murdoch
children's
research
institute

THE UNIVERSITY OF
MELBOURNE

Handling missing data when estimating causal effects with Targeted Maximum Likelihood Estimation

Ghazaleh Dashti, Katherine Lee, Julie Simpson, Ian White, John Carlin, Margarita Moreno-Betancur

VICTORIAN CENTRE FOR BIOSTATISTICS

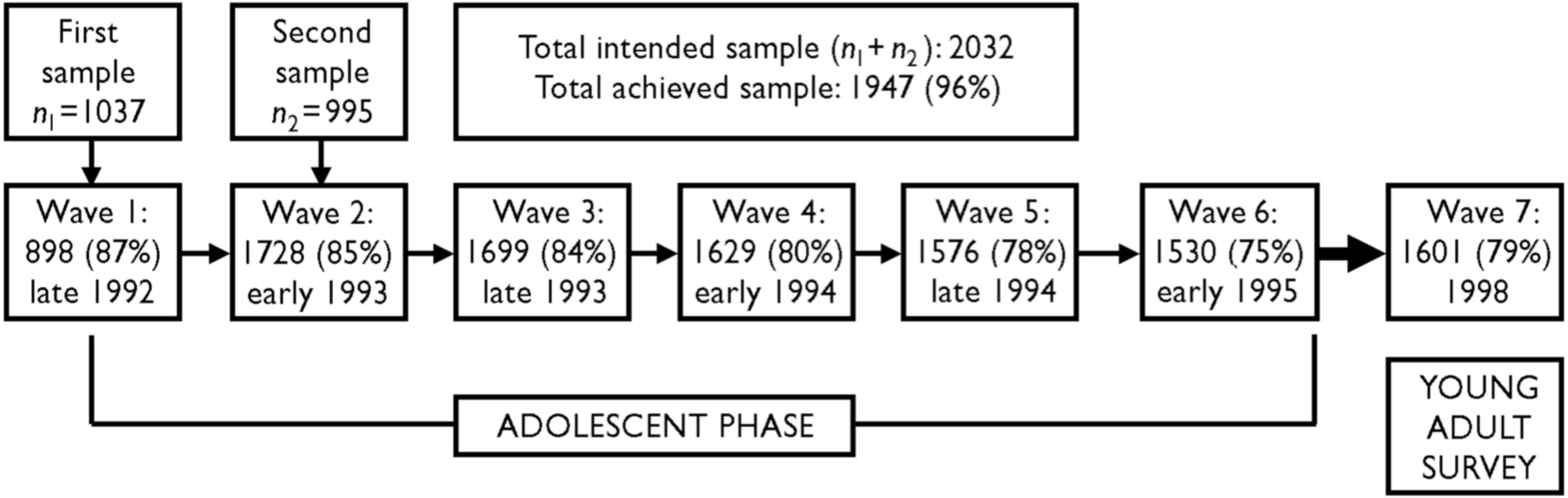
ViCBiostat

Outline

- Introduction
  - Motivating example
  - Project aim
- Simulation study
  - Generating the complete data
  - Imposing missing data
  - Missing data methods
  - Results
- Application to the VAHCS case study
- Concluding remarks

Introduction: motivating example

Based on the Victorian Adolescent Health Cohort Study



First sample $n_1 = 1037$

Second sample $n_2 = 995$

Total intended sample ($n_1 + n_2$): 2032
Total achieved sample: 1947 (96%)

Wave 1: 898 (87%) late 1992

Wave 2: 1728 (85%) early 1993

Wave 3: 1699 (84%) late 1993

Wave 4: 1629 (80%) early 1994

Wave 5: 1576 (78%) late 1994

Wave 6: 1530 (75%) early 1995

Wave 7: 1601 (79%) 1998

ADOLESCENT PHASE

YOUNG ADULT SURVEY

Introduction: motivating example

Question: What is the causal effect of frequent cannabis use in adolescent females on mental health in young adulthood?

	Variable (notation)	Collected at
Outcome	Std log(mental health score) (Y)	Wave 7
Exposure	Frequent cannabis use (X)	Wave 2-6
Confounders	Parental divorce (Z1)	
	Antisocial behavior (Z2)	
	Depression & anxiety (Z3)	
	Alcohol use (Z4)	
	Parental education (Z5)	
Auxiliary	Age (A)	Wave 2

Introduction: counterfactuals and causal parameter

Causal parameter:

Average causal effect (ACE) = $\mathrm{E}(\mathrm{Y}^{\mathrm{x}=1}) - \mathrm{E}(\mathrm{Y}^{\mathrm{x}=0})$

$\mathrm{Y}^{\mathrm{x}=1}$ and $\mathrm{Y}^{\mathrm{x}=0}$ are the potential mental health scores

when exposed $(\mathrm{x} = 1)$ & not exposed $(\mathrm{x} = 0)$

Under identifiability assumptions of exchangeability, consistency, and positivity, ACE can be identified from the observable data by:

$$\mathrm{E}(\mathrm{E}(\mathrm{Y}|\mathrm{X} = 1, \mathrm{Z}) - \mathrm{E}(\mathrm{Y}|\mathrm{X} = 0, \mathrm{Z}))$$

Introduction: identifiability assumptions

Exchangeability: $\mathrm{Y}^{\mathrm{x}} \coprod \mathrm{X}|\mathrm{Z}$ for all x

Consistency: $\mathrm{Y}^{\mathrm{x}} = \mathrm{Y}$ when X=x

Positivity: $\mathrm{P}[\mathrm{X} = \mathrm{x}|\mathrm{Z} = \mathrm{z}] > 0$ for all x and z

Introduction: estimation

- Singly robust methods (either model for outcome or exposure)
  - Outcome regression
  - G-computation
  - Inverse probability weighting

- Doubly robust methods (combine outcome and exposure models)
  - Augmented inverse probability weighting
  - **Targeted maximum likelihood estimation**

Introduction: TMLE (1)

1. Predict outcome for all when exposed and unexposed ($\widehat{\mathrm{E}}(\mathrm{Y}|\mathrm{X},\mathrm{Z})$)
2. Estimate the propensity score $\widehat{\mathrm{P}}(\mathrm{X}=1|\mathrm{Z})$
3. Incorporate information from $\widehat{\mathrm{P}}(\mathrm{X}=1|\mathrm{Z})$ to update $\widehat{\mathrm{E}}(\mathrm{Y}|\mathrm{X},\mathrm{Z})$
4. Plug in the updated predications ($\widehat{\mathrm{E}}^*(\mathrm{Y}|\mathrm{X},\mathrm{Z})$) in the G-formula

$$\widehat{\mathrm{ACE}} = \widehat{\mathrm{E}}(\widehat{\mathrm{E}}^*(\mathrm{Y}|\mathrm{X}=1,\mathrm{Z}) - \widehat{\mathrm{E}}^*(\mathrm{Y}|\mathrm{X}=0,\mathrm{Z}))$$

Introduction: TMLE (2)

Targeting step:

1. For each individual calculate a clever covariate $\widehat{\mathrm{H}}$ as a function of the $\widehat{\mathrm{P}}(\mathrm{X} = 1|\mathrm{Z})$
2. Regress residuals from the initial outcome model on $\widehat{\mathrm{H}}$ to estimate $\hat{\varepsilon}$
3. Upate the initial estimate: $\mathrm{logit}(\widehat{\mathrm{E}}^{*}(\mathrm{Y}|\mathrm{X},\mathrm{Z})) = \mathrm{logit}(\widehat{\mathrm{E}}\ (\mathrm{Y}|\mathrm{X},\mathrm{Z})) + \hat{\varepsilon}\widehat{\mathrm{H}}$

Introduction: several desirable properties of TMLE

- Double robust: consistent if either $\widehat{\mathrm{E}}(\mathrm{Y}|\mathrm{X},\mathrm{Z})$ or $\widehat{\mathrm{P}}(\mathrm{X}=1|\mathrm{Z})$ consistently estimated

- Asymptotically linear

- Asymptotically efficient if both consistently estimated (under the Donsker class condition)

- If data adaptive approaches used for exposure & outcome models:
  - Optimizes the bias-variance trade-off
  - Offers asymptotically valid confidence intervals

Introduction: target analysis

- The average causal effect estimated using TMLE with Super Learner for the exposure and outcome models

- Super Learner library included:

mean, glm, glm.interaction, bayesglm, gam, glmnet, earth, rpart, rpartPrune, ranger

