Fred Hutch Cancer Center

What's the Weight?

Estimating Controlled Outcome Differences in Complex Surveys for Health Disparities Research

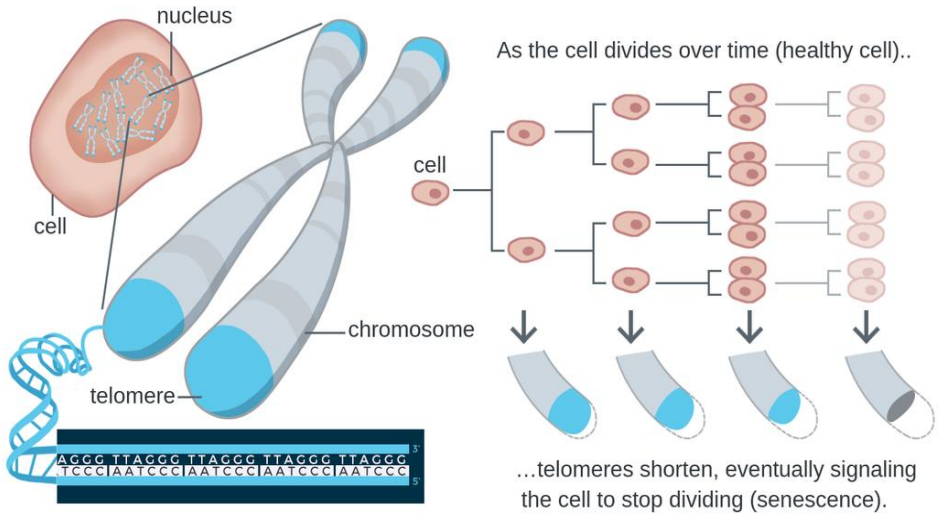
Stephen Salerno
Fred Hutchinson Cancer Center

EcoSta 2025
August 22, 2025

Paradoxical Relationship Between Telomere Length, Race, and SES

Longer telomeres in Black individuals w/ ***lower SES***, ***but*** comparable in similar SES populations

- Regions of DNA that protect against cell death
- Shortening associated w/ cardiometabolic outcomes
- Affected by age, sex, ***race/ethnicity,*** genetics, SES, environment, psychosocial stressors, …
- ***Lower SES*** often associated w/ ***shorter*** telomeres



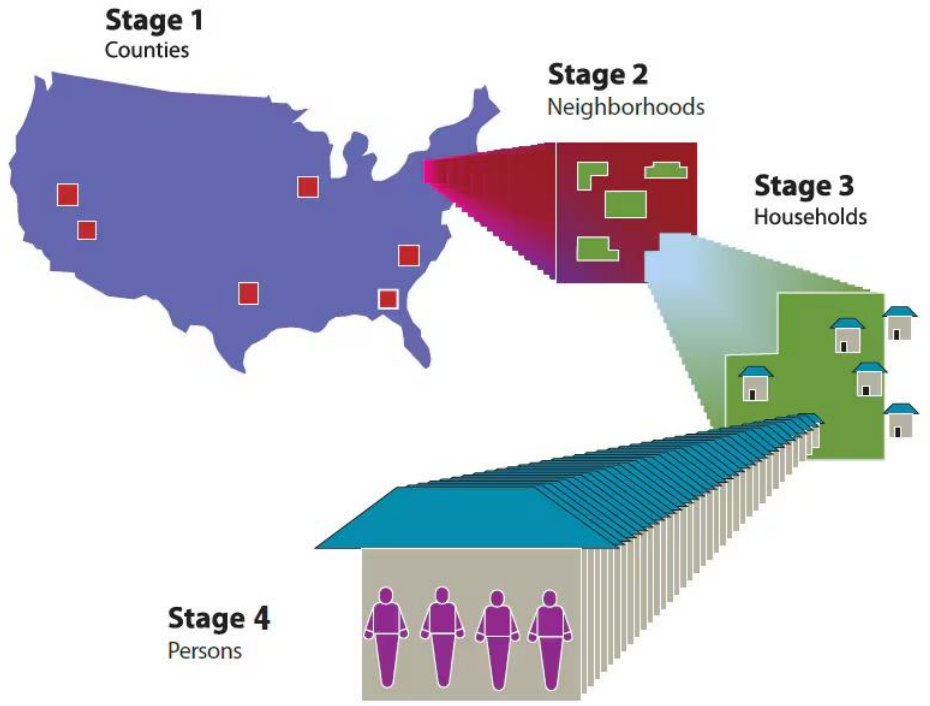
Credit: theory.labster.com/telomere-length

