Robust Causal Inference Methods for Electronic Health Record-Based Studies with Missing Eligibility

Luke Benz

Biomedical and Health Data Sciences Collaborative Winter Symposium
Tufts University
December 10, 2025

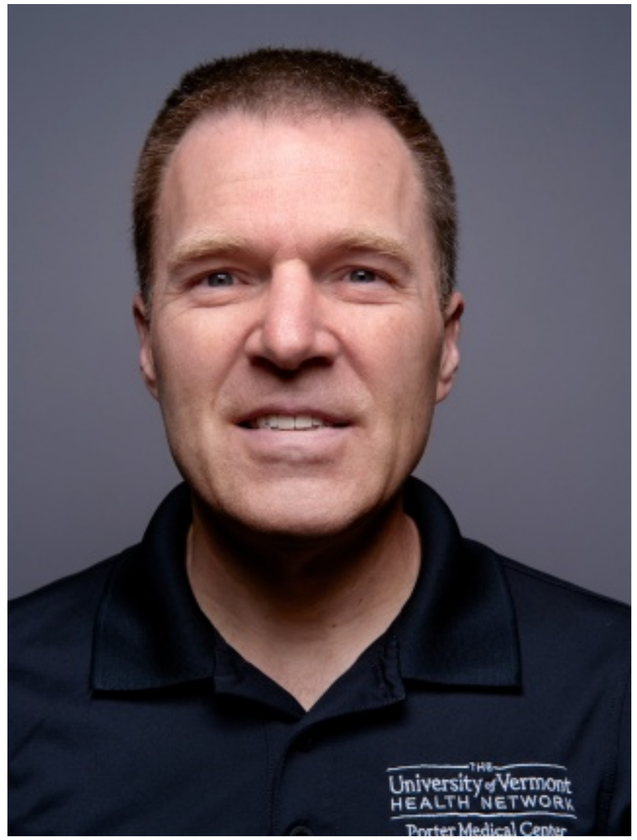
HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

Electronic Health Records Aren't Designed How a Doctor Thinks

Otherwise known as, "How I got interested in Electronic Health Records"

Eric Benz, MD
(~2005)

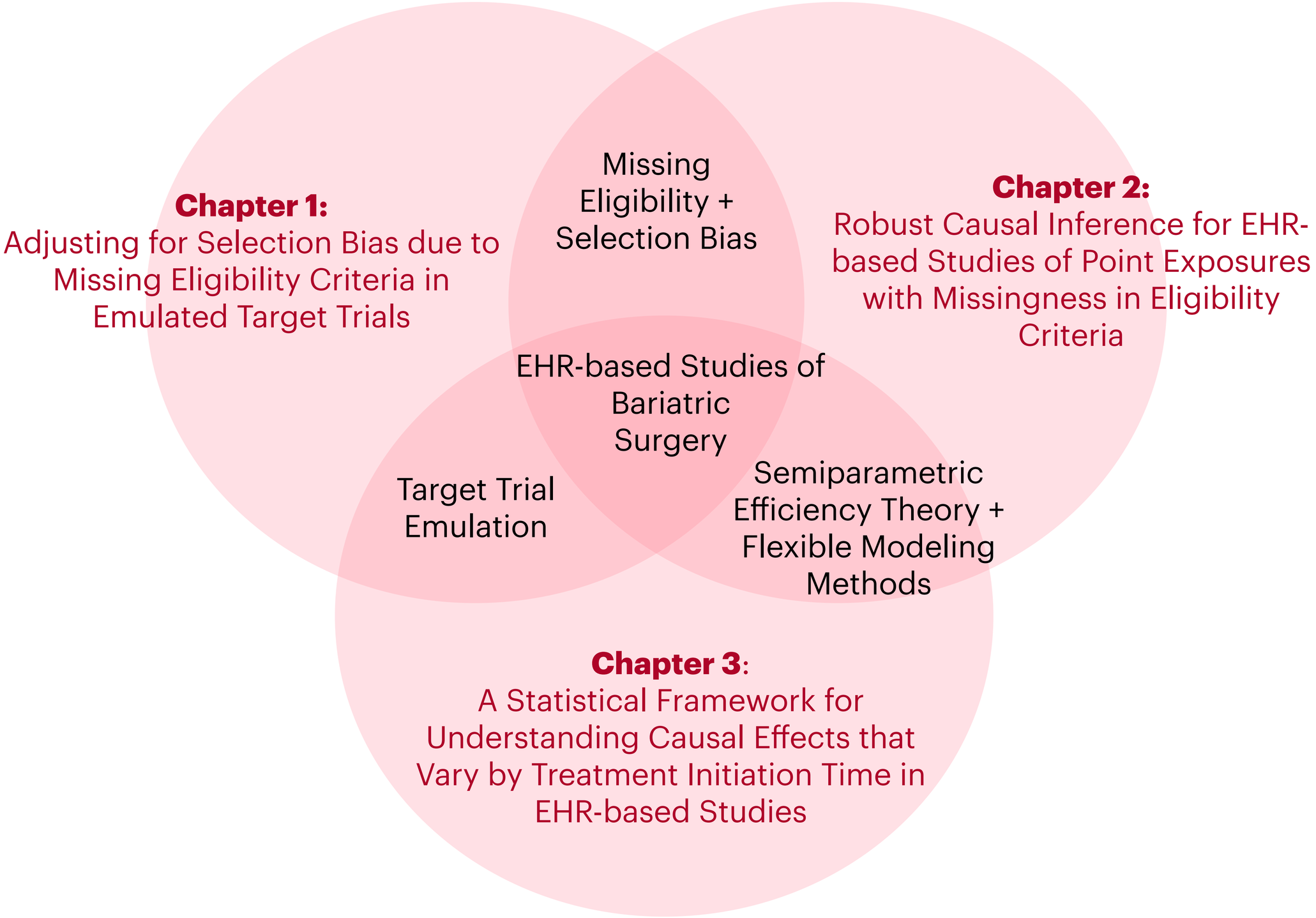
"I'll wear a tie to work again the day once EHR allows me to see more patients rather than slowing me down"



Eric Benz, MD
(~2025)

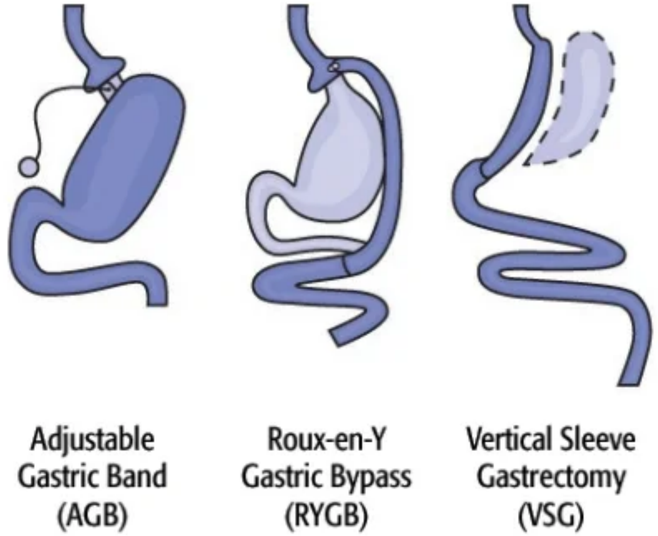
Electronic Health Records Aren't Designed How a Statistician Thinks

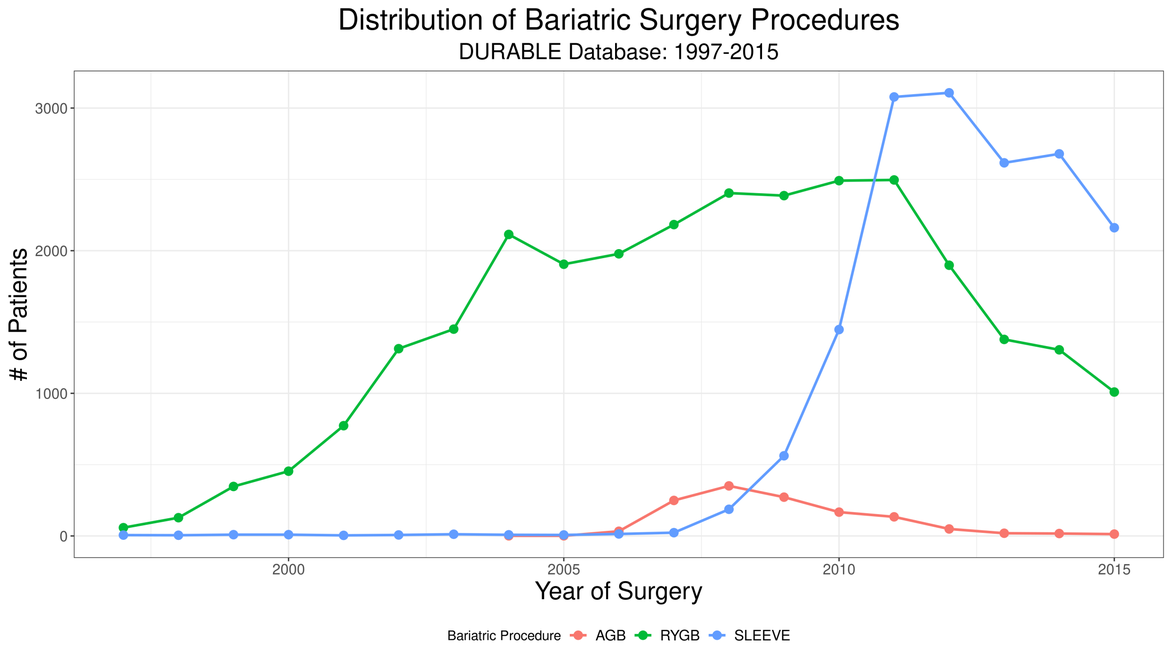
- EHR aren't designed for research purposes
  - Treatments are not random **(confounding)**
  - Information a researcher wants isn't available when they want it **(missing data, selection bias)**
  - Not always clear when to start follow-up **(immortal time bias)**
  - Treatment patterns change over time
    - Who gets treated
    - Treatment techniques or best practices can evolve
- Analysis of EHR-based studies requires statistical methods that address/acknowledge these challenges


Chapter 1:
Adjusting for Selection Bias due to Missing Eligibility Criteria in Emulated Target Trials
Missing Eligibility + Selection Bias
Chapter 2:
Robust Causal Inference for EHR-based Studies of Point Exposures with Missingness in Eligibility Criteria
EHR-based Studies of Bariatric Surgery
Target Trial Emulation
Semiparametric Efficiency Theory + Flexible Modeling Methods
Chapter 3:
A Statistical Framework for Understanding Causal Effects that Vary by Treatment Initiation Time in EHR-based Studies