- We used the TMLE package in R

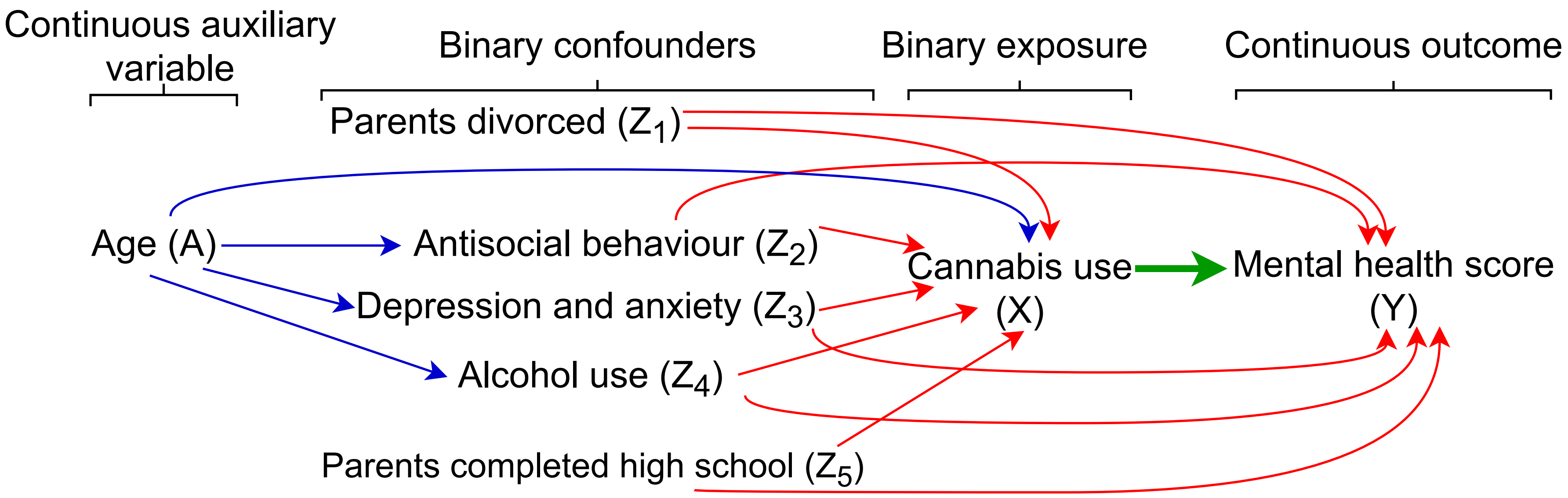
Introduction: aim

	Variable (notation)	%; mean (SD)	% with missing values
Outcome	Std mental health score (Y)	0 (1)	13
Exposure	Frequent cannabis use (X)	12	31
Confounders	Parental divorce (Z1)	22	0.1
	Antisocial behavior) (Z2)	15	27
	Depression & anxiety) (Z3)	60	14
	Alcohol use (Z4)	37	21
	Parental education (Z5)	38	3
	Any missing	**-**	**40**

Project aim: evaluate the performance of available approaches for dealing with missing data when using TMLE to estimate the ACE

Simulation study: generating the complete data (1)

- Used the following causal diagram in data generation:

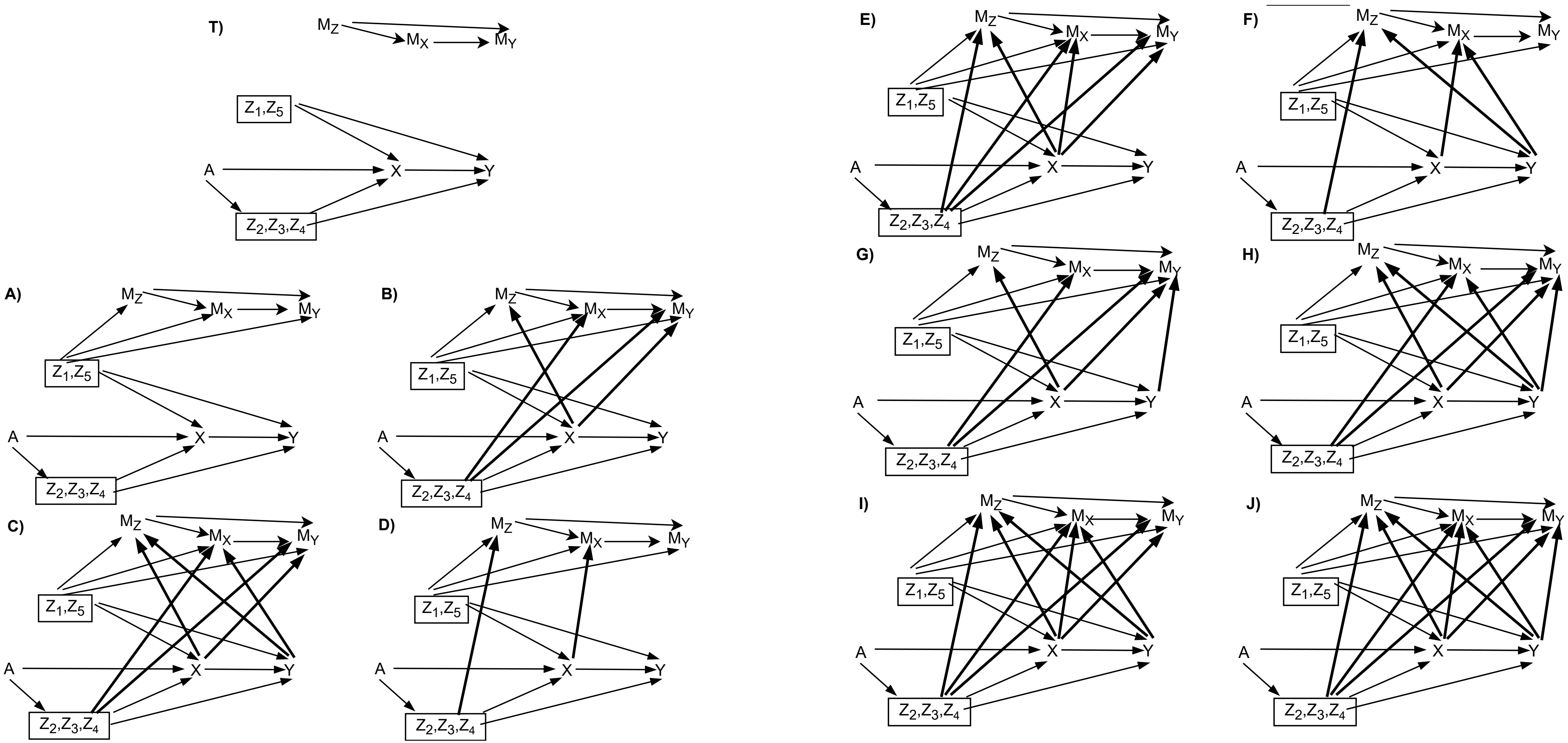


- 2000 simulations, each with 2000 records

Simulation study: generating the complete data (2)

Simple scenario	Complex scenario 1 & 2
$A \sim \text{Normal}(0,1)$	
$Z1 \sim \text{Binomial}(1, \text{invlogit}(\alpha_0))$	
$Z2 \sim \text{Binomial}(1, \text{invlogit}(\beta_0 + \beta_1 A))$	
$Z3 \sim \text{Binomial}(1, \text{invlogit}(\gamma_0 + \gamma_1 A))$	
$Z4 \sim \text{Binomial}(1, \text{invlogit}(\delta_0 + \delta_1 A))$	
$Z5 \sim \text{Binomial}(1, \text{invlogit}(\theta_0))$	
$X \sim \text{Binomial}(1, \text{logit}^{-1}(\tau_0 + \tau_1 Z1 + \tau_2 Z2 + \tau_3 Z3 + \tau_4 Z4 + \tau_5 Z5 + \tau_6 A))$	$X \sim \text{Binomial}(1, \text{logit}^{-1}(\tau_0^* + \tau_1 Z1 + \tau_2 Z2 + \tau_3 Z3 + \tau_4 Z4 + \tau_5 Z5 + \tau_6 Z1Z3 + \tau_7 Z1Z4 + \tau_8 Z1Z5 + \tau_9 Z3Z4 + \tau_{10} Z3Z5 + \tau_{11} Z4Z5))$
$Y \sim \text{Normal}(\varphi_0 + \varphi_1 X + \varphi_2 Z1 + \varphi_3 Z2 + \varphi_4 Z3 + \varphi_5 Z4 + \varphi_6 Z5, 1)$ ACE = $\varphi_1 = 0.2$	$Y \sim \text{Normal}(\varphi_0^* + \varphi_1 X + \varphi_2 Z1 + \varphi_3 Z2 + \varphi_4 Z3 + \varphi_5 Z4 + \varphi_6 Z5 + \varphi_7 Z1Z3 + \varphi_8 Z1Z4 + \varphi_9 Z1Z5 + \varphi_{10} Z3Z4 + \varphi_{11} Z3Z5 + \varphi_{12} Z4Z5 + \varphi_{13} Z1Z3Z4 + \varphi_{14} Z1Z3Z5 + \varphi_{15} Z1Z4Z5 + \varphi_{16} Z3Z4Z5 + \varphi_{17} Z1Z3Z4Z5, 1)$

Simulation study: imposing missing data (1)



- Missingness imposed on Z2, Z3, Z4, X, Y

Figure adapted from Moreno-Betancur et al, AJE (2018)

Simulation study: imposing missing data (2)

- $\mathrm{M}_{\mathrm{Z}_2} \sim \mathrm{Binomial}\left(1, \mathrm{logit}^{-1}(\iota_0 + \iota_1 \mathrm{Z}_1 + \iota_2 \mathrm{Z}_5 + \iota_3 \mathrm{Z}_2 + \iota_4 \mathrm{X} + \iota_5 \mathrm{Y})\right)$
- $\mathrm{M}_{\mathrm{Z}_3} \sim \mathrm{Binomial}(1, \mathrm{logit}^{-1}(\kappa_0 + \kappa_1 \mathrm{Z}_1 + \kappa_2 \mathrm{Z}_5 + \kappa_3 \mathrm{Z}_3 + \kappa_4 \mathrm{X} + \kappa_5 \mathrm{Y} + \kappa_6 \mathrm{M}_{\mathrm{Z}_2}))$
- $\mathrm{M}_{\mathrm{Z}_4} \sim \mathrm{Binomial}(1, \mathrm{logit}^{-1}(\lambda_0 + \lambda_1 \mathrm{Z}_1 + \lambda_2 \mathrm{Z}_5 + \lambda_3 \mathrm{Z}_4 + \lambda_4 \mathrm{X} + \lambda_5 \mathrm{Y} + \lambda_6 \mathrm{M}_{\mathrm{Z}_2} + \lambda_7 \mathrm{M}_{\mathrm{Z}_3}))$
- $\mathrm{M}_{\mathrm{X}} \sim \mathrm{Binomial}(1, \mathrm{logit}^{-1}(\nu_0 + \nu_1 \mathrm{Z}_1 + \nu_2 \mathrm{Z}_5 + \nu_3 \mathrm{Z}_2 + \nu_4 \mathrm{Z}_3 + \nu_5 \mathrm{Z}_4 + \nu_6 \mathrm{X} + \nu_7 \mathrm{Y} + \nu_8 \mathrm{M}_{\mathrm{Z}_2} + \nu_9 \mathrm{M}_{\mathrm{Z}_3} + \nu_{10} \mathrm{M}_{\mathrm{Z}_4}))$
- $\mathrm{M}_{\mathrm{Y}} \sim \mathrm{Binomial}(1, \mathrm{logit}^{-1}(\xi_0 + \xi_1 \mathrm{Z}_1 + \xi_2 \mathrm{Z}_5 + \xi_3 \mathrm{Z}_2 + \xi_4 \mathrm{Z}_3 + \xi_5 \mathrm{Z}_4 + \xi_6 \mathrm{X} + \xi_7 \mathrm{Y} + \xi_8 \mathrm{M}_{\mathrm{Z}_2} + \xi_9 \mathrm{M}_{\mathrm{Z}_3} + \xi_{10} \mathrm{M}_{\mathrm{Z}_4} + \xi_{11} \mathrm{M}_{\mathrm{X}}))$