"If we could hypothetically ***balance*** SES between Black and White individuals in a ***nationally representative sample***, would we still see significant Black/White ***differences*** in telomere length?"

National Health and Nutrition Examination Survey (NHANES)

Nationally Representative Survey by the CDC with a Stratified, Clustered Complex Design

- ***Primary sampling units*** (counties)
- Drawn from ***demographic-specific strata:***
  - Oversamples participants living ***≤ 130% FPL***
  - Oversamples ***non-Hispanic Black*** participants
- ***Rich data*** from interviews, physical exams, lab tests, …



Credit: cdc.gov/nchs/nhanes/

Confounding + Selection Bias in a Complex Survey Design

How to weight when selection depends on the group variable under comparison (i.e., race)?

- Within sample ***confounding, covariate imbalance***
- Need to ***generalizing*** results to target population (U.S.)
- ***Statistically*** challenging because:
  - ***SES*** is associated with ***race***
  - ***Both*** impact ***telomere length*** and ***selection***

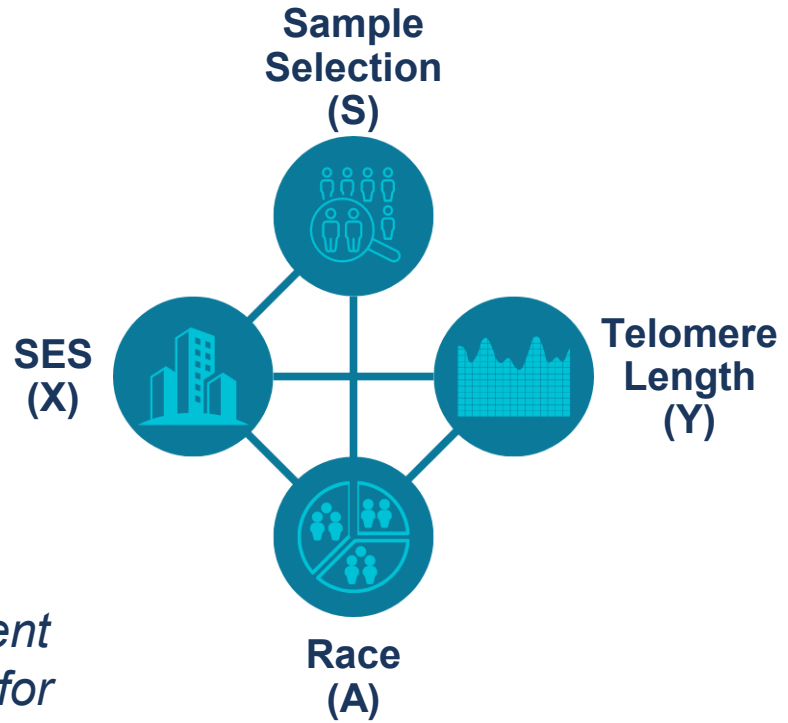
Notation and Target of Inference

Want to Identify a Descriptive Target: Population Average Controlled Difference (ACD)

$$ACD = E_X[\, E(\, Y \mid A = 1, X\,) - E(\, Y \mid A = 0, X\,)\,]$$

- *A*: ***Groups*** of Interest (race; 1 = Black, 0 = White)
- *X*: ***Confounders*** (SES)
- *Y*: ***Outcome*** (log telomere length)
- *S*: Sample ***Selection Indicator***

Can identify a descriptive or causal target (same estimand, different assumptions), but confounding and selection must be accounted for



G-Formula and Inverse Probability Weighting Approaches

Some Questions: Do you survey weight the propensity model? How to weight the outcome?

