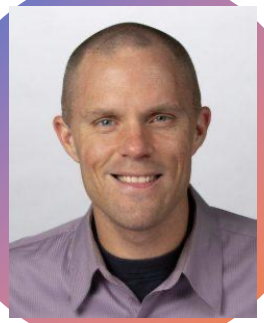
Apply Weights to Longitudinal Survey Data

Network for Innovative Methods in Longitudinal Aging Studies (NIMLAS) Workshop

Yajuan Si

Univ. of Michigan

Sept 15, 2022



Congratulations on the Launch of NIMLAS!!!

- *More information is available: https://nimlas.isr.umich.edu*

Outline

1. What is the weight?
2. How do we weight?
3. To weight or not to weight?
4. Alternatives to weighting
5. Case study

Outline

1. What is the weight?
2. How do we weight?
3. To weight or not to weight?
4. Alternatives to weighting
5. Case study

Longitudinal survey data

- Longitudinal surveys (aka panel studies) collect rich data about individual characteristics and enable the estimation of trajectory modeling, a fundamental approach to understanding the developmental course and lifespan.
- Examples include the Monitoring the Future (MTF) panel study, the Panel Study of Income Dynamics (PSID), the Health and Retirement Study (HRS), and the Midlife in the United States (MIDUS) Series.
- However, longitudinal studies are subject to selection and nonresponse bias.

Complex survey sampling design

- Sample selection procedures for longitudinal studies are often complex in nature.
  1. Many probability survey data sets are collected from analytic units randomly selected according to a stratified, multistage sample design with unequal probabilities.
  2. Nonprobability samples need to adjust auxiliary variables that predict survey outcomes and affect inclusion (selection and response).

Example: PSID design (Morgan & Smith, 1969)

SUMMARY OF CHARACTERISTICS AND DESIGN OF SRC NATIONAL SAMPLE OF DWELLINGS[a]

Multistage	"area"	Probability	sample	of dwellings and	their occupants
1st	County, county group, or SMSA as primary sampling unit (psu)	Probability proportionate to size (pps)	1 psu from each of 74 homogeneous strata	Universe of dwellings in 48 states and D.C., exclusive of du's on military reservations (population in non-du's excluded)	Households, families, spending units, or persons
2nd	Cities, towns, census tracts, minor civil divisions, rural areas	pps	3-10 selections and average of five selections per psu		
3rd	Blocks, census ED's, chunks, city directory pages	pps or equal probability	1 selection per expected cluster of 4 du's, for 3,000 interview survey		
4th	Segments or clusters of dwellings	Equal probability	6-20 segments (av.=8) of 4 du's each per psu, for 3,000 interview survey		
5th	Families, spending units, or adults	Equal or unequal probability	All families or spending units, or one adult per household		
...	...	Over-all sampling fraction = prod. of selection probabilities over the several stages	...	...	...

3. Sample size determined: (1) to yield acceptable reliability for estimates of means, percentages, comparisons; (2) to be administratively feasible: 8-10 traveling supervisors; permanent, local, part-time interviewers in 74 locations (12 largest SMSA's, 62 other 1st stage units).
4. Stratification within psu's: (1) Central cities of SMSA's; (2) other cities of 50,000 or more; (3) urban places 2,500-49,999; (4) other urbanized areas; (5) remaining areas.
5. Sample dwellings for specific study selected from: (1) area segments of about 4 du's; OR (2) clusters of du's (about 4 per cluster) at city directory addresses--supplemented by area sample to include du's at addresses NOT in directory.
6. Probability sampling permits calculation of sampling error from the sample itself.
7. The measurement of nonsampling errors, an important contribution to total error in survey data, is not readily estimated from the sample alone; often the measurement is dependent upon information obtained from independent sources.

[a]From The Survey Research Center's National Sample of Dwellings, op. cit.

-22-

Design features

- Complex samples generally feature **stratification**, **cluster** sampling and survey **weights** that are publicly released for sampling error computation.
  - Valliant, Dever, & Kreuter (2018) provide detailed steps in the construction of probability survey weights in practice.
  - Elliott (2013) develops pseudo-weights for nonprobability samples.

Longitudinal nonresponse

- Longitudinal studies are subject to two types of nonresponse:
  1. Unit nonresponse (i.e., attrition) when individuals fail to provide data for entire waves, and
  2. Item nonresponse when they only respond to particular questions of the survey.
- With declining response rates in surveys, study respondents could be systematically different from nonrespondents in terms of survey outcome distributions, causing potential bias in estimates.

Response patterns: Monotone vs. intermittent

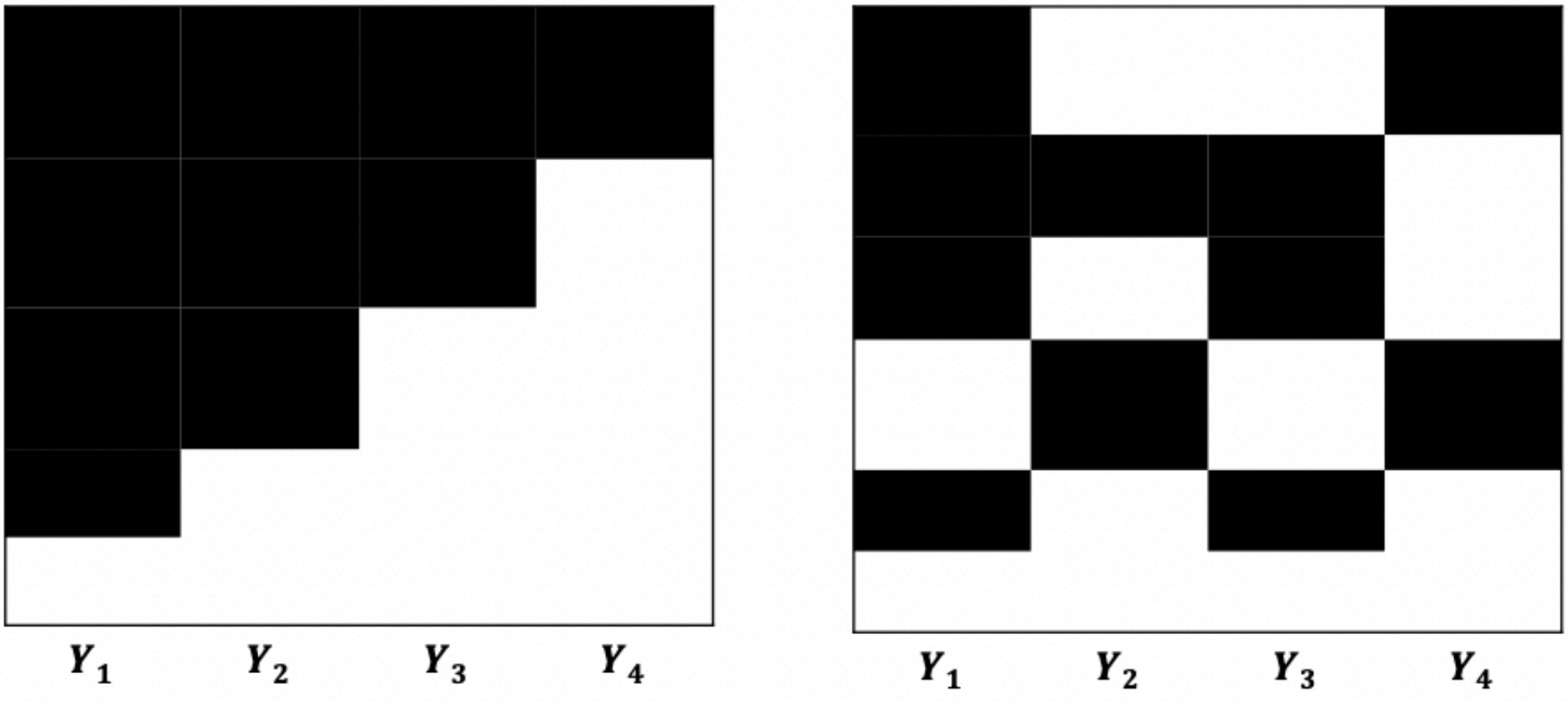
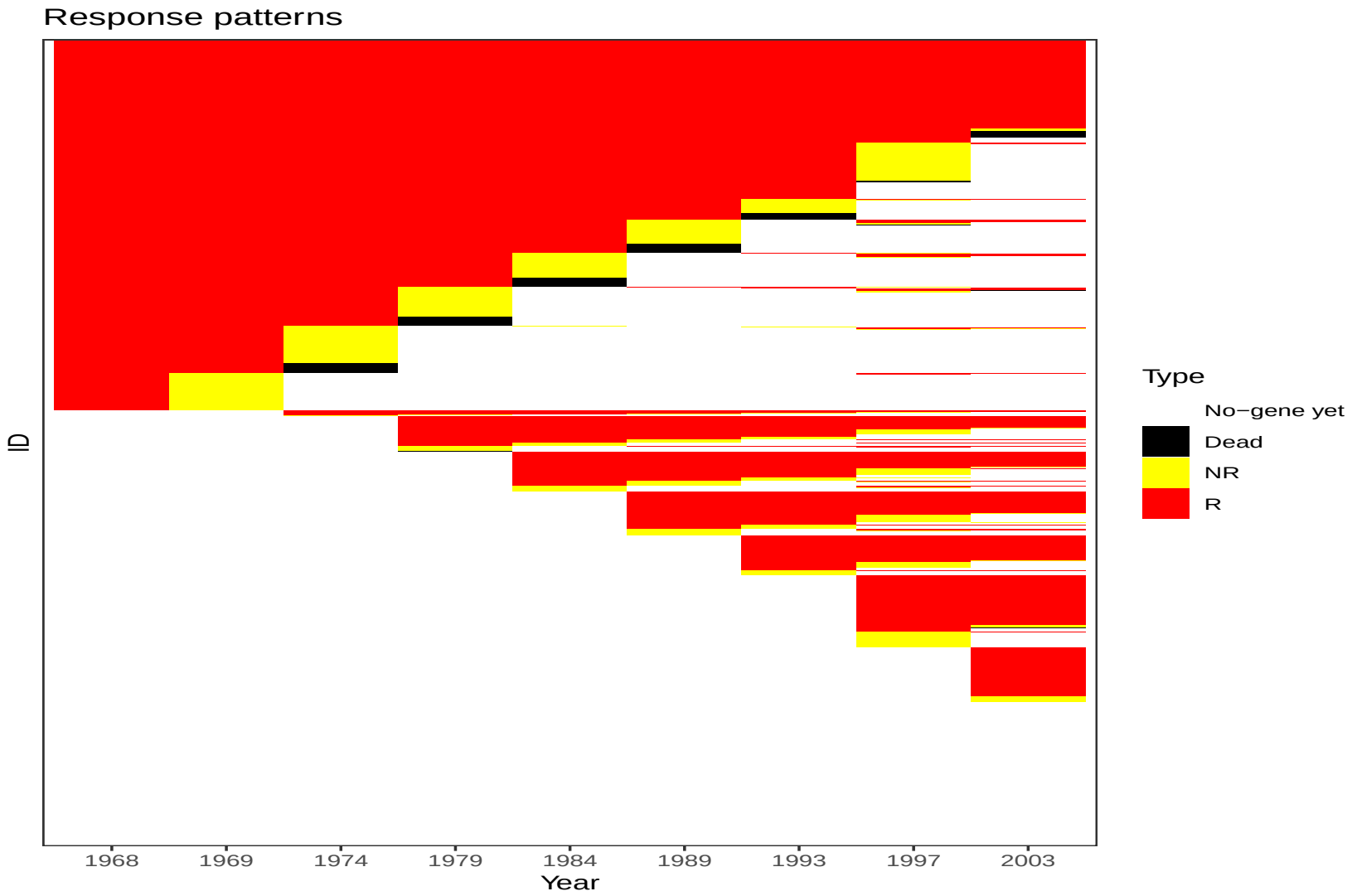


