

# Statistical Analysis with Missing Data Module 11

Missing data in clinical  
trials and other problems



# Outline

- Summarize findings of NRC study on treatment of missing data in clinical trials
- Analysis principles and methods
- Choice of estimand
- Sensitivity analysis

# Outline

- Summarize findings of NRC study on treatment of missing data in clinical trials
- Analysis principles and methods
- Choice of estimand
- Sensitivity analysis

# Defining Missing Data 1

- Missing data are unrecorded values that, if recorded, would be meaningful for analysis.
- Outcomes that are not defined for some participants are not considered by the panel as missing data
  - See e.g. Little and Rubin (2002)

# Key Take-Home Messages 1

- Missing data undermines randomization, the lynchpin of inferences in confirmatory trials
- Limiting missing data should be a major consideration when weighing **alternative study designs**
  - Analysis methods come with unverifiable assumptions, and limiting these assumptions is crucial

# Key Take-Home Messages 2

- Careful attention to avoiding missing data in **trial conduct** can greatly limit the scope of the problem
- **Analysis methods** need to be driven by plausible scientific assumptions
- With substantial missing data, **sensitivity analyses** to assess robustness to alternative analysis models are needed
  - difficulties in specifying these analyses reinforces the need to limit missing data in trial design and conduct

# Trial Conduct Strategies to Reduce Missing Data

- Limit participant burden
  - Reduce the number of visits and assessments
  - Allow a relatively large time window for each follow-up assessment
- Set maximal acceptable rates of missing data, and monitor during the trial
- Provide incentives for investigators and participants to stay in the try, subject to ethical guidelines
- Continuous update of contact information
- Educate study staff on importance of limiting missing data

# Design to reduce the occurrence of missingness

1. Run-in periods before randomization to identify who can tolerate or respond to the study treatment
2. Flexible-dose (titration) studies
3. Restrict trial to target population for whom treatment is indicated
4. Reduce length of follow-up period
5. Allow rescue medication in the event of poor response
6. Define outcomes that can be ascertained in a high proportion of participants

Benefits of these options need to be weighed against costs




# Outline

- Summarize findings of NRC study on treatment of missing data in clinical trials
- **Analysis principles and methods**
- Choice of estimand
- Sensitivity analysis

# Analysis Methods: Principles

1. Missing data: missingness hides a true underlying value that is meaningful for analysis
2. Formulate the analysis for inference about an appropriate and well-defined causal estimand
3. Document, to the degree possible, the reasons for missing data, and incorporate in the analysis  
Some may be MAR, others not
4. Decide on a defensible primary set of assumptions about the missing data mechanism
5. Conduct a statistically valid analysis under the primary missing data assumptions
6. Assess the robustness of the treatment effect inferences by prespecified sensitivity analyses.

# Some missing-data analysis methods

- Complete-case analysis
  - Single imputation methods, including LOCF, BOCF
  - Inverse probability-weighted methods, simple and augmented
  - Likelihood – based methods
    - Maximum likelihood, Bayes, Multiple imputation
- 
- Preferred methods

# Complete-case analysis

Deletion of cases with missing data

- Generally inappropriate for a regulatory setting.
- Essentially requires MCAR within treatment groups – very strong assumption.
- Furthermore, loss of efficiency from discarding of information.
- Methods based on MAR are better, since they make use of the information from the incomplete data to reduce bias from deviations from MCAR.

# Single imputation: LOCF

Generally not recommended by the panel.

LOCF is based on the strong assumption that the outcome of a participant does not change after dropout.

LOCF is mistakenly considered to be valid under MCAR or MAR, but it is generally only valid under the above (MNAR) assumption.

Sometimes justified as being conservative, but this is not necessarily true either.

Even if scientifically reasonable (unlikely), it does not propagate uncertainty and so fails in general to provide valid tests and confidence intervals.

Similar comments can be made about BOCF.

# Inverse Probability Weighting (IPW)

- Assign a missingness weight to the complete cases to make them more representative of all cases.
- Fit a model for  $\pi(X, V, \theta) = P_{\theta}(R = 1 | X, V)$
- Estimate the mean of Y using the weighted average:

$$\hat{\mu} = (1/n) \sum_i \frac{R_i Y_i}{\pi(X_i, V_i, \hat{\theta})}$$

- Standard errors can be estimated analytically or using the bootstrap.
- A useful refinement is augmented IPW: create predictions from a model, and add IPW residuals for robustness to model misspecification. More efficient, double robustness property

# Multiple Imputation (MI)

- Create multiple filled-in data sets using draws from predictive distribution, and apply MI combining rules
- An important advantage is that the imputation model and analysis model can differ
  - In particular, auxiliary variables  $V$  that are not included in the final analysis model can be used in the imputation model, weakening the MAR assumption.
- Other advantages and disadvantages:
  - Software available
  - MI propagates imputation uncertainty, and is more efficient than single imputation of draws
  - Imputations (but not necessarily the complete-data inference) rely on parametric assumptions.
  - Marked incompatibility between the data model and imputation model may be a concern.

# Defining Missing Data 2

- Clinical trials record and compare *treatment-specific* outcomes. So it is important to distinguish between
  - *Treatment discontinuation*: Treatment-specific outcomes are not recorded when participants discontinue their assigned treatments (lack of efficacy, lack of tolerability,...)
  - *Analysis dropouts*: Missing data arising from inability to record outcomes, e.g. from missed clinic visits, attrition. Participants may or may not be off protocol
  - Important issue: should we attempt to record outcomes after discontinuation? Depends on setting, choice of estimand



# Analysis dropouts

- Individuals do not discontinue assigned treatment, but outcome values are missing
  - E.g. administrative censoring, missed clinical visits
- This is like a standard missing data problem
- Usual methods and concepts apply
  - MAR vs MNAR models
  - Likelihood or Augmented IPW methods
  - Sensitivity analysis

# Treatment discontinuation

- Individuals discontinue assigned treatment for reasons potentially associated with that treatment
  - Side effects, ineffectiveness, advice of physician
- Outcome data may continue to be recorded
  - But might be outcomes under a treatment different from the treatment assigned
- Can view treatment discontinuation as a form of noncompliance, and invoke the causal literature on noncompliance

# Complier-average causal effect (CACE)

- The CACE measures the treatment effect in the subgroup of principal compliers -- individuals who would comply under either treatment (e.g. Angrist, Imbens and Rubin 1996, Little, Long and Lin 2006)
- Special case of principal stratification (Frangakis and Rubin 2002)
- Need to “impute” compliance under treatment(s) not assigned
- Similarly we can define “Completer-average causal effect”, and apply instrumental variable methods
- My focus here on “Intention to treat” estimands, which apply to the whole population that is randomized

# Outline

- Summarize findings of NRC study on treatment of missing data in clinical trials
- Analysis principles and methods
- **Choice of estimand**
- Sensitivity analysis

# Trial Outcomes and Estimands

- The choice of estimand involves the outcome measure of interest, the relevant population under study, and the period of measurement, and is a key starting point for the design of a clinical trial
- Alternative choices of estimand may have important implications for trial design and implementation and on the rate of missingness
- For instance, the estimand tells one whether or not it is important to collect outcome data after treatment discontinuation

# Examples of Estimands

Treatment difference in average outcome improvement (change from baseline, area under curve)...

1) In all randomized participants.

2) In tolerators to treatment.

3) If all subjects tolerated and adhered to treatment.

4) During adherence to treatment

4) is an on-treatment summary – avoids need to impute outcomes after dropout (Kang and Little 2015)

# Example: ATLAS ACS 2 TIMI 51 Trial

- Large clinical trial that assessed Rivaroxaban for its ability to reduce the risk of cardiovascular death, myocardial infarction or stroke in patients with acute coronary syndrome (ACS) (Mega et al. 2012)
- 15,526 patients randomized into three treatment groups: rivaroxaban 2.5 mg b.i.d., rivaroxaban 5 mg b.i.d and placebo.
- Primary analysis: Cox proportional hazards model
- Study showed a statistically significant reduction in the primary efficacy outcome: the composite of cardiovascular (CV) death, myocardial infarction (MI) and stroke for the combined rivaroxaban doses compared to placebo (Hazard Ratio (HR) and 95% CI 0.84 (0.74-0.96))

# ATLAS Trial

- There were concerns about 5-10% who dropped out prior to final endpoint – what if dropouts had worse than expected outcomes (informative censoring) that biased the treatment comparison (differential informative censoring)?
- Two estimands: time to primary outcome for
- intent-to treat (ITT)
  - included all events occurring up until the end of study.
- modified intent-to-treat (mITT, primary)
  - events of all randomized participants up to the earlier of: (a) the end of study, (b) 30 days after the last study treatment, or (c) 30 days after randomization for those who had not received any study medication.
- Modified intent-to-treat has much less missing data



# Example: Insulin Trial

- Eli Lilly study of a new oral anti-hyperglycemic medication for patients with type 2 diabetes (T2DM), compared to the standard therapies based on injected insulin Glargine.
- Randomized, parallel-group study of individuals experiencing lack of control of glucose levels, as measured by the HbA1c laboratory test (low is good).
- 3 treatment arms:
  - New (n = 221)
  - Glargine (IG, n = 213)
  - Combined (new and IG) arm (n = 115)

# Lilly T2DM Study

- Measures of HbA1c were obtained at baseline and at weeks 1, 2, 4, 6, 8, 10, 12 and 24.
- Primary analysis in protocol is by ITT, change in HbA1c from baseline to 24 weeks, including all randomized patients with a baseline and at least one follow-up measure after baseline.
- Rescue medications were allowed for IG and New treatments, and HbA1C values while on rescue medications were included
- Thus, “treatment” was in fact a “treatment protocol” which includes any effects of the rescue medications
- No measures after subsequent treatment discontinuation: missing data imputed by last observation carried forward.