Bariatric Surgery

- Bariatric surgery is a weight-loss surgery
  - Typical candidates have BMI $\geq 35 \text{kg/m}^2$
- Sleeve Gastrectomy (SG) is a newer procedure than Roux-en-Y Gastric Bypass (RYGB)
  - SG surpassed RYGB in popularity in late 2000s/ early 2010s
  - Less invasive and technically less complex
- DURABLE: NIH funded study of long-term outcomes following bariatric surgery, particularly in relation to non-surgical patients
  - Kaiser Permanente (Washington, N. California, S. California)
  - 1997-2015; ~45,000 surgical patients and 1.7 million non surgical patients



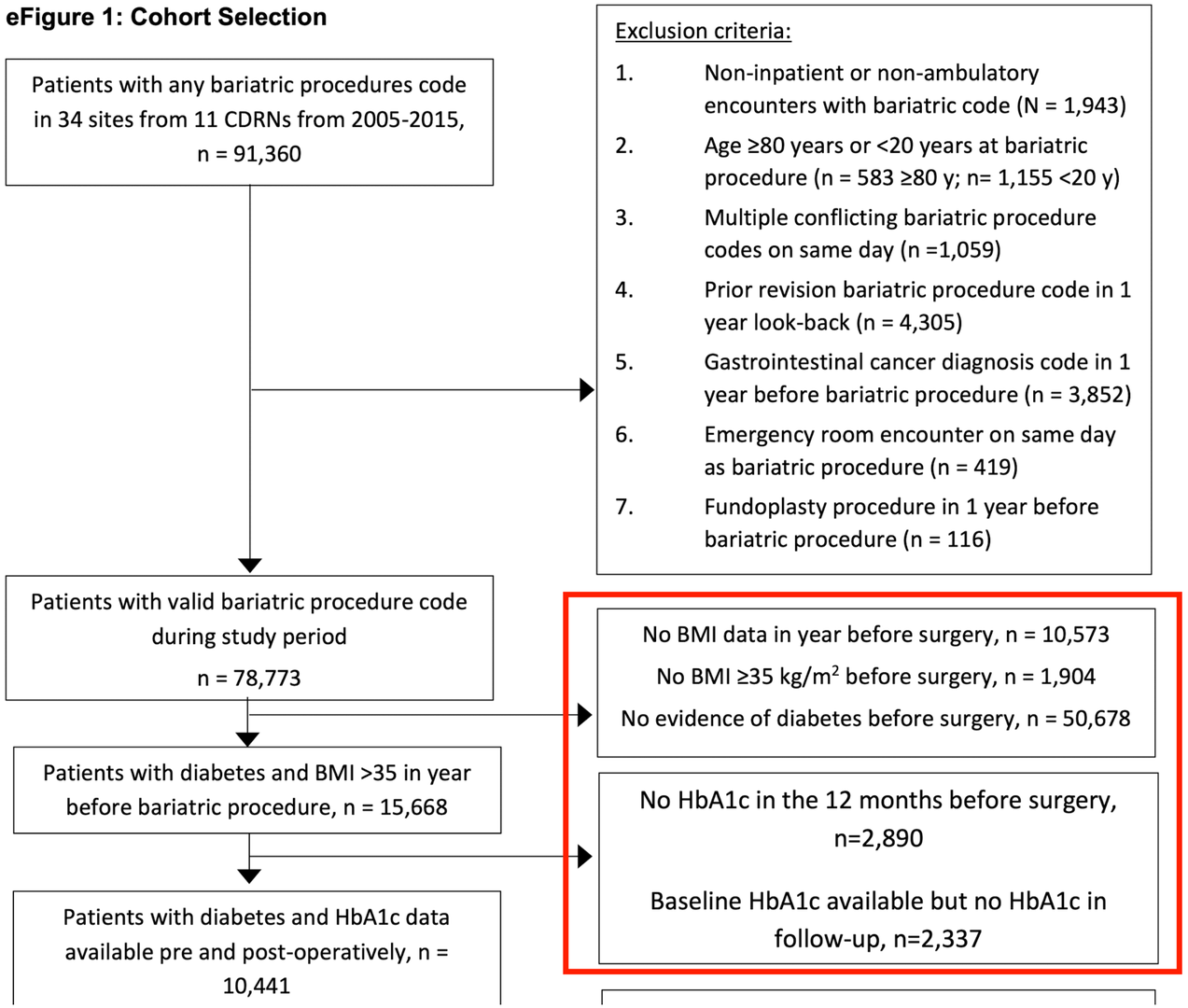


Comparisons of RYGB and SG

- Rise in popularity of SG outpaced the rate at which long-term evidence was generated for diabetic populations
    - American Diabetes Association recommends consideration of bariatric surgery for patients with T2DM
    - Doesn't make any recommendations about choice of procedure
- (McTigue et al., 2020) compared RYGB and SG among T2DM patients:
    - RYGB: better rates of remission at 3 years (84.3% vs. 81.5%)
    - RYGB: greater weight loss at 3 years (26.2% vs. 19.2%)

Cohort Flowchart

eFigure 1: Cohort Selection



Missing Information on Eligibility in EHR-based Studies

- Defining eligibility is clearly a critical component of any study design
  - Defines the ***target population***
- In practice, subjects with missing or incomplete eligibility information are often excluded from analysis
  - Handled as if they are ineligible
- Depending on the context and the approach taken there is the potential for
  - Loss of power
  - Selection bias

Selection Bias due to Missing Eligibility Criteria

12 time NBA All-Star, Isaiah Thomas, on not making the 1992 United States Olympic Team (The Dream Team) in *The Last Dance*

His comments also summarize why we might worry about selection bias when excluding patients with missing eligibility criteria

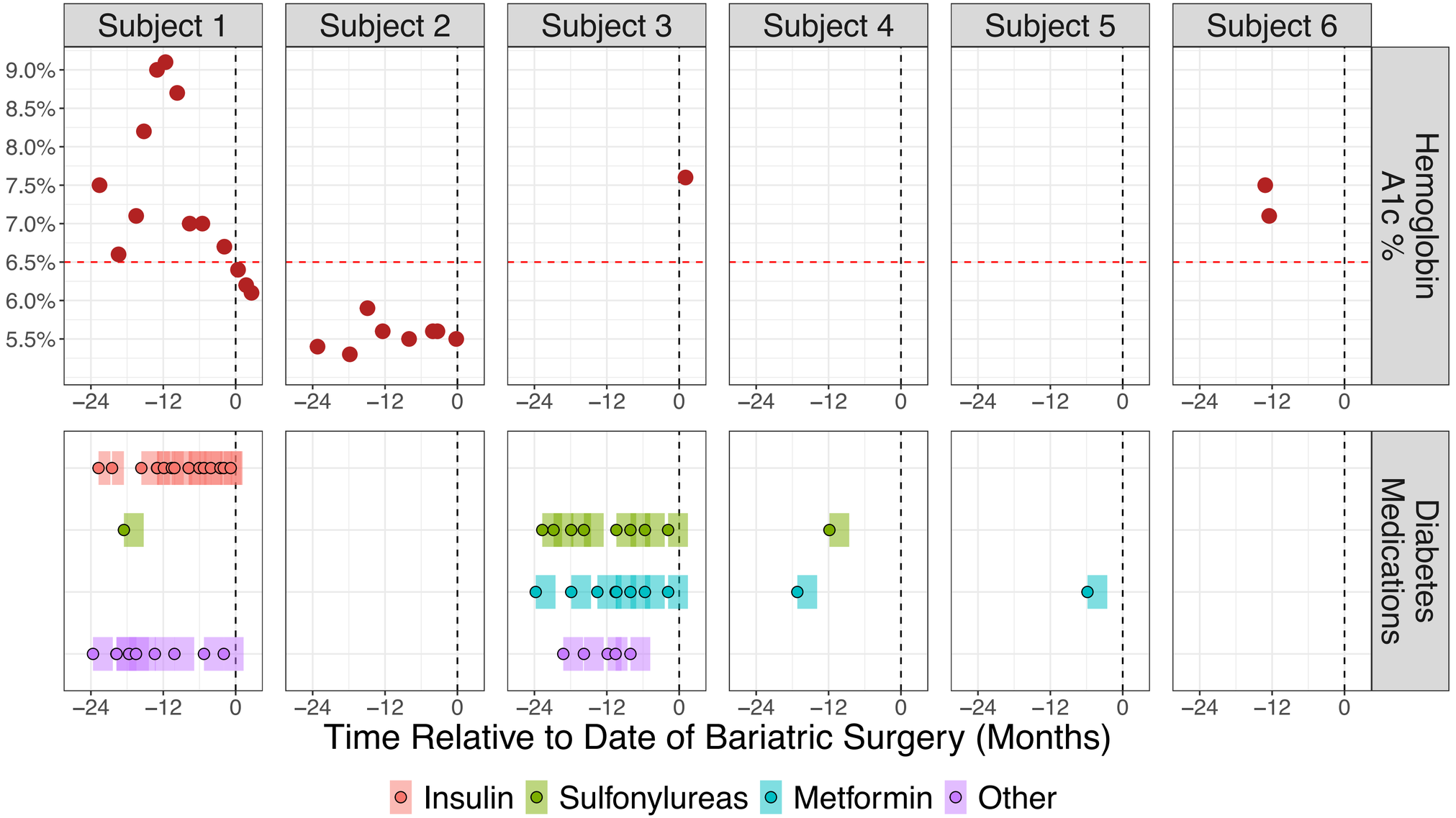
Missing Information on Eligibility in EHR-based Studies

- Superficially, missing eligibility criteria may seem like a standard missing data problem.
- A key difference however, is that we are in a position to make concessions
    - But need to be wary of trade offs
- For example, one may be in a position to control the “lookback window”
    - How far prior to study baseline to query eligibility
- Longer lookbacks
    - Reduce missingness amount → increase sample size, enhance power
    - Lower quality information?
    - Change the nature of missingness assumptions being invoked (often implicitly) → less plausible?