Answer: Depends on factorization of the joint probability, Pr(S = 1, A = a | X)

(1) $$\mathbb{E}_X\left[\mathbb{E}[Y \mid A=a, S=1, X] \cdot \frac{\Pr(S=1)}{\Pr(S=1 \mid X)} \mid S=1\right]$$

(2) $$\mathbb{E}_X\left[\frac{AY}{\Pr(A=a \mid X)} \cdot \frac{\Pr(S=1)}{\Pr(S=1 \mid A=a, X)} \mid S=1\right]$$

(3) $$\mathbb{E}_X\left[\frac{AY}{\Pr(A=a \mid S=1, X)} \cdot \frac{\Pr(S=1)}{\Pr(S=1 \mid X)} \mid S=1\right]$$

Estimate via ***g-formula*** (1) or ***inverse probability weighting*** (2, 3)

Either we ***weight*** our ***propensity score*** and specifically take ***selection given A = a***

Or we fit a ***within-sample*** propensity score and ***marginalize A out*** of the selection probability

Assumptions for Identifying ACD versus PATE

Pop. average treatment effect (PATE) potential outcome means, $E[Y^a]$, w/ stronger assumptions

Assumption	Definition	ACD	PATE
Positivity	$Pr(A=a \mid X=x) > 0 \; \forall a \in A$, every x s.t. $f_X(x) > 0$	☑	☑
Selection Positivity	$Pr(S=1 \mid A=a, X=x) > 0$ for every a, x s.t. $f_{A,X}(a, x) > 0$	☑	☑
Weak Selection Exchangeability	$E[Y \mid A = a, X] = E[Y \mid A = a, S = 1, X]$ or $E[Y^a \mid A = a, X] = E[Y^a \mid A = a, S = 1, X]$	☑	☑
Stable Unit Treatment Value Assumption	$Y_i = Y_i^a \; \forall i, A_i = a \in A$		☑
Weak Treatment Exchangeability	$E[Y^a \mid X] = E[Y^a \mid A=a, X]$		☑

Estimation and Inference

Proposed G-formula (OM) estimator and IPW-based estimators (IPW1, IPW2)

- ***OM*** is consistent of the outcome model is correctly specified, ***IPW1/2*** are if the propensity model is

$$\hat{\mu}_{\text{OM}}(a) = \frac{1}{n}\sum_{i=1}^{n} \hat{g}_a(\boldsymbol{X}_i)\frac{\Pr(S_i = 1)}{\Pr(S_i = 1 \mid \boldsymbol{X}_i)} \qquad \hat{\mu}_{\text{IPW1}}(a) = \frac{1}{n}\sum_{i=1}^{n} \frac{I(A_i = a)Y_i Pr(S_i = 1)}{\hat{e}_a^w(\boldsymbol{X}_i)Pr(S_i = 1 \mid A_i = a, \boldsymbol{X}_i)} \qquad \hat{\mu}_{\text{IPW2}}(a) = \frac{1}{n}\sum_{i=1}^{n} \frac{I(A_i = a)Y_i Pr(S_i = 1)}{\hat{e}_a(\boldsymbol{X}_i)Pr(S_i = 1 \mid \boldsymbol{X}_i)}$$

- ***Inference*** for *i*th ind. in *j*th clust. in *k*th strat. via empirical sandwich estimator with influence func., ϕ:

$$\hat{V}(\hat{\theta}) = \sum_{k=1}^{K} \frac{J_k}{J_k - 1} \sum_{j=1}^{J_k} \left\{ \phi_{\cdot jk}(Y, \hat{\theta}, \mathcal{P}) - \bar{\phi}_{\cdot\cdot k}(Y, \hat{\theta}, \mathcal{P}) \right\} \left\{ \phi_{\cdot jk}\left(Y, \hat{\theta}, \mathcal{P}\right) - \bar{\phi}_{\cdot\cdot k}(Y, \hat{\theta}, \mathcal{P}) \right\}^{\top}$$