## Example. Lilly T2DM Trial: Discontinuation and Missing Data in Study Groups by Reason

	Type	Combined	IG	New
0	completed	25	48	42
1	subject	10	12	33
2	physician	5	3	7
3	Protocol viol	1	4	3
4	Adverse event	2	0	4
5	Death	1	0	1
6	Sponsor terminated	68	146	119
7	Lost follow-up	3	10	13

1-4: treatment discontinuation + missing data

6-7: missing data

5: not missing data, since dead people do not have Hba1c levels

## Example. Lilly T2DM Trial: Discontinuation and Missing Data in Study Groups by Reason

	Type	Combined	IG no rescue	IG rescue	New no rescue	New rescue
0	completed	25	47	1	24	18
1	subject	10	12	0	30	3
2	physician	5	3	0	6	1
3	Protocol viol	1	4	0	3	0
4	Adverse event	2	0	0	4	0
5	Death	1	0	0	1	0
6	Sponsor terminated	68	137	9	94	25
7	Lost follow-up	3	10	0	13	0

Rescue for “IG” is minor, rescue for “New” is sizeable: should rescue values be included?

# Treatment Discontinuation: ITT options

- If (as here) the primary estimand involves measures after discontinuation, methods need to in effect impute (explicitly or implicitly, e.g. by weighting) these measures
- Imputation method needs to be appropriate for the estimand, which involves assumptions about treatments after discontinuation
- The standard ITT approach measures the effect of randomization to treatment  $Y(t_{\text{true}})$ , where  $t_{\text{true}}$  is set of actual treatments received after discontinuation
  - measured if follow-up measures are obtained after treatment
  - includes effects of any treatments between discontinuation and end of study (so these should be specified in the protocol)
  - But are we interested in including effects of these in the analysis?

# Treatment Discontinuation: ITT options

We could conceive of other estimands, for the ITT population:

- Estimand under Assigned Treatment  $Y(t_a)$  estimand if discontinuers had continued to take the assigned treatment
- Estimand under Control Treatment  $Y(t_c)$ : estimand if discontinuers had taken control (or reference) treatment
  - Latter two are counterfactual and need to be imputed (see Little and Yau 1996, Ratitch et al., 2013)
  - In particular,  $Y(t_c)$  is estimated by methods that impute by jumping to reference after discontinuation. Whether these alternatives make sense varies according to context
- In most protocols, the method of imputation is described without stating the estimand

# Imputation for Lilly T2DM Example

- The Lilly protocol specified LOCF imputation
- Which of the outcomes  $Y(t_{\text{true}})$ ,  $Y(t_a)$  or  $Y(t_c)$ , is being imputed by LOCF? Here, as in many trials, this is left unspecified and unclear. This clouds the nature of the treatment
- All of these estimands are problematic in the new treatment arm:
- $Y(t_{\text{true}})$  and  $Y(t_c)$ , both include the effects of Glargine administered after drop-out, not the new treatment
- $Y(t_a)$  is counter-factual, and there are no data to impute this since, for safety of participants, HbA1c levels need to be brought under control
- Suggest that there is a better alternative: on-treatment summaries

# On-treatment summaries

- A measure of the effectiveness of a treatment that only uses information while individuals are on the assigned treatment. Examples:
  - Dropout as failure. Define a binary measure for success or failure, and treat discontinuers as failures
  - Area under curve (measured relative to baseline value) while on treatment. Dropout is penalized in that area is restricted to time while on assigned treatment.
  - Change from baseline to min(dropout, end of study). This estimate is the same change from baseline to end with LOCF for dropouts, but it avoids (unreasonable) assumption of no change after dropout
  - Impute zero change for dropouts. Estimate same as BOCF.



# On-treatment summary for Lilly study

- The protocol specified LOCF imputation, equivalent to the on-treatment summary: change in HbA1c levels between baseline and min(dropout, end of study)
- A problem with this outcome is that it does not penalize dropout prior to end of trial (though it does reflect tendency for higher HbA1c values at time of discontinuation)
- On the other hand BOCF corresponds to imputing zero change, which voids any benefit of treatment before dropout.
- An alternative on-treatment summary measure is
- *$P = \text{proportion of 24 weeks where individual was on treatment and HbA1c levels were under control}$*
- This penalizes early discontinuation appropriately
- Other more quantitative summaries might be developed

# Lilly T2DM Study: administrative censoring

- Analysis of  $P$  still requires imputation for cases incomplete because of administrative censoring. A (better?) alternative to LOCF is:
- (a) Impute discontinuation indicator and (if 1) time of discontinuation for censored cases, given their history up to censoring time
- (b) Impute  $P$  given the discontinuation time imputed in (a)
- Repeat (a) and (b) and apply MI combining rules to propagate imputation uncertainty
- Since the administrative censoring is plausibly missing at random – unrelated to individual outcome measures – this is a defensible approach, and a sensitivity analysis for deviations from MAR seems unnecessary.

# Three ITT ANCOVA Analyses of Diabetes Data

	A. Outcome = Change from Baseline to Week 52		B. Outcome = Transformed Proportion of 52 weeks when on Treatment and HbA1c≤7.5%			
	All Types of Missing Data Treated by LOCF Imputation		B1. Admin Censoring or Loss to Follow Up Treated by MI		B2. Admin Censoring or Loss to Follow Up Treated by LOCF	
Regressor	Estimate (95% CI)	P-Value	Estimate (95% CI)	P-Value	Estimate (95% CI)	P-Value
Intercept	-0.50 (-0.83,-0.17)	0.003	0.48 (0.38,0.59)	<0.001	0.52 (0.41,0.63)	<0.001
“Inhaled”-“IG”	-0.18 (-0.38,0.03)	0.090	0.08 (-.01,0.17)	0.068	0.09 (-.01,0.18)	0.067
“Inhaled+IG”-“IG”	-0.38 (-0.63,0.14)	0.002	0.12 (0.02,0.23)	0.021	0.14 (0.02,0.25)	0.017
Taking Insulin Secretagogue	-0.08 (-0.28,-0.13)	0.847	-0.03 (-0.12,0.06)	0.478	-0.05 (-.14,0.05)	0.333
Baseline (centered to 0)	-0.47 (-0.57,-0.37)	<0.001	-0.24 (-0.28,-0.20)	<0.001	-0.26 (-0.30,-0.21)	<0.001
Country (DF=10)	---	0.585		0.158		0.009

# Reasons to Collect Data After Discontinuation

- If the usual ITT estimand, involving  $Y(t_{\text{true}})$ , is appropriate, then it is better to follow up dropouts rather than impute values.
- For other ITT estimands, the value of measure is less clear...
- Follow up is not needed for an on-treatment summary measure (though follow-up might still be important to monitor side effects)
  - If sponsor and regulator can agree on a suitable on-treatment summary, the result might make both parties happy!

# Summary

- Follow up after discontinuation allows side effects after discontinuation to be monitored
- It is also indicated if the measures after discontinuation are needed for the primary outcome and relevant to the treatment under study – which is more likely if treatments after discontinuation are specified in the protocol as part of a treatment regimen
- Defining an on-treatment summary to measure of treatment effect deserves more consideration –
  - does not require follow-up measures discontinuation
  - limits the amount of missing data.

# Outline

- Summarize findings of NRC study on treatment of missing data in clinical trials
- Analysis principles and methods
- Choice of estimand
- **Sensitivity analysis**

# Sensitivity Analysis

- Parameters of MNAR models cannot be reliably estimated – identifiability requires structural assumptions that are often questionable
- Varying certain parameters in a sensitivity analysis is the preferred approach
- In many (not all) situations, it would be reasonable to choose an MAR primary model, and look at MNAR models via a sensitivity analysis to assess plausible deviations from MAR

# Simple example

Consider first two treatments  $T = 1, 2$ , single outcome  $Y$ , no auxiliary data

Let  $R = 1$  if  $Y$  is observed,  $R = 0$  otherwise

Problem is to estimate mean in each treatment arm based on data from subjects in each arm, where some of the  $Y$ 's are missing.

MAR would assume that within each treatment arm, the distribution of  $Y$  for respondents was the same as for nonrespondents.



# Simple example

For treatment  $t$ , let

$$\mu_{1t} = E(Y \mid T = t, R = 1), \mu_{0t} = E(Y \mid T = t, R = 0)$$

Goal is to estimate  $\mu_t = \pi_t \mu_{t1} + (1 - \pi_t) \mu_{t0}$

There is information about respondent means

$\{\mu_{t1}\}$  and response rates  $\{\pi_t\}$

No information about nonrespondent means  $\{\mu_{t0}\}$

# Simple pattern-mixture model for sensitivity analysis

$$[Y \mid R = 1, T = j] \sim N(\mu_{1j}, \sigma_j^2)$$

$$[Y \mid R = 0, T = j] \sim N(\mu_{1j} + \Delta_j, \sigma_j^2)$$

$$\Pr(R = 1 \mid T = j) = \pi_j$$

$$\text{MAR: } \Delta_1 = \Delta_2 = 0$$

MNAR: For prespecified plausible values of  $\Delta_1, \Delta_2$ :

generate inferences about  $\mu_1, \mu_2$  and  $\mu_2 - \mu_1$

by replacing  $\mu_{1j}, \sigma_j^2$  and  $\pi_j$  by sample estimates,  
with associated standard errors.