Simulation study: imposing missing data (3)

Imposed missingness so that:

- 50% with missing data for any variable in the target analysis
- 40% with missing data for any confounder or exposure
- 30% with missing data for exposure

Simulation study: missing data methods (1)

Non-multiple imputation approaches:

1- Complete case analysis

- exclude 50% with missing any data

Simulation study: missing data methods (1)

Non-multiple imputation approaches:

1- Complete case analysis

- exclude 50% with missing any data

2- Extended TMLE to handle missing outcome data

- exclude 40% with missing confounder or exposure data
- incorporates $\hat{P}(M_y = 0|X, Z)$ in the TMLE estimation procedure
- Extended exchangeability assumption: $\mathrm{Y}^{\mathrm{x}} \coprod \mathrm{X}|\mathrm{Z}$ & $\mathrm{Y}^{\mathrm{x}} \coprod \mathrm{M}_{Y}|\mathrm{X}, \mathrm{Z}$ for $\mathrm{x} = 0{,}1$

Simulation study: missing data methods (1)

Non-multiple imputation approaches:

1- Complete case analysis

- exclude 50% with missing any data

2- Extended TMLE to handle missing outcome data

- exclude 40% with missing confounder or exposure data
- incorporates $\hat{P}(M_y = 0|X, Z)$ in the TMLE estimation procedure
- Ext exchangeability assumption: $\mathrm{Y^x} \coprod \mathrm{M}_Y|\mathrm{X, Z}$ and $\mathrm{Y^x} \coprod \mathrm{X}|\mathrm{Z}$ for $\mathrm{x} = 0{,}1$

3- Like 2 & missing covariate missing indicator approach (MCMI) to handle missing confounder data

- Exclude 30% with missing exposure data
- Missingness indicators for each confounder are included in the analysis models
- Ext exchangeability assumption: $\mathrm{Y^x} \coprod \mathrm{X}|\mathrm{Z, M}$ for $\mathrm{x} = 0{,}1$ & $\mathrm{X} \coprod \mathrm{Z_{miss}}|\mathrm{Z_{obs}, M}$ OR $\mathrm{Y^x} \coprod \mathrm{Z_{miss}}|\mathrm{Z_{obs}, M}$

Simulation study: missing data methods (2)

Multiple imputation by chained equations (MICE) approaches

- Simultaneously handle missing outcome, confounder, and exposure data
- Include all analysis variables & auxiliary variable age in the imputation models

4- Use linear regression to impute Y, logistic for X, Z2, Z3, Z4

5- Use instead predictive mean matching (PMM) to impute Y

6- Like 5, additionally include all 2×2 interactions in the imputation models

7- Like 6, additionally include all 3- & 4-way confounder-confounder interactions

8- Use classification and regression trees (CART)

9- Use random forest (RF)

MI with machine learning algorithms
Both available in the MICE package in R

Simulation study: TMLE performance in complete data

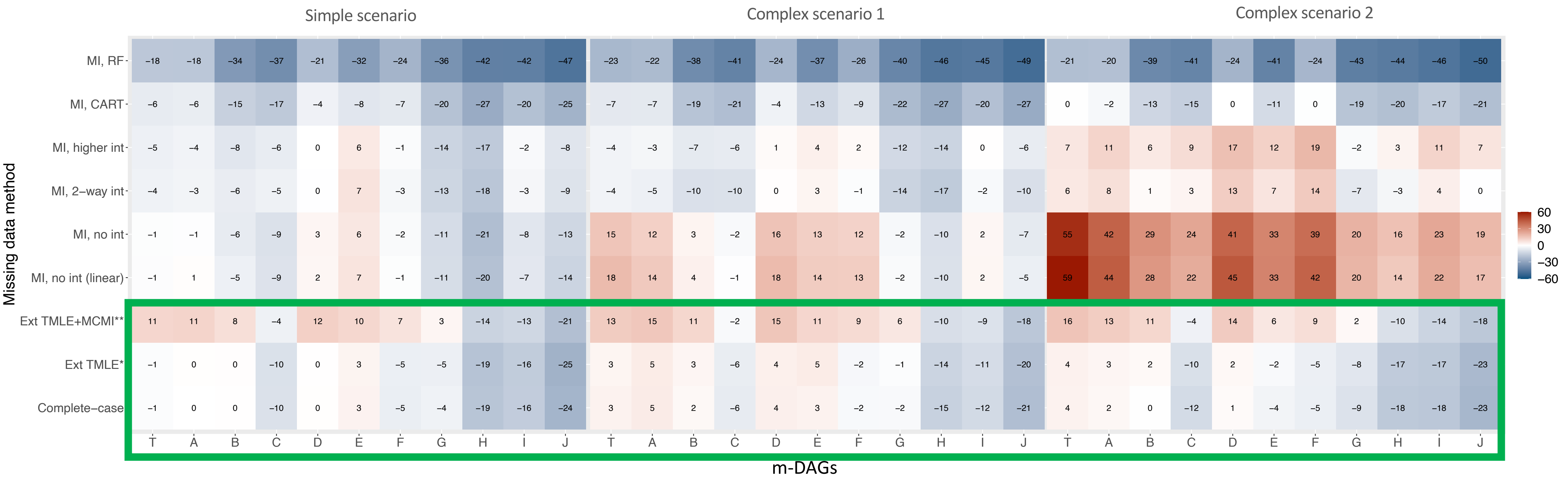
	$\widehat{ACE}$	Absolute bias		Rel bias (%)	Coverage		Bias Eliminated Coverage		Empirical SE		Average model SE		Relative % error in model SE	
		Est	MCSE		Est	MCSE	Est	MCSE	Est	MCSE	Est	MCSE	Est	MCSE
Simple scenario														
Regression	0.20	0.00	0.00	0.03	0.95	0.00	0.95	0.00	0.07	0.00	0.07	0.00	0.06	1.58
G-comp	0.20	0.00	0.00	0.02	0.95	0.00	0.95	0.00	0.07	0.00	0.07	0.00	0.41	1.59
TMLE	0.20	0.00	0.00	0.21	0.93	0.01	0.93	0.01	0.08	0.00	0.08	0.00	-5.70	1.49
Complex scenario 1														
Regression	0.26	0.06	0.00	31.63	0.87	0.01	0.96	0.00	0.07	0.00	0.07	0.00	4.30	1.65
G-comp	0.20	0.00	0.00	0.46	0.96	0.00	0.96	0.00	0.07	0.00	0.07	0.00	4.01	1.65
TMLE	0.20	0.00	0.00	1.99	0.93	0.01	0.93	0.01	0.08	0.00	0.08	0.00	-5.68	1.49
Complex scenario 2														
Regression	0.41	0.21	0.00	106.63	0.22	0.01	0.96	0.00	0.08	0.00	0.08	0.00	3.38	1.64
G-comp	0.18	-0.02	0.00	-8.34	0.93	0.01	0.94	0.01	0.07	0.00	0.07	0.00	-1.66	1.56
TMLE	0.20	0.00	0.00	0.51	0.87	0.01	0.87	0.01	0.09	0.00	0.07	0.00	-21.06	1.25

Simulation study: TMLE performance in complete data

	$\widehat{ACE}$	Absolute bias		Rel bias (%)	Coverage		Bias Eliminated Coverage		Empirical SE		Average model SE		Relative % error in model SE	
		Est	MCSE		Est	MCSE	Est	MCSE	Est	MCSE	Est	MCSE	Est	MCSE
							Simple scenario							
Regression	0.20	0.00	0.00	0.03	0.95	0.00	0.95	0.00	0.07	0.00	0.07	0.00	0.06	1.58
G-comp	0.20	0.00	0.00	0.02	0.95	0.00	0.95	0.00	0.07	0.00	0.07	0.00	0.41	1.59
TMLE	0.20	0.00	0.00	0.21	0.93	0.01	0.93	0.01	0.08	0.00	0.08	0.00	-5.70	1.49
							Complex scenario 1							
Regression	0.26	0.06	0.00	31.63	0.87	0.01	0.96	0.00	0.07	0.00	0.07	0.00	4.30	1.65
G-comp	0.20	0.00	0.00	0.46	0.96	0.00	0.96	0.00	0.07	0.00	0.07	0.00	4.01	1.65
TMLE	0.20	0.00	0.00	1.99	0.93	0.01	0.93	0.01	0.08	0.00	0.08	0.00	-5.68	1.49
							Complex scenario 2							
Regression	0.41	0.21	0.00	106.63	0.22	0.01	0.96	0.00	0.08	0.00	0.08	0.00	3.38	1.64
G-comp	0.18	-0.02	0.00	-8.34	0.93	0.01	0.94	0.01	0.07	0.00	0.07	0.00	-1.66	1.56
TMLE	0.20	0.00	0.00	0.51	0.87	0.01	0.87	0.01	0.09	0.00	0.07	0.00	-21.06	1.25