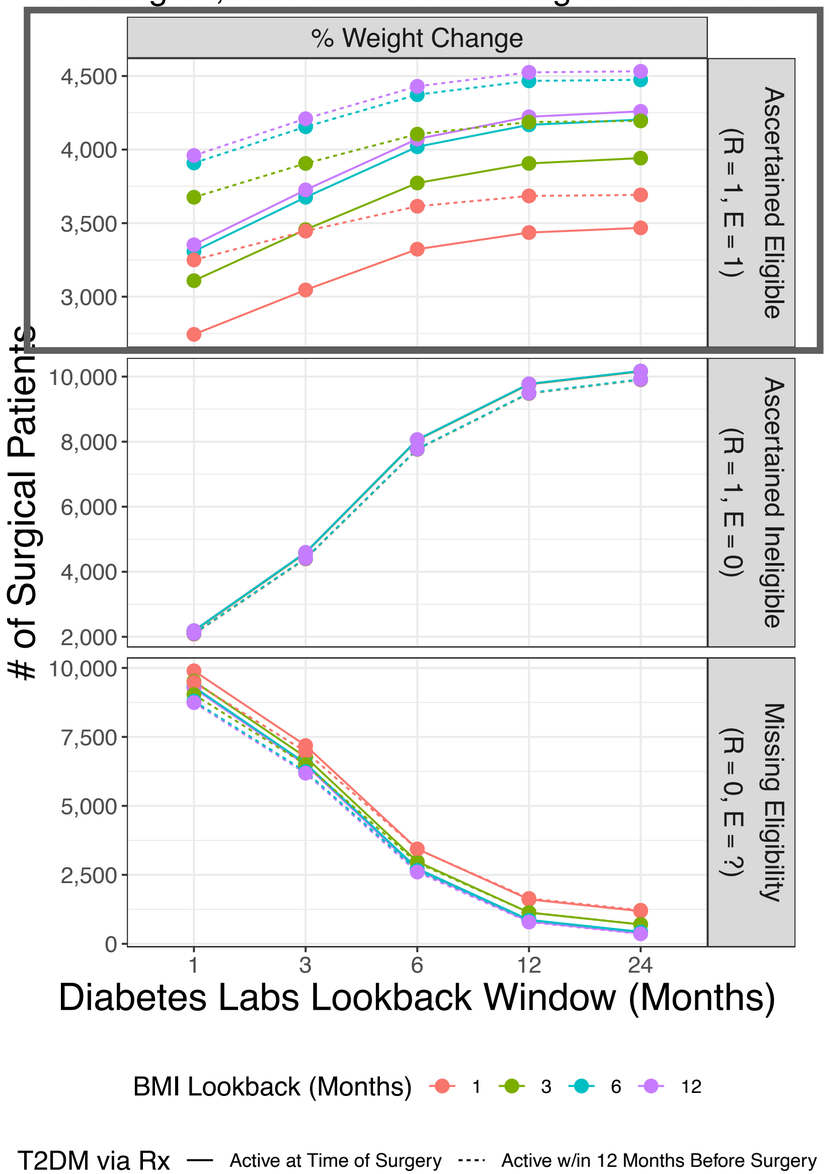
RYGB vs. SG using DURABLE

- 14,809 patients undergoing RYGB or SG
  - Between 2008-2011 (surge in popularity of SG)
- Outcomes
  - 3 Year %-Weight Change
  - Remission of T2DM at any point w/in 3 years of surgery
- Eligibility Criteria:
  - BMI $\geq 35 kg/m^2$
  - T2DM
  - Age 19-79
  - DiaRem ≥ 3 (remission outcome)
    - This additional eligibility restriction places particular emphasis on patients less likely to experience remission (e.g., with greater disease severity), a population where differences between RYGB and SG on T2DM remission rates may be more substantial

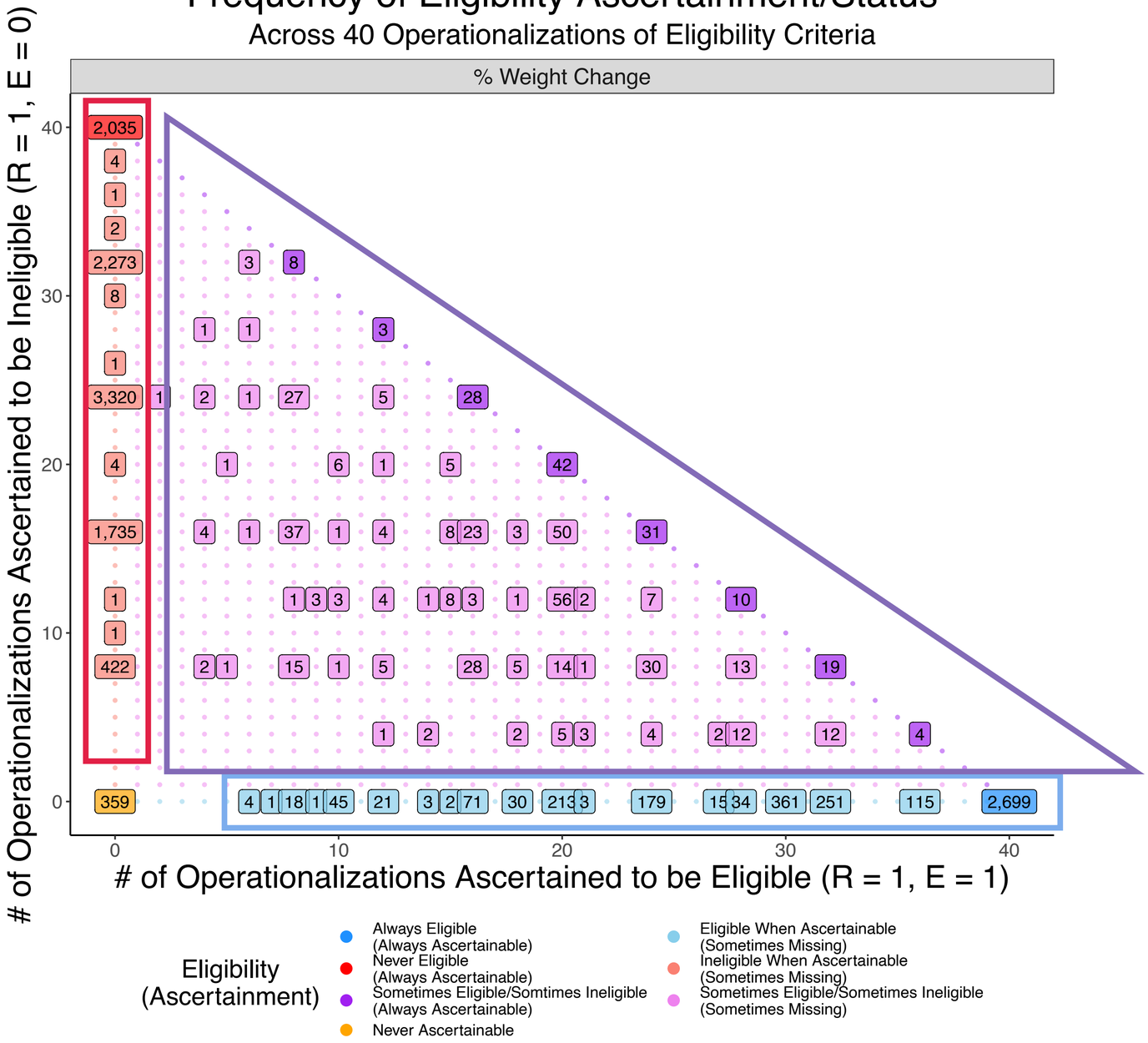
Ascertainment of T2DM



A)


Eligibility Ascertainment Distribution
Among 14,809 Patients Receiving RYGB or SG
% Weight Change
Ascertained Eligible (R = 1, E = 1)
Ascertained Ineligible (R = 1, E = 0)
Missing Eligibility (R = 0, E = ?)
# of Surgical Patients
Diabetes Labs Lookback Window (Months)
BMI Lookback (Months) 1 3 6 12
T2DM via Rx Active at Time of Surgery Active w/in 12 Months Before Surgery

B)


Frequency of Eligibility Ascertainment/Status
Across 40 Operationalizations of Eligibility Criteria
% Weight Change
# of Operationalizations Ascertained to be Ineligible (R = 1, E = 0)
# of Operationalizations Ascertained to be Eligible (R = 1, E = 1)
Eligibility (Ascertainment)
Always Eligible (Always Ascertainable)
Never Eligible (Always Ascertainable)
Sometimes Eligible/Somtimes Ineligible (Always Ascertainable)
Never Ascertainable
Eligible When Ascertainable (Sometimes Missing)
Ineligible When Ascertainable (Sometimes Missing)
Sometimes Eligible/Sometimes Ineligible (Sometimes Missing)

Goal for this Work

- Mechanisms in EHR are hard to model because we don't always know the structure of how decisions are made or how data arises
  - How **treatment** is selected
  - Why some patients have **observed information** and others don't
  - Complex health **outcomes**
- Might be tempted to use flexible, machine learning methods to model these mechanisms
  - Don't converge at standard $\sqrt{n}$ rates like classical parametric models
  - But parametric models may not be sufficient for the complexity of how data arises in EHR-based settings
- Use **semi parametric theory** to efficiently estimate causal contrasts
  - Minimize reliance on "correct specification" of nuisance functions
  - Enable use of ML tools while retaining desirable properties ("multiply robust")

Notation and Problem Set-Up

- A: binary treatment (1 = RYGB, 0 = SG)
- Y: outcome (3-year weight change, T2DM remission)
- $\boldsymbol{L}$: baseline covariates sufficient for 3 purposes
    1. Control confounding
    2. Define study eligibility
    3. Predict missingness in eligibility
- $\boldsymbol{L}^e$: Eligibility defining covariates (BMI, Age, A1c, T2DM medications, DiaRem)
    - $\boldsymbol{L}_m^e \subseteq \boldsymbol{L}^e$: Covariates needed to define eligibility subject to missingness
- $\boldsymbol{L}^* = \boldsymbol{L} \setminus \boldsymbol{L}_m^e$: Everything else
- E: Binary eligibility indicator
    - $E = g(\boldsymbol{L}^e, A)$ for some fixed eligibility rule $g(\cdot)$
- R: Complete case indicator for $\boldsymbol{L}_m^e$

Notation and Problem Set-Up

- Estimand of interest

$$\theta^{\text{elig}}_{\text{ATT}} := \mathbb{E}_P[Y(1) - Y(0) \mid A = 1, E = 1]$$

- What complete case implicitly targets

$$\theta^{\text{CC}}_{\text{ATT}} := \mathbb{E}_P[Y(1) - Y(0) \mid A = 1, E = 1, R = 1]$$