Examples of choices of  $\Delta_1, \Delta_2$  :

$$\Delta_j = k\sigma_j, \quad k = 0.2, 0.5$$

# Choosing sensitivity parameters

Setting  $\Delta_j = k\sigma_j$  for fixed  $k$  (e.g.  $k = 0.2, 0.5$ ) "rewards"  
strong covariates by a reduction in size of  $\sigma_j, \Delta_j$

To "penalize" dropout in treatment group, might set

$\Delta_1 = 0$  for control group  $T = 1$

$\Delta_2 > 0$  for treatment group  $T = 2$

"Tipping point" = value of  $\Delta_2$  where statistical significance  
of treatment effect is lost

If tipping point is unrealistically high, treatment effect is robust

# Sensitivity analysis for selection models

- For selection models, a sensitivity analysis can be conducted by making an assumption about how the odds of nonresponse change with the values of the outcome  $Y$ .
- For example, one can assume that the log odds of nonresponse differs by  $\alpha$  for those who differ by one unit on  $Y$ , that is

$$\text{logit} \{P[R = 0 \mid V, Y = y]\} = h(V) + \alpha y.$$

This is a selection model since it models the probability of nonresponse as a function of the outcome and auxiliary variables  $V$ .

# Sensitivity analysis for selection models

Adopting a value of  $\alpha$  is equivalent to adopting a known link between the distribution of the respondents and the nonrespondents.

A sensitivity analysis consists of repeating the inference for  $\mu$  at different plausible values of  $\alpha$  so as to assess the sensitivity of inferences about  $\mu$  to deviations from MAR.

I prefer the pattern-mixture approach since it is easier to implement and explain to clinicians, and  $\alpha$  has a complex interpretation

# Application: ATLAS ACS 2 TIMI 51 Trial

- Large clinical trial that assessed Rivaroxaban for its ability to reduce the risk of cardiovascular death, myocardial infarction or stroke in patients with acute coronary syndrome (ACS) (Mega et al. 2012)
- 15,526 patients randomized into three treatment groups: rivaroxaban 2.5 mg b.i.d., rivaroxaban 5 mg b.i.d and placebo.
- Primary analysis: Cox proportional hazards model
- Study showed a statistically significant reduction in the primary efficacy outcome: the composite of cardiovascular (CV) death, myocardial infarction (MI) and stroke for the combined rivaroxaban doses compared to placebo (Hazard Ratio (HR) and 95% CI 0.84 (0.74-0.96))

# Sensitivity analysis

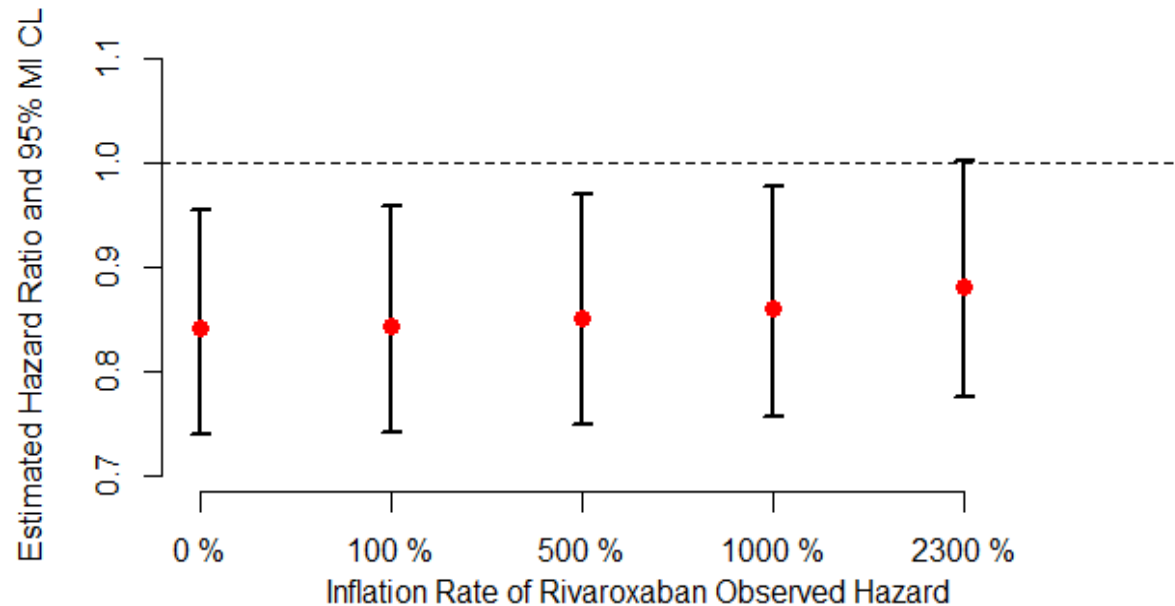
- There were concerns about 5-10% who dropped out prior to final endpoint – what if dropouts had worse than expected outcomes (informative censoring) that biased the treatment comparison (differential informative censoring)?
- Sensitivity analysis was applied to assess the impact of deviations from non-informative censoring on two key analyses:
- intent-to treat (ITT)
  - included all events occurring up until the end of study.
- modified intent-to-treat (mITT, primary)
  - events of all randomized participants up to the earlier of: (a) the end of study, (b) 30 days after the last study treatment, or (c) 30 days after randomization for those who had not received any study medication.

# Overview of method

- Estimate hazard for each dropout at time of dropout, under non-informative censoring
- Differentially increase the hazard of the primary outcome in the rivaroxaban treatment groups,
- Multiply-impute events between drop-out and the end of the study, assuming Weibull distribution
- Combine results using MI combining rules
- Tipping point F: increase in hazard at which significance is lost
  - (Little, Wang, Sun et al. 2015)



Hazard Ratio and 95% confidence interval for combined rivaroxaban vs. placebo, mITT analysis of primary outcome. Sensitivity analysis, inflating the individually estimated hazard in the rivaroxaban groups by known factors. Tipping Point = 2300%



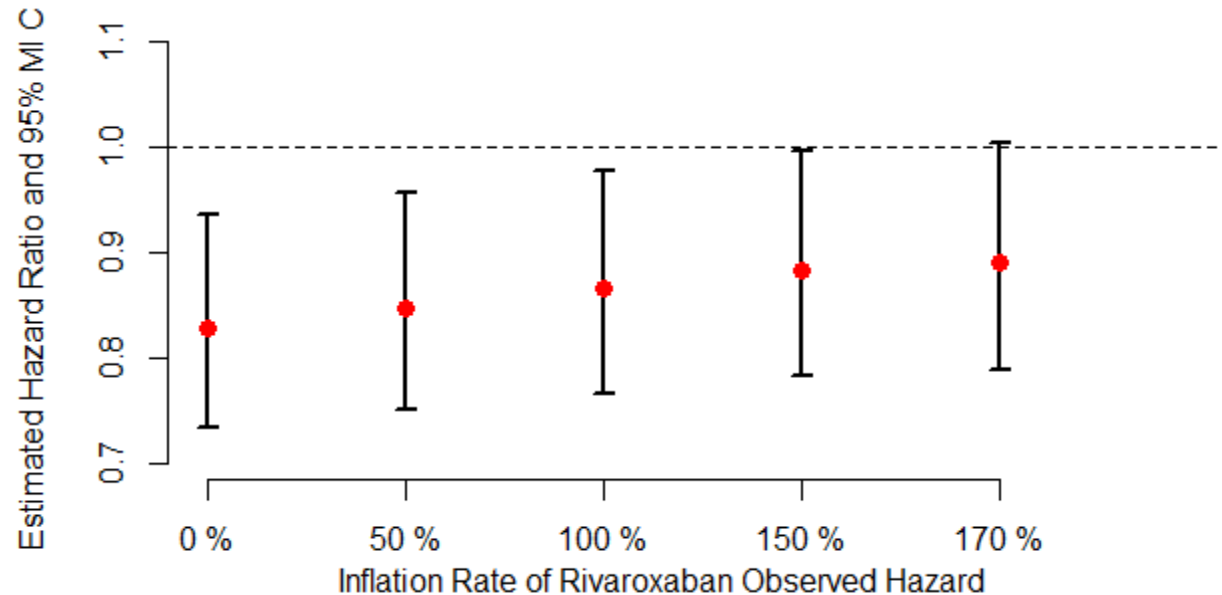
Mean Number of Imputed Events:

Rivar	3	4	10	17	34
Placebo	2	2	2	2	2

Missing Data in Clinical Trials

Hazard Ratio and 95% confidence interval for combined rivaroxaban vs. Placebo, ITT analysis of primary outcome.

Sensitivity analysis, inflating the individually estimated hazard in the rivaroxaban groups by known factors. Tipping point = 160%.



Mean Number of Imputed Events:

Rivar	41	58	73	88	94
Placebo	21	21	21	21	21

# Summary of sensitivity analysis

- Sensitivity analysis is a scientific way of attempting to reflect uncertainty arising from potentially MNAR missing data
- Deciding on how to implement and interpret a sensitivity analysis in the regulatory setting is challenging
- The need and importance of sensitivity analysis increases with the amount of potentially MNAR missing data
- This reinforces the need to limit missing data in the design and implementation stage
  - Avoiding substantial amounts of missing data is key!

