Sentinel

Characterizing Missing Data Processes in EHR Data

Janick Weberpals, RPh, PhD

Instructor in Medicine

FDA Sentinel Innovation Center at Harvard Medical School

Disclosures

- Janick Weberpals reports prior employment by Hoffmann-La Roche and previously held shares in Hoffmann-La Roche
- This project was supported by Task Order 75F40119F19002 under Master Agreement 75F40119D10037 from the U.S. Food and Drug Administration (FDA)

Knowledge Gaps and Objectives

Missing data in confounding factors are frequent

- <u>Examples:</u> Labs (e.g., HbA1c), Vitals (e.g., ejection fraction), Physician assessments (e.g., ECOG)
- **Mechanisms**: Missing completely at random (**MCAR**), at random (**MAR**) and not at random (**MNAR**)
- **Patterns**: Monotone, Non-monotone

Unresolved challenges for causal inference

- In **an empirical study**, it is usually unclear which of the missing data mechanisms and patterns are dominating.
- How do any of these mechanisms relate to **bias in a given RWD study**, given the strength of correlations between exposure, covariates and outcomes in high-dimensional covariate spaces (e.g., database linkages)?

Objectives

- **Develop a framework and tools to assess the structure of missing data processes in EHR studies**
- **Connect this with the most appropriate analytical approach, followed by sensitivity analyses**

- Rubin DB. Inference and Missing Data. *Biometrika*. 1976;63(3):581-592. doi:10.2307/2335739
- Mitra, R., McGough, S.F., Chakraborti, T. et al. Learning from data with structured missingness. Nat Mach Intell 5, 13–23 (2023)
- Mohan K, Pearl J, Tian J. Graphical models for inference with Missing data. In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'13. Curran Associates Inc.; 2013:1277-1285.

Sentinel Causal Inference Work Stream

Development site

Data sources

EHR

Mass General Brigham

Deterministic linkage

Claims

Products

Identify the most promising approaches to **assess missing confounder data in EHR** studies and develop a **Toolkit** to routinely implement these methods in future studies