- Selection Bias

$$\theta^{\text{elig}}_{\text{ATT}} \neq \theta^{\text{CC}}_{\text{ATT}}$$

- Observed data:

$$O_1, \ldots, O_n \sim P, \text{ where } O = (\boldsymbol{L}^*, A, Y, R, R\boldsymbol{L}^e_m)$$

Assumptions

1. **Consistency:** $Y(A) = Y$ when $E = 1$, almost surely
2. **Positivity:** $0 < \epsilon \leq P(A = 1 \mid \boldsymbol{L}, E = 1) \leq 1 - \epsilon < 1$, almost surely
3. **No Unmeasured Confounding:** $Y(a) \perp\!\!\!\perp A \mid \boldsymbol{L}, E = 1$ for $a \in \{0,1\}$
4. **MAR**: $R \perp\!\!\!\perp (Y, \boldsymbol{L}_m^e) \mid \boldsymbol{L}^*, A$
5. **Complete Case Positivity:** $0 < \epsilon \leq P(R = 1 \mid \boldsymbol{L}^*, A)$

Likelihood Factorization + Identification

$$p(O) = p(\boldsymbol{L}^*) \times p(A \mid \boldsymbol{L}^*) \times p(R \mid \boldsymbol{L}^*, A) \times p(\boldsymbol{L}_m^e \mid \boldsymbol{L}^*, A, R = 1)^R \times p(Y \mid \boldsymbol{L}^*, \boldsymbol{L}_{\boldsymbol{m}}^{\boldsymbol{e}}, A, R = 1)$$

- No component conditions on covariate stratum w/ incomplete information
  - Each component is readily estimable from **observed** data
- Components are variational independent
  - Any set of valid choices results in a valid joint density
  - Choose any modeling strategy

Identification Result

Under Assumptions 1-5, $\theta^{\text{elig}}_{\text{ATT}}$ is identified by the functional $\theta(P) = \dfrac{\beta(P)}{\alpha(P)}$ where

$$\beta(P) = \mathbb{E}_P\left[\frac{ARE}{\eta(\boldsymbol{L}^*,1)}\Big(Y - \mu_0(\boldsymbol{L}^*, \boldsymbol{L}^e_m)\Big)\right]$$

$$\alpha(P) = \mathbb{E}_P\left[\frac{ARE}{\eta(\boldsymbol{L}^*,1)}\right]$$

- $\eta(\boldsymbol{L}^*, A) = P(R = 1 \mid \boldsymbol{L}^*, A)$
  - Ascertainment Probability (Complete-Case Probability)
- $\mu_a(\boldsymbol{L}^*, \boldsymbol{L}^e_m) = \mathbb{E}_P[Y \mid A = a, \boldsymbol{L}^*, \boldsymbol{L}^e_m, R = 1]$
  - Outcome Regression

Estimation Strategies

$$\widehat{\theta}_{\text{CC}} = \frac{\mathbb{P}_n\left[ARE\left(Y - \widehat{\mu}_0(\boldsymbol{L}^*, \boldsymbol{L}_m^e)\right)\right]}{\mathbb{P}_n[ARE]}$$

$$\widehat{\theta}_{\text{IWOR}} = \frac{\mathbb{P}_n\left[\frac{ARE}{\widehat{\eta}(\boldsymbol{L}^*,1)}\left(Y - \widehat{\mu}_0(\boldsymbol{L}^*, \boldsymbol{L}_m^e)\right)\right]}{\mathbb{P}_n\left[\frac{ARE}{\widehat{\eta}(\boldsymbol{L}^*,1)}\right]}$$

$$\widehat{\theta}_{\text{IF}} = \frac{\mathbb{P}_n[\dot{\beta}'_{\hat{P}}(O)]}{\mathbb{P}_n[\dot{\alpha}'_{\hat{P}}(O)]}$$

$$\widehat{\theta}_{\text{EIF}} = \frac{\mathbb{P}_n[\dot{\beta}_{\hat{P}}(O)]}{\mathbb{P}_n[\dot{\alpha}_{\hat{P}}(O)]}$$

- $\widehat{\theta}_{\text{CC}}$ complete case estimator of ATT
  - Average of $Y - \widehat{\mu}_0(\boldsymbol{L})$ among treated, eligible complete-case population
  - Most similar to (McTigue et al., 2020)
- $\widehat{\theta}_{\text{IWOR}}$ plug-in estimator
- $\widehat{\theta}_{\text{EIF}}$ and $\widehat{\theta}_{\text{IF}}$ are one-step estimators
  - $\dot{\beta}'_P(O)$ and $\dot{\alpha}'_P(O)$ (uncentered) nonparametric influence functions
  - $\dot{\beta}_P(O)$ and $\dot{\alpha}_P(O)$ (uncentered) efficient influence functions
    - Projection onto the tangent space restricted by assumption 4 (MAR)

Nuisance Functions

Function	Definition	$\widehat{\theta}_{CC}$	$\widehat{\theta}_{IWOR}$	$\widehat{\theta}_{IF}$	$\widehat{\theta}_{EIF}$
$\pi(\boldsymbol{L}^*)$	$P(A=1 \mid \boldsymbol{L}^*)$				
$\eta(\boldsymbol{L}^*, A)$	$P(R=1 \mid \boldsymbol{L}^*, A)$		✓	✓	✓
$\lambda_a(\boldsymbol{L}_m^e; \boldsymbol{L}^*)$	$P(\boldsymbol{L}_m^e \mid \boldsymbol{L}^*, A=a, R=1)$				
$\mu_a(\boldsymbol{L}^*, \boldsymbol{L}_m^e)$	$\mathbb{E}[Y \mid \boldsymbol{L}^*, \boldsymbol{L}_m^e, A=a, R=1]$	✓	✓	✓	✓
$u(\boldsymbol{L}^*, \boldsymbol{L}_m^e)$	$P(A=1 \mid \boldsymbol{L}^*, \boldsymbol{L}_m^e, R=1)$			✓	✓
$\varepsilon_a(\boldsymbol{L}^*, Y)$	$P(E=1 \mid \boldsymbol{L}^*, Y, A=a, R=1)$				✓
$\xi(\boldsymbol{L}^*, Y)$	$\mathbb{E}\left[E\mu_0(\boldsymbol{L}^*, \boldsymbol{L}_m^e) \mid \boldsymbol{L}^*, Y, A=1, R=1\right]$				✓
$\gamma(\boldsymbol{L}^*, Y)$	$\mathbb{E}\left[E\frac{u(\boldsymbol{L}^*, \boldsymbol{L}_m^e)}{1-u(\boldsymbol{L}^*, \boldsymbol{L}_m^e)} \mid \boldsymbol{L}^*, Y, A=0, R=1\right]$				✓
$\chi(\boldsymbol{L}^*, Y)$	$\mathbb{E}\left[E\frac{u(\boldsymbol{L}^*, \boldsymbol{L}_m^e)}{1-u(\boldsymbol{L}^*, \boldsymbol{L}_m^e)}\mu_0(\boldsymbol{L}^*, \boldsymbol{L}_m^e) \mid \boldsymbol{L}^*, Y, A=0, R=1\right]$				✓
$\nu(\boldsymbol{L}^*)$	$\mathbb{E}\left[E(Y-\mu_0(\boldsymbol{L}^*, \boldsymbol{L}_m^e)) \mid \boldsymbol{L}^*, A=1, R=1\right]$			✓	
$\omega_a(\boldsymbol{L}^*)$	$P(E=1 \mid \boldsymbol{L}^*, A=a, R=1)$			✓	

- Tied to likelihood decomposition
- Propensity score among complete cases
  - Avoids conditional density estimation
- P(Eligible | Complete Case)
- Nested nuisance functions
  - Expectation of eligibility indicator × additional nuisance function(s)

Efficient Influence Functions of $\alpha(P)$ and $\beta(P)$

$$\dot{\alpha}_P^*(O) = A\left(1 - \frac{R}{\eta(\boldsymbol{L}^*,1)}\right)\varepsilon_1(\boldsymbol{L}^*, Y) + \frac{ARE}{\eta(\boldsymbol{L}^*,1)} - \alpha(P)$$

$$\begin{aligned}\dot{\beta}_P^*(O) = {} & \frac{AR}{\eta(\boldsymbol{L}^*,1)}\left[\Big(E - \varepsilon_1(\boldsymbol{L}^*, Y)\Big)Y - \Big(E\mu_0(\boldsymbol{L}^*, \boldsymbol{L}_m^e) - \xi(\boldsymbol{L}^*, Y)\Big)\right] + A\Big(\varepsilon_1(\boldsymbol{L}^*, Y)Y - \xi(\boldsymbol{L}^*, Y)\Big) \\ & -\frac{(1-A)R}{\eta(\boldsymbol{L}^*,1)}\left[E\frac{u(\boldsymbol{L}^*, \boldsymbol{L}_m^e)}{1 - u(\boldsymbol{L}^*, \boldsymbol{L}_m^e)}\Big(Y - \mu_0(\boldsymbol{L}^*, \boldsymbol{L}_m^e)\Big) - \Big(\gamma(\boldsymbol{L}^*, Y)Y - \chi(\boldsymbol{L}^*, Y)\Big)\right] \\ & -(1-A)\frac{\eta(\boldsymbol{L}^*,0)}{\eta(\boldsymbol{L}^*,1)}\Big(\gamma(\boldsymbol{L}^*, Y)Y - \chi(\boldsymbol{L}^*, Y)\Big) - \beta(P)\end{aligned}$$

Theoretical Properties of $\widehat{\theta}_{\text{EIF}}$

Theorem 3 *If* $\|\dot{\alpha}_{\widehat{P}} - \dot{\alpha}_P\| = o_P(1)$, $\|\dot{\beta}_{\widehat{P}} - \dot{\beta}_P\| = o_P(1)$, $\alpha(P) > 0$, *and* $P\Big[\big|\mathbb{P}_n\big(\dot{\alpha}_{\widehat{P}}(O)\big)\big| \geq \epsilon\Big] = 1$ *for some* $\epsilon > 0$, *then*

$$\widehat{\theta}_{\text{EIF}} - \theta(P) = \mathbb{P}_n[\dot{\theta}^*_P(O)] + O_P\Big(R_\alpha(\widehat{P}, P) + R_\beta(\widehat{P}, P)\Big) + o_P(n^{-1/2})$$

