IV is associated with the treatment
  - IV affects the outcome only through the treatment
  - IV is not associated with unmeasured confounders after conditioning on measured confounders.
- Provider preference is often used as an IV in pharmacoepidemiology studies comparing treatment A vs. treatment B.


Measured Confounders (baseline depression)
Outcome (depression after 6 months)
X
2
Treatment (anti-depressants or not)
1
IV (provider)
3
X
Unmeasured Confounders (social support)

Two Stage Least Squares IV estimation

1. Regress treatment on IV(s) and measured covariates.
2. Regress outcome on predicted treatment and measured covariates.

- Two stage least squares provides consistent estimate of average treatment effect under an assumption that the IV is valid, treatment effect is homogeneous and measured covariates have no missing data.
- When treatment effect is not homogeneous, two stage least squares estimates a particular weighted average of treatment effects (Hernán and Robins, 2006).

Missing Data Problem

- Providers typically see a different mix of patients.
- Provider is only a plausible IV after conditioning on patient mix variables that affect outcomes (e.g., baseline outcomes).
  - Providers who see more depressed patients may appear to prefer anti-depressants but this may just reflect their patient mix rather than genuine preference.
- Patient mix variables are often missing for some patients in health care databases.
- Ignoring missing values in patients may bias the result
- Missing covariate data is particularly problematic when the probability that a value is missing is related to the value itself, non-ignorable missingness (baseline depression may be less likely to be measured and more likely to be missing if the provider does not think person is depressed).
- In such cases, imputation methods based on missing at random assumptions are biased.

Assumptions

A.1 The missingness probability can be decomposed into two multiplicative components

- one is the missingness at the provider level that cannot be related to the unmeasured confounders and the treatment prescribed.
- the other is at the individual level that can depend on individuals' characteristics (measured or unmeasured)

A.2 The effect of unmeasured confounders on the choice of treatment does not vary by providers, e.g., assuming a logit model for treatment, the log odds curve of probability of treatment given the unmeasured confounders and the measured confounders is parallel for the different providers.

A.3 The intercepts for the log odds curves (genuine provider preferences) are independent of the measured and unmeasured confounders.

Proposed Method

We propose a two-step procedure to estimate the treatment effect in the presence of non-ignorable missingness.

1. Complete-case analysis: Fit a linear mixed effects model to predict the treatment among patients without missing data that includes
   - measured patient mix variables, and
   - a random intercept for each provider ID.

   The estimated random effect is considered as PP IV (genuine provider preference).
2. Estimate the treatment effect using two-stage least squares (2SLS) on all patients:
   - Regress the treatment on the PP IV and covariates with no missing values.
   - Regress the outcome on the predicted values of treatment and covariates with no missing values.

Intuition behind method:

1. Under A.1 and A.2, step 1 provides genuine provider preferences.
2. Under A.3, if we use genuine provider preferences as IV, we do not need to control for measured confounders with missing values.

Simulation Study

- 200 providers.
- For the ith patient of physician j, X_{ji} is a baseline variable $\sim N(\mu_j, 2^2)$; $\mu_j \sim N(0, .5^2)$.
- Treatment D_{ji} is generated as $P(D_{ji} = 1) = \text{expit}(-1 + b_j + U_{ji} + X_{ji})$ where $b_j \sim N(0, 1)$ is genuine provider preference for provider j.
- Outcome $Y_{ji} = .5X_{ji} + D_{ji} + 2U_{ji} + N(0, .5^2)$.
- Probability that X_{ji} is missing is $\text{expit}(2 + X_{ji} + U_{ji} + .5Y^*_{ji})\text{expit}(2 + V_j + V_j X_{ji})$, where Y^* is standardized outcome and $V_j \sim \text{Uniform}[-2, 2]$ reflects effect of provider j on missingnness.
- Missingness mechanism is non-ignorable because it depends on an unmeasured covariate and the outcome.
- Missingness rate varies across providers, and in general, patients with higher values of X have higher rates of missingness.

Simulation Study Results

Table: Simulation Study.