Simulation study: Performance of missing data methods

Relative bias



	Simple scenario										
	T	A	B	C	D	E	F	G	H	I	J
MI, RF	−18	−18	−34	−37	−21	−32	−24	−36	−42	−42	−47
MI, CART	−6	−6	−15	−17	−4	−8	−7	−20	−27	−20	−25
MI, higher int	−5	−4	−8	−6	0	6	−1	−14	−17	−2	−8
MI, 2−way int	−4	−3	−6	−5	0	7	−3	−13	−18	−3	−9
MI, no int	−1	−1	−6	−9	3	6	−2	−11	−21	−8	−13
MI, no int (linear)	−1	1	−5	−9	2	7	−1	−11	−20	−7	−14
Ext TMLE+MCMI**	11	11	8	−4	12	10	7	3	−14	−13	−21
Ext TMLE*	−1	0	0	−10	0	3	−5	−5	−19	−16	−25
Complete−case	−1	0	0	−10	0	3	−5	−4	−19	−16	−24

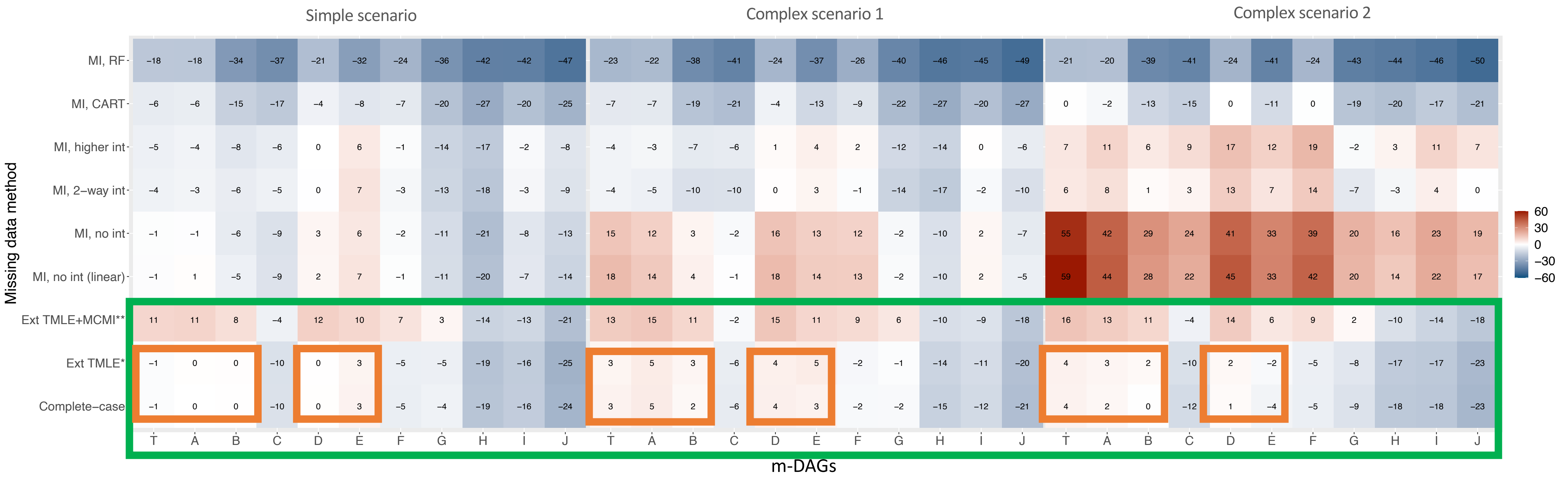
	Complex scenario 1										
	T	A	B	C	D	E	F	G	H	I	J
MI, RF	−23	−22	−38	−41	−24	−37	−26	−40	−46	−45	−49
MI, CART	−7	−7	−19	−21	−4	−13	−9	−22	−27	−20	−27
MI, higher int	−4	−3	−7	−6	1	4	2	−12	−14	0	−6
MI, 2−way int	−4	−5	−10	−10	0	3	−1	−14	−17	−2	−10
MI, no int	15	12	3	−2	16	13	12	−2	−10	2	−7
MI, no int (linear)	18	14	4	−1	18	14	13	−2	−10	2	−5
Ext TMLE+MCMI**	13	15	11	−2	15	11	9	6	−10	−9	−18
Ext TMLE*	3	5	3	−6	4	5	−2	−1	−14	−11	−20
Complete−case	3	5	2	−6	4	3	−2	−2	−15	−12	−21

	Complex scenario 2										
	T	A	B	C	D	E	F	G	H	I	J
MI, RF	−21	−20	−39	−41	−24	−41	−24	−43	−44	−46	−50
MI, CART	0	−2	−13	−15	0	−11	0	−19	−20	−17	−21
MI, higher int	7	11	6	9	17	12	19	−2	3	11	7
MI, 2−way int	6	8	1	3	13	7	14	−7	−3	4	0
MI, no int	55	42	29	24	41	33	39	20	16	23	19
MI, no int (linear)	59	44	28	22	45	33	42	20	14	22	17
Ext TMLE+MCMI**	16	13	11	−4	14	6	9	2	−10	−14	−18
Ext TMLE*	4	3	2	−10	2	−2	−5	−8	−17	−17	−23
Complete−case	4	2	0	−12	1	−4	−5	−9	−18	−18	−23

The Monte Carlo standard errors for absolute bias ranged from 0.001 to 0.004 in the simple scenario, 0.001 to 0.003 in the complex scenario 1, and 0.002 to 0.003 in the complex scenario 2.

Simulation study: Performance of missing data methods

Relative bias



Simple scenario

Missing data method	T	A	B	C	D	E	F	G	H	I	J
MI, RF	−18	−18	−34	−37	−21	−32	−24	−36	−42	−42	−47
MI, CART	−6	−6	−15	−17	−4	−8	−7	−20	−27	−20	−25
MI, higher int	−5	−4	−8	−6	0	6	−1	−14	−17	−2	−8
MI, 2−way int	−4	−3	−6	−5	0	7	−3	−13	−18	−3	−9
MI, no int	−1	−1	−6	−9	3	6	−2	−11	−21	−8	−13
MI, no int (linear)	−1	1	−5	−9	2	7	−1	−11	−20	−7	−14
Ext TMLE+MCMI**	11	11	8	−4	12	10	7	3	−14	−13	−21
Ext TMLE*	−1	0	0	−10	0	3	−5	−5	−19	−16	−25
Complete−case	−1	0	0	−10	0	3	−5	−4	−19	−16	−24

Complex scenario 1

Missing data method	T	A	B	C	D	E	F	G	H	I	J
MI, RF	−23	−22	−38	−41	−24	−37	−26	−40	−46	−45	−49
MI, CART	−7	−7	−19	−21	−4	−13	−9	−22	−27	−20	−27
MI, higher int	−4	−3	−7	−6	1	4	2	−12	−14	0	−6
MI, 2−way int	−4	−5	−10	−10	0	3	−1	−14	−17	−2	−10
MI, no int	15	12	3	−2	16	13	12	−2	−10	2	−7
MI, no int (linear)	18	14	4	−1	18	14	13	−2	−10	2	−5
Ext TMLE+MCMI**	13	15	11	−2	15	11	9	6	−10	−9	−18
Ext TMLE*	3	5	3	−6	4	5	−2	−1	−14	−11	−20
Complete−case	3	5	2	−6	4	3	−2	−2	−15	−12	−21

Complex scenario 2

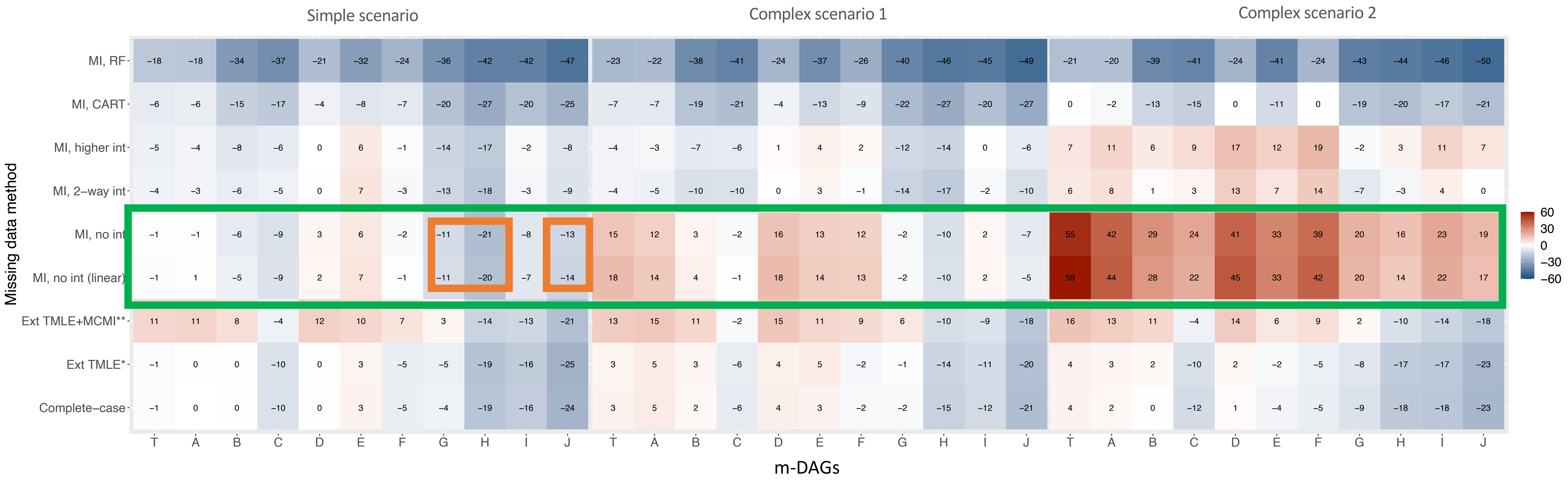
Missing data method	T	A	B	C	D	E	F	G	H	I	J
MI, RF	−21	−20	−39	−41	−24	−41	−24	−43	−44	−46	−50
MI, CART	0	−2	−13	−15	0	−11	0	−19	−20	−17	−21
MI, higher int	7	11	6	9	17	12	19	−2	3	11	7
MI, 2−way int	6	8	1	3	13	7	14	−7	−3	4	0
MI, no int	55	42	29	24	41	33	39	20	16	23	19
MI, no int (linear)	59	44	28	22	45	33	42	20	14	22	17
Ext TMLE+MCMI**	16	13	11	−4	14	6	9	2	−10	−14	−18
Ext TMLE*	4	3	2	−10	2	−2	−5	−8	−17	−17	−23
Complete−case	4	2	0	−12	1	−4	−5	−9	−18	−18	−23