Augmented Inverse Probability Weighting

Can combine the g-formula (1) and IPW formula (3) to form a doubly robust estimator

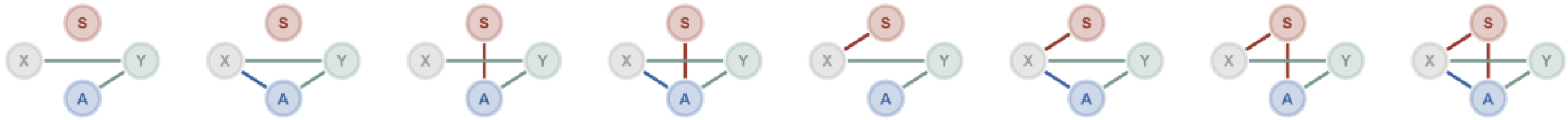
- G-formula most efficient if correctly specified, but
- AIPW Consistent if either the outcome model, g(), or the propensity model, e(), is correctly specified

$$\hat{\mu}_{DR}(a) = \frac{1}{n}\sum_{i=1}^{n} \frac{\Pr(S=1)}{\Pr(S=1 \mid \boldsymbol{X})} \left\{ \hat{g}_a(\boldsymbol{X}_i) + \frac{1(A_i = a)}{\hat{e}_a(\boldsymbol{X}_i)} (Y_i - \hat{g}_a(\boldsymbol{X}_i)) \right\}$$

Simulation Study Overview

Compared existing methods to proposed over a range of possible dependence settings

- Generated stratified, clustered population-level confounders, X
- Specified propensity model (A | X), selection model (S | A, X), and outcome model (Y | A, X)
- Systematically varied relationships between X, A, S, and Y:



	Setting 1	Setting 2	Setting 3	Setting 4	Setting 5	Setting 6	Setting 7	Setting 8
Propensity Model	$\tau_0 = -1$ $\tau_X = 0$	$\tau_0 = -1$ $\tau_X = 1$	$\tau_0 = -1$ $\tau_X = 0$	$\tau_0 = -1$ $\tau_X = 1$	$\tau_0 = -1$ $\tau_X = 0$	$\tau_0 = -1$ $\tau_X = 1$	$\tau_0 = -1$ $\tau_X = 0$	$\tau_0 = -1$ $\tau_X = 1$
Selection Model	$\beta_0 = -4.5$ $\beta_A = 0$ $\beta_X = 0$	$\beta_0 = -4.5$ $\beta_A = 0$ $\beta_X = 0$	$\beta_0 = -4.5$ $\beta_A = 1$ $\beta_X = 0$	$\beta_0 = -4.5$ $\beta_A = 1$ $\beta_X = 0$	$\beta_0 = -4.5$ $\beta_A = 0$ $\beta_X = 1$	$\beta_0 = -4.5$ $\beta_A = 0$ $\beta_X = 1$	$\beta_0 = -4.5$ $\beta_A = 1$ $\beta_X = 1$	$\beta_0 = -4.5$ $\beta_A = 1$ $\beta_X = 1$
Outcome Model	$\gamma_0 = 1$ $\gamma_A = 1$ $\gamma_X = 1$ $\gamma_{AX} = 0.1$	$\gamma_0 = 1$ $\gamma_A = 1$ $\gamma_X = 1$ $\gamma_{AX} = 0.1$	$\gamma_0 = 1$ $\gamma_A = 1$ $\gamma_X = 1$ $\gamma_{AX} = 0.1$	$\gamma_0 = 1$ $\gamma_A = 1$ $\gamma_X = 1$ $\gamma_{AX} = 0.1$	$\gamma_0 = 1$ $\gamma_A = 1$ $\gamma_X = 1$ $\gamma_{AX} = 0.1$	$\gamma_0 = 1$ $\gamma_A = 1$ $\gamma_X = 1$ $\gamma_{AX} = 0.1$	$\gamma_0 = 1$ $\gamma_A = 1$ $\gamma_X = 1$ $\gamma_{AX} = 0.1$	$\gamma_0 = 1$ $\gamma_A = 1$ $\gamma_X = 1$ $\gamma_{AX} = 0.1$
Bias Expected	None	Confounding	None	Confounding	Selection	Confounding + Selection	Selection	Confounding + Selection