Toolkit deployment

Testing site 1

EHR

DukeHealth

Deterministic linkage

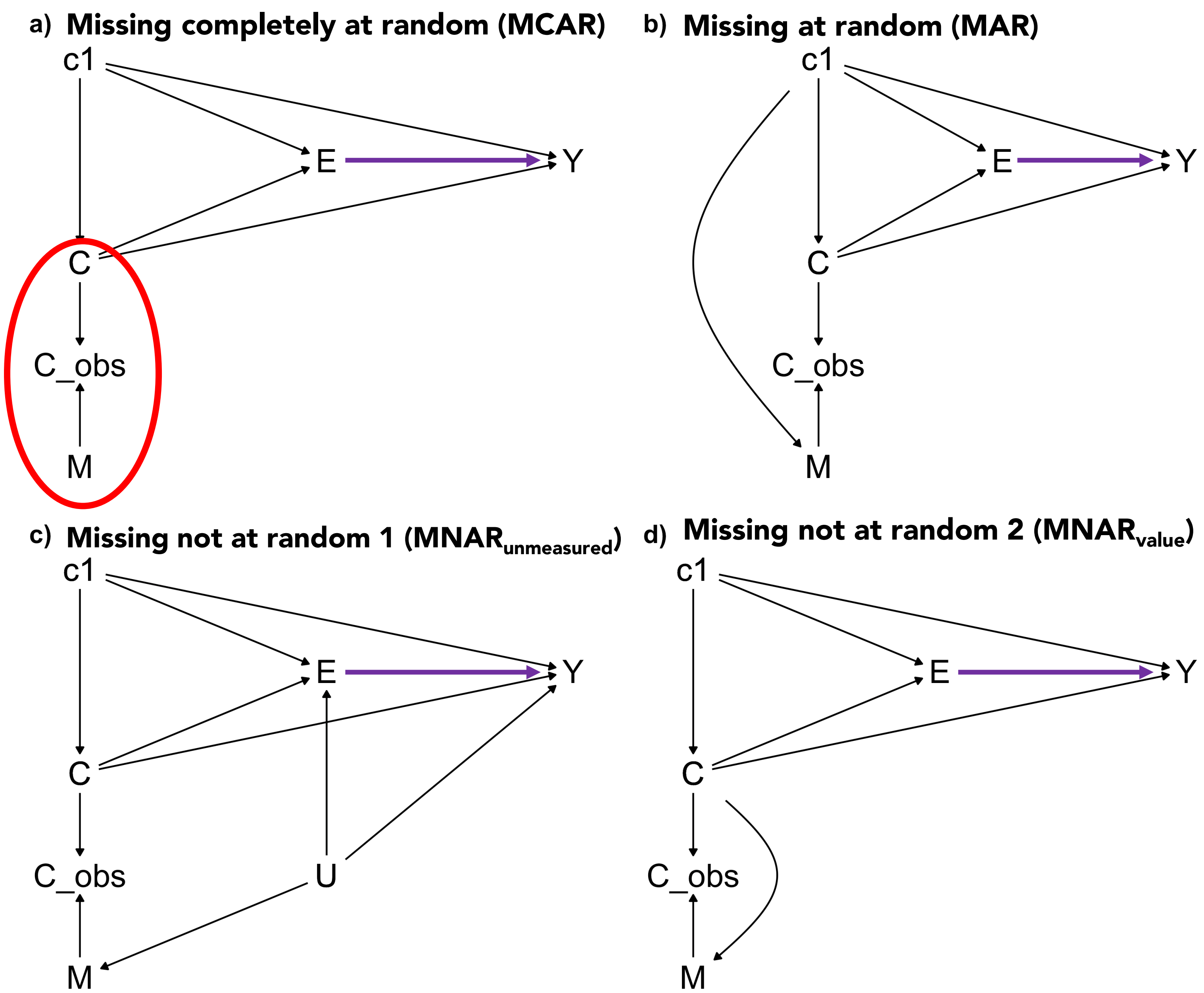
Claims

Test Toolkit and compare performance with development site

Assumed causal missingness structures

Causal diagrams/M-graphs provide a more natural way to understand the assumptions regarding missing (**confounder**) data for a given research question

E	Exposure/treatment
Y	Outcome
C	Confounder of interest
C_obs	Observed portion of C
M	Missingness of C (M=0 fully observed and M=1 fully missing)
c1	Covariates associated with outcome and missingness
c0	Auxiliary covariates
U	Unmeasured covariate/confounder



- Choi J, Dekkers OM, le Cessie S. A comparison of different methods to handle missing data in the context of propensity score analysis. Eur J Epidemiol. 2019 Jan;34(1):23-36.
- Mohan K, Pearl J, Tian J. Graphical models for inference with Missing data. In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'13. Curran Associates Inc.; 2013:1277-1285.

Empirical Diagnostics to Characterize Missingness Mechanisms

Group 1 Diagnostics

	Absolute standardized mean difference (ASMD)	P-value Hoteling/Little
Purpose	Comparison of distributions between patients with vs w/o observed value of the partially observed covariate	
Example value	ASMD = 0.1	p-value <0.001
Interpretation	**<0.1***: missingness is not associated with other observed covariates may be completely at random **>0.1***: missingness differs between patients and observed covariates can explain difference * Equivalent to propensity score-based balance measures (Austin PC, Multivariate Behavioral Research, 46:3, 399-424 (2011)	**Low p-values:** Indicate differences in covariate distributions and null hypothesis would be rejected (≠**MCAR**) Hotelling H. Ann Math Stat. 2(3):360-378. (1931) & Little RJA. J Am Stat Assoc. 83(404):1198-1202. doi:10.2307/2290157 (1988)