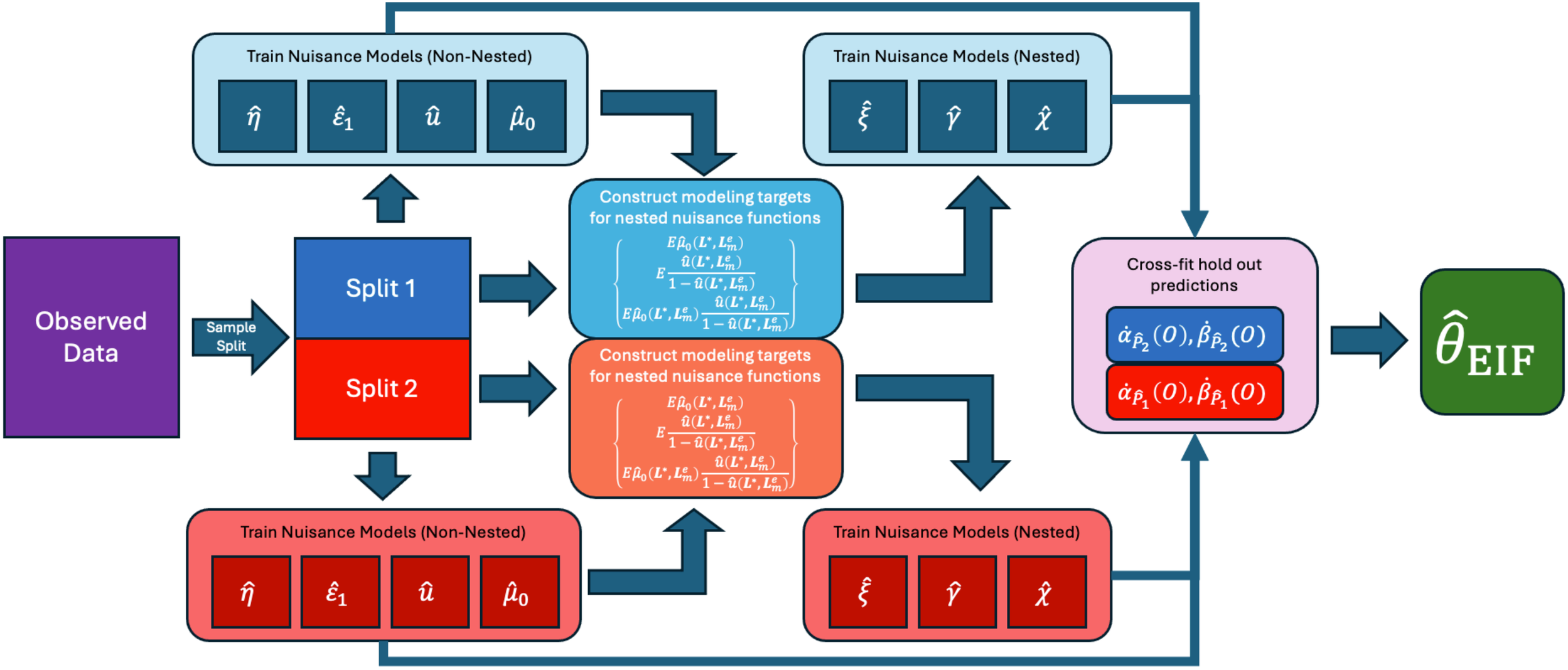
Moreover, if $R_\alpha(\widehat{P}, P) + R_\beta(\widehat{P}, P) = o_P(n^{-1/2})$ *then* $\sqrt{n}\big(\widehat{\theta}_{\text{EIF}} - \theta(P)\big) \xrightarrow{d} \mathcal{N}\big(0, \text{Var}_P[\dot{\theta}^*_P(O)]\big)$, *whereby* $\widehat{\theta}_{\text{EIF}}$ *attains the semiparametric efficiency bound induced by Assumption 4.*

Corollary 3.1 *Under the conditions of Theorem 3 and assuming that* $P(\delta \leq 1 - \widehat{u} \leq 1 - \delta) = 1$ *for some* $\delta > 0$, $P(\widehat{\eta}_1 > \epsilon) = 1$ *for some* $\epsilon > 0$, *and* $\mathbb{E}_P[Y^2] \leq M < \infty$,

$$R_\alpha(\widehat{P}, P) + R_\beta(\widehat{P}, P) = O_P\Big(\|\widehat{\mu}_0 - \mu_0\|\|\widehat{u} - u\| + \|\widehat{\eta}_1 - \eta_1\|\Big\{\|\widehat{\varepsilon}_1 - \varepsilon_1\| + \|\widehat{\xi} - \xi\|\Big\} + \|\widehat{\eta}_0 - \eta_0\|\Big\{\|\widehat{\gamma} - \gamma\| + \|\widehat{\chi} - \chi\|\Big\}\Big)$$

- Nice asymptotic behavior to facilitate valid inference, even using machine learning models for nuisances
- Something “right” for standard ATT nuisances
- Something “right” for missingness in eligibility
  - If we can model ascertainment probability well, nested nuisances don’t matter so much

Estimation of $\hat{\theta}_{\text{EIF}}$



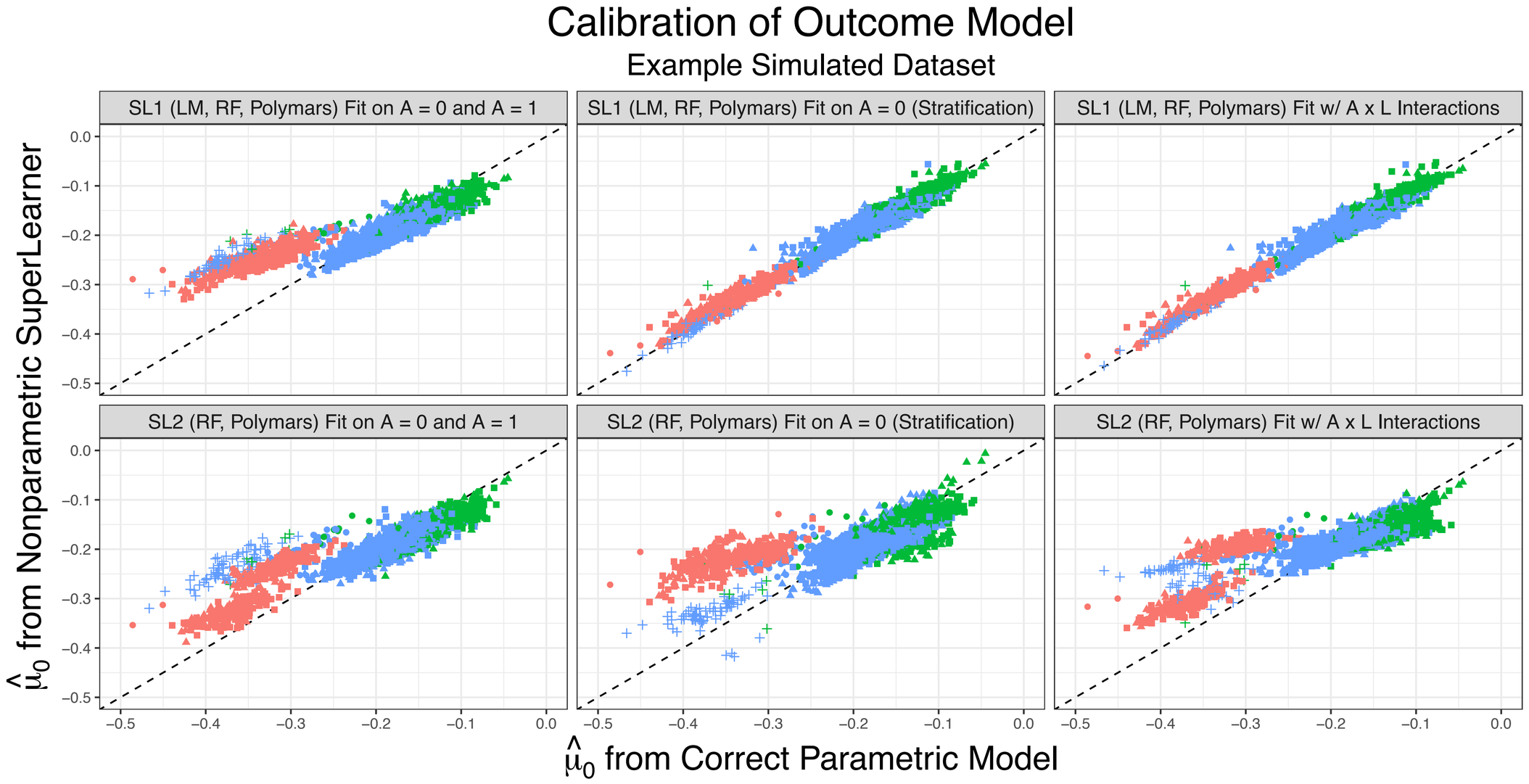
Estimator	Strategy	SL Libs[a]	μ_0 Strategy[b]	True μ/η[c]	$n = 10,000$ Patients %-Bias	$n = 10,000$ Patients SD	$n = 10,000$ Patients Coverage	$n = 25,000$ Patients %-Bias	$n = 25,000$ Patients SD	$n = 25,000$ Patients Coverage
$\widehat{\theta}_{\text{CC}}$	Parametric	—	1	μ	15.4	3.72e-03	—	15.7	2.39e-03	—
				η/μ	0.0	4.35e-03	—	0.1	2.72e-03	—
	Parametric	—	1	μ	16.5	3.77e-03	—	16.8	2.43e-03	—
				η	23.8	3.76e-03	—	24.0	2.43e-03	—
$\widehat{\theta}_{\text{IWOR}}$	Nonparametric	SL1	1	—	19.1	4.30e-03	—	17.9	3.18e-03	—
			2		3.7	4.56e-03	—	1.8	2.82e-03	—
			3		3.8	4.50e-03	—	2.2	2.86e-03	—
		SL2	1		7.9	5.99e-03	—	11.0	4.53e-03	—
			2		17.1	4.78e-03	—	11.7	3.34e-03	—
			3		17.9	5.26e-03	—	13.3	4.78e-03	—
$\widehat{\theta}_{\text{IF}}$	Nonparametric	SL1	1	—	-0.2	5.40e-03	95.0	0.1	3.21e-03	95.6
			2		-0.2	4.88e-03	93.8	0.1	2.92e-03	94.7
			3		-0.1	4.92e-03	93.3	0.1	2.91e-03	94.3
		SL2	1		0.3	5.34e-03	95.1	0.2	3.21e-03	94.8
			2		0.4	5.40e-03	93.8	0.3	3.16e-03	93.8
			3		0.5	5.36e-03	94.5	0.5	3.20e-03	94.6
$\widehat{\theta}_{\text{EIF}}$	Nonparametric	SL1	1	—	-0.4	4.89e-03	94.2	-0.2	2.83e-03	93.8
			2		-0.1	4.36e-03	94.7	-0.1	2.59e-03	95.4
			3		-0.1	4.34e-03	94.7	-0.1	2.58e-03	95.0
		SL2	1		0.0	5.05e-03	93.6	-0.1	2.92e-03	94.8
			2		0.0	5.19e-03	94.1	0.1	2.88e-03	94.5
			3		0.2	5.16e-03	94.5	0.1	2.96e-03	93.6