# Sensitivity Analysis

- Parameters of MNAR models cannot be reliably estimated – identifiability requires structural assumptions that are often questionable
- Varying certain parameters in a sensitivity analysis is the preferred approach
- In many (not all) situations, it would be reasonable to choose an MAR primary model, and look at MNAR models via a sensitivity analysis to assess plausible deviations from MAR
- Xiang Sun presents such a sensitivity analysis in her talk at this workshop

# Other problems as missing data

- 1. Measurement Error
- 2. Combining Information from Multiple Data Sources
- 3. Bayesian inference for finite population
- 4. Disclosure limitation
- 5. Causal inference

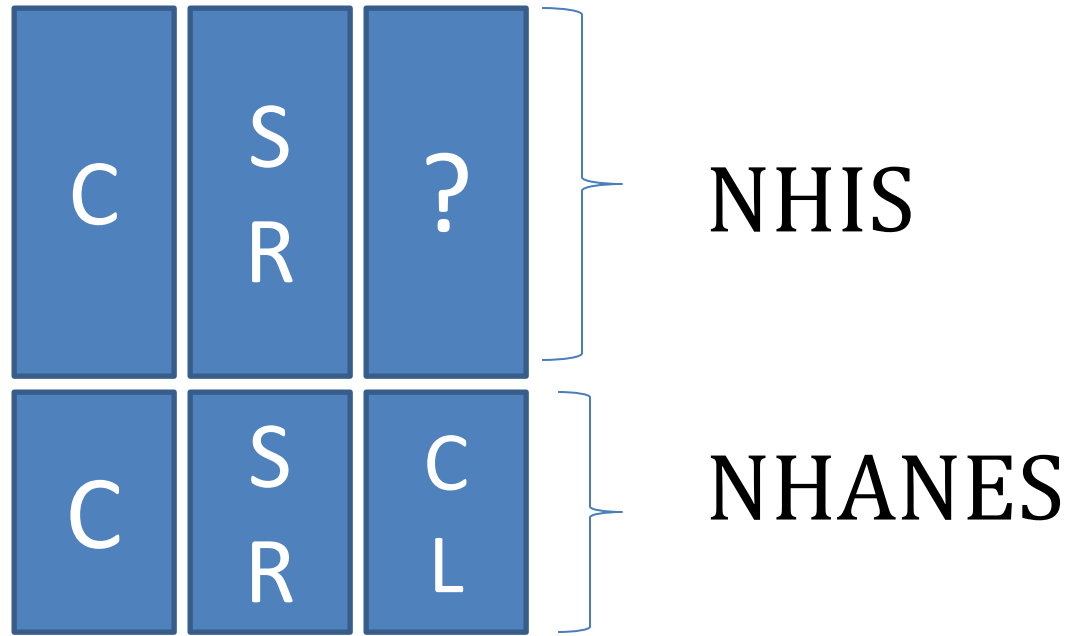
# 1&2: Combining NHIS and NHANES data

- Combining information from an examination survey and an interview survey to improve on analyses of self-reported data (Raghu)
  - National Health Interview Survey collects self-report (SR) disease conditions
    - May underestimate prevalence rate
    - Bias regression coefficients
  - National Health and Nutritional Examination Survey collects clinical (CL) as well as self-report (SR) disease conditions
  - Both surveys have some core common items

# Results (NHANES)

Group	Hypertension		Diabetes		Obesity	
Education	SR	CL	SR	CL	SR	CL
< HS	29.5	38.1	12.4	16.2	28.2	31.2
HS	24.5	31.8	6.3	9.2	29.2	32.2
> HS	18.3	24.0	4.2	6.1	22.7	26.8
Race/Ethnicity						
White	14.1	20.5	8.5	10.6	26.6	29.8
Black	30.9	38.7	9.2	12.3	33.7	37.1
Hispanic	22.3	28.8	5.6	8.2	24.3	28.0

# SETUP



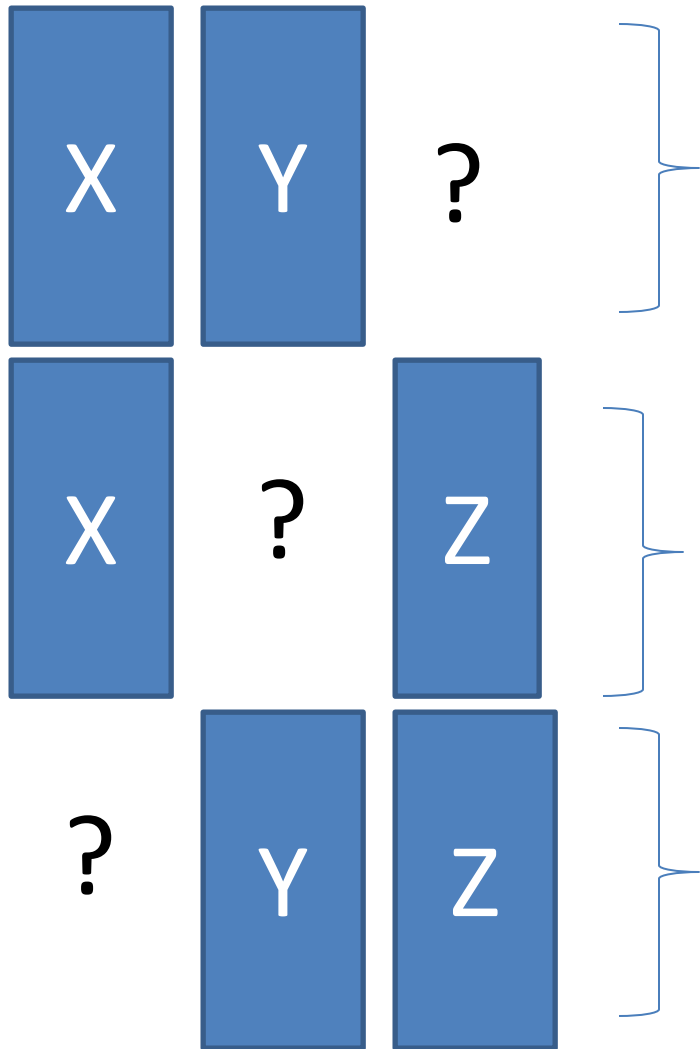
C = covariates, SR = Self Report, CL = Clinical  
Multiply Impute “?” after appending the data from  
the two surveys



# Results NHIS (Before and after imputation)

Group	Hypertension		Diabetes		Obesity	
Education	SR	MICL	SR	MICL	SR	MICL
< HS	30.9	39.5	11.1	14.2	25.7	30.1
HS	22.9	30.1	6.6	8.8	23.5	28.1
> HS	16.5	22.8	4.2	6.5	18.7	23.1
Race/Ethnicity						
White	14.1	20.8	6.9	9.7	23.2	28.2
Black	26.7	35.1	8.8	11.3	29.9	34.8
Hispanic	20.8	27.6	5.6	7.9	19.8	23.1

# Extension-COMBINING INFORMATION



Epidemiologic study 1:  
Relationship between  
Disease Y and Biological risk  
factor X

Omnibus survey NHANES:  
Data on Biological (X) and  
Social (Z) risk factors

Epidemiologic study 2:  
Relationship between  
Disease Y and Social risk  
factor Z

# Extension-2

Study	Disease	Social Factors	Genetic Markers	Biological Factors
1	X	X		X
2	X		X	X
3		X		X
4			X	X
New Study				

# 3. Survey Inference

$Y = (Y_1, \dots, Y_N)$  = population values

$Q = Q(Y)$  = finite population quantity

$I = (I_1, \dots, I_N)$  = Sample Inclusion Indicators

$I_i = \begin{cases} 1, & \text{unit included in sample} \\ 0, & \text{otherwise} \end{cases}$

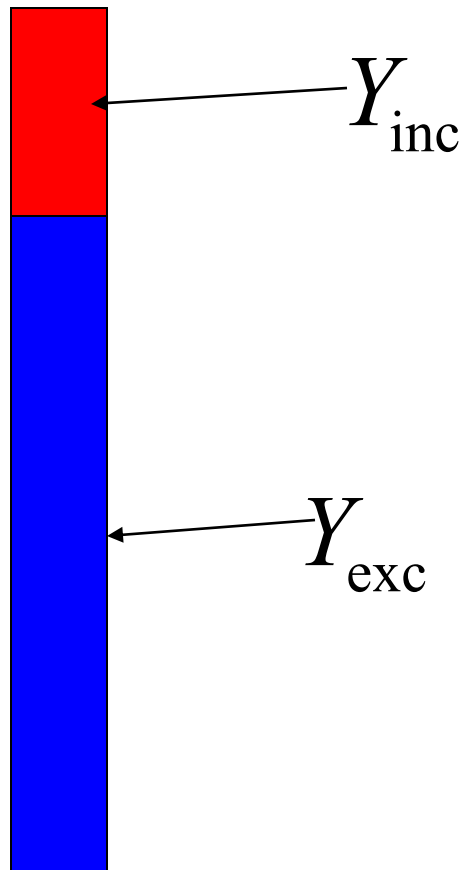
$Y_{\text{inc}}$  = part of  $Y$  included in the survey