Illustration of missing patterns and the implied response indicator matrices with four variables. (black areas indicate response, and white areas indicate nonresponse). The left plot is a monotone "staircase" pattern, and the right shows a general "swiss-cheese" pattern of missingness.

More complex patterns: A subset of PSID



Response indicators: Monotone vs. intermittent

Binary indicator of response (R=1) and nonresponse (NR, R=0).

R_1	R_2	R_3	R_4
1	1	1	1
1	1	1	0
1	1	0	0
1	0	0	0

R_1	R_2	R_3	R_4
1	0	0	1
1	1	1	0
1	0	1	0
0	1	0	1
1	0	1	0

Stochastic nonresponse mechanisms (Rubin, 1976)

Every unit has a response probability/propensity, denoted by $\psi_i = \Pr(R_i = 1)$.

1. Missing completely at random (MCAR): respondents are a simple random subsample of the base samples, every unit of which has the same response probability.
2. **Missing at random (MAR)**: the probability of response depends on observed information.
3. Missing not at random (MNAR): the probability of response depends on the missing values themselves.

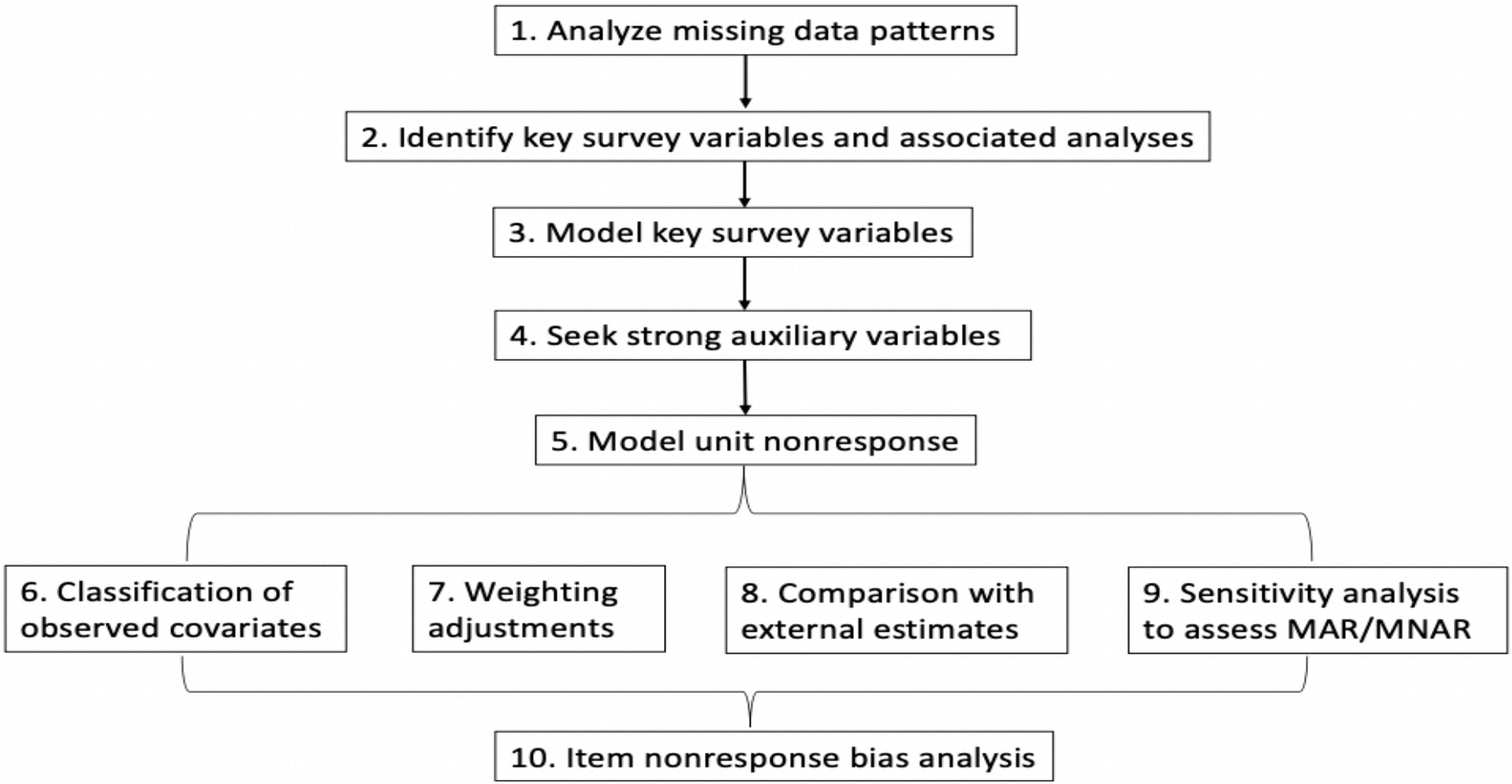
Nonresponse bias

- The nonresponse bias of the respondent mean $\bar{y}_R$ is approximately

$$Bias(\bar{y}_R) \approx \frac{1}{\bar{\psi}} Cov(Y, \psi).$$

- Here $Cov(Y, \psi) = \frac{1}{N} \sum_{i=1}^{N} (Y_i - \bar{Y})(\psi_i - \bar{\psi})$ is the population covariance between a given survey variable Y and the response propensity ψ, and $\bar{\psi} = \sum_{i=1}^{N} \psi_i / N$ is the overall population response propensity mean.
- Response rates only provide partial evidence of the risk of nonresponse bias.
- Nonresponse bias can be further explored when other auxiliary variables, denoted by Z, are available for all samples.

Nonresponse bias analysis (Si, Little, Mo, & Sedransk, 2022a)



Propensity score weighting

- We fit a model to predict response propensities based on available predictors.
- Approaches to modeling the response propensities include parametric models such as logistic regressions and machine learning algorithms.
  1. Use the inverse of predicted response propensity as the weight: $w_i = 1/\hat{\psi}_i = \widehat{\Pr}^{-1}(R_i = 1 \mid Z_i)$.
  2. Construct cells/weighting classes, such as using quintiles of predicted propensity scores, compute response rate within each cell, and treat the inverse value as the weight
- The key is the existence of auxiliary variables Z: **strongly related to survey outcomes** and available for both respondents and nonrespondents.

Key to success: Highly predictive auxiliary variables

- The nonresponse bias depends on the **covariance** between the survey variable and response propensity, the **response rates**, and the **difference in estimates between respondents and nonrespondents**.

- The main approaches to bias adjustment are nonresponse weighting, where responding units are weighted by the inverse of an estimate of the response probability, and imputation, where missing values are imputed by predictions based on observed variables.

- To reduce bias, auxiliary variables X (we use Z) in the nonresponse adjustment must be predictive of the survey variable of interest Y (Little & Vartivarian, 2005).

