Advances in Research on Participant Attrition from Prevention Intervention Studies

John W. Graham[1], Lauren E. Connell[1], & Michael L. Hecht[2]

The Prevention Research Center
[1]Department of Biobehavioral Health
[2]Communication Arts & Sciences
Penn State University

jgraham@psu.edu

EUSPR, Palma de Mallorca, Spain
October 17, 2014

Causes of Missingness

- Ignorable
  - MCAR: Missing Completely At Random
  - MAR: Missing At Random
- Non-Ignorable
  - NMAR: Not Missing At Random

NMAR Causes

- The recommended analysis methods (**multiple imputation** and **FIML**) assume missingness is MAR

- But what if the cause of missingness is not MAR?

- Should these methods be used when MAR assumptions not met?

  . . .

YES! These Methods Work!

- It's not like other methods
  - where there are better methods when assumptions not met
- MI and ML methods work better than “old” methods (listwise deletion)
- Multiple causes of missingness
  - Only small part of missingness may be NMAR

Conventional Wisdom

- One CAN know if MCAR holds
- One canNOT know whether missingness is MAR or NMAR
- Some truth to latter statement
- BUT with longitudinal data, there is much that CAN be known
- This paper shows how you can know

What if the cause of missingness is NMAR?

Problems with this statement

- ***The*** cause of missingness is ***never*** purely NMAR or MAR
- Better to think of MAR and NMAR as forming a continuum
- MAR vs NMAR ***NOT*** even the dimension of interest

MAR vs NMAR: What IS the Dimension of Interest?

- How much **ESTIMATION BIAS**?
  - when cause of missingness cannot be included in the model

Bottom Line ...

- All missing data situations are partly MAR and partly NMAR

- Sometimes it matters ...
  - bias affects statistical conclusions

- Often it does not matter
  - bias has tolerably little effect on statistical conclusions

(Collins, Schafer, & Kam, *Psych Methods*, 2001)

Collins, Schafer, & Kam (2001; CSK) CSK Paradigm

- CSK study
- Graham et al. (2008; 2013)

- Example model of interest:

$$T \rightarrow Y \quad (Z)$$

Creating Missingness NMAR (Linear)

- Create NMAR missing with IF statements:
  - **if Z=1** then prob(Ymissing) = **.20**
  - **if Z=2** then prob(Ymissing) = **.40**
  - **if Z=3** then prob(Ymissing) = **.60**
  - **if Z=4** then prob(Ymissing) = **.80**

% missing= average of probabilities

50% missing

Quartiles

Creating Missingness
Relevant Quantities

- % Missing
- r_{TY}
  - **treatment effect size**
- r_{YZ}
- r_{ZR}

Creating Missingness

Relevant Quantities: r_{ZR}

- Z = cause of missingness
- R = missingness (observed=1; missing=0)

- r_{ZR} related to IF statements
  - **if Z=1** then prob(Ymissing) = **.20**
  - **if Z=2** then prob(Ymissing) = **.40**
  - **if Z=3** then prob(Ymissing) = **.60**
  - **if Z=4** then prob(Ymissing) = **.80**

Range = .60

- r_{ZR} = range x constant*
  - with 50% missing, range = .60 means r_{ZR} = .45

Yardsticks for Measuring Bias

Standardized Bias < 40 is tolerable

(average parameter est) – (population value)
--- X 100
Standard Error (SE)

- |bias| < 40 considered small enough to be tolerable (Collins et al., 2001)
- t-value off by 0.4

- Relative Bias < .10 is tolerable
  - parameter estimate off by 10% of true value
- Best when both rules are met

Research Results
with NMAR Linear Missingness

- % missing
  - Less missing means less bias
- r_{TY} (effect size)
  - Choose values from empirical research
  - r_{TY} = .60 unrealistic
  - r_{TY} = .20 has 75% less bias!
- r_{YZ}
  - r = .50 very realistic (with longitudinal data)
  - AND with effect size (r_{TY}) = .20, no scenario produces appreciable bias when r_{YZ} = .50

Research Results for r_{ZR} (range)

- r_{ZR} (range) more difficult
  - Cannot be estimated directly
- But range = **.60**
  very unusual in prevention research
- Range = **.20** much more common

bottom line …

- Scenario studied by CSK …
  - Not a problem in typical prevention research

- But this scenario is only part of the story

A Taxonomy of Attrition
Causes of Attrition on Y (main DV)

Case **1**: not T, not Y, not TY interaction (MCAR)

Case **2**: T only (MAR)

Case **3**: Y only (CSK scenario)