$p(Y)$  = prior distribution for  $Y$

$p(Q(Y) | Y_{\text{inc}})$  = posterior predictive distribution of  $Q$  given  $Y_{\text{inc}}$

# Schematic Display

- Use model to “fill-in” nonsampled values in the population



Model :

$$\Pr(Y_{\text{inc}}, Y_{\text{exc}})$$

$$= \int \Pr(Y_{\text{inc}}, Y_{\text{exc}} \mid \theta) \Pr(\theta) d\theta$$

$$= \int \left( \prod_{i=1}^N \Pr(Y_i \mid \theta) \right) \Pr(\theta) d\theta$$

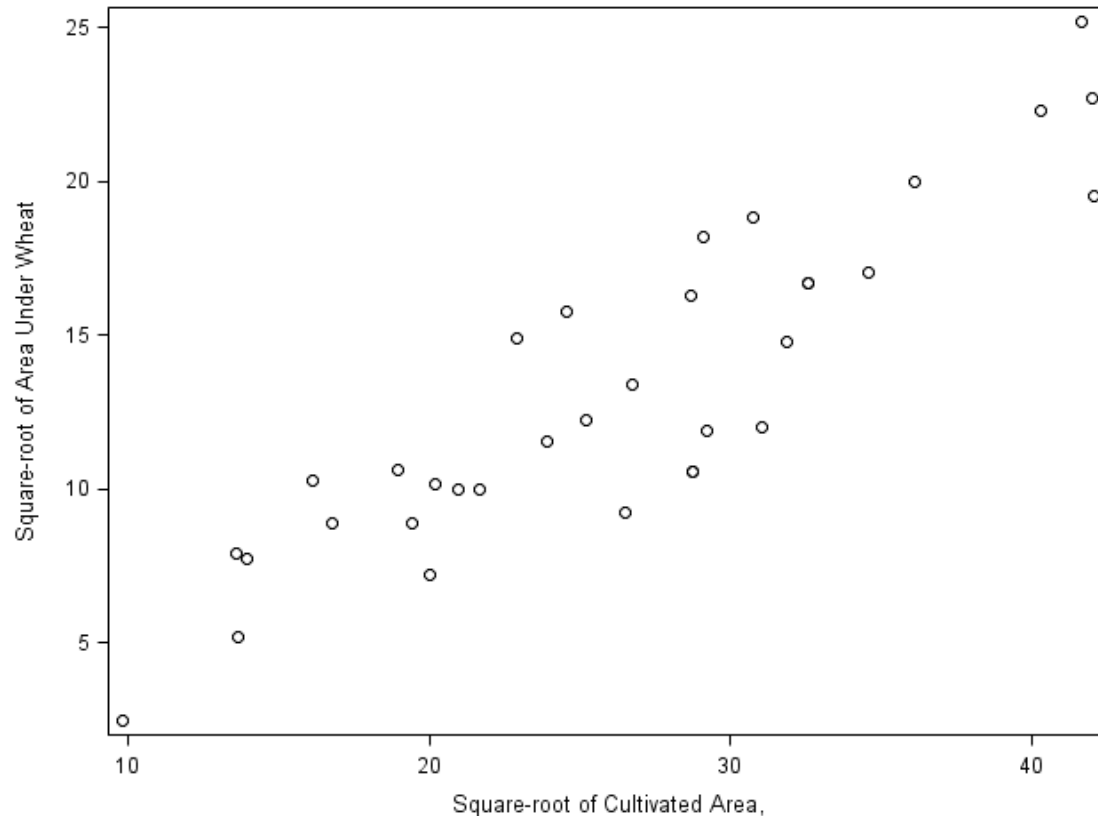
$$\Pr(Y_{\text{exc}} \mid Y_{\text{inc}} = y_{\text{inc}})$$

$$\Pr(Q(Y_{\text{exc}}, y_{\text{inc}}) \mid Y_{\text{inc}} = y_{\text{inc}})$$

# Example

- Population: 170 villages in Lucknow subdivision in India
- Sample: Probability proportional to size (cultivated area, known for all 170 villages)
- Sample size: 34 villages
- Outcome: Area under wheat production
- Goal: Infer about population total area under wheat
- Inference approach: Multiply impute area under wheat production for the 136 nonsampled villages

# Scatter Plot & Model



$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$$\varepsilon \mid X \sim N(0, \sigma^2)$$

$$\Pr(\beta_0, \beta_1, \sigma) \propto \sigma^{-1}$$

$Y$  = Square-root of Area Under Wheat

$X$  = Square-root of Size

# Inference

- Multiply imputed missing values for 136 villages using IVEware
- Compute the population total from each filled-in population
- Note that Within-variance is 0 as the entire population is filled-in
- Approximate posterior distribution

$$(Q - \bar{Q}_M) / \sqrt{(1 + M^{-1})B_M} \mid Data \sim t_{M-1}$$

$$B_M = \sum_{l=1}^M (Q_l - \bar{Q}_M)^2 / (M - 1)$$

$$\bar{Q}_M = \sum_l Q_l / M$$



# Measuring departures from prob sampling

- Little et al. (2020) derive measures based on proxy pattern-mixture model for nonresponse in Andridge and Little (2011)

$$(X, Y \mid S = j, U) \sim N_2 \left( (\mu_X^{(j)}, \mu_Y^{(j)}), \begin{pmatrix} \sigma_{XX}^{(j)} & \sigma_{XY}^{(j)} \\ \sigma_{XY}^{(j)} & \sigma_{YY}^{(j)} \end{pmatrix} \right)$$

$$\Pr(S = j) = g \left( ((1 - \phi)X^* + \phi Y), U \right)$$

$X^*$  = best auxiliary proxy for  $Y$

$U$  = other variables in  $Z$  orthogonal to  $X^*$

- $\Pr(S = 1)$  is allowed to depend on both  $X^*$  and  $Y$
- If  $\phi = 0$ , then selection is SAR
- If  $\phi = 1$ , then selection depends only on  $Y$  only
- The  $\phi$  parameter (unknown) enables **sensitivity analysis**

# Approach, cont'd

$$\hat{\mu}_Y(\phi) = \bar{y}_n + \frac{\phi + (1-\phi)r_{XY}}{\phi r_{XY} + (1-\phi)} \sqrt{\frac{s_{YY}}{s_{XX}}} (\bar{X}_N - \bar{x}_n)$$

- This leads to a **standardized measure of unadjusted bias (SMUB)**:

$$\text{SMUB}(\phi) = \frac{\phi + (1-\phi)r_{XY}}{\phi r_{XY} + 1 - \phi} \frac{(\bar{x}_n - \bar{X}_N)}{\sqrt{s_{XX}}}$$

and standardized measure of adjusted bias (SMAB), which assesses bias from deviations from SAR:

$$\text{SMAB}(\phi) = \text{SMUB}(\phi) - \text{SMUB}(0) = \frac{\phi(1-r_{XY}^2)}{\phi r_{XY} + 1 - \phi} \frac{(\bar{x}_n - \bar{X}_N)}{\sqrt{s_{XX}}}$$

(But this is strongly tied to assumptions of the model)

# SMUB

- Proposed index is **simple**: it only depends on  $\phi$ ; means, standard deviations, and correlations from the observed non-probability sample; and the population mean for  $X$

$$\text{SMUB}(0) = r_{XY} \frac{(\bar{x}_n - \bar{X}_N)}{\sqrt{s_{XX}}} \quad (\text{Same as regression measure})$$

$$\text{SMUB}(0.5) = \frac{(\bar{x}_n - \bar{X}_N)}{\sqrt{s_{XX}}} \quad (\text{Treats Best Proxy as } Y)$$

$$\text{SMUB}(1) = \frac{1}{r_{XY}} \frac{(\bar{x}_n - \bar{X}_N)}{\sqrt{s_{XX}}}$$