	Association with outcome	
Association with nonresponse	Low	High
Low	Cell 1 Bias: --- Var: ---	Cell 3 Bias: --- Var: ↓
High	Cell 2 Bias: --- Var: ↑	Cell 4 Bias: ↓ Var: ↓

Table 1. Effect of weighting adjustment on bias and variance of a mean, by strength of association of the adjustment cell variables with nonresponse and outcome (Little & Vartivarian, 2005).

Outline

1. What is the weight?
2. How do we weight?
3. To weight or not to weight?
4. Alternatives to weighting
5. Case study

Options

1. Whether to use the complete cases (CCs) that respond to all waves (including those who do not answer all questions) or the available cases (ACs) that respond to any wave.
2. Different weighting adjustments for attrition, depending on the response patterns and available information.

Example: MTF panel study

- The MTF panel study follows a sample of U.S. 12th graders with modal age of 18 every one or two years to study changes in the beliefs, attitudes, and behaviors regarding substance use and other health risks.
- The longitudinal follow-ups permit examination of developmental changes within cohorts.
- We analyzed four cohorts of MTF panel data to examine trajectories of substance use from late adolescence through young adulthood (Si, West, et al., 2022).
- The covariates were selected based on the substance use literature where the conditional interpretations adjusting other variables will be substantively meaningful.

Baseline and time-varying measures

- The longitudinal outcomes (Y_t) were repeated measures of the different substance use behaviors, such as binge drinking.
- The time-varying covariates (X_t) included follow-up wave indicators, full-time four-year college attendance, and marital status.
- Baseline characteristics (X_0, Y_0) included cohort indicators (2002-2005), age in months, sex, race/ethnicity, high school grades, parental education, baseline measures of corresponding substance outcomes, and reported intent to attend a four-year college, the total number of which is ~ 70.
- Design features (Z) include drug use reporting, sex, and geographical strata of schools.

Longitudinal weighting

- Suppose the study collects baseline measures (X_0, Y_0) and the number of T follow-up measures of the study variables (X_t, Y_t), for $t = 1, \cdots, T$.
- Suppose we are constructing weights for wave t and assume MAR applies to all nonresponse mechanisms at each wave.

1. A standard set of weights based on the baseline covariates (Z_i, X_{i0}, Y_{i0}) are inverses of the estimated response propensity **given the baseline information**,

$$w_{it}^{base} = \hat{\Pr}^{-1}(R_{it} = 1 \mid Z_i, X_{i0}, Y_{i0}). \tag{1}$$

Longitudinal weighting cont.

2. An alternative approach, which makes better use of available data, expresses the response probability as a product of conditional probabilities **given response at previous waves**.

These conditional probabilities can then condition on all available information in earlier waves. That is,

$$w_{i,1}^{seq} = \hat{\Pr}^{-1}(R_{i1} = 1 \mid Z_i, X_{i0}, Y_{i0})$$

$$w_{i2}^{seq} = w_{i,1}^{seq}\hat{\Pr}^{-1}(R_{i2} = 1 \mid Z_i, X_{i0}, Y_{i0}, X_{i1}, Y_{i1}).$$

Generally, for wave t,

$$w_{it}^{seq} = w_{i,t-1}^{seq}\hat{\Pr}^{-1}(R_{it} = 1 \mid Z_i, X_{i0}, Y_{i0}, X_{i1}, Y_{i1}, \cdots, X_{i,t-1}, Y_{i,t-1}). \quad (2)$$

This approach cannot use all data under an intermittent response pattern.

Alternative weighting methods

- Form cells with predicted response propensities: calculate the nonresponse adjustment factor within each cell:
  1. The inverse of average estimated propensity.
  2. The inverse of median estimated propensity.
  3. The inverse of observed response rate.

Machine learning algorithms

- Goal is to classify units as R or NR based on covariates available for all sample cases.
- Input data are the same as those for propensity modeling.
- The approach is an automatic black box.
  - High-order interactions of covariates are handled automatically.
  - Selection of covariates and associated interactions is done automatically.
  - Variable values, whether categorical or continuous, are combined (grouped) automatically.
- Form classes so that we have MAR, i.e., given the covariates that define classes, all units have the same response probability.

Algorithms used for propensity score prediction (Kern, Weib, & Kolb, 2021)

1. Penalized Logistic Regression: Logistic regression with a (lasso or ridge) penalty on the vector of regression coefficients.
2. Decision Trees: Recursive partitioning technique, repeatedly splits predictor space into homogeneous subgroups.
3. Random Forest: Tree-based ensemble method, grows decorrelated decision trees based on bootstrap samples.
4. Extremely Randomized Trees (ExtraTrees): Tree-based ensemble method, grows trees based on randomly sampled predictors and cut points at each node with the full training data.
5. Extreme Gradient Boosting (XGBoost): Sum of trees method, builds a sequence of trees akin to optimization via gradient descent with each tree attempting to improve over the result of its predecessor.

Quality check of weights

1. Produce summary statistics, e.g., mean, median, minimum, maximum, and design effect.
   - The variability of weights: the **design effect** due to weighting, $1 + cv^2(w)$, where the coefficient of variation $cv(w) = sd(w)/mean(w)$, provides an upper bound of precision loss.
2. Extreme values of weights: trimming or winsorization.
3. Compute weighted frequencies of key survey variables.
4. Compare weighted frequencies to reliable external totals.
5. Check balance on key covariates.

Outline

1. What is the weight?
2. How do we weight?
3. To weight or not to weight?
4. Alternatives to weighting
5. Case study

Analysis goals

When considering approaches to the analysis of longitudinal survey data, it is quite important to make the analysis objectives very clear:

- Descriptive estimation at a single wave?
- Estimation of change across two waves?
- Trajectory estimation based on three or more waves?

Example: Substance use trajectory modeling

- We considered marginal (or population-averaged) logistic regression models for a binary outcome variable Y, where $Y_{it} = 1$ if individual i indicated substance use at wave t, and otherwise, $Y_{it} = 0$.

$$logit(\Pr(Y_{it} = 1)) = \beta_0 + \vec{\beta}_1 x_{i0} + \vec{\beta}_2 x_{it} + \vec{\beta}_3 wave_{it} + \vec{\beta}_4 wave_{it} * coll_{it}. \quad (3)$$

- Here the marginal model includes the primary coefficients of interest that identify possible differences in trajectories based on college status: the coefficients $\vec{\beta}_4$ for the interaction $wave_{it} * coll_{it}$ between college attendance indicator ($coll_{it}$) and wave ($wave_{it}$).

Estimation methods

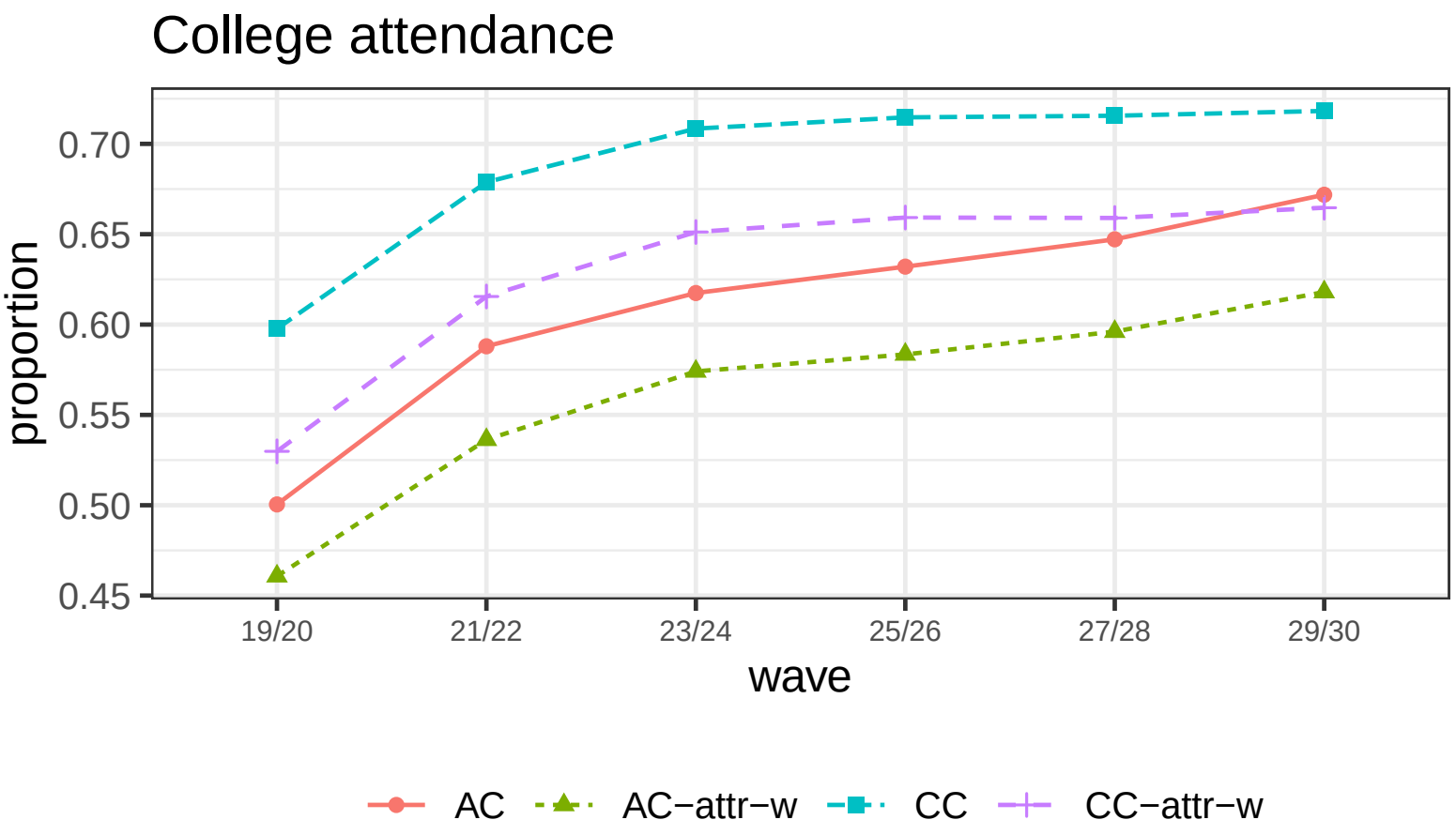
1. CC analysis or AC analysis.
2. Accounting for the correlation of repeated measures on the same individual, i.e., clustering via the individual numeric identifiers (MTF ID).
   - Generalized estimating equations (GEE)
3. Whether to apply weighting adjustments for attrition.