[a] SL Libs = `SuperLearner` libraries: S1 = Random Forest, LM/GLM, GAM, Polymars; SL2 = Random Forest, GAM, Polymars
[b] μ_0 Strategy: (1) Fit single $\widehat{\mu}$ on $A = 0, 1$ together (2) Fit $\widehat{\mu}_0$ on $A = 0$ only (stratification) (3) Specify all $A \times \boldsymbol{L}$ interactions in design matrix for $\widehat{\mu}$
[c] Correctly specified parametric models for μ and/or η

- Simulate data from factorization of observed likelihood
  - Simulate from outcome model with complex interaction structure $(A \times \boldsymbol{L}^*,\ A \times \boldsymbol{L}_m^e,\ L^* \times \boldsymbol{L}_m^e)$

- Plug-in bias
  - No robustness to partial model misspecification
- Unbiased estimation for $\widehat{\theta}_{\text{EIF}}$, $\widehat{\theta}_{\text{IF}}$
  - Robustness under partial misspecification
- Nominal coverage
- $\text{Var}[\widehat{\theta}_{\text{EIF}}] < \text{Var}[\widehat{\theta}_{\text{IF}}]$
  - Confidence intervals 4-14% wider for $\widehat{\theta}_{\text{IF}}$
- $\text{Var}[\widehat{\theta}_{\text{EIF}}] \approx \text{Var}[\widehat{\theta}_{\text{IWOR}}]$ with **correct** parametric models

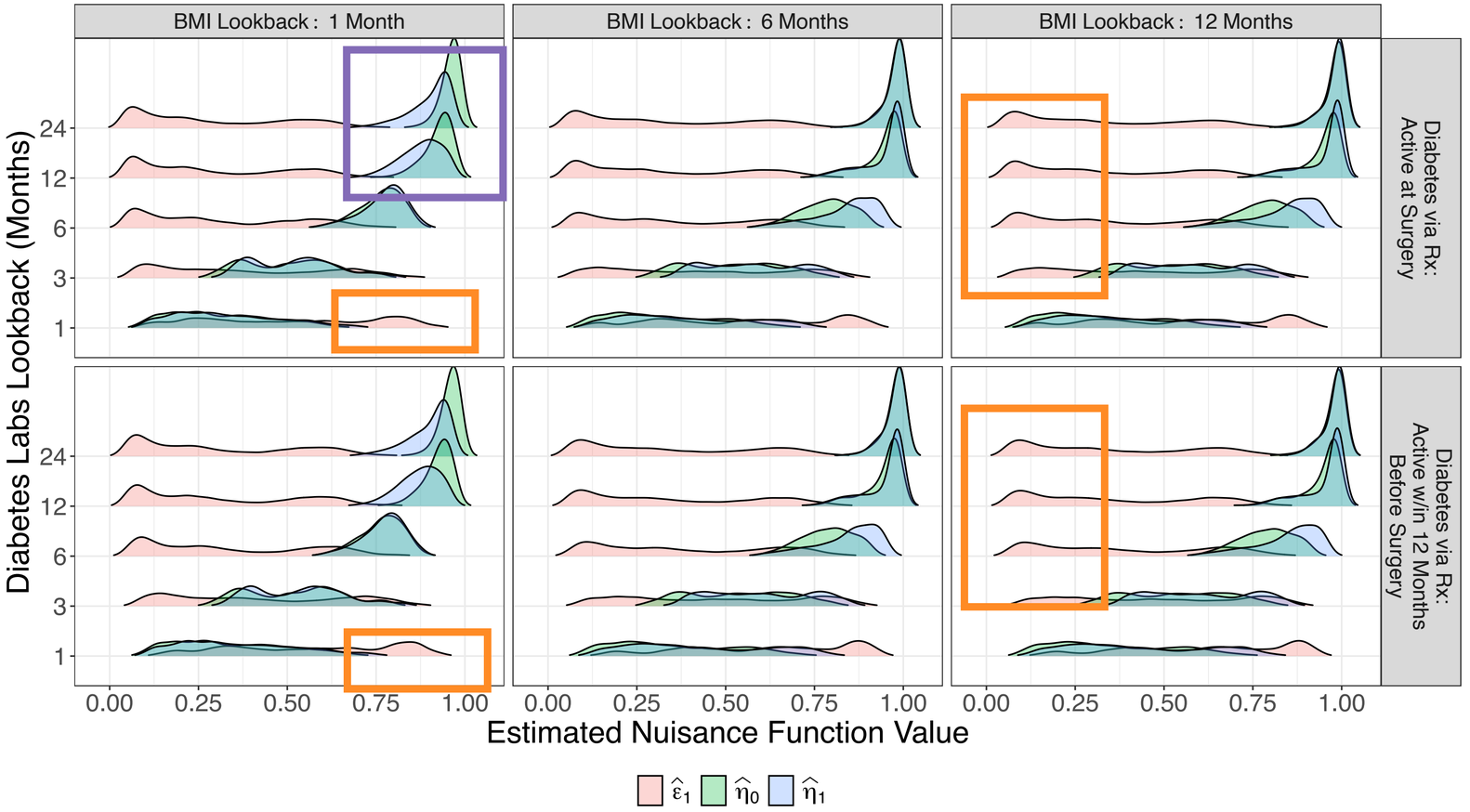
Robustness to Various Degrees of Model Misspecification



Return to Data Application

RYGB vs. SG for 3-Year % Weight Loss and T2DM Remission

- $\boldsymbol{L}^*$: surgery site, race, sex, baseline age, eGFR, smoking-status, hypertension, dyslipidemia, calendar year of surgery
- $\boldsymbol{L}_m^e$: BMI, A1c, T2DM medication usage, DiaRem
- Estimation
  - Linear model for outcome regression $\widehat{\theta}_{\text{CC}}$ and $\widehat{\theta}_{\text{IWOR}}$
    - various $A \times \boldsymbol{L}$ interactions based on those explored by (McTigue et al., 2020)
  - Logistic regression for ascertainment probabilities in $\widehat{\theta}_{\text{IWOR}}$
  - `{SuperLearner}` for all nuisance functions used by $\widehat{\theta}_{\text{IF}}$ and $\widehat{\theta}_{\text{EIF}}$
    - Appropriate sample splitting and cross fitting
    - Component libraries: random forest (`{ranger}`), Polymars (`{polspline}`), GAM, GLM

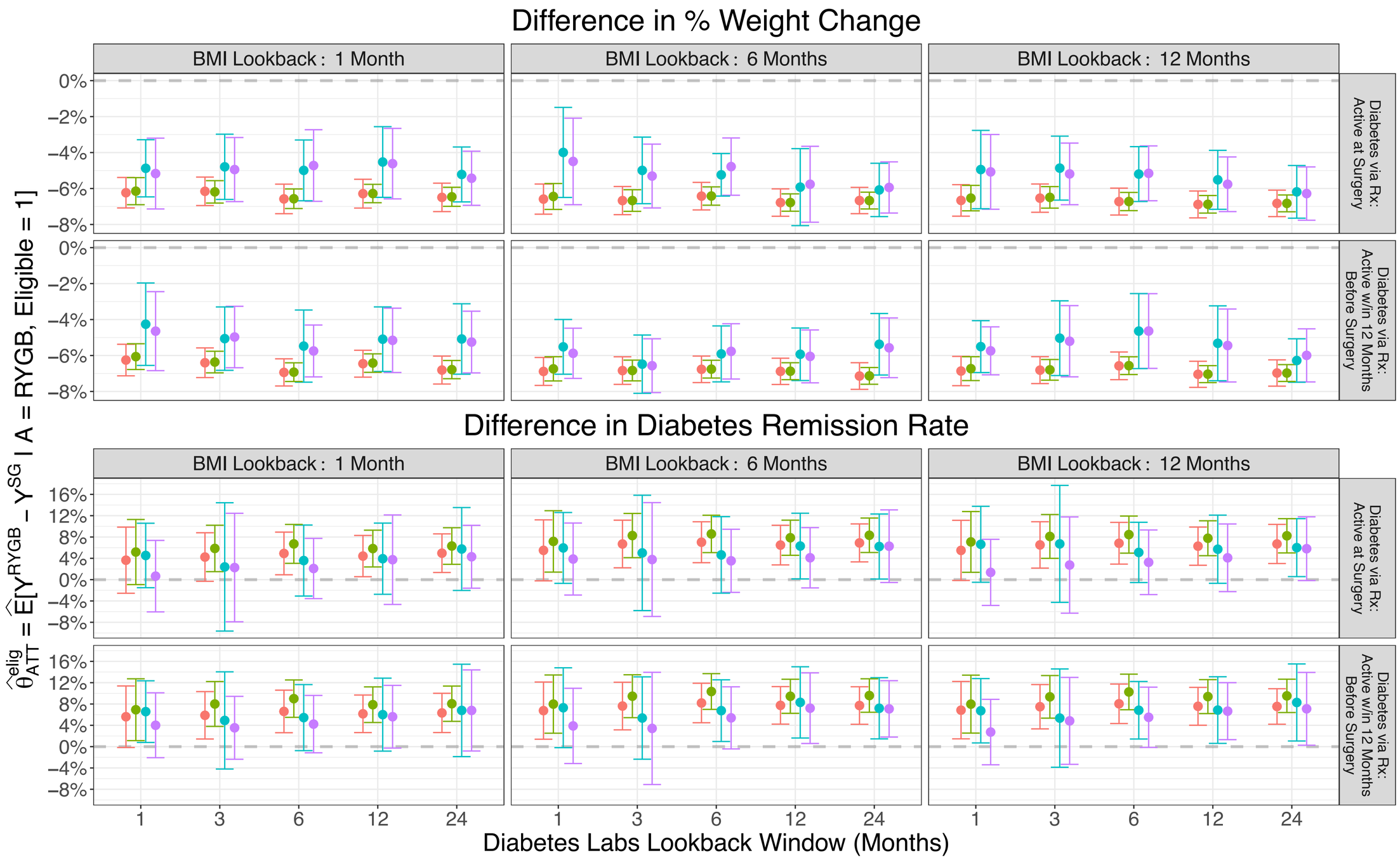


$$\hat{\eta}_a(\boldsymbol{L}^*) = \hat{P}(R = 1 \mid \boldsymbol{L}^*, A = a)$$

$$\hat{\varepsilon}_a(\boldsymbol{L}^*, Y) = \hat{P}(E = 1 \mid \boldsymbol{L}^*, Y, A = a, R = 1)$$

- Plausibility of MAR varies by lookback window?
  - Increased interaction w/ doctor prior to surgery
  - Plausible MAR more likely to hold w/in 6 months of surgery?
  - Less likely to hold 1-2 years prior to surgery?