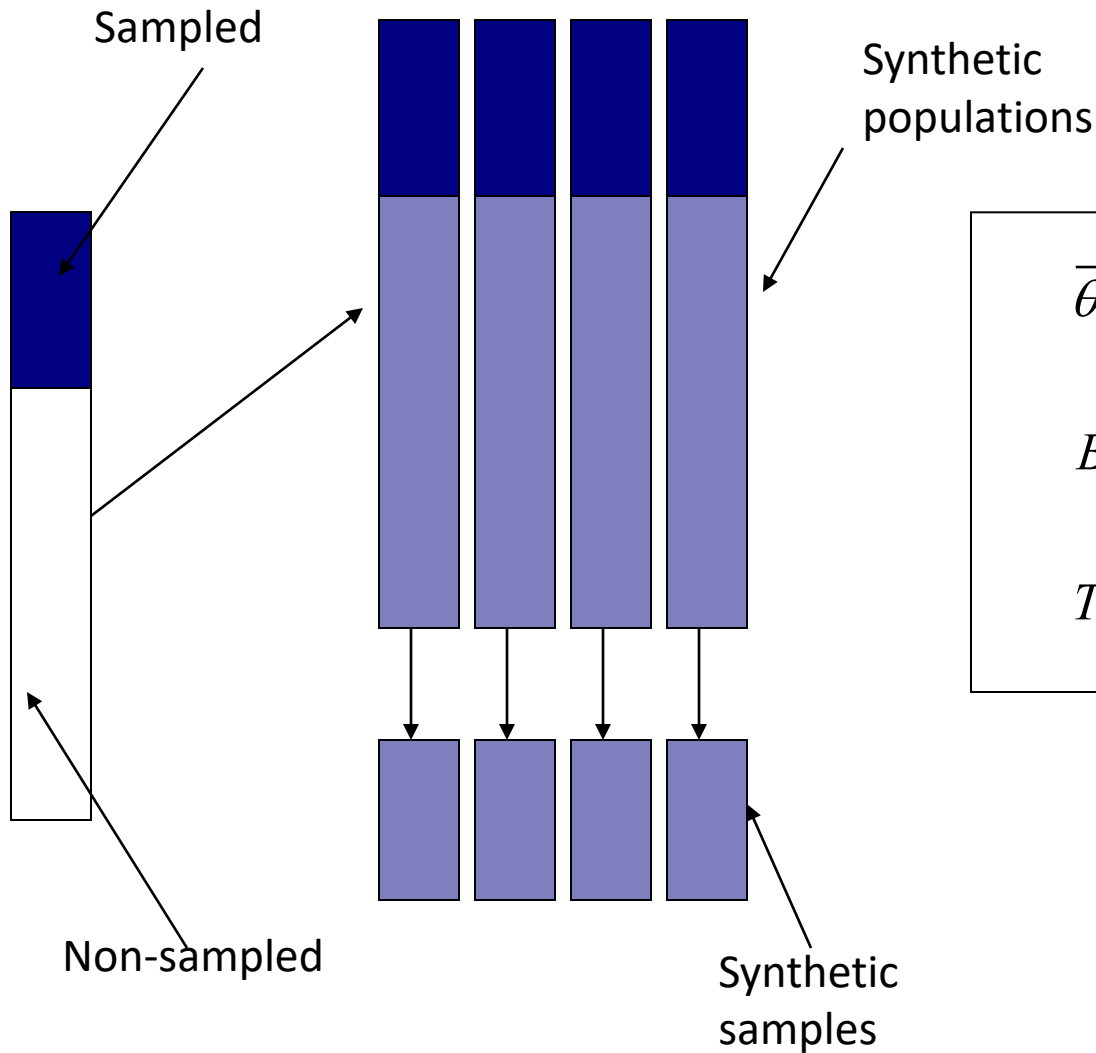
# Additional Remarks on SMUB

- SMUB(1) will become quite unstable when  $X$  is not a good predictor of  $Y$ ; in this case, bias cannot be reliably estimated
- SMUB(0.5) is similar to the Bias Effect Size proposed by Biemer and Peytchev at the 2011 Nonresponse Workshop
- We have also implemented a fully Bayesian approach to computing the index that incorporates uncertainty about all of the input parameters; the proposed interval is a recommendation for practice, but we have an R function available for the Bayesian approach
- Extensions available for a binary  $Y$  (Andridge et al. 2019) and for regression coefficients (West et al. 2021)

# 4. Multiple Imputation for Disclosure Limitation

- Confidentiality of responses in surveys has always been promised
- However:
  - Increasing demand for micro-level data
  - Linking the survey data to administrative data can increase efficiency
- Thus:
  - Potential for disclosure of responses, accidental or otherwise, may increase when micro data is released
    - Population uniques or less frequent
    - Sample uniques or less frequent
  - Possibility of disclosure (even perception) may affect quality and rates of response

# Multiple Imputation Approaches: Full synthesis

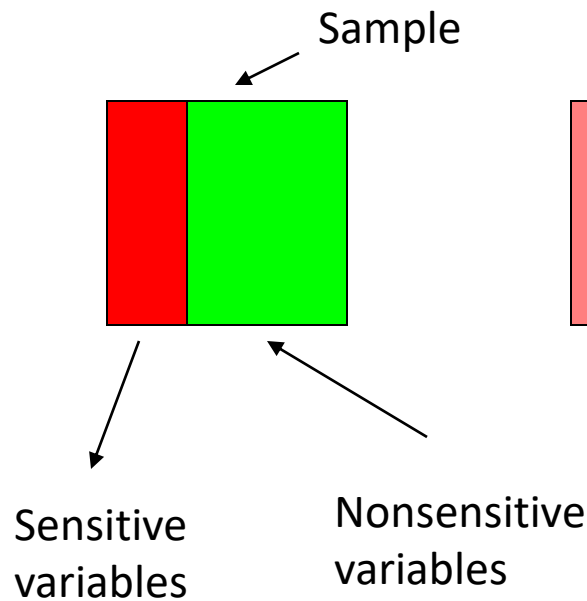


$$\bar{\theta} = \sum_{d=1}^D \hat{\theta}_d / D, \quad W = \sum_{d=1}^D \hat{V}_d / D,$$

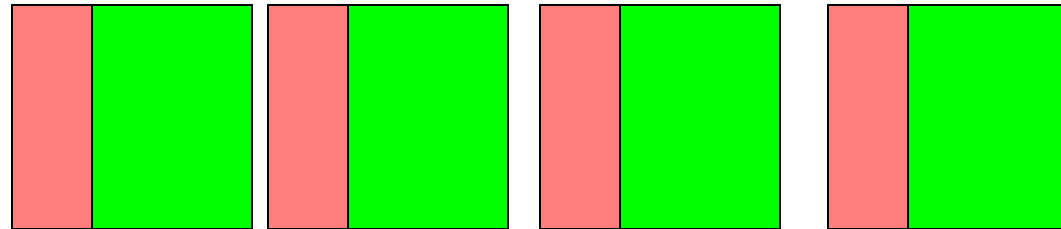
$$B = \sum_{d=1}^D (\hat{\theta}_d - \bar{\theta})^2 / (D-1)$$

$$T = (1 + \frac{1}{D})B - W$$

Rubin, 1993 (JOS); Raghunathan and Rubin (2001, ISBA);  
Raghunathan, Reiter and Rubin (2003, JOS), Reiter (2002, JOS)



# Partial synthesis



$$\bar{\theta} = \sum_{d=1}^D \hat{\theta}_d / D, \quad W = \sum_{d=1}^D \hat{V}_d / D,$$

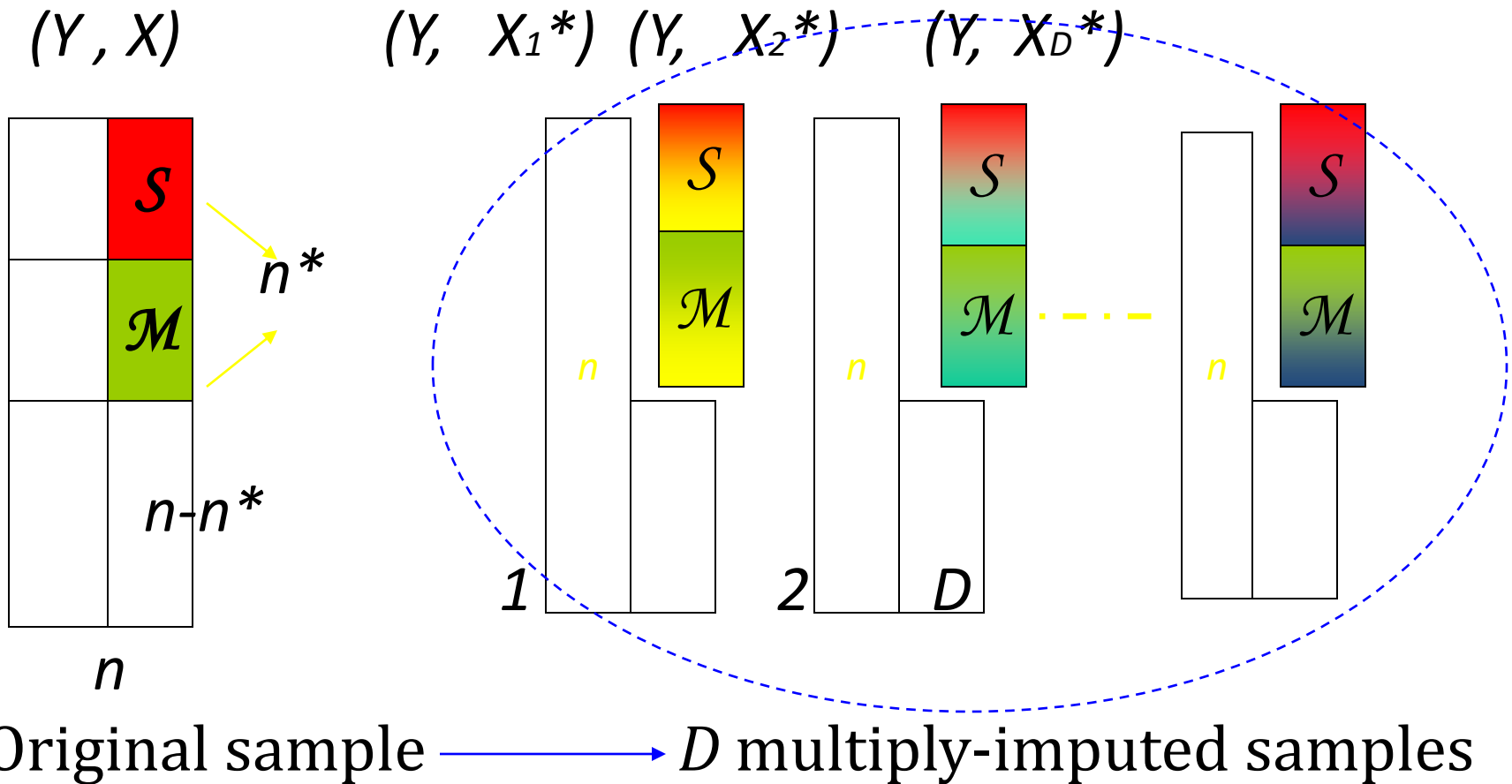
$$B = \sum_{d=1}^D (\hat{\theta}_d - \bar{\theta})^2 / (D - 1)$$

$$T = W + \frac{1}{D} B$$

- Sensitive variables
- Keys: Variables that could be used to link with administrative data
- Earnings from administrative data