Method	$n = 200$		$n = 400$	
	Bias	S.D.	Bias	S.D.
Regression	1.61	0.04	1.60	0.33
Standard IV^{CC}	0.34	0.26	0.33	0.24
Standard IV^{MI}	0.72	0.14	0.70	0.11
Standard IV^{IPW}	0.41	0.28	0.44	0.25
Proposed	0.04	0.10	0.02	0.08

Robustness to violation of Assumpion A.1

A.1 The missingness probability can be decomposed into two multiplicative components

- one is the missingness at the provider level that cannot be related to the unmeasured confounders and the treatment prescribed.
- the other is at the individual level that can depend on individuals' characteristics (measured or unmeasured)

- Violation scenario I: Covariate X is not missing with probability $\text{expit}(-1 + X_{ji} - U_{ji} + .5Y^*_{ji} + .5V_j - X_{ji}U_{ji})$
- Violation scenario II: Covariate X is not missing with probability $\text{expit}(-1 + X_{ji} - U_{ji} + .5Y^*_{ji} + .5V_j - X_{ji}U_{ji} - V_jU_{ji})$

Robustness Simulation Study Results

Table: Simulation Study for robustness to violation of Assumption A.1.

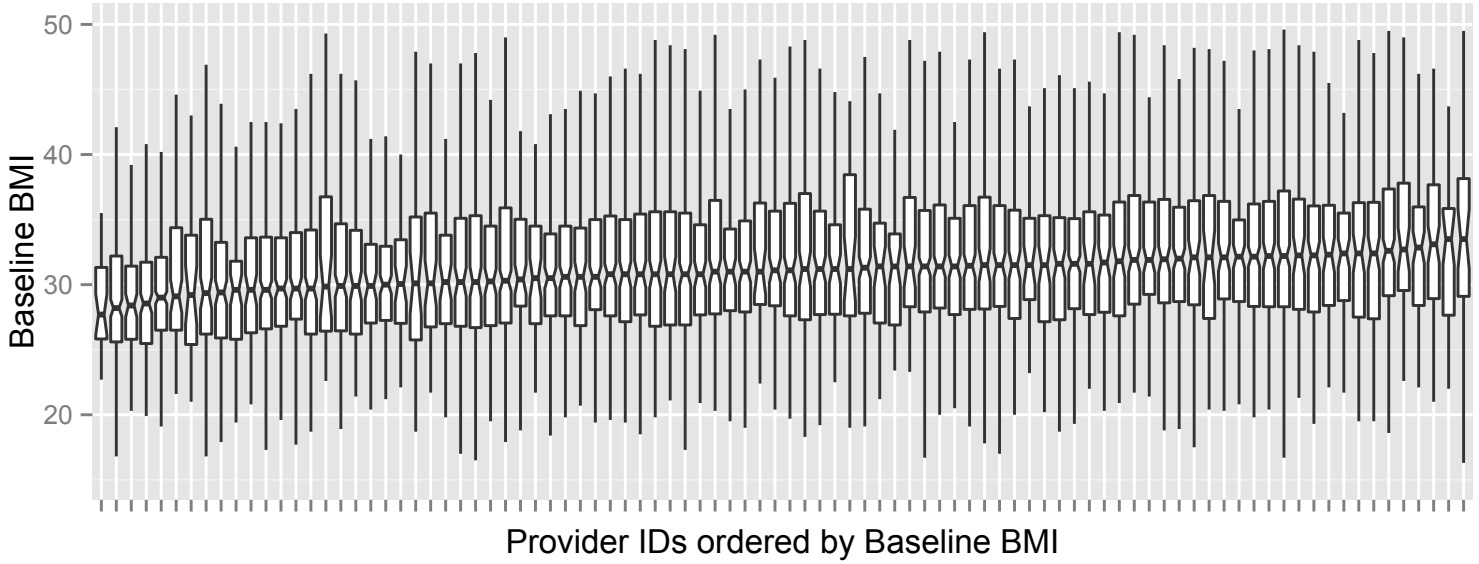
Method	$n = 200$		$n = 400$	
	Bias	S.D.	Bias	S.D.
Scenario I				
Regression	1.41	0.03	1.40	0.02
Standard IV^{CC}	0.56	0.15	0.52	0.13
Standard IV^{MI}	0.37	0.10	0.31	0.10
Standard IV^{IPW}	0.68	0.16	0.67	0.13
Proposed	0.04	0.10	0.02	0.10
Scenario II				
Regression	1.28	0.03	1.27	0.02
Standard IV^{CC}	0.22	0.29	0.21	0.28
Standard IV^{MI}	0.70	0.13	0.65	0.12
Standard IV^{IPW}	0.28	0.28	0.27	0.26
Proposed	0.05	0.09	0.02	0.07