Empirical Diagnostics to Characterize Missingness Mechanisms

	Group 1 Diagnostics		Group 2 Diagnostics
	Absolute standardized mean difference (ASMD)	**P-value Hoteling/Little**	**AUC (are under the receiver operating curve)**
Purpose	Comparison of distributions between patients with vs w/o observed value of the partially observed covariate		Assessing the ability to predict missingness based on observed covariates
Example value	ASMD = 0.1	p-value <0.001	AUC = 0.5
Interpretation	**<0.1***: missingness is not associated with other observed covariates may be completely at random **>0.1***: missingness differs between patients and observed covariates can explain difference * Equivalent to propensity score-based balance measures (Austin PC, Multivariate Behavioral Research, 46:3, 399-424 (2011)	**Low p-values:** Indicate differences in covariate distributions and null hypothesis would be rejected (≠**MCAR**) Hotelling H. Ann Math Stat. 2(3):360-378. (1931) & Little RJA. J Am Stat Assoc. 83(404):1198-1202. doi:10.2307/2290157 (1988)	**Values around 0.5:** Indicate random prediction (**MCAR**) Values meaningfully above 0.5 indicate stronger correlations between covariates (which can be determined!) and missingness (~MAR)

Empirical Diagnostics to Characterize Missingness Mechanisms

	Group 1 Diagnostics		Group 2 Diagnostics	Group 3 Diagnostics
	Absolute standardized mean difference (ASMD)	**P-value Hoteling/Little**	**AUC (are under the receiver operating curve)**	**Log HR (missingness indicator)**
Purpose	Comparison of distributions between patients with vs w/o observed value of the partially observed covariate		Assessing the ability to predict missingness based on observed covariates	Check whether missingness of a covariate is associated with the outcome (differential missingness)
Example value	ASMD = 0.1	p-value <0.001	AUC = 0.5	log HR = 0.1 (0.05 to 0.2)
Interpretation	**<0.1***: missingness is not associated with other observed covariates may be completely at random **>0.1***: missingness differs between patients and observed covariates can explain difference * Equivalent to propensity score-based balance measures (Austin PC, Multivariate Behavioral Research, 46:3, 399-424 (2011))	**Low p-values:** Indicate differences in covariate distributions and null hypothesis would be rejected (≠**MCAR**) Hotelling H. Ann Math Stat. 2(3):360-378. (1931) & Little RJA. J Am Stat Assoc. 83(404):1198-1202. doi:10.2307/2290157 (1988)	**Values around 0.5:** Indicate random prediction (**MCAR**) Values meaningfully above 0.5 indicate stronger correlations between covariates (which can be determined!) and missingness (~MAR)	**MCAR**: No association in neither crude nor adjusted model **MAR**: Association in crude but not adjusted model **MNAR**: If there was a **meaningful difference also after comprehensive adjustment (log HR)**, this may be indicative of **differential MNAR scenarios**

Diagnostics Results

- Large scale simulation revealed characteristic patterns of the diagnostic parameters matched to missing data structure
- The observed diagnostic pattern of a specific study will give insights into the likelihood of underlying missingness structures

	Group 1 Diagnostics		Group 2 Diagnostics	Group 3 Diagnostics	
Expected parameter constellations	**ASMD** (Absolute standardized mean difference)	**P-value** Hoteling/Little	**AUC** (are under the receiver operating curve)	**Log HR** (crude)	**Log HR** (adjusted)
MCAR	0.05	0.5	0.50	-0.01	0.00
MAR	0.20	<.001	0.58	0.53	0.00
$\text{MNAR}_{\text{unmeasured}}$	0.09	0.02	0.54	0.43	0.31
$\text{MNAR}_{\text{value}}$	0.06	0.10	0.53	0.04	0.10

Diagnostics Results

- Large scale simulation revealed characteristic patterns of the diagnostic parameters matched to missing data structure
- The observed diagnostic pattern of a specific study will give insights into the likelihood of underlying missingness structures

Expected parameter constellations	Group 1 Diagnostics: ASMD (Absolute standardized mean difference)	Group 1 Diagnostics: P-value Hoteling/Little	Group 2 Diagnostics: AUC (are under the receiver operating curve)	Group 3 Diagnostics: Log HR (crude)	Group 3 Diagnostics: Log HR (adjusted)
MCAR	0.05	0.5	0.50	-0.01	0.00
MAR	0.20	<.001	0.58	0.53	0.00
$MNAR_{unmeasured}$	0.09	0.02	0.54	0.43	0.31
$MNAR_{value}$	0.06	0.10	0.53	0.04	0.10

Let's have a look at some EHR examples:

Covariate	ASMD (min to max)	P-value	AUC	Log HR (crude, 95% CI)	Log HR (adjusted, 95% CI)
EGFR (cancer biomarker)	**0.24** (0.01 to 0.49)	**<.001**	**0.63**	0.06 (-0.03 to 0.15)	-0.01 (-0.10 to 0.09)

Diagnostics Results

- Large scale simulation revealed characteristic patterns of the diagnostic parameters matched to missing data structure
- The observed diagnostic pattern of a specific study will give insights into the likelihood of underlying missingness structures

	Group 1 Diagnostics		Group 2 Diagnostics	Group 3 Diagnostics	
Expected parameter constellations	**ASMD** (Absolute standardized mean difference)	**P-value** Hoteling/Little	**AUC** (are under the receiver operating curve)	**Log HR** (crude)	**Log HR** (adjusted)
MCAR	0.05	0.5	0.50	-0.01	0.00
MAR	0.20	<.001	0.58	0.53	0.00
$MNAR_{unmeasured}$	0.09	0.02	0.54	0.43	0.31
$MNAR_{value}$	0.06	0.10	0.53	0.04	0.10

Let's have a look at some EHR examples:

Covariate	ASMD (min to max)	P-value	AUC	Log HR (crude, 95% CI)	Log HR (adjusted, 95% CI)
EGFR (cancer biomarker)	**0.24** (0.01 to 0.49)	**<.001**	**0.63**	0.06 (-0.03 to 0.15)	-0.01 (-0.10 to 0.09)
ECOG (performance status)	**0.03** (0.00 to 0.07)	**0.78**	**0.51**	-0.06 (-0.16 to 0.03)	-0.06 (-0.16 to 0.03)

Diagnostics Results

- Large scale simulation revealed characteristic patterns of the diagnostic parameters matched to missing data structure
- The observed diagnostic pattern of a specific study will give insights into the likelihood of underlying missingness structures