Little (1993), Abowd and Woodcock (2002), Little and Liu (2003), Reiter (2003 Survey Methodology)

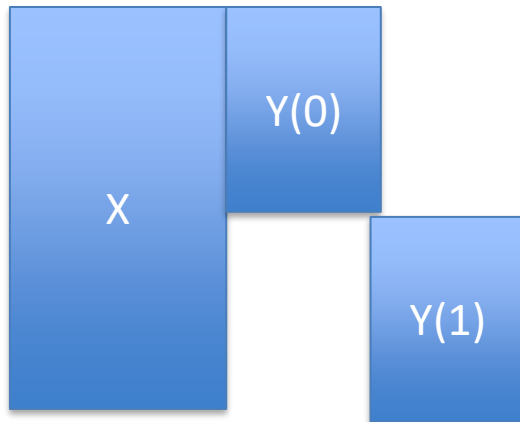
# Selective Multiple Imputation of Keys (SMIKe) Liu & Little (2002); Little & Liu (2003)





# 5. Multiple Imputation for Causal Inference

- Setup
  - Covariates,  $X$
  - Two treatments  $Z=0, 1$
  - Potential outcome framework,  $Y(0)$  &  $Y(1)$



\*Need :  $[Y(0), Y(1) | X]$

\*No information to estimate partial correlation between  $Y(0)$  and  $Y(1)$  given  $X$

But can estimate and compare marginal distributions

# Application: PENCOMP

- Zhou, T., Elliott, M.R. and Little, R.J. (2019). Penalized Spline of Propensity Methods for Treatment Comparisons (with discussion and rejoinder). *Journal of the American Statistical Association*, 114:525, 1-38.

# The building blocks of PENCOMP

- Likelihood/Bayes regression modeling and prediction
- Neyman/Rubin causal model: causal effects as a comparison of outcomes under alternative treatments, one observed, the other(s) unknown and imputed (Rubin, 1974)
  - In particular, principle stratification (Frangakis and Rubin, 2002)
- Multiple imputation, for propagating imputation uncertainty. (Rubin, 1987)
- Propensity modeling (Rosenbaum & Rubin, 1983)
  - With the propensity score used as a predictor in a regression, rather than as a weight
- Penalized Spline Modeling of propensity for robustness
  - Allowing a flexible relationship between outcome and predictor, here the propensity score

# Multiple Imputation (MI): an all-purpose tool for prediction:

- Imputes missing data as *draws*, not means, from the predictive distribution of the missing values under a model
- Creates  $D > 1$  filled-in data sets with different values imputed
- Bayesian MI combining rules (Rubin, 1987) yield valid inferences under well-specified models. Basic form is:
  - $V = W + (1 + 1/D)B$ ,  $W$  = within imputation variance,  $B$  = between imputation variance
- Standard errors reflect imputation uncertainty, and averaging of estimates over MI data sets corrects the loss of efficiency from imputing draws

# Examples of MI for missing data

- Bayes for parametric models, e.g. multivariate normal, general location model (PROC MI)
- Sequential regression/chained equations MI (IVEware, MICE, STAN,...)
- Hot deck multiple imputation (predictive mean matching)
- PSPP: a more robust regression-based method

# Penalized Spline of Propensity Prediction (PSPP) (Little & An 2004, Zhang & Little 2009, 2011).

- Estimate the propensity to respond given covariates
- Impute draws from a regression model that includes
  - Penalized spline of estimated propensity to respond
  - Parametric terms for other covariates predictive of  $Y$

Exploits the key balancing property of the propensity score (Rosenbaum and Rubin, 1983):

- Conditional on the propensity score and assuming missing at random, all covariates have same distribution for respondents and nonrespondents
- Hence misspecifying regression on other covariates does not lead to bias -- a form of double robustness

# Propensity score as covariate

- Flexible penalized spline is used to model relationship between propensity and outcome “correctly”
- Unlike standard regression models, the propensity score is not quite a known covariate, because it involves regression coefficients that need to be estimated from the data
- Uncertainty in the regression coefficients can be reflected by:
  - (a) Estimating the coefficients and a set of imputations for a particular MI data set on a bootstrap sample, or
  - (b) Drawing coefficients from their posterior distribution
- We did method (b)

# Response propensity as covariate

- Why include the response propensity as a covariate rather than as a weight?
- Weighting does not reflect the strength of relationship between propensity and outcome
  - If the propensity is weakly related to the outcome, weighting is unnecessary and adds noise
- As a covariate, the influence of the response propensity on imputations depends on its strength as a predictor
- This basic idea explains why our method can be more efficient than doubly-robust weighting methods
  - Advantage is increased with multiple time points



# Basic Neyman /Rubin causal model (Rubin, 1974)

Unit, $i$	$X$	$Z$	$Y^{(0)}$	$Y^{(1)}$
1		0		?
2		0		?
...		...		...
$n_0$		0		?
$n_0+1$		1	?	
$n_0+2$		1	?	
...		...	...	
$n_0+n_1$		1	?	

$X$  baseline covariates/confounders

$Z$  treatment indicator (0,1)

$Y^{(0)}$  outcome if given  $Z = 0$

$Y^{(1)}$  outcome if given  $Z = 1$

Prediction Problem: predict outcomes  
(?) for the treatment not assigned

With no unmeasured confounders: regress  $Y$  on  $Z, X$

Multiply impute missing  $Y^{(0)}, Y^{(1)}$  with predictions for  $X = 0, 1$

Apply MI combining rules for inference about  $\Delta$

# PENCOMP MI for basic model

- “Propensity to respond” in PSPP becomes “propensity to be selected” in PENCOMP
- Using all the data, estimate the propensity of each treatment to be selected, given covariates  $X$
- Within each treatment group, estimate a regression model of  $Y$  on  $X$  that includes
  - Penalized spline of estimated propensity that treatment is selected
  - Parametric model for other covariates predictive of  $Y$
- For  $z = 0, 1$ , impute the values of  $Y^z$  for subjects in treatment group  $1-z$  in the original data set with draws from the predictive distribution of  $Y$  given  $X$  from the regression,
- Use MI combining rules for inference about average treatment effect  $\Delta$

# Longitudinal causal inference: confounding by indication

Unit, $i$	$X_1$	$Z_1$	$X_2$	$Z_2$	$Y$	$X_1$ baseline covariates/confounders
1		0		0		$Z_1$ time 1 treatment indicator (0 or 1)
2		0		0		$X_2$ intermediate outcome
...		...		...		
$n_{00}$		0		0		$Z_2$ time 2 treatment indicator (0 or 1)
$n_{00}+1$		0		1		
$n_{00}+2$		0		1		Assume ignorable assignment mechanisms
...		...		...		
$n_0=n_{00}+n_{01}$		0		1		$Y$ outcome, 3 treatment effects:
$n_0+1$		1		0		$\Delta_{jk} = \text{mean } Y (Z_1 = j, Z_2 = k)$
$n_0+2$		1		0		$-\text{mean } Y (Z_1 = Z_2 = 0) (jk = 11, 10, 01)$
...		...		...		
$n_0+n_{10}$		1		0		
$n_0+n_{10}+1$		1		1		Regression doesn't work now:
$n_0+n_{10}+2$		1		1		$X_2$ is an outcome for $Z_1$ , confounder for $Z_2$
...		...		...		So can't just condition on $X_2$
$n=n_0+n_{10}+n_{11}$		1		1		

# Confounding by indication: Neyman/Rubin causal model to the rescue! (Frangakis and Rubin 2002)

Unit, $i$	$X_1$	$Z_1$	$X_2^{(0)}$	$X_2^{(1)}$	$Z_2$	$Y^{(00)}$	$Y^{(01)}$	$Y^{(10)}$	$Y^{(11)}$
1		0		?	0		?	?	?
2		0		?	0		?	?	?
...		...		?	...		...	...	...
$n_{00}$		0		?	0		?	?	?
$n_{00}+1$		0		?	1	?		?	?
$n_{00}+2$		0		?	1	?		?	?
...		...		...	...			...	...
$n_0=n_{00}+n_{01}$		0		?	1	?		?	?
$n_0+1$		1	?		0	?	?		?
$n_0+2$		1	?		0	?	?		?
...		...	?		...	...			...
$n_0+n_{10}$		1	?		0	?	?		?
$n_0+n_{10}+1$		1	?		1	?	?	?	
$n_0+n_{10}+2$		1	?		1	?	?	?	
...		...	...		...	...			
$n=n_0+n_{10}+n_{11}$		1	?		1	?	?	?	

So much  
missing  
data!  
But MI  
propagates  
uncertainty  
...