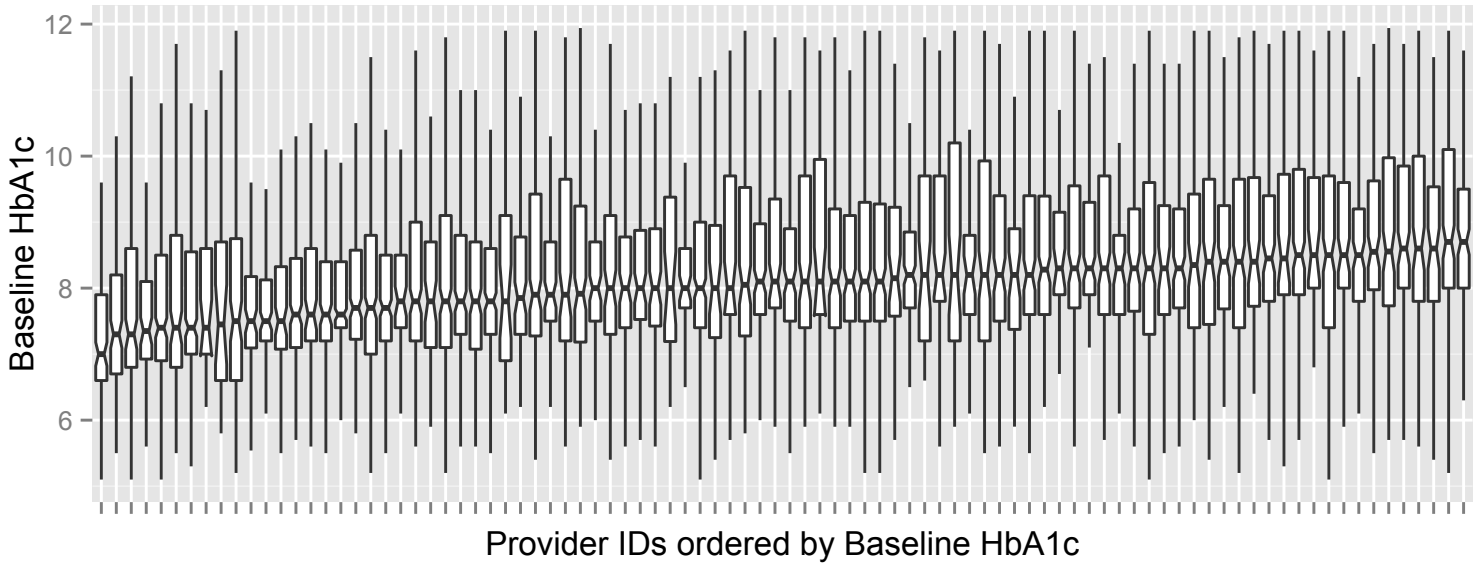
Application to healthcare data

The goal is to assess the effect of sulfonylurea vs. metformin as initial therapy on BMI among diabetic patients.