m-DAGs

The Monte Carlo standard errors for absolute bias ranged from 0.001 to 0.004 in the simple scenario, 0.001 to 0.003 in the complex scenario 1, and 0.002 to 0.003 in the complex scenario 2.

Simulation study: Performance of missing data methods

Relative bias



Simple scenario

Missing data method	T	A	B	C	D	E	F	G	H	I	J
MI, RF	−18	−18	−34	−37	−21	−32	−24	−36	−42	−42	−47
MI, CART	−6	−6	−15	−17	−4	−8	−7	−20	−27	−20	−25
MI, higher int	−5	−4	−8	−6	0	6	−1	−14	−17	−2	−8
MI, 2−way int	−4	−3	−6	−5	0	7	−3	−13	−18	−3	−9
MI, no int	−1	−1	−6	−9	3	6	−2	−11	−21	−8	−13
MI, no int (linear)	−1	1	−5	−9	2	7	−1	−11	−20	−7	−14
Ext TMLE+MCMI**	11	11	8	−4	12	10	7	3	−14	−13	−21
Ext TMLE*	−1	0	0	−10	0	3	−5	−5	−19	−16	−25
Complete−case	−1	0	0	−10	0	3	−5	−4	−19	−16	−24

Complex scenario 1

Missing data method	T	A	B	C	D	E	F	G	H	I	J
MI, RF	−23	−22	−38	−41	−24	−37	−26	−40	−46	−45	−49
MI, CART	−7	−7	−19	−21	−4	−13	−9	−22	−27	−20	−27
MI, higher int	−4	−3	−7	−6	1	4	2	−12	−14	0	−6
MI, 2−way int	−4	−5	−10	−10	0	3	−1	−14	−17	−2	−10
MI, no int	15	12	3	−2	16	13	12	−2	−10	2	−7
MI, no int (linear)	18	14	4	−1	18	14	13	−2	−10	2	−5
Ext TMLE+MCMI**	13	15	11	−2	15	11	9	6	−10	−9	−18
Ext TMLE*	3	5	3	−6	4	5	−2	−1	−14	−11	−20
Complete−case	3	5	2	−6	4	3	−2	−2	−15	−12	−21

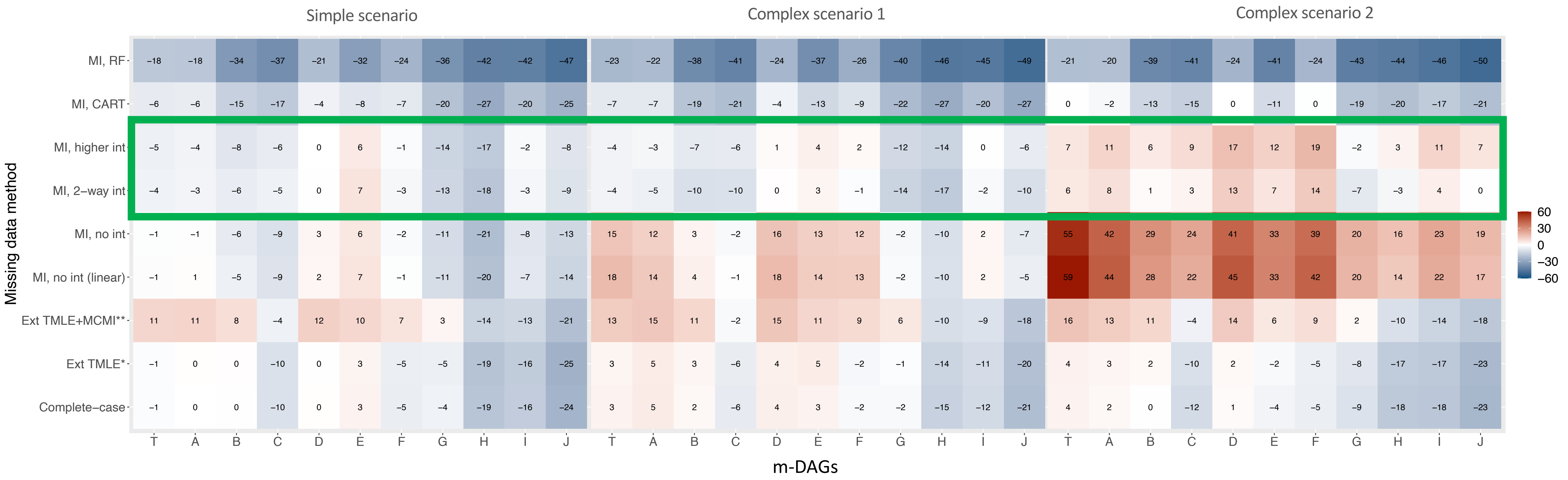
Complex scenario 2

Missing data method	T	A	B	C	D	E	F	G	H	I	J
MI, RF	−21	−20	−39	−41	−24	−41	−24	−43	−44	−46	−50
MI, CART	0	−2	−13	−15	0	−11	0	−19	−20	−17	−21
MI, higher int	7	11	6	9	17	12	19	−2	3	11	7
MI, 2−way int	6	8	1	3	13	7	14	−7	−3	4	0
MI, no int	55	42	29	24	41	33	39	20	16	23	19
MI, no int (linear)	59	44	28	22	45	33	42	20	14	22	17
Ext TMLE+MCMI**	16	13	11	−4	14	6	9	2	−10	−14	−18
Ext TMLE*	4	3	2	−10	2	−2	−5	−8	−17	−17	−23
Complete−case	4	2	0	−12	1	−4	−5	−9	−18	−18	−23

The Monte Carlo standard errors for absolute bias ranged from 0.001 to 0.004 in the simple scenario, 0.001 to 0.003 in the complex scenario 1, and 0.002 to 0.003 in the complex scenario 2.

Simulation study: Performance of missing data methods

Relative bias



Simple scenario

Missing data method	T	A	B	C	D	E	F	G	H	I	J
MI, RF	−18	−18	−34	−37	−21	−32	−24	−36	−42	−42	−47
MI, CART	−6	−6	−15	−17	−4	−8	−7	−20	−27	−20	−25
MI, higher int	−5	−4	−8	−6	0	6	−1	−14	−17	−2	−8
MI, 2−way int	−4	−3	−6	−5	0	7	−3	−13	−18	−3	−9
MI, no int	−1	−1	−6	−9	3	6	−2	−11	−21	−8	−13
MI, no int (linear)	−1	1	−5	−9	2	7	−1	−11	−20	−7	−14
Ext TMLE+MCMI**	11	11	8	−4	12	10	7	3	−14	−13	−21
Ext TMLE*	−1	0	0	−10	0	3	−5	−5	−19	−16	−25
Complete−case	−1	0	0	−10	0	3	−5	−4	−19	−16	−24

Complex scenario 1

Missing data method	T	A	B	C	D	E	F	G	H	I	J
MI, RF	−23	−22	−38	−41	−24	−37	−26	−40	−46	−45	−49
MI, CART	−7	−7	−19	−21	−4	−13	−9	−22	−27	−20	−27
MI, higher int	−4	−3	−7	−6	1	4	2	−12	−14	0	−6
MI, 2−way int	−4	−5	−10	−10	0	3	−1	−14	−17	−2	−10
MI, no int	15	12	3	−2	16	13	12	−2	−10	2	−7
MI, no int (linear)	18	14	4	−1	18	14	13	−2	−10	2	−5
Ext TMLE+MCMI**	13	15	11	−2	15	11	9	6	−10	−9	−18
Ext TMLE*	3	5	3	−6	4	5	−2	−1	−14	−11	−20
Complete−case	3	5	2	−6	4	3	−2	−2	−15	−12	−21