Multiply impute the missing data (?s)

# PENCOMP MI for confounding by indication

- (a) Take Bootstrap sample of the original data. For each bootstrap sample:
  - (b) missing values of the intermediate treatment outcomes  $X_2^{(0)}$  and  $X_2^{(1)}$  and are imputed using the PSPP model
  - (c) Conditional on the values of  $X_1, Z_1$  and the observed or imputed values of  $X_2$ , the propensity that  $Z_2 = 1$  given  $X_1, Z_1$  and  $X_2$  is estimated based on a logistic regression  $Z_2$  on  $X_1, Z_1$  and  $X_2$
  - (d) missing values of  $Y^{(jk)}$  are then imputed as draws based on the regression  $Y^{(jk)}$  on  $X_1, Z_1$  and  $X_2$  for a model that includes a spline on the propensity from (c); a distinct regression model is fitted for each outcome  $Y^{(jk)}$
- (e) Apply MI combining rules for inference about average treatment effects

# Alternative methods: IPTW and AIPTW

## Inverse Probability of Treatment Weighting (IPTW)

Weight subjects by the inverse of estimate of  $\Pr(Z | X)$ -- in effect creates a pseudo-population that is free of treatment confounders.

Consistent if the treatment assignment mechanism is correctly specified. But weights can be highly variable, leading to poor efficiency

## Augmented Inverse Probability of Treatment Weighting (AIPTW)

Doubly robust: consistent if the treatment assignment mechanism is correctly specified, or the prediction model is correctly specified.

# Application

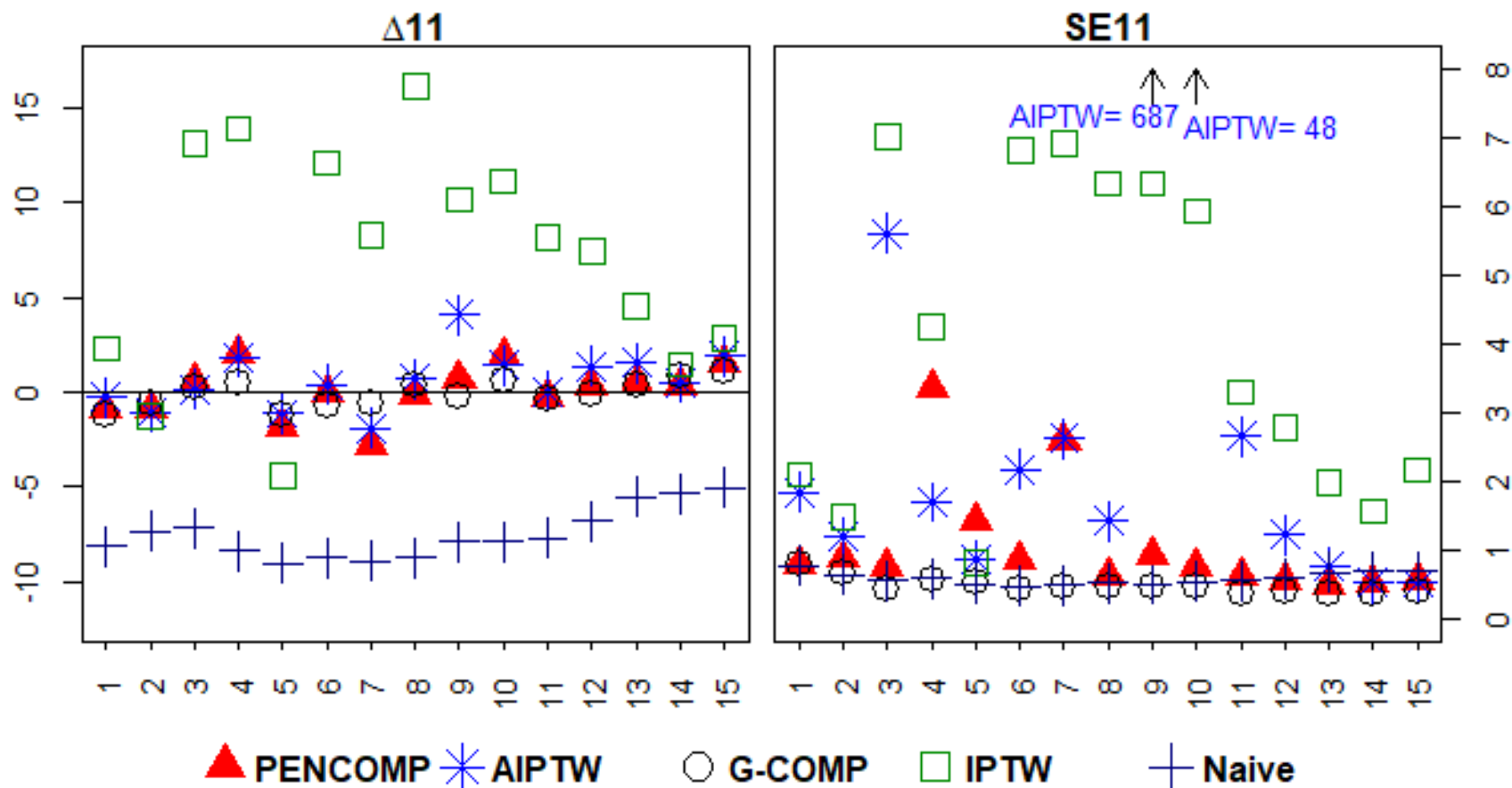
We applied our method to the Multicenter AIDS Cohort study (MACS) to analyze the effect of antiretroviral treatment on CD4 counts for HIV+. (Kaslow et al, 1987)

CD4 count is an intermediate outcome of past treatments and confounds the next treatment.

Restrict our analyses to the period between visit 6 and 21, when zidovudine was approved and available for use and before the advent of highly active antiretroviral therapy (HAART).

We estimate the short-term (1 year) effects of using antiretroviral treatment for HIV+ subjects during this period, for each of the three-visit moving windows 1,  $\dots$ , 14.

# Estimates of Delta11 and SEs for five methods





# Summary of simulation comparisons

- When the confounding is low or moderate, the weights are more stable, PENCOMP and AIPTW perform similarly, and are both superior to IPTW.
- PENCOMP has slightly larger (but still negligible) bias than AIPTW when the prediction model is misspecified and weights are variable.
- But PENCOMP tends to outperform AIPTW in RMSE, coverage probability and efficiency.

# Why does PENCOMP do better?

- When propensities (and their inverse, the selection weights) are highly variable:
  - weighting can be very inefficient
  - regressing on the weight is less inefficient – if the weight is not associated with the outcome, it receives low estimated regression coefficients
  - This explains the improved performance of PENCOMP over weighting methods in this setting
  - Also, we may be far from “asymptotia”, where theoretical concepts like semi-parametric efficiency apply

# Avoiding extrapolation of splines beyond range of data

- Positivity requires that cases have a propensity to be assigned to any of the compared treatments that lie between zero and one.
- Causal effects for cases with estimated propensities close to zero or one are very poorly determined, and rely excessively on extrapolation of the prediction model outside the range of estimated propensities.

# The case of multiple time points

- With many time points, the number of treatment combinations become very large
  - 10 repeated measures,  $2^{10}$  possible treatment combinations
  - Need to restrict causal inferences to subset of treatments for which there are “sufficient data”
  - Disparities in propensity score distributions become acute – this is a serious problem
  - Also need to define estimands to avoid the excessive extrapolation issue noted above
- This problem is if anything even worse for weighting methods

# Addressing disparities in propensity distribution

- Modify the estimand to restrict to subpopulations where propensities are not close to zero or one
  - modification cannot involve estimated propensities, for estimand to be well defined
  - Truncate tails of the true propensity distribution, with increasingly severe truncation for smaller sample sizes
  - Various alternatives, e.g. Li, Morgan and Zaslavsky (2017), although interpretation perhaps trickier
  - Can easily extend PENCOMP to estimate alternative estimand

# Addressing disparities in propensity distribution

- Covariates associated with the selection propensity but are not associated with the outcome are not confounders
- Our simulations suggest that weeding out or penalizing such variables in the propensity model can reduce the disparity in the estimated propensity distributions and greatly improve efficiency
- Rubin (2007) opposes this, arguing by analogy with randomized studies that variables in the propensity should be chosen *a priori* to avoiding “cheating”

# Avoiding “cheating”

- Prespecification of methods for weeding out spurious variables in the propensity model
- Select out such variables using regressions that do not include the treatment indicators, so we don't know how selection influences the treatment effect

# Avoiding “cheating”

- Pre-specified “automatic” penalized regression methods can reduce the risk of “model shopping”
  - Adaptive LASSO (Zhou 2006) for propensity and outcome models separately or jointly (Shortreed and Ertefaie 2017)
  - Two-stage (only use selected variable from propensity model to estimate outcome or vice versa)
    - Use “bagging” (Efron 2014) to account for uncertainty in model selection across bootstrap samples
- Simulation study shows two-stage has greatest efficiency gains



# Conclusion

- Multiple Imputation – a useful all-purpose tool for prediction
- PENCOMP – regression models for Rubin Causal Model that include spline of propensity as predictor
  - Conceptually simple, robustness properties
  - Propensity treated as a covariate rather than a weight, improving stability and avoiding problems with highly-variable weights
  - Issues with disparity in propensity distributions – defining the estimand, and variable selection for the propensity to weed out non-confounders

# Summary and Conclusions

- Many applications can be set as a missing data problem
- Multiple imputation provides a framework for handling a variety of such problems
- Software availability makes it easy to implement
- Imputations under different models can be used to perform sensitivity analysis
- Imputations of observed values may be used for model checking, detecting outliers
- By no means, imputation task is easy. It needs expertise in statistical modeling, primary focus of statistics anyway (or should be)!!