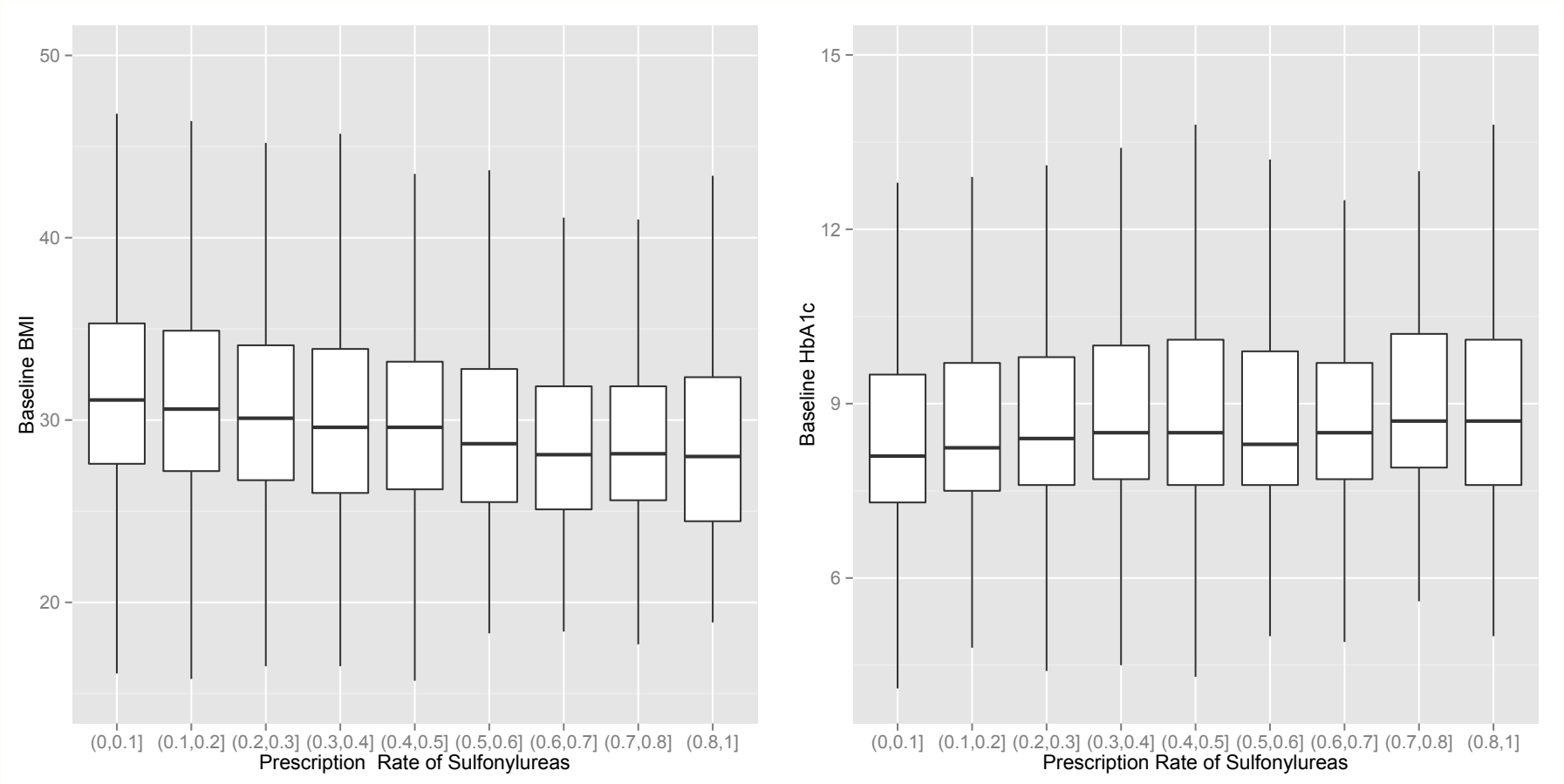
- We identified 141,080 patients using The Health Improvement Network (THIN), EMR database.
- Outcome: The first measurement of BMI after two years of follow-up.
- Missing values: 40% in baseline BMI and 46% in baseline Hba1c. In total 61% of patients had missing values for either of these baseline measurements (85,471).

There is Clustering of BMI and HbA1c by Provider





Does prescription rate of sulfonylureas depend on baseline BMI and Hba1c?



Possible analyses

- *Standard IV*^{1} : does not include baseline BMI and HbA1c.

- Complete-case analysis:
  - *Standard IV*^{CC2} : includes only the baseline BMI.
  - *Standard IV*^{CC3} : includes both baseline BMI and HbA1c.
  - *Regression*: regress the outcome on treatment, baseline BMI and baseline HbA1c.

- Considering missingness under missing at random assumption:
  - *Standard IV*^{MI} : imputes the missing values and fits the same model as *Standard IV*^{CC3}.
  - *Standard IV*^{IPW} : adjusts for missing mechanism by estimating the missing probabilities using the observed covariates and fits the same model as *Standard IV*^{CC3}.
- Considering non-ignorable missingness:
  - *Proposed Method*.

Estimating Treatment effect of sulfonylureas vs. metformin on BMI

Table: THIN Data.

Method	Covariates	Est.	95% CI.	S.D.
Regression	BMI, HbA1c	0.64	(0.58,0.70)	0.03
Standard IV[1]	–	-2.28	(-3.22,-1.34)	0.47
Standard IV^{CC2}	BMI	0.56	(0.26,0.86)	0.15
Standard IV^{CC3}	BMI, HbA1c	0.34	(0.06,0.62)	0.14
Standard IV^{MI}	BMI, HbA1c	-0.01	(-0.75,0.73)	0.37
Standard IV^{IPW}	BMI, HbA1c	0.60	(0.32,0.88)	0.14
Proposed	BMI, HbA1c	1.03	(0.15,1.91)	0.44

Summary

- We have proposed a method to handle a specific type of non-ignorable missingness in studies using provider preference as an IV.
- Standard IV methods are biased for this setting.
- Simulation studies show that the proposed method performs well under its assumptions and the method is robust to some violations of one of its assumptions, A.1.

Acknowledgments

This work was supported in part by NSF grant SES-1260782 and CPeRT.

Thank you!

dsmall@wharton.upenn.edu