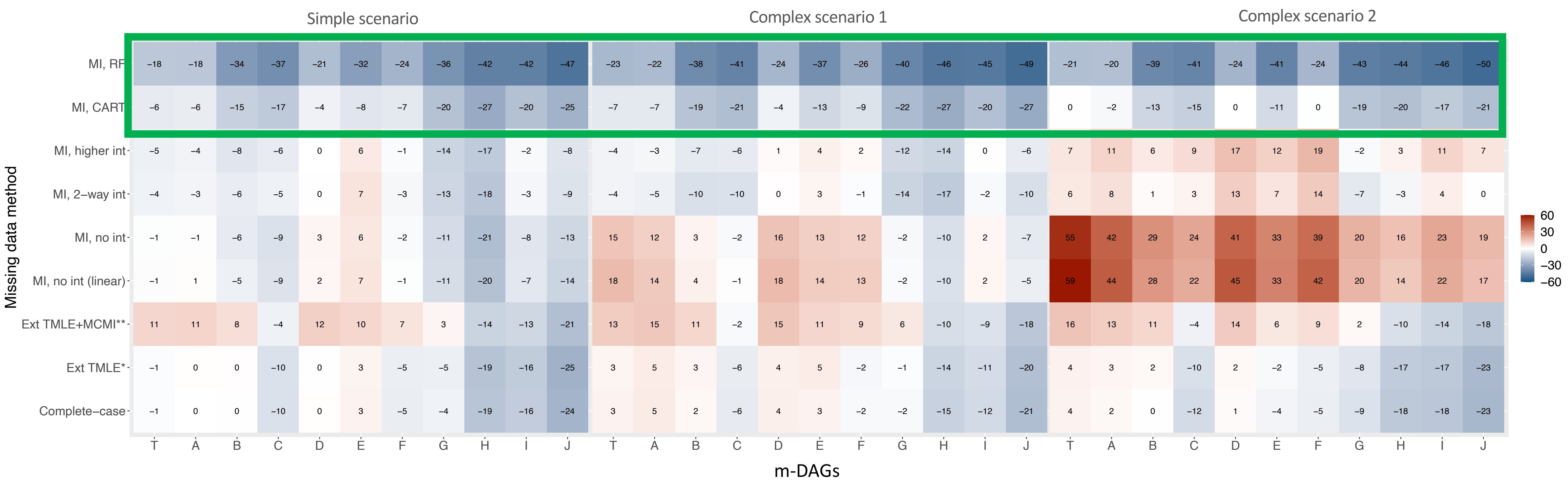
Complex scenario 2

Missing data method	T	A	B	C	D	E	F	G	H	I	J
MI, RF	−21	−20	−39	−41	−24	−41	−24	−43	−44	−46	−50
MI, CART	0	−2	−13	−15	0	−11	0	−19	−20	−17	−21
MI, higher int	7	11	6	9	17	12	19	−2	3	11	7
MI, 2−way int	6	8	1	3	13	7	14	−7	−3	4	0
MI, no int	55	42	29	24	41	33	39	20	16	23	19
MI, no int (linear)	59	44	28	22	45	33	42	20	14	22	17
Ext TMLE+MCMI**	16	13	11	−4	14	6	9	2	−10	−14	−18
Ext TMLE*	4	3	2	−10	2	−2	−5	−8	−17	−17	−23
Complete−case	4	2	0	−12	1	−4	−5	−9	−18	−18	−23

The Monte Carlo standard errors for absolute bias ranged from 0.001 to 0.004 in the simple scenario, 0.001 to 0.003 in the complex scenario 1, and 0.002 to 0.003 in the complex scenario 2.

Simulation study: Performance of missing data methods

Relative bias

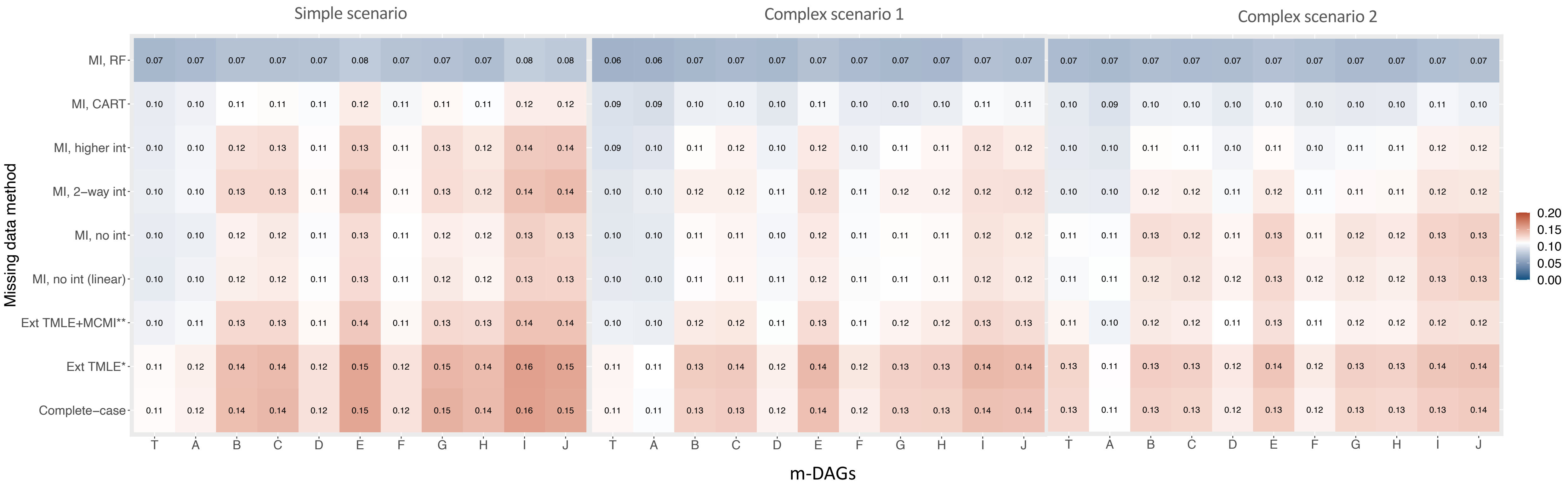


Missing data method	Simple scenario											Complex scenario 1											Complex scenario 2										
	T	A	B	C	D	E	F	G	H	I	J	T	A	B	C	D	E	F	G	H	I	J	T	A	B	C	D	E	F	G	H	I	J
MI, RF	−18	−18	−34	−37	−21	−32	−24	−36	−42	−42	−47	−23	−22	−38	−41	−24	−37	−26	−40	−46	−45	−49	−21	−20	−39	−41	−24	−41	−24	−43	−44	−46	−50
MI, CART	−6	−6	−15	−17	−4	−8	−7	−20	−27	−20	−25	−7	−7	−19	−21	−4	−13	−9	−22	−27	−20	−27	0	−2	−13	−15	0	−11	0	−19	−20	−17	−21
MI, higher int	−5	−4	−8	−6	0	6	−1	−14	−17	−2	−8	−4	−3	−7	−6	1	4	2	−12	−14	0	−6	7	11	6	9	17	12	19	−2	3	11	7
MI, 2–way int	−4	−3	−6	−5	0	7	−3	−13	−18	−3	−9	−4	−5	−10	−10	0	3	−1	−14	−17	−2	−10	6	8	1	3	13	7	14	−7	−3	4	0
MI, no int	−1	−1	−6	−9	3	6	−2	−11	−21	−8	−13	15	12	3	−2	16	13	12	−2	−10	2	−7	55	42	29	24	41	33	39	20	16	23	19
MI, no int (linear)	−1	1	−5	−9	2	7	−1	−11	−20	−7	−14	18	14	4	−1	18	14	13	−2	−10	2	−5	59	44	28	22	45	33	42	20	14	22	17
Ext TMLE+MCMI**	11	11	8	−4	12	10	7	3	−14	−13	−21	13	15	11	−2	15	11	9	6	−10	−9	−18	16	13	11	−4	14	6	9	2	−10	−14	−18
Ext TMLE*	−1	0	0	−10	0	3	−5	−5	−19	−16	−25	3	5	3	−6	4	5	−2	−1	−14	−11	−20	4	3	2	−10	2	−2	−5	−8	−17	−17	−23
Complete–case	−1	0	0	−10	0	3	−5	−4	−19	−16	−24	3	5	2	−6	4	3	−2	−2	−15	−12	−21	4	2	0	−12	1	−4	−5	−9	−18	−18	−23

The Monte Carlo standard errors for absolute bias ranged from 0.001 to 0.004 in the simple scenario, 0.001 to 0.003 in the complex scenario 1, and 0.002 to 0.003 in the complex scenario 2.

Simulation study: Performance of missing data methods

Empirical standard error



Missing data method	Simple scenario										
	T	A	B	C	D	E	F	G	H	I	J
MI, RF	0.07	0.07	0.07	0.07	0.07	0.08	0.07	0.07	0.07	0.08	0.08
MI, CART	0.10	0.10	0.11	0.11	0.11	0.12	0.11	0.11	0.11	0.12	0.12
MI, higher int	0.10	0.10	0.12	0.13	0.11	0.13	0.11	0.13	0.12	0.14	0.14
MI, 2–way int	0.10	0.10	0.13	0.13	0.11	0.14	0.11	0.13	0.12	0.14	0.14
MI, no int	0.10	0.10	0.12	0.12	0.11	0.13	0.11	0.12	0.12	0.13	0.13
MI, no int (linear)	0.10	0.10	0.12	0.12	0.11	0.13	0.11	0.12	0.12	0.13	0.13
Ext TMLE+MCMI**	0.10	0.11	0.13	0.13	0.11	0.14	0.11	0.13	0.13	0.14	0.14
Ext TMLE*	0.11	0.12	0.14	0.14	0.12	0.15	0.12	0.15	0.14	0.16	0.15
Complete–case	0.11	0.12	0.14	0.14	0.12	0.15	0.12	0.15	0.14	0.16	0.15