Eight approaches investigated

Method	CC	AC	Selection features	Correlation	Attrition-adjusted weight
CC	√		√		
AC		√	√		
CC-gee	√		√	√	
AC-gee		√	√	√	
CC-ID cluster	√		√	√	
AC-ID cluster		√	√	√	
CC-attr-w	√		√	√	√
AC-attr-w		√	√	√	√

CC: complete case analysis; AC: available case analysis; CC-gee: GEE with complete cases; AC-gee: GEE with available cases; CC-ID cluster: complete case analysis accounting for the cluster structure by individuals; AC-ID cluster: available case analysis accounting for the cluster structure by individuals; CC-attr-w: attrition-adjusted weighted analysis accounting for the cluster structure by individuals with complete cases; and AC-attr-w: attrition-adjusted weighted analysis accounting for the cluster structure by individuals with available cases.

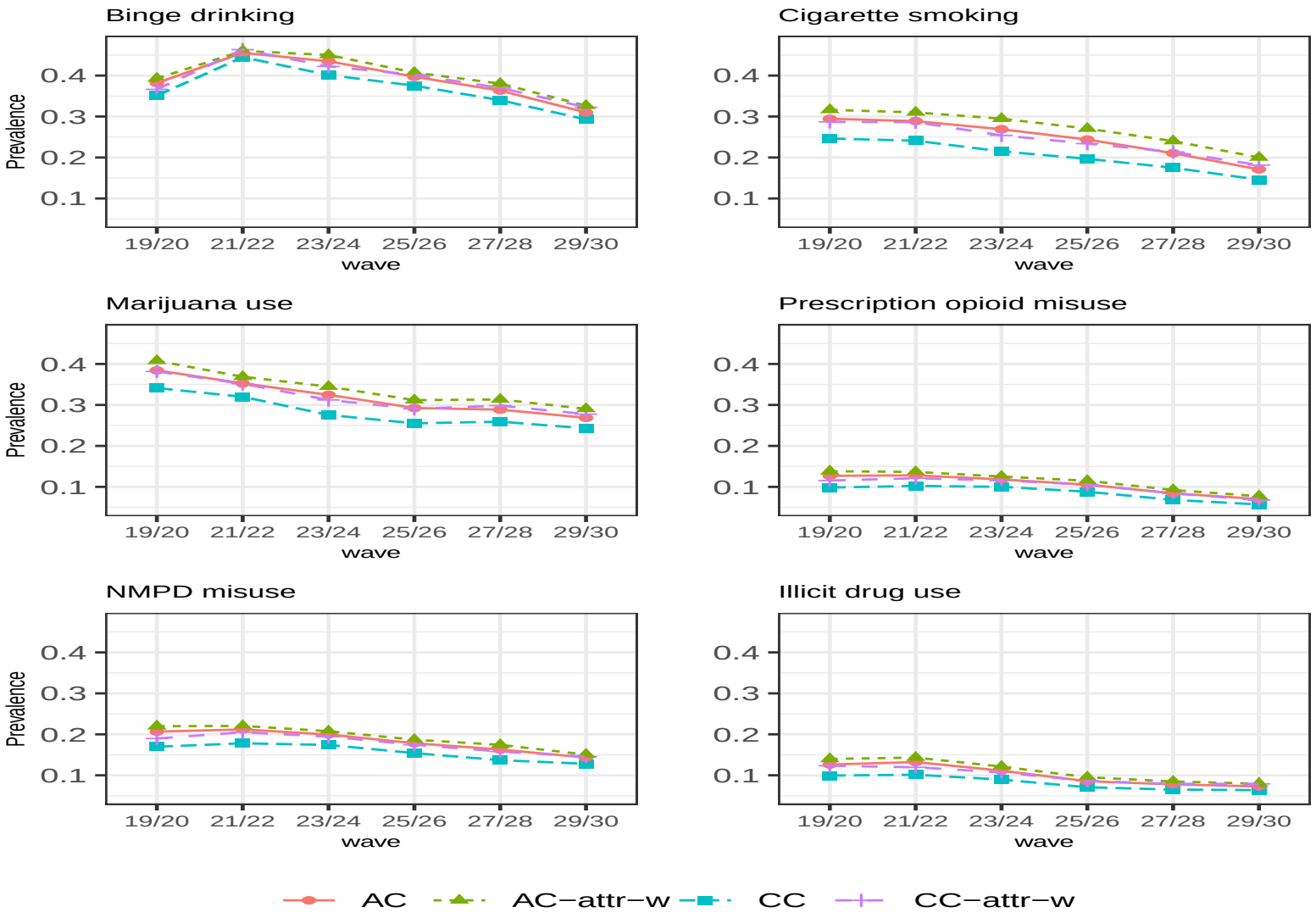
Weights change descriptive summaries



Findings: (I)

- The CC analysis yielded the highest estimated proportions of individuals attending college, around 60-72%, while the AC analysis yielded estimates around 50-67%.
- The attrition-adjusted, weighted analyses reduced both estimates, showing that the attritors were less likely to attend college across young adulthood.

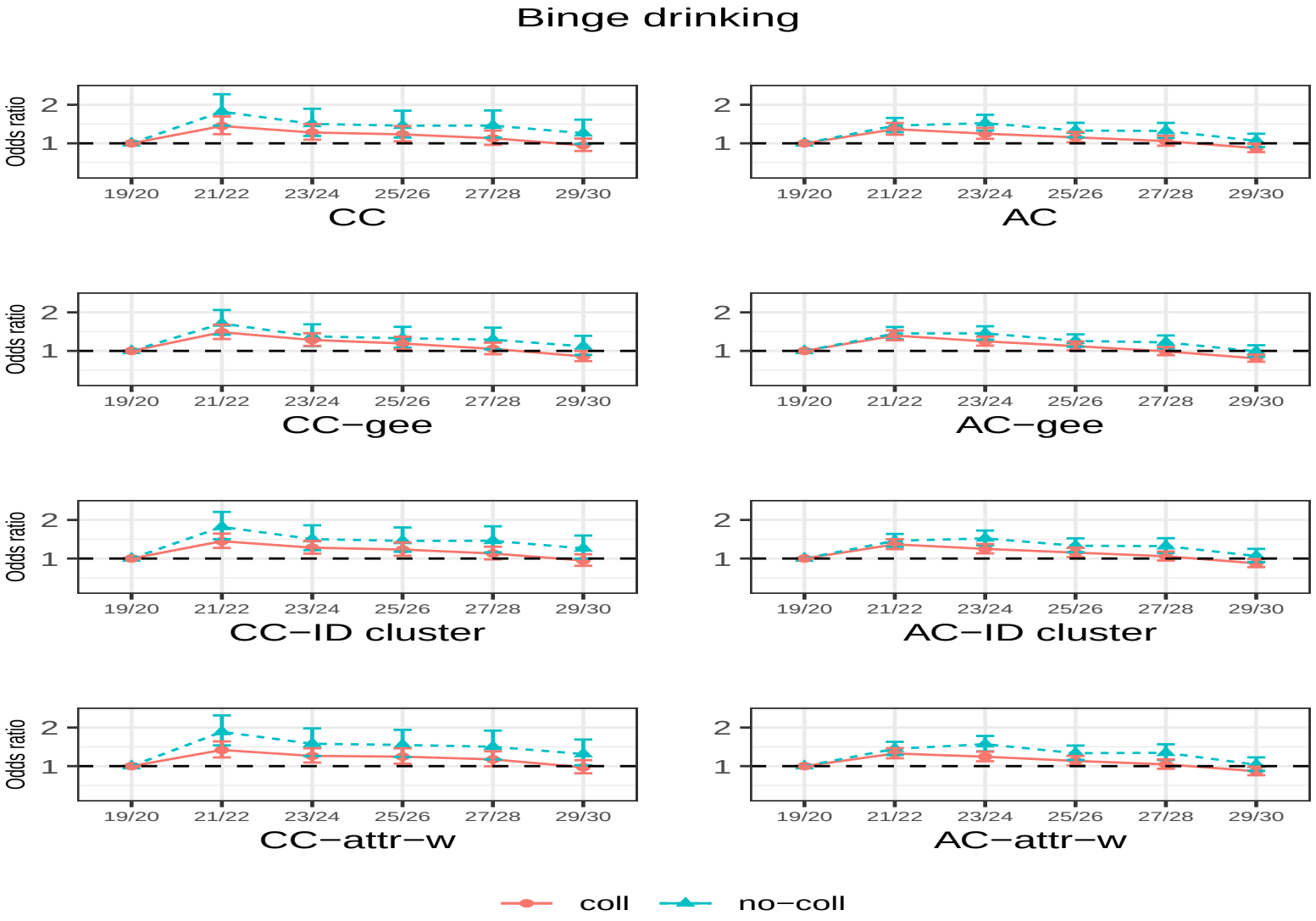
Weights change descriptive summaries cont.