Simulation Results – No Confounding or Selection Bias

Facilitates a fair comparison – All perform comparably w.r.t. bias, MSE, SE, coverage

Method	ACD	Est. ACD	Rel. Bias	ASE	MCSE	MSE	Cov.
Setting 1: No Bias Expected ($\tau_X = 0, \beta_A = 0, \beta_X = 0$)							
Oracle Estimator	1.003	1.000	-0.003	0.109	0.107	0.012	0.964
Simple Reg.	1.003	1.004	0.002	0.204	0.206	0.042	0.948
Multiple Reg.	1.003	1.001	-0.002	0.109	0.107	0.012	0.958
IPTW Estimator	1.003	1.008	0.005	0.110	0.263	0.069	0.586
Survey-Weighted Multiple Reg.	1.003	1.001	-0.002	0.109	0.108	0.012	0.956
IPTW Multiple Reg.	1.003	1.001	-0.002	0.109	0.108	0.012	0.958
IPTW + Survey-Weighted Multiple Reg.	1.003	1.001	-0.002	0.109	0.108	0.012	0.956
Weighted IPTW + Survey-Weighted Multiple Reg.	1.003	1.001	-0.002	0.109	0.108	0.012	0.956
Outcome Modeling and Direct Standardization	1.003	1.001	-0.002	0.109	0.107	0.012	0.962
Inverse Probability Weighting 1	1.003	1.001	-0.002	0.110	0.108	0.012	0.956
Inverse Probability Weighting 2	1.003	1.002	-0.001	0.113	0.110	0.012	0.960
Augmented Inverse Probability Weighting	1.003	1.001	-0.002	0.110	0.108	0.012	0.958

Simulation Results – Confounding + Selection Bias

Assumed relationship in our data – Proposed estimators outperform all current approaches

Method	ACD	Est. ACD	Rel. Bias	ASE	MCSE	MSE	Cov.
Setting 8: Both Confounding and Selection Bias Expected ($\tau_X = 1, \beta_A = 1, \beta_X = 1$)							
Oracle Estimator	1.001	0.991	-0.010	0.097	0.093	0.009	0.956
Simple Reg.	1.001	2.210	1.207	0.117	0.110	1.474	0.000
Multiple Reg.	1.001	1.232	0.230	0.073	0.079	0.059	0.126
IPTW Estimator	1.001	3.877	2.872	0.173	0.172	8.301	0.000
Survey-Weighted Multiple Reg.	1.001	0.994	-0.007	0.118	0.134	0.018	0.920
IPTW Multiple Reg.	1.001	1.233	0.231	0.074	0.081	0.060	0.144
IPTW + Survey-Weighted Multiple Reg.	1.001	1.016	0.015	0.109	0.123	0.015	0.912
Weighted IPTW + Survey-Weighted Multiple Reg.	1.001	0.998	-0.004	0.122	0.145	0.021	0.912
Outcome Modeling and Direct Standardization	1.001	0.993	-0.008	0.114	0.095	0.009	0.982
Inverse Probability Weighting 1	1.001	0.996	-0.005	0.157	0.155	0.024	0.954
Inverse Probability Weighting 2	1.001	1.017	0.016	0.226	0.264	0.070	0.968
Augmented Inverse Probability Weighting	1.001	0.992	-0.010	0.154	0.158	0.025	0.946