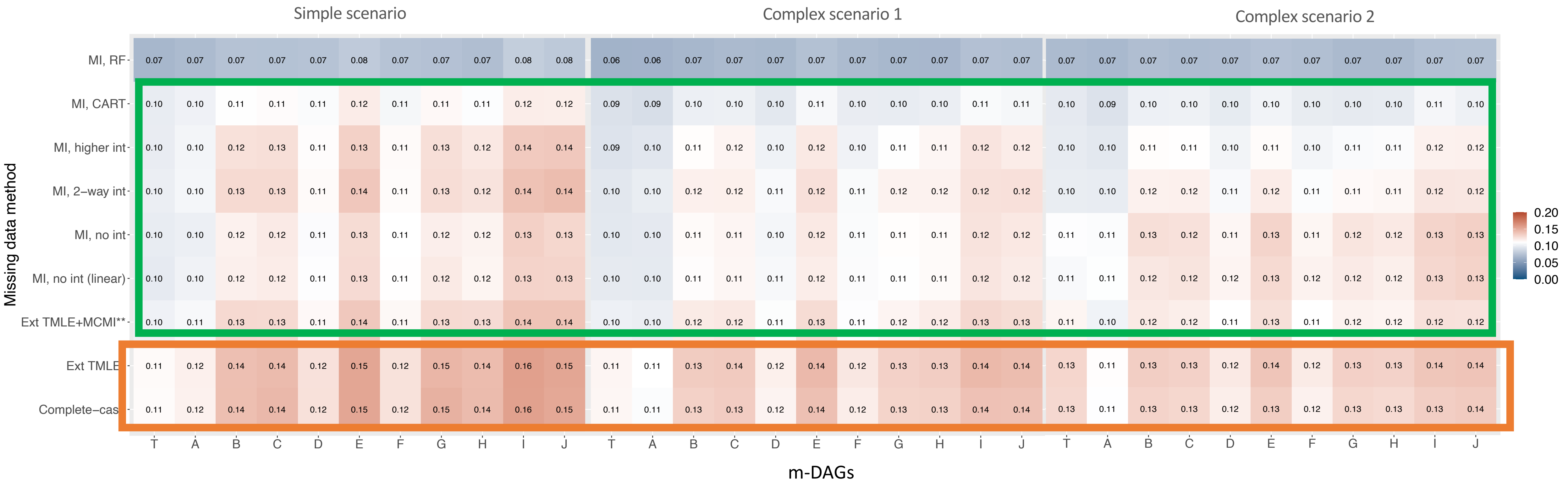
Missing data method	Complex scenario 1										
	T	A	B	C	D	E	F	G	H	I	J
MI, RF	0.06	0.06	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07
MI, CART	0.09	0.09	0.10	0.10	0.10	0.11	0.10	0.10	0.10	0.11	0.11
MI, higher int	0.09	0.10	0.11	0.12	0.10	0.12	0.10	0.11	0.11	0.12	0.12
MI, 2–way int	0.10	0.10	0.12	0.12	0.11	0.12	0.11	0.12	0.12	0.12	0.12
MI, no int	0.10	0.10	0.11	0.11	0.10	0.12	0.11	0.11	0.11	0.12	0.12
MI, no int (linear)	0.10	0.10	0.11	0.11	0.11	0.12	0.11	0.11	0.11	0.12	0.12
Ext TMLE+MCMI**	0.10	0.10	0.12	0.12	0.11	0.13	0.11	0.12	0.12	0.13	0.13
Ext TMLE*	0.11	0.11	0.13	0.14	0.12	0.14	0.12	0.13	0.13	0.14	0.14
Complete–case	0.11	0.11	0.13	0.13	0.12	0.14	0.12	0.13	0.13	0.14	0.14

Missing data method	Complex scenario 2										
	T	A	B	C	D	E	F	G	H	I	J
MI, RF	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07
MI, CART	0.10	0.09	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.11	0.10
MI, higher int	0.10	0.10	0.11	0.11	0.10	0.11	0.10	0.11	0.11	0.12	0.12
MI, 2–way int	0.10	0.10	0.12	0.12	0.11	0.12	0.11	0.11	0.11	0.12	0.12
MI, no int	0.11	0.11	0.13	0.12	0.11	0.13	0.11	0.12	0.12	0.13	0.13
MI, no int (linear)	0.11	0.11	0.12	0.12	0.12	0.13	0.12	0.12	0.12	0.13	0.13
Ext TMLE+MCMI**	0.11	0.10	0.12	0.12	0.11	0.13	0.11	0.12	0.12	0.12	0.12
Ext TMLE*	0.13	0.11	0.13	0.13	0.12	0.14	0.12	0.13	0.13	0.14	0.14
Complete–case	0.13	0.11	0.13	0.13	0.12	0.13	0.12	0.13	0.13	0.13	0.14

The Monte Carlo standard errors ranged from 0.001 to 0.002 for all scenarios.

Simulation study: Performance of missing data methods

Empirical standard error



	Simple scenario										
	T	A	B	C	D	E	F	G	H	I	J
MI, RF	0.07	0.07	0.07	0.07	0.07	0.08	0.07	0.07	0.07	0.08	0.08
MI, CART	0.10	0.10	0.11	0.11	0.11	0.12	0.11	0.11	0.11	0.12	0.12
MI, higher int	0.10	0.10	0.12	0.13	0.11	0.13	0.11	0.13	0.12	0.14	0.14
MI, 2–way int	0.10	0.10	0.13	0.13	0.11	0.14	0.11	0.13	0.12	0.14	0.14
MI, no int	0.10	0.10	0.12	0.12	0.11	0.13	0.11	0.12	0.12	0.13	0.13
MI, no int (linear)	0.10	0.10	0.12	0.12	0.11	0.13	0.11	0.12	0.12	0.13	0.13
Ext TMLE+MCMI**	0.10	0.11	0.13	0.13	0.11	0.14	0.11	0.13	0.13	0.14	0.14
Ext TMLE	0.11	0.12	0.14	0.14	0.12	0.15	0.12	0.15	0.14	0.16	0.15
Complete–cas	0.11	0.12	0.14	0.14	0.12	0.15	0.12	0.15	0.14	0.16	0.15

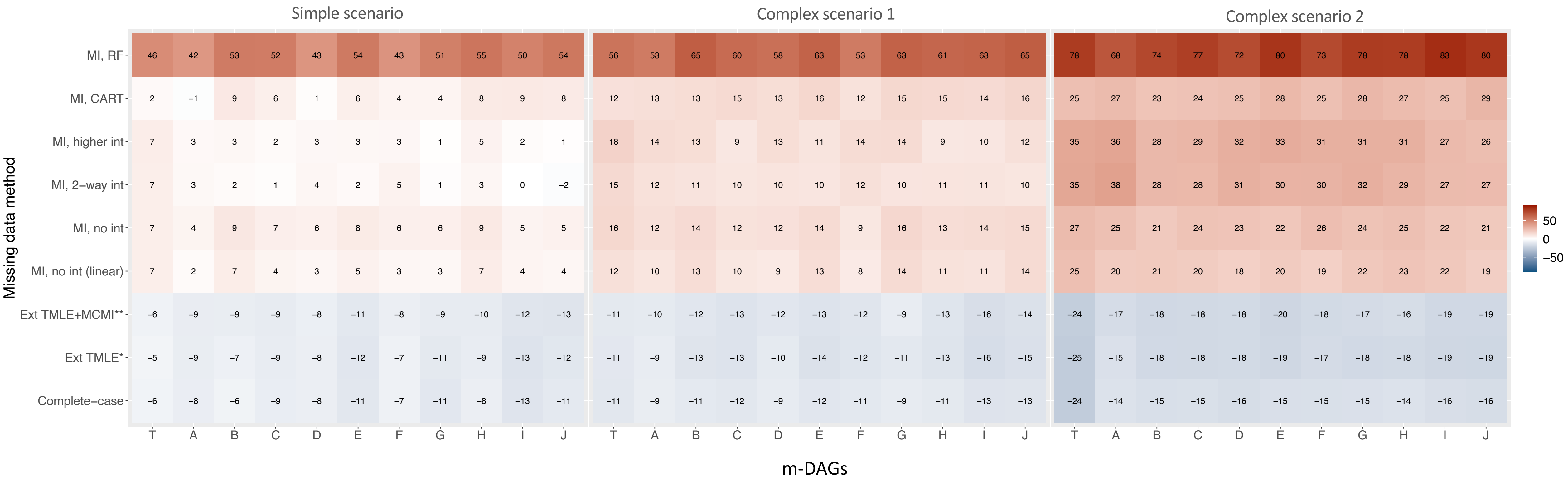
	Complex scenario 1										
	T	A	B	C	D	E	F	G	H	I	J
MI, RF	0.06	0.06	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07
MI, CART	0.09	0.09	0.10	0.10	0.10	0.11	0.10	0.10	0.10	0.11	0.11
MI, higher int	0.09	0.10	0.11	0.12	0.10	0.12	0.10	0.11	0.11	0.12	0.12
MI, 2–way int	0.10	0.10	0.12	0.12	0.11	0.12	0.11	0.12	0.12	0.12	0.12
MI, no int	0.10	0.10	0.11	0.11	0.10	0.12	0.11	0.11	0.11	0.12	0.12
MI, no int (linear)	0.10	0.10	0.11	0.11	0.11	0.12	0.11	0.11	0.11	0.12	0.12
Ext TMLE+MCMI**	0.10	0.10	0.12	0.12	0.11	0.13	0.11	0.12	0.12	0.13	0.13
Ext TMLE	0.11	0.11	0.13	0.14	0.12	0.14	0.12	0.13	0.13	0.14	0.14
Complete–cas	0.11	0.11	0.13	0.13	0.12	0.14	0.12	0.13	0.13	0.14	0.14

	Complex scenario 2										
	T	A	B	C	D	E	F	G	H	I	J
MI, RF	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07
MI, CART	0.10	0.09	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.11	0.10
MI, higher int	0.10	0.10	0.11	0.11	0.10	0.11	0.10	0.11	0.11	0.12	0.12
MI, 2–way int	0.10	0.10	0.12	0.12	0.11	0.12	0.11	0.11	0.11	0.12	0.12
MI, no int	0.11	0.11	0.13	0.12	0.11	0.13	0.11	0.12	0.12	0.13	0.13
MI, no int (linear)	0.11	0.11	0.12	0.12	0.12	0.13	0.12	0.12	0.12	0.13	0.13
Ext TMLE+MCMI**	0.11	0.10	0.12	0.12	0.11	0.13	0.11	0.12	0.12	0.12	0.12
Ext TMLE	0.13	0.11	0.13	0.13	0.12	0.14	0.12	0.13	0.13	0.14	0.14
Complete–cas	0.13	0.11	0.13	0.13	0.12	0.13	0.12	0.13	0.13	0.13	0.14

The Monte Carlo standard errors ranged from 0.001 to 0.002 for all scenarios.