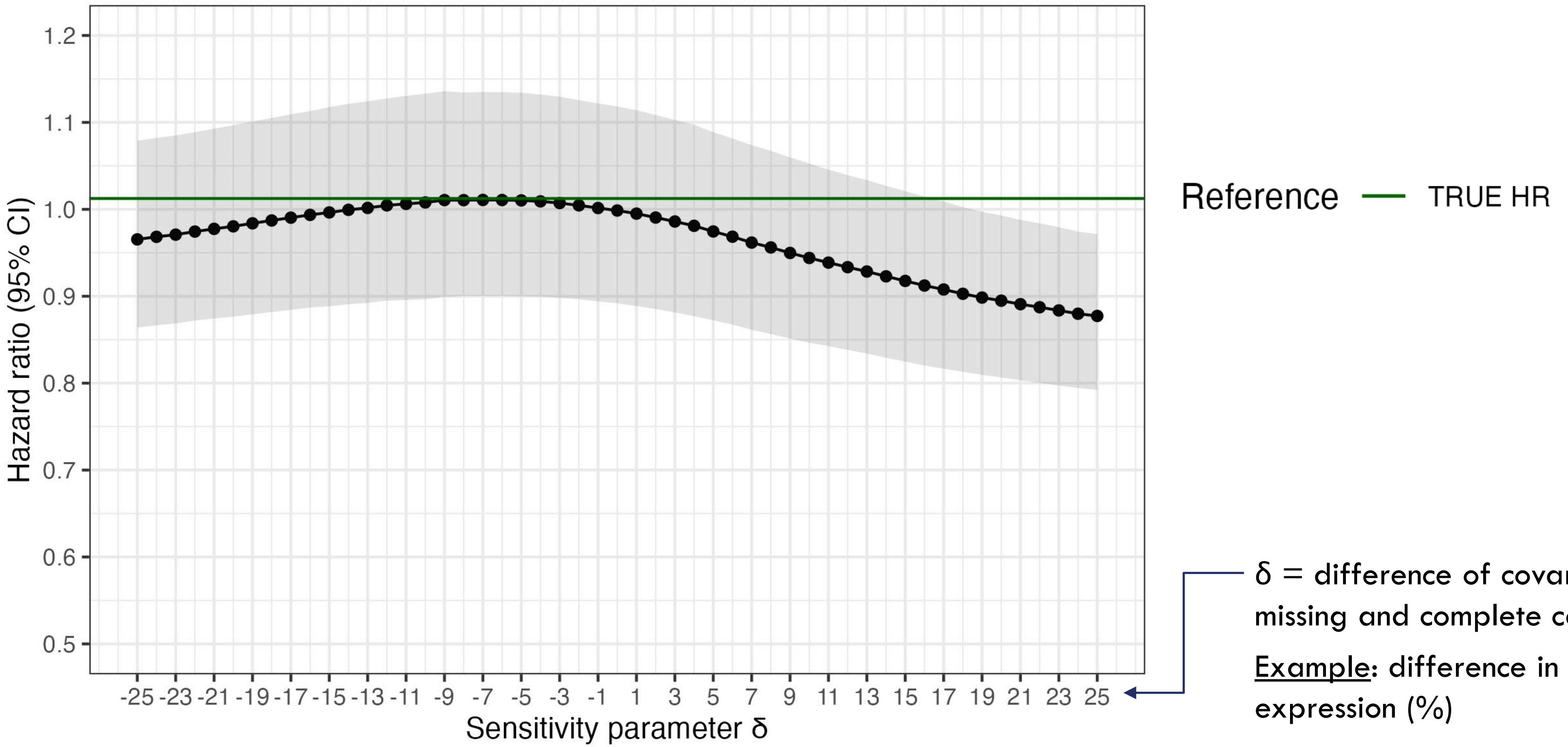
	Group 1 Diagnostics		Group 2 Diagnostics	Group 3 Diagnostics	
Expected parameter constellations	**ASMD (Absolute standardized mean difference)**	**P-value Hoteling/Little**	**AUC (are under the receiver operating curve)**	**Log HR (crude)**	**Log HR (adjusted)**
MCAR	0.05	0.5	0.50	-0.01	0.00
MAR	0.20	<.001	0.58	0.53	0.00
$MNAR_{unmeasured}$	0.09	0.02	0.54	0.43	0.31
$MNAR_{value}$	0.06	0.10	0.53	0.04	0.10

Let's have a look at some EHR examples:

Covariate	ASMD (min to max)	P-value	AUC	Log HR (crude, 95% CI)	Log HR (adjusted, 95% CI)
EGFR (cancer biomarker)	**0.24** (0.01 to 0.49)	**<.001**	**0.63**	0.06 (-0.03 to 0.15)	-0.01 (-0.10 to 0.09)
ECOG (performance status)	**0.03** (0.00 to 0.07)	**0.78**	**0.51**	-0.06 (-0.16 to 0.03)	-0.06 (-0.16 to 0.03)
PD-L1 (cancer biomarker)	**0.06** (0.02 to 0.34)	**<.001**	**0.52**	0.12 (0.01 to 0.23)	0.11 (-0.00, 0.22)

Sensitivity Analysis

- **Missing Not At Random (MNAR$_{value}$) typically leads to strongest bias**
- Since **key diagnostic parameters remain unobservable,** we cannot determine the amount of bias caused by MNAR$_{value}$
- Sensitivity tipping point analysis: How sensitive are results to a departure from MAR?



δ = difference of covariate dist^{n} in the missing and complete cases

Example: difference in mean PD-L1 expression (%)

Tompsett DM, Leacy F, Moreno-Betancur M, Heron J, White IR. Stat Med. 2018 Jul 10;37(15):2338-2353

Toolkit - R Package

Easy implementation of **routine** **structural missing data investigations (smdi)**

- Selected functions:
  - ***smdi_diagnose()*** – flagship function that will return all three group diagnostics evaluated in simulation study
  - *smdi_summarize()* & *smdi_vis()* – easy and quick visualization of proportion missingness as (variables can be specified; if not specified, all variables with NA will be displayed)
  - More…
- Disclaimer: Package is currently in beta testing and will be validated at DukeHealth testing site