Findings: (II)

- Overall, across the six different substance use outcomes, the AC estimates were higher than the CC estimates with varying trajectories over time;
- Attrition-adjusted weighting increased both estimates, and the effect was larger for CC than for AC.
- This indicates that the attriters who missed follow-up waves tended to be substance users, and the CC analysis underestimated prevalence as a result.

Weights negligibly affect trajectory modeling



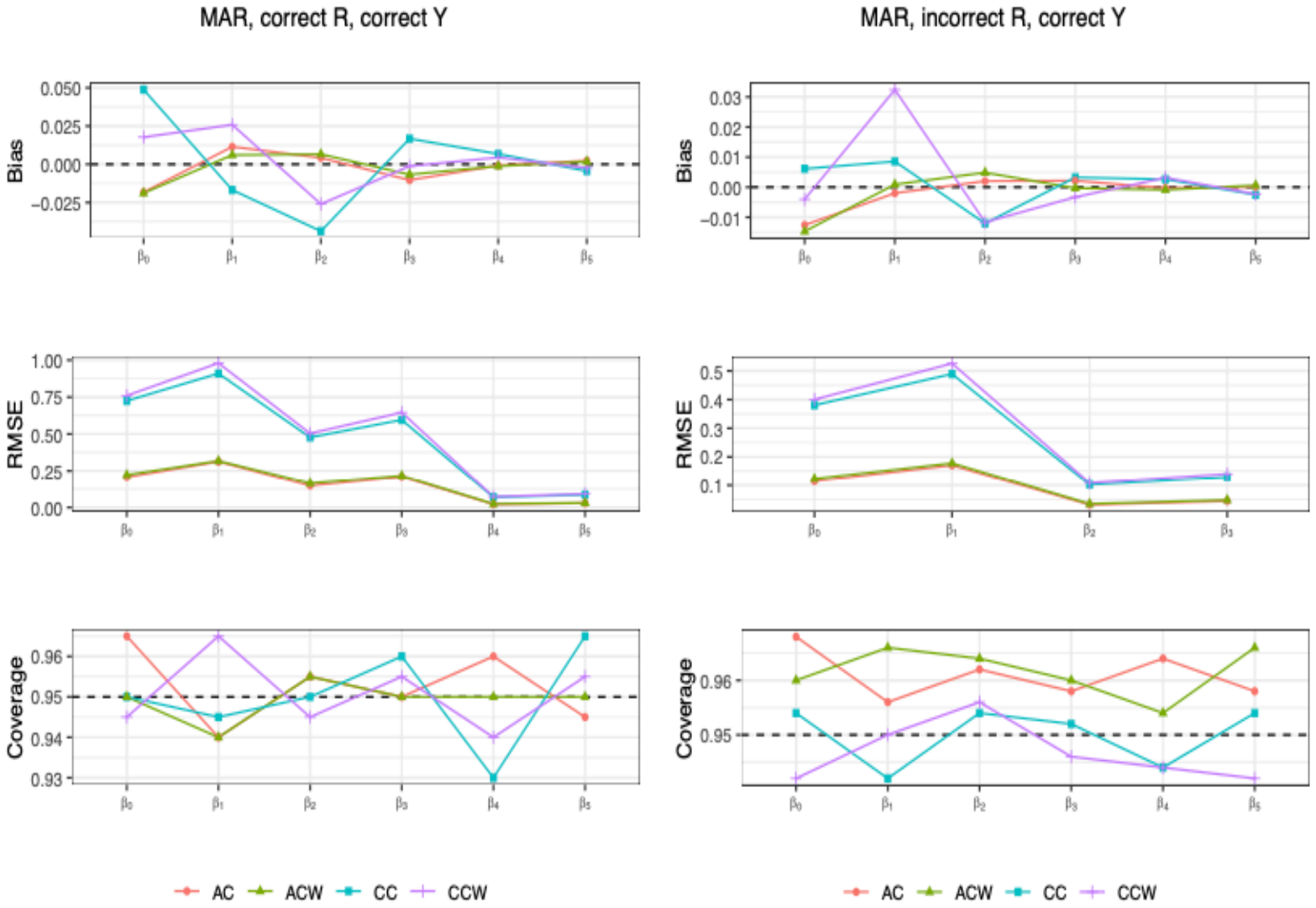
Findings: (III)

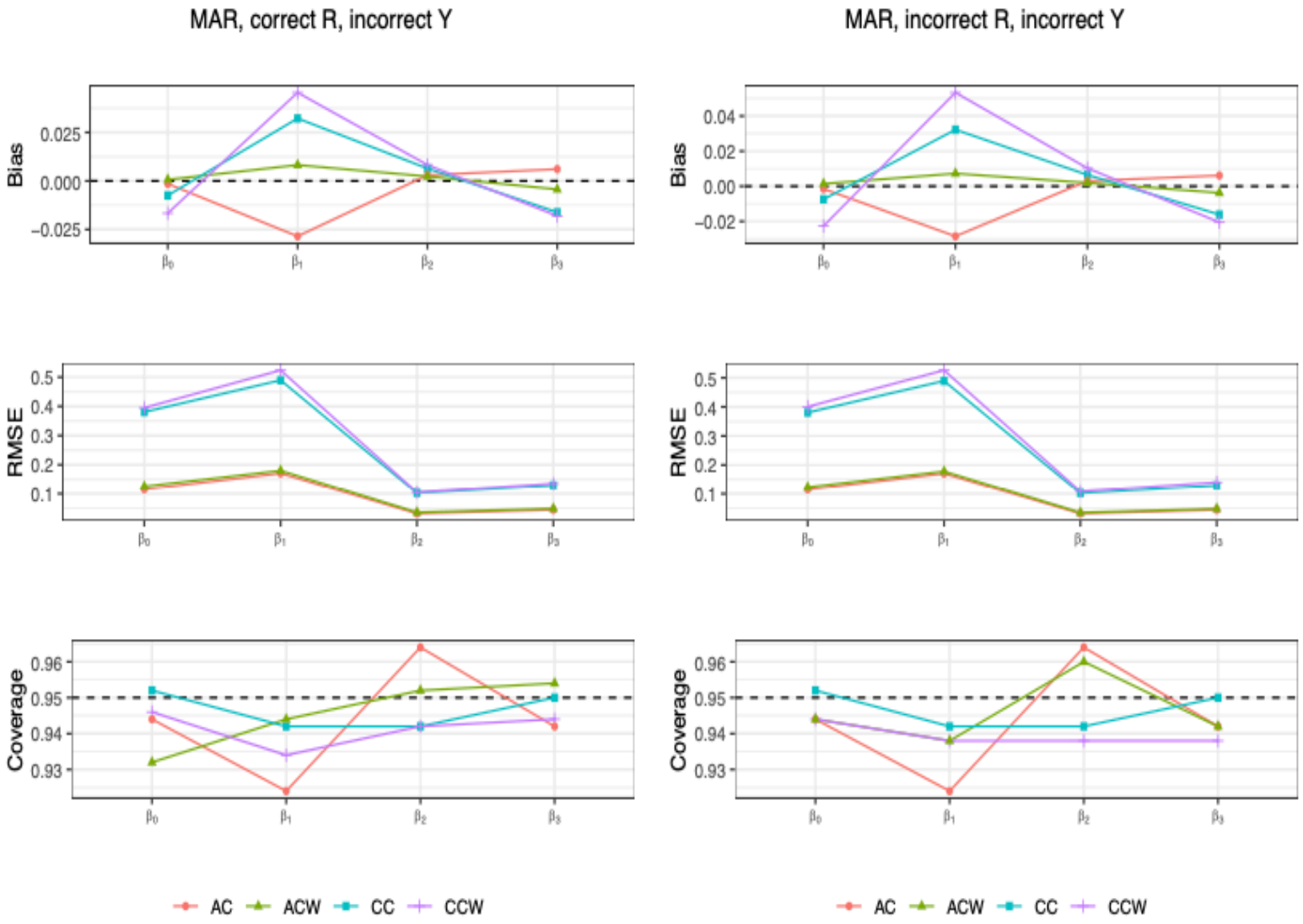
- The prevalence of binge drinking followed a non-linear trend, first increasing through age 23/24 in the AC analyses, and then decreasing after that through age 29/30.
- The attrition-adjusted AC analysis showed that college attendance reduced the prevalence of binge drinking at age 23/24 and age 27/28.
- All CC analyses showed that the peak was at age 21/22 and did not display any substantial differences between college attenders and non-attenders.
- The AC analyses yielded more efficient trajectory estimates of binge drinking with smaller variances than the CC analyses, and weighting slightly changed the estimates.