Simulation study: Performance of missing data methods

Relative % error in model SE



Simple scenario

Missing data method	T	A	B	C	D	E	F	G	H	I	J
MI, RF	46	42	53	52	43	54	43	51	55	50	54
MI, CART	2	−1	9	6	1	6	4	4	8	9	8
MI, higher int	7	3	3	2	3	3	3	1	5	2	1
MI, 2−way int	7	3	2	1	4	2	5	1	3	0	−2
MI, no int	7	4	9	7	6	8	6	6	9	5	5
MI, no int (linear)	7	2	7	4	3	5	3	3	7	4	4
Ext TMLE+MCMI**	−6	−9	−9	−9	−8	−11	−8	−9	−10	−12	−13
Ext TMLE*	−5	−9	−7	−9	−8	−12	−7	−11	−9	−13	−12
Complete−case	−6	−8	−6	−9	−8	−11	−7	−11	−8	−13	−11

Complex scenario 1

Missing data method	T	A	B	C	D	E	F	G	H	I	J
MI, RF	56	53	65	60	58	63	53	63	61	63	65
MI, CART	12	13	13	15	13	16	12	15	15	14	16
MI, higher int	18	14	13	9	13	11	14	14	9	10	12
MI, 2−way int	15	12	11	10	10	10	12	10	11	11	10
MI, no int	16	12	14	12	12	14	9	16	13	14	15
MI, no int (linear)	12	10	13	10	9	13	8	14	11	11	14
Ext TMLE+MCMI**	−11	−10	−12	−13	−12	−13	−12	−9	−13	−16	−14
Ext TMLE*	−11	−9	−13	−13	−10	−14	−12	−11	−13	−16	−15
Complete−case	−11	−9	−11	−12	−9	−12	−11	−9	−11	−13	−13

Complex scenario 2

Missing data method	T	A	B	C	D	E	F	G	H	I	J
MI, RF	78	68	74	77	72	80	73	78	78	83	80
MI, CART	25	27	23	24	25	28	25	28	27	25	29
MI, higher int	35	36	28	29	32	33	31	31	31	27	26
MI, 2−way int	35	38	28	28	31	30	30	32	29	27	27
MI, no int	27	25	21	24	23	22	26	24	25	22	21
MI, no int (linear)	25	20	21	20	18	20	19	22	23	22	19
Ext TMLE+MCMI**	−24	−17	−18	−18	−18	−20	−18	−17	−16	−19	−19
Ext TMLE*	−25	−15	−18	−18	−18	−19	−17	−18	−18	−19	−19
Complete−case	−24	−14	−15	−15	−16	−15	−15	−15	−14	−16	−16

The Monte Carlo standard errors ranged from 1.40 to 2.54 in the simple scenario, 1.36 to 2.74 in the complex scenario 1, and 1.22 to 3.02 in the complex scenario 2.

Application to the VAHCS case study

Method	Difference in the mean	Standard error	95% confidence interval	Time to run
Complete-case	0.09	0.12	-0.14, 0.32	11.9 sec
Ext TMLE	0.12	0.11	-0.09, 0.33	8.0 sec
Ext TMLE+MCMI	0.13	0.13	-0.13, 0.39	15.5 sec
MI, no int (linear)	0.19	0.15	-0.11, 0.50	5.6 min
MI, no int (PMM)	0.20	0.15	-0.09, 0.49	5.4 min
MI, 2-way int	0.14	0.17	-0.20, 0.49	6.1 min
MI, higher int	0.16	0.16	-0.14, 0.47	6.2 min
MI, CART	0.13	0.15	-0.18, 0.43	12.2 min
MI, RF	0.14	0.16	-0.17, 0.46	14.8 min

Application to the VAHCS case study

Method	Difference in the mean	Standard error	95% confidence interval	Time to run	Corrected SE
Complete-case	0.09	0.12	-0.14, 0.32	11.9 sec	0.14
Ext TMLE	0.12	0.11	-0.09, 0.33	8.0 sec	0.13
Ext TMLE+MCMI	0.13	0.13	-0.13, 0.39	15.5 sec	0.15
MI, no int (linear)	0.19	0.15	-0.11, 0.50	5.6 min	0.13
MI, no int (PMM)	0.20	0.15	-0.09, 0.49	5.4 min	0.14
MI, 2-way int	0.14	0.17	-0.20, 0.49	6.1 min	0.15
MI, higher int	0.16	0.16	-0.14, 0.47	6.2 min	0.14
MI, CART	0.13	0.15	-0.18, 0.43	12.2 min	0.14
MI, RF	0.14	0.16	-0.17, 0.46	14.8 min	0.11

Concluding remarks

- We used simulation study to evaluate available approaches for handling missing data when estimating the ACE using TMLE with data adaptive approaches
- Under simple and complex scenarios, data generation was fairly simple
- Key observations:
  - No approach performs well in general
  - Consideration of missingness mechanism could be helpful
  - If missingness in no variable depends on the outcome: CC, Ext-TMLE (small bias & loss of precision)
  - Correctly specified parametric MI perform well
  - MI-CART might be a useful alternative

References

1- Patton GC, Coffey C, Carlin JB, Degenhardt L, Lynskey M, Hall W. Cannabis use and mental health in young people: cohort study. BMJ. 2002;325(7374):1195-1198

2- Hernán MA RJ. Causal Inference: What If. Boca Raton: Chapman & Hall/CRC. 2020

3- Schuler MS, Rose S. Targeted Maximum Likelihood Estimation for Causal Inference in Observational Studies. Am J Epidemiol. 2017;185(1):65-73

4- Moreno-Betancur M, Lee KJ, Leacy FP, White IR, Simpson JA, Carlin JB. Canonical Causal Diagrams to Guide the Treatment of Missing Data in Epidemiologic Studies. Am J Epidemiol. 2018;187(12):2705-2715

5- Blake HA, Leyrat C, Mansfield KE, Seaman S, Tomlinson LA, Carpenter J, et al. Propensity scores using missingness pattern information: a practical guide. Stat Med. 2020;39(11):1641-57

6-Naimi AI, Mishler AE, Kennedy EH. Challenges in Obtaining Valid Causal Effect Estimates with Machine Learning Algorithms. arXiv e-prints. 2020.

7-Balzer LB, Westling T. Demystifying Statistical Inference When Using Machine Learning in Causal Research. Am J Epidemiol. 2021.

8- Bartlett JW, Hughes RA. Bootstrap inference for multiple imputation under uncongeniality and misspecification. Stat Methods Med Res. 2020;29(12):3533-46.

Acknowledgements

Katherine Lee, Julie Simpson, Ian White, John Carlin, Margarita Moreno-Betancur

Pre-print: arxiv.org/abs/2112.05274

@ghazalehdashti

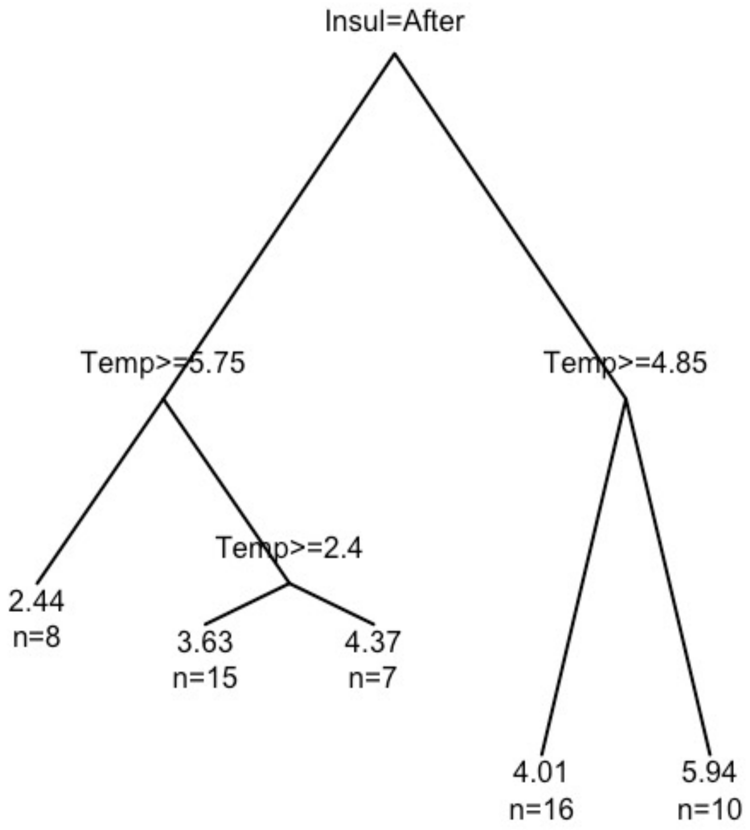
Simulation study: missing data methods (3)

MI using classification and regression trees (CART):

- Fit a tree for variable with missing data
- Identify the donor leaf for each record with missing value
- Randomly select a value from the donor leaf

MI using random forests (RF):

- Draw a bootstrap sample from the data
- Fit a tree for variable with missing data for each
- Randomly select a value from all donor leaves



Regression tree for predicting gas consumption

Image taken from S. van Buuren, Flexible Imputation of Missing Data, 2018