Case **4**: T and Y only

Case **5**: TY interaction only

Case **6**: T + TY interaction only

Case **7**: Y + TY interaction only

Case **8**: T + Y + TY interaction

Graham, J. W. (2012). *Chapter 1*

Studying the 8 Cases is Complex

- Design & Monte Carlo simulation utility
  - Built around IF statements

IF Z=1 then prob(Ymissing) = .20

IF Z=2 then prob(Ymissing) = .40

IF Z=3 then prob(Ymissing) = .60

IF Z=4 then prob(Ymissing) = .80

Z is cause of missingness on Y

Design & Simulation Utility

- IF statements for Cases 4-8
  - Treatment Group
    - IF Z=1 then prob(Ymissing) = .20
    - IF Z=2 then prob(Ymissing) = .40
    - IF Z=3 then prob(Ymissing) = .60
    - IF Z=4 then prob(Ymissing) = .80
  - Control Group
    - IF Z=1 then prob(Ymissing) = .10
    - IF Z=2 then prob(Ymissing) = .20
    - IF Z=3 then prob(Ymissing) = .30
    - IF Z=4 then prob(Ymissing) = .40

Design & Simulation Utility

- IF statements for Cases 4-8
  - Treatment Group
    - IF Z=1 then prob(Ymissing) = .20
    - IF Z=2 then prob(Ymissing) = .40
    - IF Z=3 then prob(Ymissing) = .60
    - IF Z=4 then prob(Ymissing) = .80
  - Control Group
    - IF Z=1 then prob(Ymissing) = .10
    - IF Z=2 then prob(Ymissing) = .20
    - IF Z=3 then prob(Ymissing) = .30
    - IF Z=4 then prob(Ymissing) = .40

50% missing

Average missing
37.5%

25% missing

Design & Simulation Utility

- IF statements for Cases 4-8
  - Treatment Group
    - IF Z=1 then prob(Ymissing) = .20
    - IF Z=2 then prob(Ymissing) = .40
    - IF Z=3 then prob(Ymissing) = .60
    - IF Z=4 then prob(Ymissing) = .80
  - Control Group
    - IF Z=1 then prob(Ymissing) = .10
    - IF Z=2 then prob(Ymissing) = .20
    - IF Z=3 then prob(Ymissing) = .30
    - IF Z=4 then prob(Ymissing) = .40

range = .60

Y Effect
avg. range = .45

range = .30

Design & Simulation Utility

- IF statements for Cases 4-8
  - Treatment Group
    - IF Z=1 then prob(Ymissing) = .20
    - IF Z=2 then prob(Ymissing) = .40
    - IF Z=3 then prob(Ymissing) = .60
    - IF Z=4 then prob(Ymissing) = .80
  - Control Group
    - IF Z=1 then prob(Ymissing) = .10
    - IF Z=2 then prob(Ymissing) = .20
    - IF Z=3 then prob(Ymissing) = .30
    - IF Z=4 then prob(Ymissing) = .40

50% missing

T Effect
25% difference

25% missing

Design & Simulation Utility

- IF statements for Cases 4-8
  - Treatment Group
    - IF Z=1 then prob(Ymissing) = .20
    - IF Z=2 then prob(Ymissing) = .40
    - IF Z=3 then prob(Ymissing) = .60
    - IF Z=4 then prob(Ymissing) = .80
  - Control Group
    - IF Z=1 then prob(Ymissing) = .10
    - IF Z=2 then prob(Ymissing) = .20
    - IF Z=3 then prob(Ymissing) = .30
    - IF Z=4 then prob(Ymissing) = .40

range = .60

TY Interaction Effect
range diff = .30

range = .30

Design & Simulation Utility …

- Prompts for these quantities
- Writes SAS code
- Performs Monte Carlo Simulation
  - with e.g., 1000 replications
- Writes out bias estimates
- All automatically

Using the Design/MCsim Utility with Empirical Data

- Drug Resistance Strategies Rural (DRSR) Project (*keepin' it REAL* program) (Colby, Hecht et al., 2013)
  - 39 Rural schools in Pennsylvania & Ohio
  - Implement in 7th grade
  - 4 waves of meaurement
    - (7a) 7th grade early (pretest)
    - (7b) 7th grade late (immediate posttest)
    - (8) 8th grade late
    - (9) 9th grade late

Estimating Various Quantities in Empirical Data

- % missing ... easy
- r_{TY} (effect size) ... pretty easy
- r_{ZY} ... pretty easy

- r_{ZR} (range) ... more difficult
  - Must estimate $r_{\text{Drugs}\mathbf{9},\text{Missingness}\mathbf{9}}$
  - Must use regressions with:
    - Drugs**7a**, Missingness**9**
    - Drugs**7b**, Missingness**9**
    - Drugs**8**, Missingness**9**

Estimating r_{ZR} (range) in Empirical Data

- r_{ZR} is $r_{\text{Drugs}\mathbf{9},\text{Missing}\mathbf{9}}$
- But $r_{\text{Drugs}\mathbf{9},\text{Missing}\mathbf{9}}$ is not estimable
- Must use proxy correlations:
  - Drugs**7a**, Missing**9**
  - Drugs**7b**, Missing**9**
  - Drugs**8**, Missing**9**