- Missingness in BMI
  - 16.2% based on 1 month window
  - 1.5% based on 6 month window
- Missingness in A1c
  - 68.5% based on 1 month window
  - 19.2% based on 6 month window
  - T2DM main driver of missingness in E
- Differential missingness by procedure
  - $\mathbb{P}_n[\hat{\eta}_1(\boldsymbol{L}^*) - \hat{\eta}_0(\boldsymbol{L}^*)]$ ranged between -6.1% to 7.6%
  - More prominent when lookback window length differed
- $\mathbb{P}_n[\hat{\varepsilon}_1(\boldsymbol{L}^*, Y)]$ ranged from 61% in most stringent settings to 30% in most relaxed
  - Higher proportion of subjects with complete information judged to be eligible w/ narrower lookback windows
  - Potential “danger” associated with going back as far in time as possible to reduce missingness?


Difference in % Weight Change
BMI Lookback : 1 Month
BMI Lookback : 6 Months
BMI Lookback : 12 Months
Diabetes via Rx: Active at Surgery
Diabetes via Rx: Active w/in 12 Months Before Surgery
0%
−2%
−4%
−6%
−8%
Difference in Diabetes Remission Rate
16%
12%
8%
4%
0%
−4%
−8%
1
3
6
12
24
Diabetes Labs Lookback Window (Months)
$\hat{\theta}^{elig}_{ATT} = \hat{E}[Y^{RYGB} - Y^{SG} \mid A = RYGB, Eligible = 1]$
$\hat{\theta}_{CC}$
$\hat{\theta}_{IWOR}$
$\hat{\theta}_{IF}$
$\hat{\theta}_{EIF}$

Results Summary

- Complete-case analyses provided evidence that eligible subjects undergoing RYGB experienced better outcomes, with values of $\widehat{\theta}_{CC}$ ranging between
  - -7.1% to -6.2% for 3-year weight change
  - 3.7% to 8.2% for remission of T2DM w/in 3 years
  - Consistent with (McTigue et al. 2020)
- $\widehat{\theta}_{EIF}$ estimates still provide evidence of better outcomes for RYGB, but attenuated (relatively) towards the null by an average of
  - 19.5% for 3-year weight change
  - 34.1% for T2DM remission
- Estimated standard errors are on average 5.3% larger for $\widehat{\theta}_{IF}$ than corresponding standard errors for $\widehat{\theta}_{EIF}$
- Biggest differences in short T2DM lookback windows
  - Where MAR is more plausible?
  - Where there is greater consistency with definition of T2DM remission in DURABLE studies (Coleman et al. 2016)

Talk Summary

- Propose identification results and estimations strategies for ATT among those eligible, in the presence of missing eligibility data.
- Desirable properties
    - Various paths to robustness
    - Statistical efficiency
    - Variational independence
    - Choice in the use of your favorite ML tools for nuisance functions
- Highlight some practical considerations and trade-offs that analysts will have to contend with
    - Lookback windows
    - Power vs. selection bias vs. MAR
    - Missing eligibility as a 2-sided problem
- **Key Message:** Many considerations to be aware of and better to think about them before conducting analysis

Thank You

Pre-Print

GitHub Repository

Slides

Adjusting for selection bias due to missing eligibility criteria in emulated target trials

Luke Benz, Rajarshi Mukherjee, Rui Wang, David Arterburn, Heidi Fischer, Catherine Lee, Susan M Shortreed, Sebastien Haneuse

American Journal of Epidemiology, Volume 194, Issue 11, November 2025, Pages 3126–3139, https://doi.org/10.1093/aje/kwae471

Published: 26 December 2024 **Article history**

PDF Split View Cite Permissions Share

Additional work with lower barrier of entry

Appendix

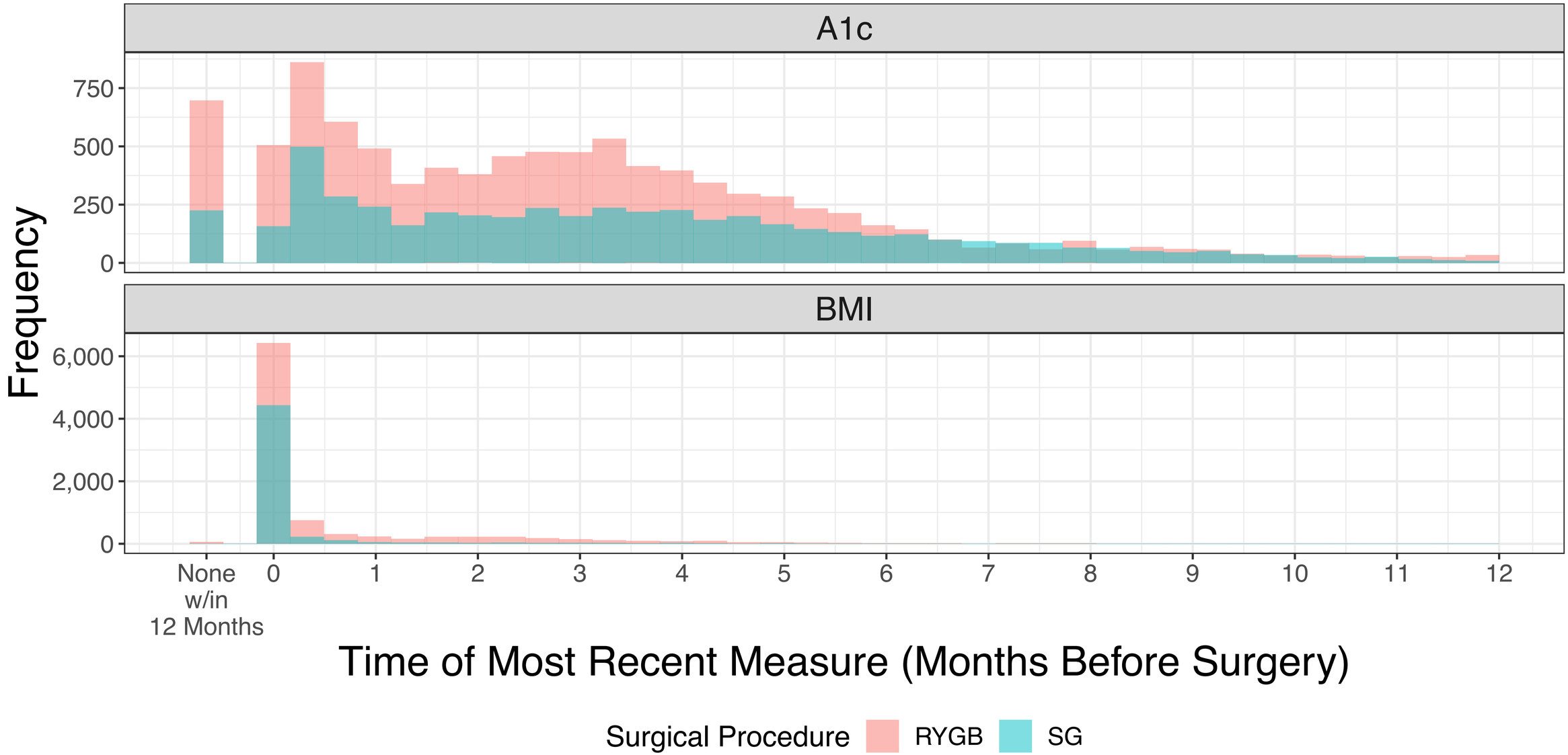
HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

Additional Missing Data Considerations

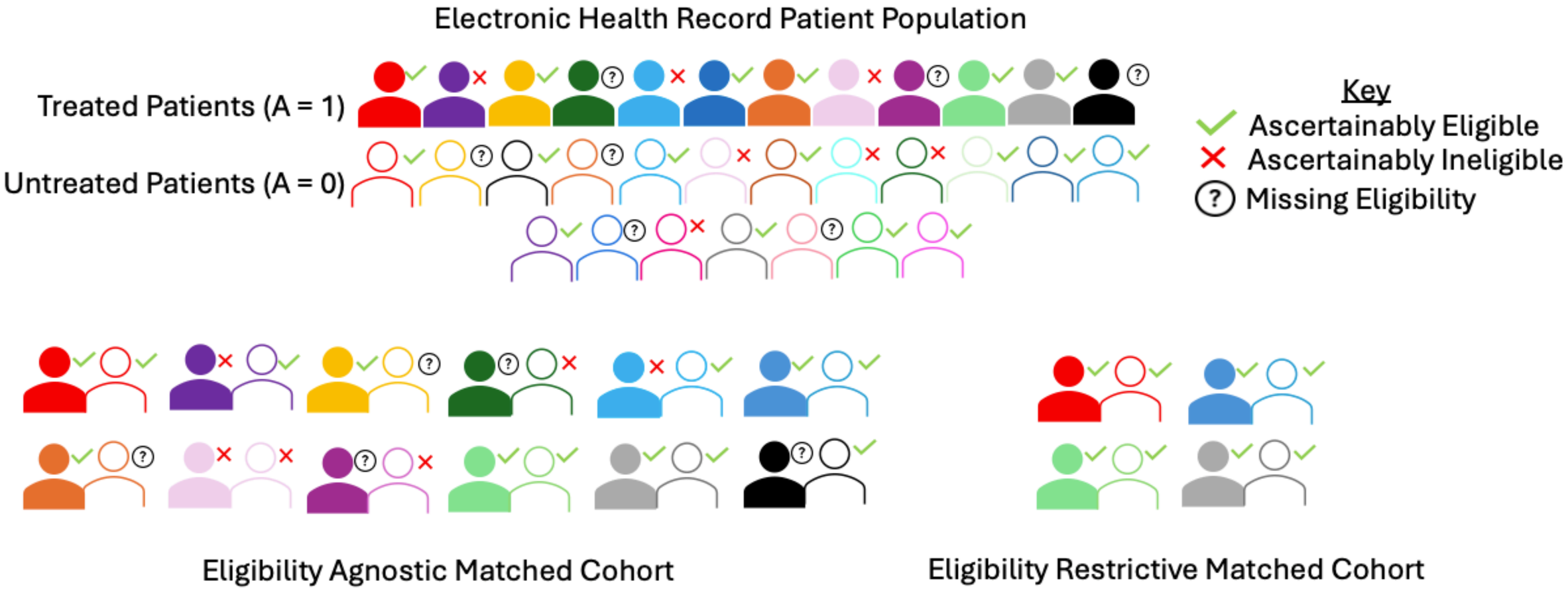
- In cases where Y or $\boldsymbol{L}^*$ are missing with less frequency than $\boldsymbol{L}^e$, imputation might be a reasonable approach
  - This is what we did in our study of RYGB vs. SG
- In cases when one can determine the E in-spite of missingness in some component of $\boldsymbol{L}^e$, but we want to adjust for that component as a confounder
  - Ex: T2DM could be established w/out baseline A1c
  - But we want to adjust for baseline A1c as a (potential) confounder
  - Impute A1c **among those who are eligible but missing** , then use as confounder
    - Importantly, $\boldsymbol{L}^e$ only used in nuisance models among eligible subjects
- When missing outcomes are of concern, can use $\widehat{\theta}_{\text{IF}}$ rather than $\widehat{\theta}_{\text{EIF}}$
  - Because Y is not in the conditioning set of any nuisance function for $\widehat{\theta}_{\text{IF}}$
- Future work: monotone missingness in $(Y, \boldsymbol{L}^e)$
  - Ex: %-Weight Change
  - BMI can be missing when determining eligibility
  - Can later be missing (for defining outcome) even if someone is eligible