Race + Telomere Length Data from NHANES

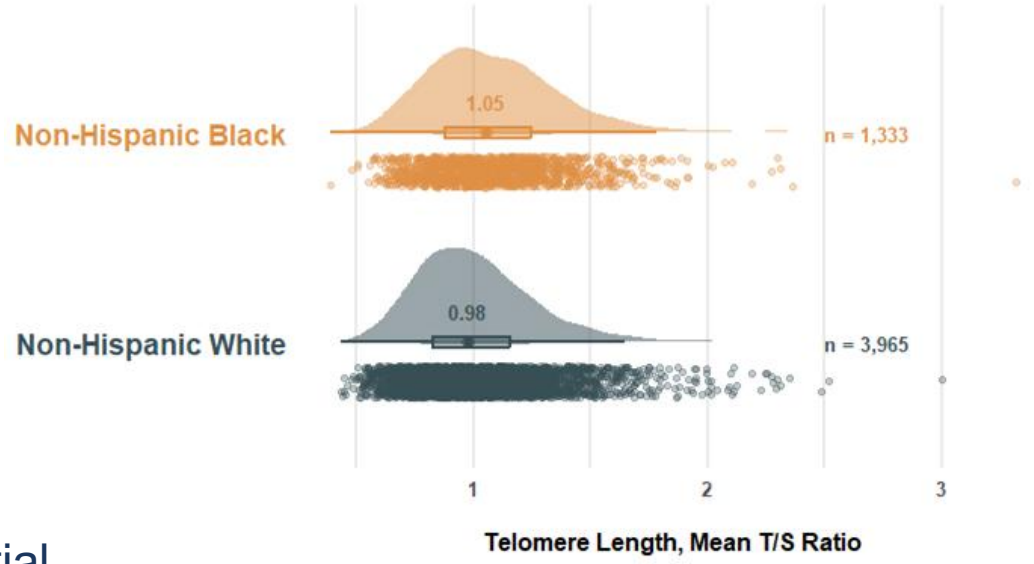
5,298 non-Hispanic Black/White adults from 1999 to 2002 with measured telomere length

- **12 socio-demographic indicators**:

  Education, Marital Status, Household Size, Household Income, Home Ownership, Home Type, Poverty-Income Ratio, Employment Status, Occupational Category, Insurance Status, Food Security Status, WIC Utilization

- ***8 adjustment covariates***:

  Age, Sex, White Blood Cell Count, 5-Part Differential



Descriptive Statistics

Univariate tests suggest Black/White differences across all socioeconomic indicators

Characteristic	Overall, N = 5,298[1]	Non-Hispanic White, N = 3,965[1]	Non-Hispanic Black, N = 1,333[1]	p-value[2]
Telomere Length, Mean T/S Ratio	1.00 (0.84, 1.18)	0.98 (0.83, 1.16)	1.05 (0.88, 1.25)	<0.001
Age, Years	50 (35, 67)	52 (35, 70)	45 (34, 62)	<0.001
Sex				0.4
Male	2,574 (49%)	1,939 (49%)	635 (48%)	
Female	2,724 (51%)	2,026 (51%)	698 (52%)	
Education				
High School or GED	2,621 (49%)	1,790 (45%)	831 (62%)	
Some College	1,448 (27%)	1,102 (28%)	346 (26%)	
College Graduate	1,222 (23%)	1,068 (27%)	154 (12%)	
Refused/Unknown	7 (0.1%)	5 (0.1%)	2 (0.2%)	
Marital Status				<0.001
Never Married	754 (14%)	436 (11%)	318 (24%)	
Widowed/Divorced/Separated	1,157 (22%)	775 (20%)	382 (29%)	
Married/Living with Partner	3,144 (59%)	2,574 (65%)	570 (43%)	
Refused/Unknown	243 (4.6%)	180 (4.5%)	63 (4.7%)	
Household Size				<0.001
…	…	…	…	…

Results From Our Study

ACD ***attenuates*** as we make ***appropriate adjustments*** for confounding + selection bias

- SES ***imbalance*** across race and ***oversampling*** not appropriately accounted for by other methods
- Methods which ***properly*** incorporate ***design*** have more ***conservative SEs***