- Estimation strategy outlined in my book

Regressions: Treatment Group

Model	R^2	R^2 - Imp	R
drugs7a ➔ Miss9	.0313	.0313	**.177**
+drugs7b ➔ Miss9	.0329	.0016	**.040**
+drugs8 ➔ Miss9	.0515	.0186	**.136**
drugs9 ➔ Miss9			**???**

Regressions: Control Group

Model	R^2	R^2 - Imp	R
drugs7a ➔ Miss9	.0126	.0126	**.112**
+drugs7b ➔ Miss9	.0147	.0021	**.046**
+drugs8 ➔ Miss9	.0219	.0072	**.085**
drugs9 ➔ Miss9			**???**

Predicting $r_{drugs9,M9}$ (range)

Model	R		avg / diff
	Treatment	Control	
drugs7a ➔ Miss9	.177	.112	
+drugs7b ➔ Miss9	.040	.046	
+drugs8 ➔ Miss9	.136	.085	
Predicted $r_{drugs9,Miss9}$ (range)			
Use $r_{Drugs8,Miss9}$	.136 (.154)	.085 (.096)	.125/.058
Quadratic Trend	.465 (.526)	.229 (.259)	.393/.267
Linear Trend waves 2&3 only	**.232 (.262)**	**.124 (.140)**	**.201 / .122**

Miss9 = missingness at 9th grade

Summary of Empirical Info

- %missing
  - average: 19.4% (**.194**) (%missing)
  - difference: 0.6% (**.006**) (T effect)

	Y effect (range)	
	avg	diff
Linear Trend (2,3)	**.201**	**.122**

- Y effect (range)
  - Linear Trend (2,3)

Standardized and Relative Bias

	Standardized Bias	Relative Bias
Real data (19% missing, .006 T effect) Linear Trend (based on waves 2,3)	**-24.0**	**.064**

Blue = bias tolerably low
Pink = borderline
Red = bias can affect statistical conclusions

Standardized and Relative Bias

	Standardized Bias	Relative Bias
Real data (19% missing, .006 T effect) Linear Trend (based on waves 2,3)	**-24.0**	**.064**
Same, except 40% missing	**-38.0**	**.116**

Blue = bias tolerably low
Pink = borderline
Red = bias can affect statistical conclusions

Standardized and Relative Bias

	Standardized Bias	Relative Bias
Real data (19% missing, .006 T effect) Linear Trend (based on waves 2,3)	-24.0	.064
Same, except 40% missing	-38.0	.116
Same, except 40% missing + .10 T Effect	-46.0	.142

Blue = bias tolerably low
Pink = borderline
Red = bias can affect statistical conclusions

Auxiliary Variables

- Restores some power lost due to attrition
- Reduces attrition bias
  - Variables that predict attrition

Value of Attrition Related Auxiliary Variables

- Predict missingness at 9th grade
  - Drug use variables (from all three 7th–8th grade waves)
    - **R^2 = .037**
  - Attrition-relevant Auxiliary Variables
    - **R^2 = .197**

Standardized and Relative Bias with Attrition-relevant Auxiliary Variables

	Standardized Bias	Relative Bias
Real data Linear Trend (based on waves 2,3)	-24.0 ➔ **-22.2**	.064 ➔ **.058**

Blue = bias tolerably low
Pink = borderline
Red = bias can affect statistical conclusions

Standardized and Relative Bias with Attrition-relevant Auxiliary Variables

	Standardized Bias	Relative Bias
Real data Linear Trend (based on waves 2,3)	-24.0 → **-22.2**	.064 → **.058**
Same, except 40% missing	-38.0 → **-29.1**	.116 → **.091**

Blue = bias tolerably low
Pink = borderline
Red = bias can affect statistical conclusions

Standardized and Relative Bias with Attrition-relevant Auxiliary Variables

	Standardized Bias	Relative Bias
Real data Linear Trend (based on waves 2,3)	-24.0 → **-22.2**	.064 → **.058**
Same, except 40% missing	-38.0 → **-29.1**	.116 → **.091**
Same, except 40% missing + .10 T Effect	-46.0 → **-30.1**	.142 → **.093**

Blue = bias tolerably low
Pink = borderline
Red = bias can affect statistical conclusions

Conclusions

- Attrition CAN be bad for internal validity
- But often it's NOT nearly as bad as often feared

- Don't rush to conclusions, even with rather substantial attrition
- Examine evidence before drawing conclusions
  - **We CAN know some things about bias**

- Use MI and ML missing data procedures!
- Use good auxiliary variables to minimize impact of attrition

END
Relevant Work:

- **Graham,** J.W., **(2009).** Missing data analysis: making it work in the real world. *Annual Review of Psychology*, *60*, 549-576.
- Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*, 330_351.
- Hedeker, D., & Gibbons, R.D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies, *Psychological Methods*, *2*, 64-78.
- **Graham**, J.W., **(2012)**. *Missing Data: Analysis and Design*. New York: Springer.