Distribution of Measurement Times

Eligibility Defining Covariates



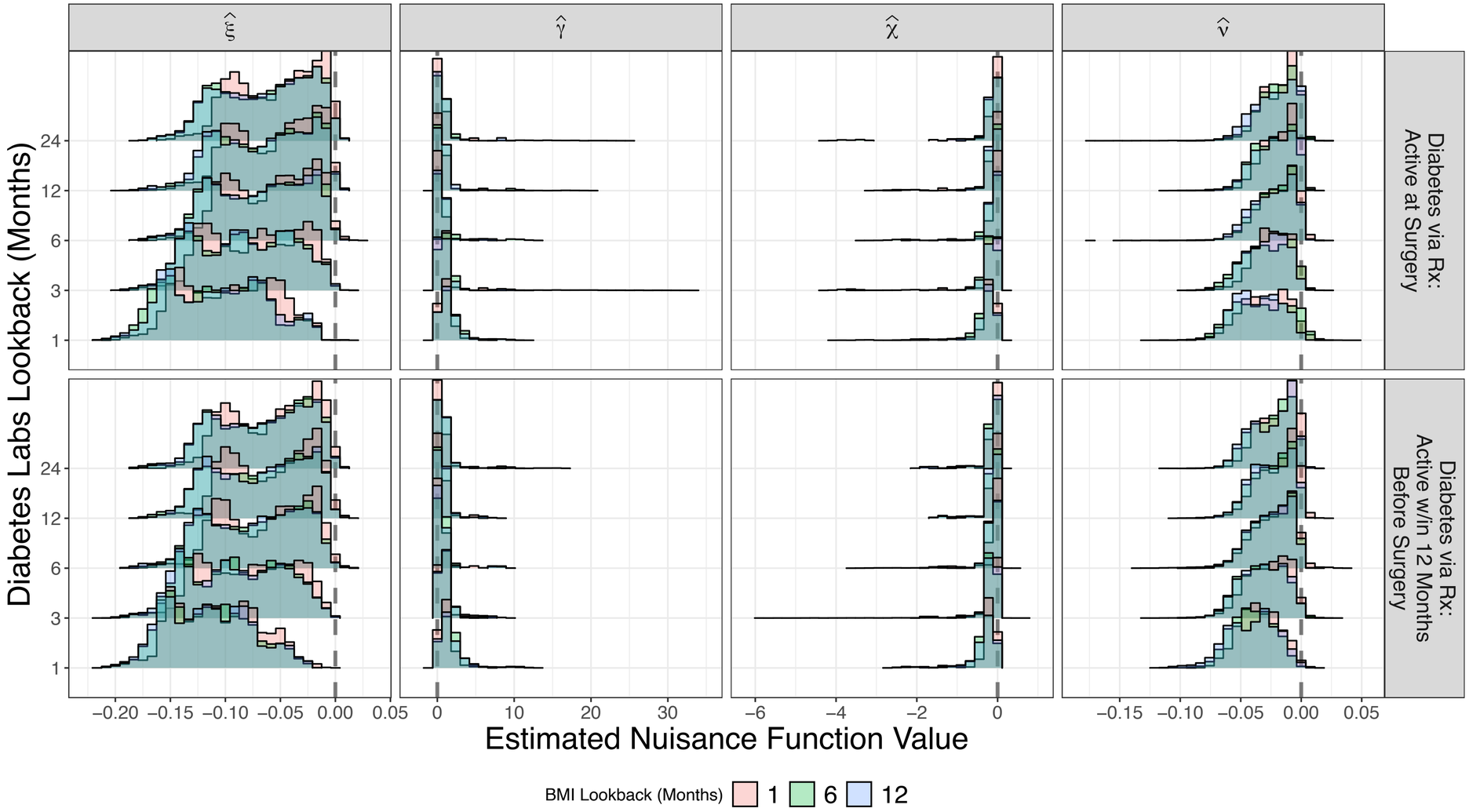
Eligibility Agnostic Matched Cohort



Variance of $\hat{\theta}_{\text{EIF}}$

- Perhaps counterintuitively, confidence intervals for $\hat{\theta}_{\text{EIF}}$ and $\hat{\theta}_{\text{IF}}$ don't monotonically get tighter with less stringent operationalizations of study eligibility
- There is a complex interplay between
  - Increasing the number of patients with complete information
  - Complexity of modeling certain nuisance functions
- Less stringent lookbacks
  - Increase P(R = 1) and increase the total number of eligible patients
  - Decrease P(E = 1 | R = 1)
  - Nested nuisance functions may simultaneously have to contend with
    - Greater zero-inflation
    - Greater range of values in modeling targets

Nested Nuisance Functions



Alternative Covariate Partitioning

- $\boldsymbol{L}_m^{e,\bar{c}}$: Eligibility-defining covariates subject to missingness, but not confounders
- $\boldsymbol{L}_m^{e,c}$: Eligibility-defining covariates subject to missingness, which are confounders
- $\boldsymbol{L}^{*,\bar{c}}$: Completely observed covariates which are not confounders
- $\boldsymbol{L}^{*,c}$: Completely observed covariates which are confounders
- The set of covariates to satisfy **No Unmeasured Confounding** is no longer sufficient to either define eligible or satisfy **MAR**

1. **Consistency:** $Y(A) = Y$ when $E = 1$, almost surely
2. **Positivity:** $0 < \epsilon \leq P(A = 1 \mid \boldsymbol{L}^{*,c}, \boldsymbol{L}_m^{e,c}, E = 1) \leq 1 - \epsilon < 1$, almost surely
3. **No Unmeasured Confounding:** $Y(a) \perp\!\!\!\perp A \mid \boldsymbol{L}^{*,c}, \boldsymbol{L}_m^{e,c}, E = 1$ for $a \in \{0,1\}$
4. **MAR**: $R \perp\!\!\!\perp (Y, \boldsymbol{L}_m^e) \mid \boldsymbol{L}^*, A$
5. **Complete Case Positivity:** $0 < \epsilon \leq P(R = 1 \mid \boldsymbol{L}^*, A)$

Alternative Covariate Partitioning

Theorem S1 *Under Assumptions 1, 2', 3', 4, and 5,* θ^{elig}_{ATT} *is identified by* $\frac{\zeta(P)}{\alpha(P)}$ *where*

$$\zeta(P) = \mathbb{E}_P\left[REY\left\{\frac{A}{\eta(\boldsymbol{L}^*,1)} - \frac{(1-A)}{\eta(\boldsymbol{L}^*,0)}\cdot\frac{\kappa(\boldsymbol{L}^{*,c})\rho(\boldsymbol{L}^{e,c}_m,\boldsymbol{L}^{*,c})\big(1-\sigma(\boldsymbol{L}^{*,c})\big)}{\big(1-\kappa(\boldsymbol{L}^{*,c})\big)\big(1-\rho(\boldsymbol{L}^{e,c}_m,\boldsymbol{L}^{*,c})\big)\sigma(\boldsymbol{L}^{*,c})}\right\}\right]$$

$$\alpha(P) = \mathbb{E}_P\left[\frac{ARE}{\eta(\boldsymbol{L}^*,1)}\right]$$

$$\Lambda_a(\boldsymbol{L}^{e,c}_m;\boldsymbol{L}^{*,c}) = p(E=1,\boldsymbol{L}^{e,c}_m \mid A=a,\boldsymbol{L}^{*,c},R=1)$$

$$\delta_a(\boldsymbol{L}^{e,c}_m;\boldsymbol{L}^{*,c}) = \int_{\mathcal{L}^{*,\bar{c}}} p(\boldsymbol{L}^{e,c}_m \mid A=a,\boldsymbol{L}^{*,c},\boldsymbol{\ell}^{*,\bar{c}},R=1)d\boldsymbol{\ell}^{*,\bar{c}}$$

$$\kappa(\boldsymbol{L}^{*,c}) = P(A=1 \mid \boldsymbol{L}^{*,c})$$

$$\rho(\boldsymbol{L}^{e,c}_m,\boldsymbol{L}^{*,c}) = P(A=1 \mid E=1,\boldsymbol{L}^{e,c}_m,\boldsymbol{L}^{*,c},R=1)$$

$$\sigma(\boldsymbol{L}^{*,c}) = P(A=1 \mid \boldsymbol{L}^{*,c},R=1)$$

- Can get identification result in terms of easily estimable quantities
  - Various ratios of treatment probabilities
- Avoided more difficult conditional density ratios + integrals over the entire covariate distribution
  - These nuisances or nested versions would likely pop-up in (efficient) influence-function based estimators