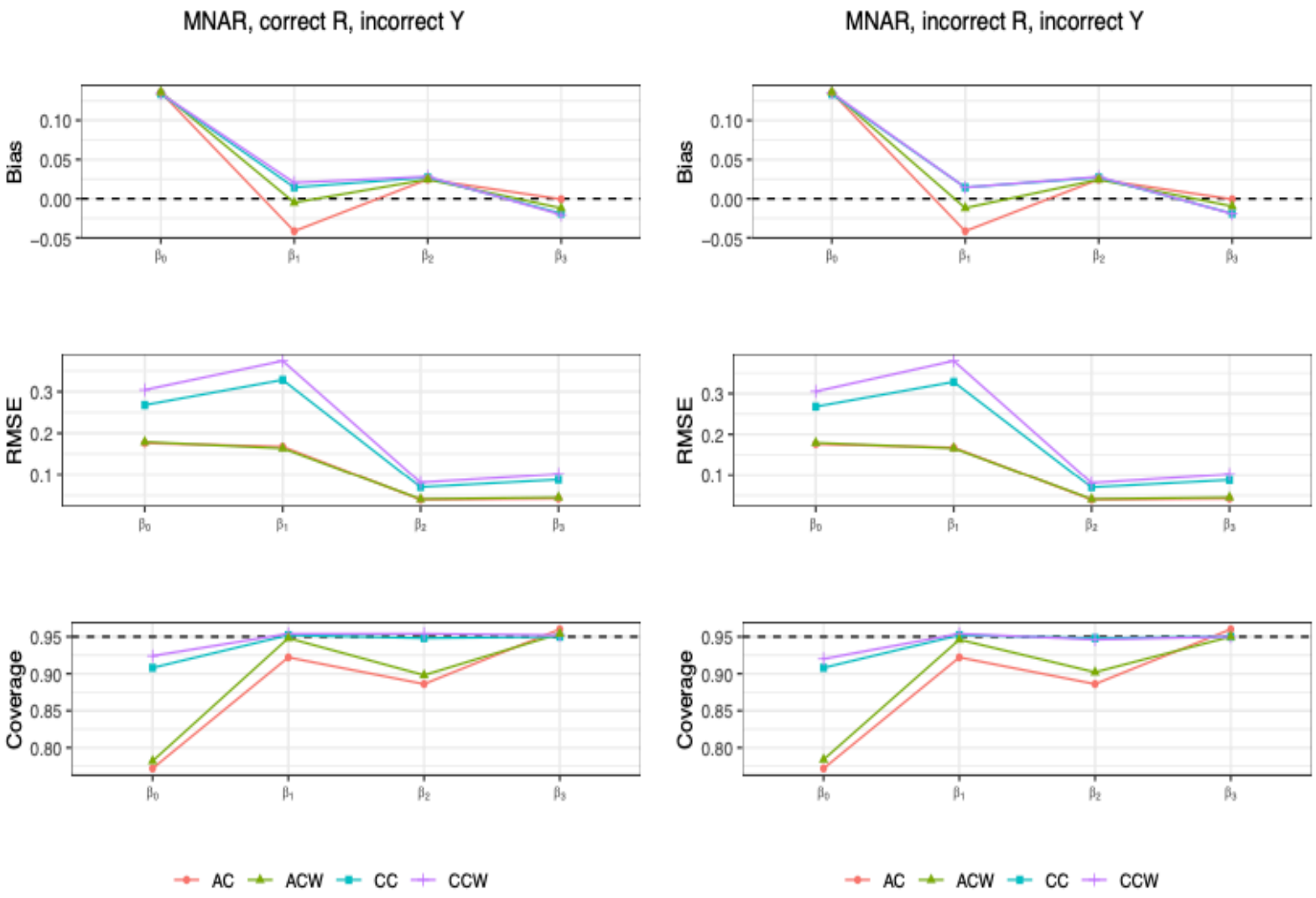
Simulation studies

Empirical results cannot be validated, so we use simulation studies to confirm the findings given true values.

1. MNAR vs. MAR: whether or not the missing data mechanism is non-ignorable;
2. Correct or incorrect trees for nonresponse adjustment; and
3. Correct or incorrect specification of the substantive model for the outcome Y.


MAR, correct R, correct Y
MAR, incorrect R, correct Y
Bias
RMSE
Coverage
AC
ACW
CC
CCW


MAR, correct R, incorrect Y
MAR, incorrect R, incorrect Y
Bias
RMSE
Coverage
AC
ACW
CC
CCW


MNAR, correct R, incorrect Y
MNAR, incorrect R, incorrect Y
Bias
RMSE
Coverage
AC
ACW
CC
CCW

Summary

- Overall, we recommend AC analysis and account for clustering in longitudinal trajectory modeling.
- Analyst effort should be dedicated to improving the specification of the model for the survey outcome of interest.
- With correctly specified models, weights will only reduce precision without changing the estimates
- With a misspecified model, the role of the weights depends.

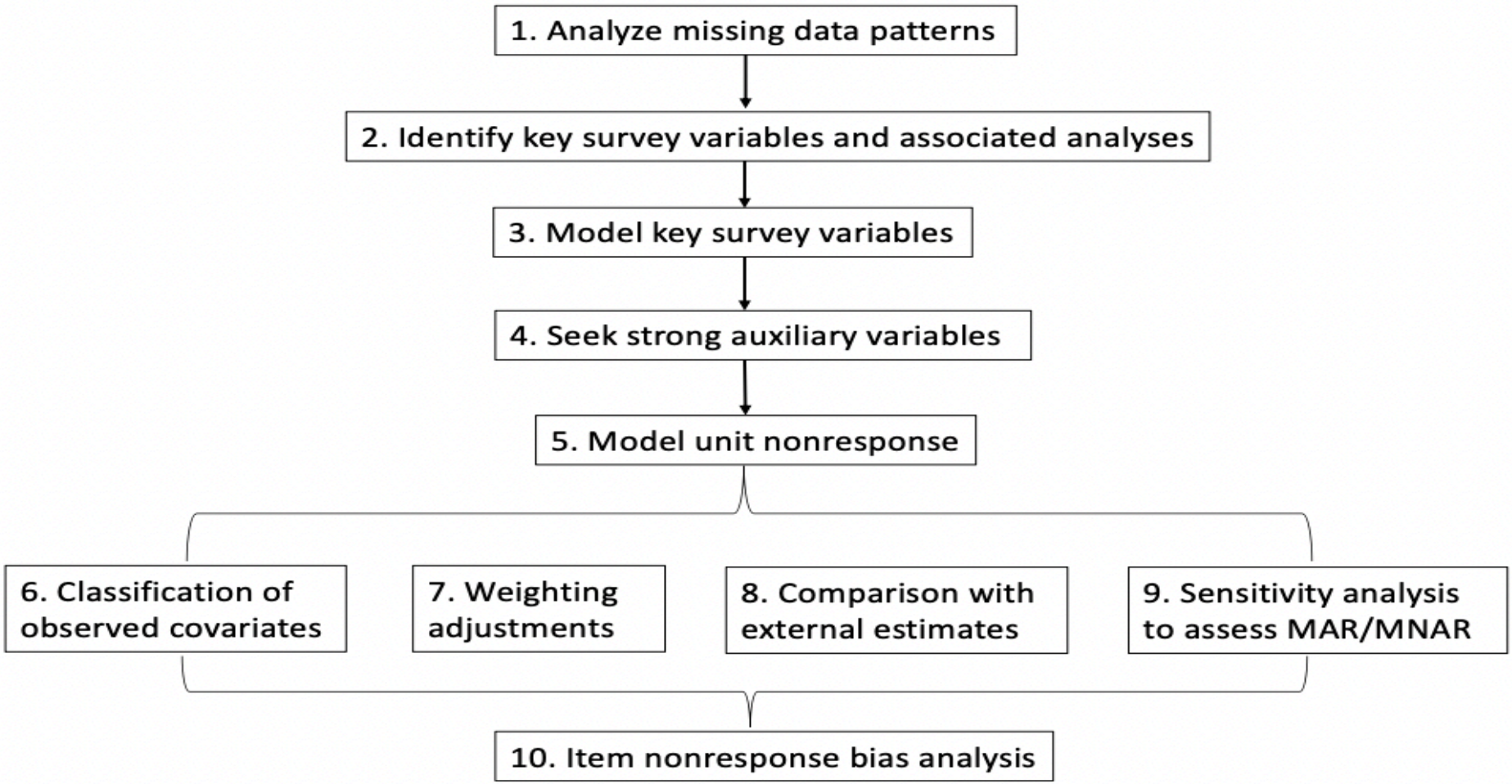
Compare weighted and unweighted results.

- Large changes in parameter estimates suggest model misspecification, and that at the very least weights should be used to obtain robust estimates. Recommend further investigations.
- Small changes in parameter estimates and large changes in standard errors would suggest an appropriate specification, and minimal problems with using unweighted models.
- Check sensitivity of results to different approaches.
- In practice, fit an unweighted model that includes all predictors of interest, and add 1) the weight, and 2) all two-way interactions between the weights and other predictors.
    - Perform a Wald test of the additional coefficients: Significant-weights important; not significant-ignore weights.