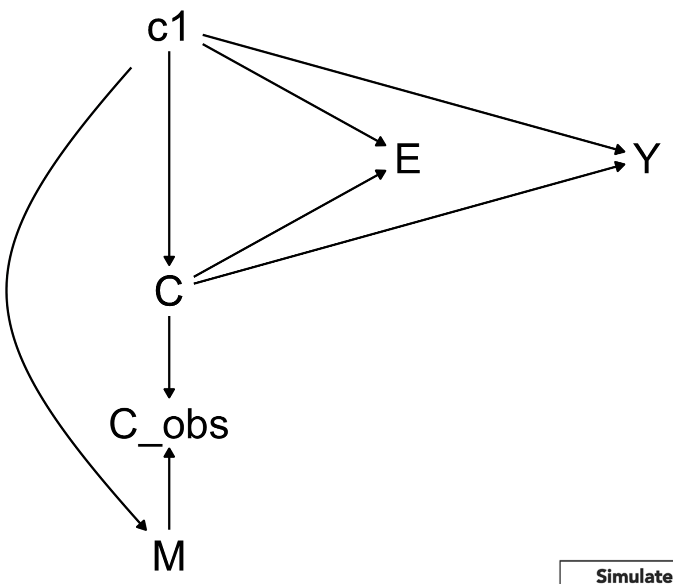
janickweberpals.gitlab-pages.partners.org/smdi

Take Home Message

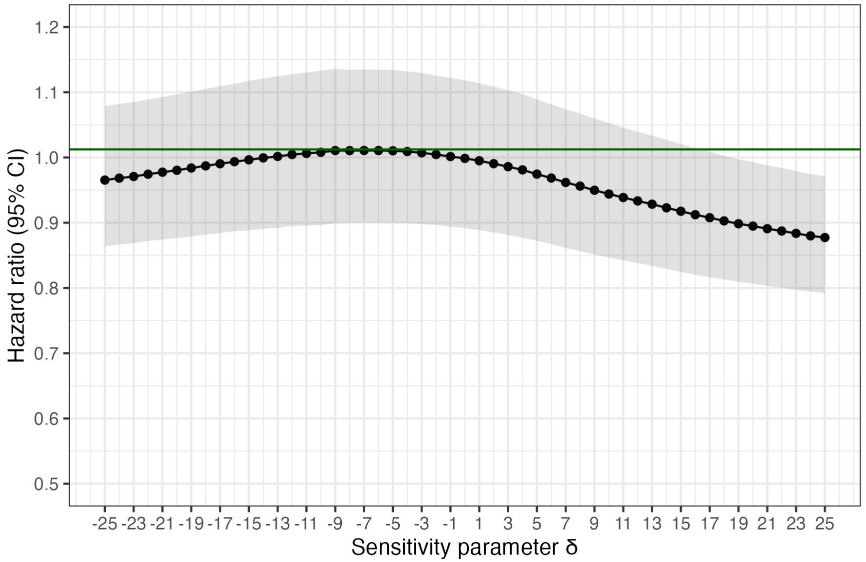
Outline your assumptions

Integrate routine diagnostics already at study design level

Check robustness of assumptions in sensitivity analyses



smdi
structural missing data investigations



Simulated mechanism	ASMD (Absolute standardized mean difference)	P-value Hoteling/Little	AUC (are under the receiver operating curve)	Log HR (crude)	Log HR (adjusted)
MCAR	0.05	0.5	0.50	-0.01	0.00
MAR	0.20	<.001	0.58	0.53	0.00
$MNAR_{unmeasured}$	0.09	0.02	0.54	0.43	0.31
$MNAR_{value}$	0.06	0.10	0.53	0.04	0.10

Team & Acknowledgements

- **Mass General Brigham**
  - Rishi Desai
  - Bob Glynn
  - Shamika More
  - Luke Zabotka
- **Duke**
  - Sudha Raman
  - Brad Hammill
- **Kaiser WA**
  - Pamela Shaw
- **Harvard Pilgrim/SOC**
  - Darren Toh
  - John Connolly
  - Kimberly J. Dandreo Gegear
- **FDA**
  - Fang Tian
  - Wei Liu
  - Hana Lee
  - Jenni Li
  - Jose Hernandez