Method	ACD Estimate	95% Confidence Interval
Multiple Regression	0.0265	0.0106, 0.0424
IPTW Estimator	0.0262	0.0079, 0.0445
Survey-Weighted Multiple Regression	0.0298	-0.0010, 0.0606
IPTW Multiple Regression	0.0183	0.0053, 0.0312
IPTW + Survey-Weighted Multiple Regression	0.0220	-0.0088, 0.0527
Weighted IPTW + Survey-Weighted Multiple Regression	0.0186	-0.0126, 0.0498
Proposed Outcome Modeling and Direct Standardization	0.0176	-0.0029, 0.0381
Proposed Inverse Probability Weighting 1	0.0141	-0.0134, 0.0416
Proposed Inverse Probability Weighting 2	0.0131	-0.0081, 0.0342
Proposed Augmented Inverse Probability Weighting	0.0122	-0.0077, 0.0321

Conclusions

Some thoughts on the approach and our results

- Approach for estimating ***controlled outcome differences*** when the group variable of interest and its confounders both influence ***sample selection***
- Proposal ***minimizes bias*** and achieves ***correct inference*** compared to standard analysis methods
- Context of studying ***racial disparities*** presents these challengess in such a way that should be ***rigorously studied*** for ***best practice*** recommendations

Conclusions

Some thoughts on future work

- Focus on ***complex surveys,*** but concepts readily extend to other ***observational settings***
- Areas of interest for ***future work*** include:
  - ***Electronic health record*** data with unknown sampling probabilities
  - Expanded ***relationship diagrams***
  - Extending this framework to ***two-stage sampling*** or ***sequential designs***
  - Time-to-event outcomes, AI/ML assisted surveys

Acknowledgements

Emily Roberts

University of Iowa
Department of Biostatistics

Fan Li

Yale University
Department of Biostatistics
Department of Cardiovascular Medicine

Belinda Needham

University of Michigan
Department of Epidemiology

Bhramar Mukherjee

Yale University
Department of Biostatistics
Department of Epidemiology
Department of Statistics and Data Science

Tyler McCormick

University of Washington
Department of Statistics
Department of Sociology

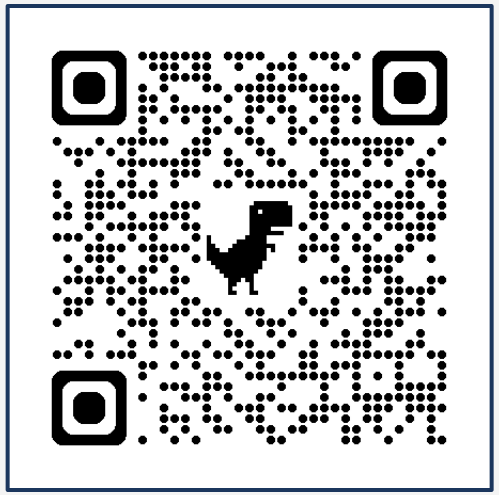
Xu Shi

University of Michigan
Department of Biostatistics

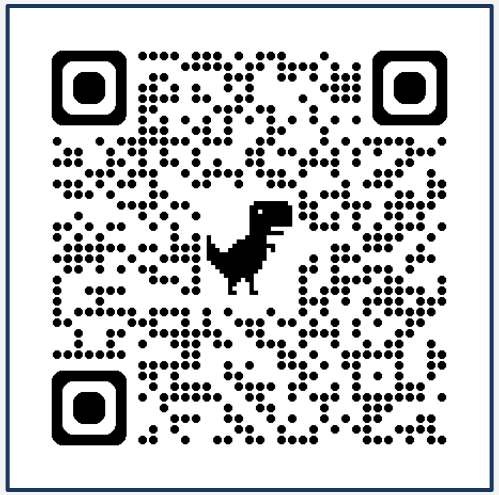
Fred Hutch Cancer Center

Thank You!

Fred Hutch Cancer Center



Paper



R Package