Outline

1. What is the weight?
2. How do we weight?
3. To weight or not to weight?
4. Alternatives to weighting
5. Case study

Nonresponse bias analysis (Si, Little, et al., 2022a)



Multiple imputation (MI)

- Weight construction requires that covariates used in the response propensity modeling are fully observed; however, in practice both the outcome and covariates are subject to item nonresponse.
- Imputation of missing data allows these variables to be included in the weights, and is an attractive approach.
- To propagate the uncertainty of the imputation model, MI can be applied to simultaneously handle unit and item nonresponse.
- MI applies either joint or sequential conditional imputation models to generate multiple completed datasets and use MI combining rules for inferences about parameters or population quantities (e.g., Raghunathan, Lepkowski, van Hoewyk, & Solenberger (2001))

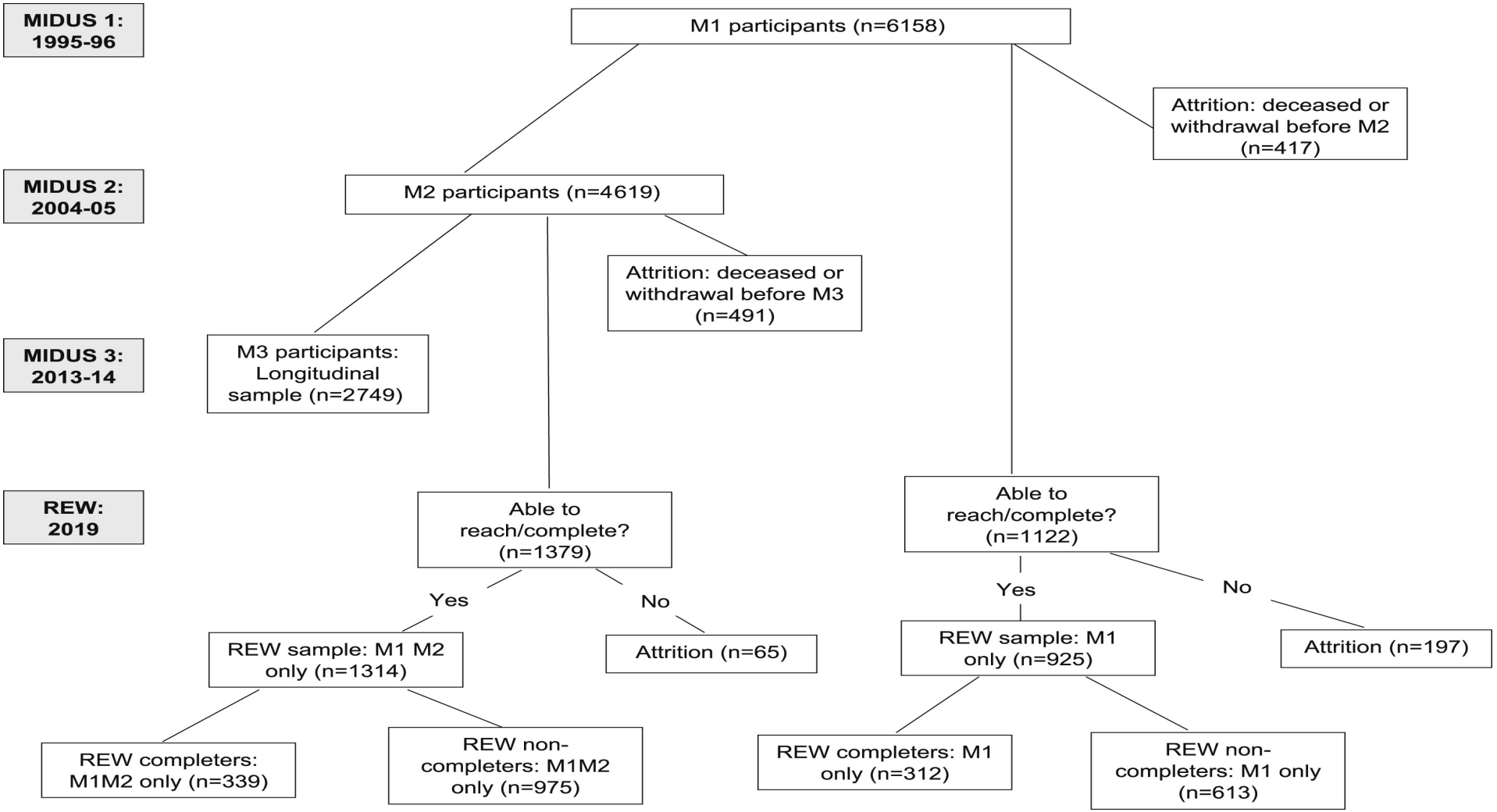
	Association of X with Outcome Y			
Association of X with Response R	**Propensity: Low Other Xs: Low**	**Propensity: Low Other Xs: High**	**Propensity: High Other Xs: Low**	**Propensity: High Other Xs: High**
Low	Cell LLL IPW MI Bias: --- --- Var: --- ---	Cell LLH IPW MI Bias --- --- Var --- ↓	Cell LHL IPW MI Bias: --- --- Var: ↓ ↓	Cell LHH IPW MI Bias: --- --- Var: ↓ ↓↓
High	Cell HLL IPW MI Bias: --- --- Var: ↑ ---	Cell HLH IPW MI Bias: --- --- Var: ↑ ↓	Cell HHL IPW MI Bias: ↓ ↓ Var: ↓ ↓	Cell HHH IPW MI Bias: ↓ ↓ Var: ↓ ↓↓

Table 2. Bias and Variance of Multiple Imputation (MI) and Inverse Propensity Weighting (IPW) Relative to Complete Case (CC) analysis for Estimating a Mean, by Strength of Association of the Auxiliary Variables with Response (R) and Outcome (Y). The X is split into the propensity, which is the best predictor of R in the regression of R on X, i.e., Propensity, and components of X orthogonal to the propensity, i.e., Other Xs (Little, Carpenter, & Lee, 2022).

Sensitivity analysis

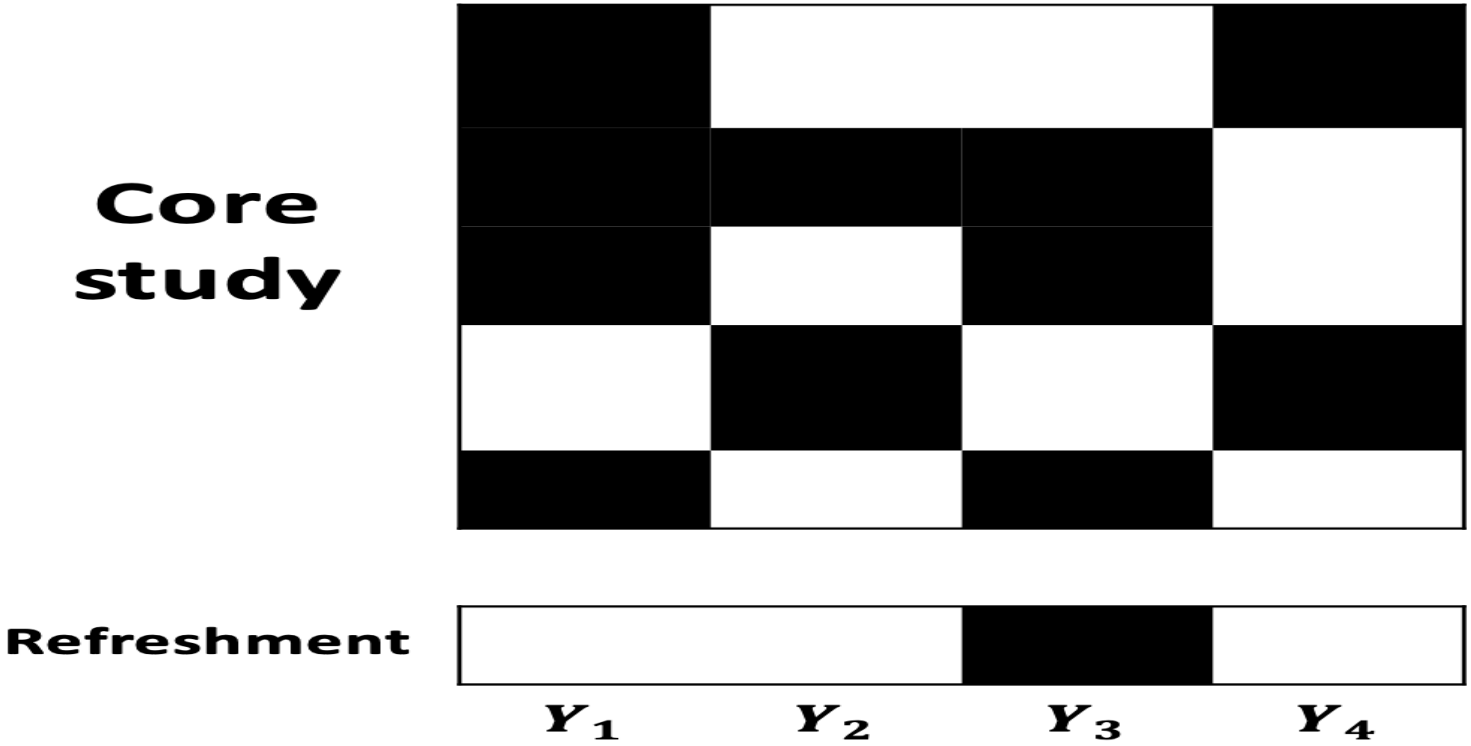
- The methods discussed so far generally assume MAR, and MNAR models require untestable assumptions to identify parameter estimates.
- Therefore, we recommend a sensitivity analysis to assess the impact of MNAR deviations from MAR.
  1. Apply an offset to MAR imputations to model departures from MAR (e.g., Si, Little, Mo, & Sedransk, 2022b).
  2. Proxy pattern mixture models, e.g., Little, West, Boonstra, & Hu (2020); West et al. (2021).

Retention New Warning samples in the Midlife in the U.S.(MIDUS) study (Song et al., 2021)



Refreshment samples

- Collect a new random sample from the population in one of or multiple subsequent waves



- E.g., Si (2012); Deng, Hillygus, Reiter, Si, & Zheng (2013); Si, Reiter, & Hillygus (2015); Si, Reiter, & Hillygus (2016).

Outline

1. What is the weight?
2. How do we weight?
3. To weight or not to weight?
4. Alternatives to weighting
5. Case study

HRS example

- The Health and Retirement Study (HRS) data file contains the sample design information in the baseline wave (2006), including the survey weights, KWGTR, the stratum variable, STRATUM, and the sampling error computation unit variable, SECU, for each household head, HHID.
- One measure of interest in each of the four waves (2006, 2008, 2010, and 2012) is the indicator of DIABETES (1: yes, 0: no).
- The AGE variable represents the age (in years) of the household head across the four waves.
- Also available for responding financial reporters in 2006: gender (1: male, 2: female), years of education (1: 0-11, 2: 12, 3: 13-15, 4: 16+) and race/ethnicity (1: Other, 2: Non-Hispanic White, 3: Non-Hispanic Black).

Key steps

1. Identify response patterns
2. Predict wave-specific response propensities

```
AC_3=glm(R_12 ~ X, data = subset(hrs_sub, R_08==1 & R_10==1), family = "binomial")
ac3_lg_p=predict(AC_3, type = "response")
```

3. Account for design features in analysis

```
ac_svy=svydesign(strata=~STRATUM, id=~HHID, weights=~AC_W, data = hrs_long)
ac_svy_aw=svyglm(MODEL-Y, design=ac_svy, family=quasibinomial)
```

R code

- Refer to the file with step-by-step illustrations: HRScasestudy.R

Other software

- Stata:
  - **svyset HHID [pweight = AC_W], strata(STRATUM)**
  - svy:logistic DIABETES i.GENDER i.EDCAT i.RACECAT AGE i.WAVE
- SAS:
  - proc surveylogistic data=hrs_long ;
  - class GENDER EDCAT RACECAT WAVE ;
  - **strata STRATUM** ;
  - **cluster HHID** ;
  - **weight AC_W** ;
  - model DIABETES = GENDER EDCAT RACECAT AGE WAVE / solution deff;
  - run ;

Resources

- Applied Survey Data Analysis, by Heeringa, West and Berglund, http://www.isr.umich.edu/src/smp/asda/#Survey_Data_Analysis_Publications
- Survey software list: https://www.hcp.med.harvard.edu/statistics/survey-soft/#Packages

Use weights (and every other statistical approach!!!) with caution

Statistical Science
2007, Vol. 22, No. 2, 153–164
DOI: 10.1214/088342306000000691
© Institute of Mathematical Statistics, 2007

Struggles with Survey Weighting and Regression Modeling[1]

Andrew Gelman

Abstract. The general principles of Bayesian data analysis imply that models for survey responses should be constructed conditional on all variables that affect the probability of inclusion and nonresponse, which are also the variables used in survey weighting and clustering. However, such models can quickly become very complicated, with potentially thousands of poststratification cells. It is then a challenge to develop general families of multilevel probability models that yield reasonable Bayesian inferences. We discuss in the context of several ongoing public health and social surveys. This work is currently open-ended, and we conclude with thoughts on how research could proceed to solve these problems.

Key words and phrases: Multilevel modeling, poststratification, sampling weights, shrinkage.

1. BACKGROUND

Survey weighting is a mess. It is not always clear how to use weights in estimating anything more complicated than a simple mean or ratios, and standard er-

the motivation for hierarchical regression, combined with poststratification, as a strategy for correcting for differences between sample and population. We sketch some directions toward a practical solution, which unfortunately has not yet been reached.

Challenge: Multilevel/mixed-effects models

- We have considered marginal approaches, e.g., weighted GEE, for population inference in correlated data modeling.
- Weighted maximum likelihood (ML) estimates maximize the pseudo-likelihood function, which weights each individual contribution to the likelihood by the power of the inverse of its estimated response propensity.
- Rabe-Hesketh & Skrondal (2006) show that the weighted likelihood function for a multilevel model requires the inverse-probability weights at each level of the data hierarchy (e.g., Level 1 weights and Level 2 weights.).
- However, the ML computation is challenging and hard to converge.
- Recently new developments have been using Bayesian approaches.

Our ongoing work

1. Multilevel regression and poststratification (MrP, Si, Pillai, & Gelman (2015))
2. Account for weights in multilevel models: an MrP solution
3. Why wait to weight? Use weights to inform adaptive sample recruitment (HEALthy Brain and Child Development Study: https://hbcdstudy.org)

Acknowledgements

- Collaborators: R. Little, B. West, S. Heeringa, S. McCabe, et al.
- Funding support from
  - National Institutes of Health
  - National Science Foundation
  - National Institute of Statistical Sciences
  - National Center for Education Statistics
  - U.S. Department of Agriculture

Q & A

yajuan@umich.edu

References I

Deng, Y., Hillygus, D. S., Reiter, J. P., Si, Y., & Zheng, S. (2013). Handling attrition in longitudinal studies: The case for refreshment samples. Statistical Science, 22, 238–256.

Elliott, M. R. (2013). Combining data from probability and non-probability samples using pseudo-weights. Survey Practice, 2(6).

Kern, C., Weib, B., & Kolb, J.-P. (2021). Predicting nonresponse in future waves of a probability-based mixed-mode panel with machine learning. Journal of Survey Statistics and Methodology, smab009.

Little, R. J. A., Carpenter, J., & Lee, K. (2022). A comparison of three popular methods for handling missing data: Complete-case analysis, weighting and multiple imputation. Under review.

Little, R. J. A., & Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means? Survey Methodology, 31(2), 161–168.

Little, R. J. A., West, B. T., Boonstra, P. S., & Hu, J. (2020). Measures of the degree of departure from ignorable sample selection. Journal of Survey Statistics and Methodology, 8(5), 932–964.

Morgan, J. N., & Smith, J. D. (1969). A PANEL STUDY OF INCOME DYNAMICS: STUDY DESIGN, PROCEDURES, AND FORMS. Institute for Social Research, the University of Michigan, Ann Arbor, Michigan. Retrieved from https://psidonline.isr.umich.edu/Data/Documentation/pdf_doc/psid68w1.pdf

Rabe-Hesketh, S., & Skrondal, A. (2006). Multilevel modelling of complex survey data. Journal of the Royal Statistical Society Series A, 169(4), 805–827.

References II

Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Methodology, 27, 85–96.

Rubin, D. B. (1976). Inference and missing data (with discussion). Biometrika, 63, 581–592.

Si, Y. (2012). Nonparametric Bayesian methods for multiple imputation of large scale incomplete categorical data in panel studies (PhD thesis). Duke University, https://stat.duke.edu/node/14480.

Si, Y., Little, R. J. A., Mo, Y., & Sedransk, N. (2022a). A case study of nonresponse bias analysis in educational assessment surveys. https://arxiv.org/abs/2104.04432.

Si, Y., Little, R., Mo, Y., & Sedransk, N. (2022b). Nonresponse bias analysis in longitudinal educational assessment studies. https://arxiv.org/abs/2204.07105.

Si, Y., Pillai, N. S., & Gelman, A. (2015). Bayesian nonparametric weighted sampling inference. Bayesian Analysis, 10(3), 605–625.

Si, Y., Reiter, J. P., & Hillygus, D. S. (2015). Semi-parametric selection models for potentially non-ignorable attrition in panel study with refreshment sample. Political Analysis, 23, 92–112.

Si, Y., Reiter, J. P., & Hillygus, D. S. (2016). Bayesian latent pattern mixture models for handling attrition in panel studies with refreshment samples. The Annals of Applied Statistics, 10, 118–143.

References III

Si, Y., West, B. T., Veliz, P., Patrick, M. E., Schulenberg, J. E., Kloska, D. D., . . . McCabe, S. E. (2022). An empirical evaluation of alternative approaches to adjusting for attrition when analyzing longitudinal survey data on young adults' substance use trajectories. International Journal of Methods in Psychiatric Research, 31(3), e1916.

Song, J., Radler, B. T., Lachman, M. E., Mailick, M. R., Si, Y., & Ryff, C. D. (2021). Who returns? Understanding varieties of longitudinal participation in MIDUS. Journal of Aging and Health, 33(10), 896–907.

Valliant, R., Dever, J. A., & Kreuter, F. (2018). Practical tools for designing and weighting survey samples (2nd ed.). Springer, New York.

West, B. T., Little, R. J. A., Andridge, R. R., Boonstra, P. S., Ware, E. B., Pandit, A., & Alvarado-Leiton, F. (2021). Assessing selection bias in regression coefficients estimated from non-probability samples, with applications to genetics and demographic surveys. Annals of Applied Statistics, 15(3), 1556–1581.