UNC
GILLINGS SCHOOL OF
GLOBAL PUBLIC HEALTH

Combining estimators in the pursuit of robustness

Michele Jonsson Funk, PhD
UNC Chapel Hill

September 6-7, 2012
ACF/OPRE Methodological Advancement Meeting
Innovative Directions in Estimating Impact

Funding and Disclosure

- Current work on doubly robust estimators funded by the Agency for Healthcare Research & Quality (AHRQ) (K02HS017950)
- This materials does not represent the views of Agency for Healthcare Research & Quality.

Methods for estimating causal effects

A. Propensity scores used for inverse probability of treatment weighting (IPTW)
B. G-computation
C. Standardized mortality/morbidity ratio (SMR) weights
D. All of the above *= Doubly robust estimator*

Propensity Score (PS)

- Rosenbaum & Rubin, 1983
- Probability of treatment (or exposure), given a set of characteristics/conditions
- Balances risk of the outcome between treated and untreated groups
- Estimated from the observed data

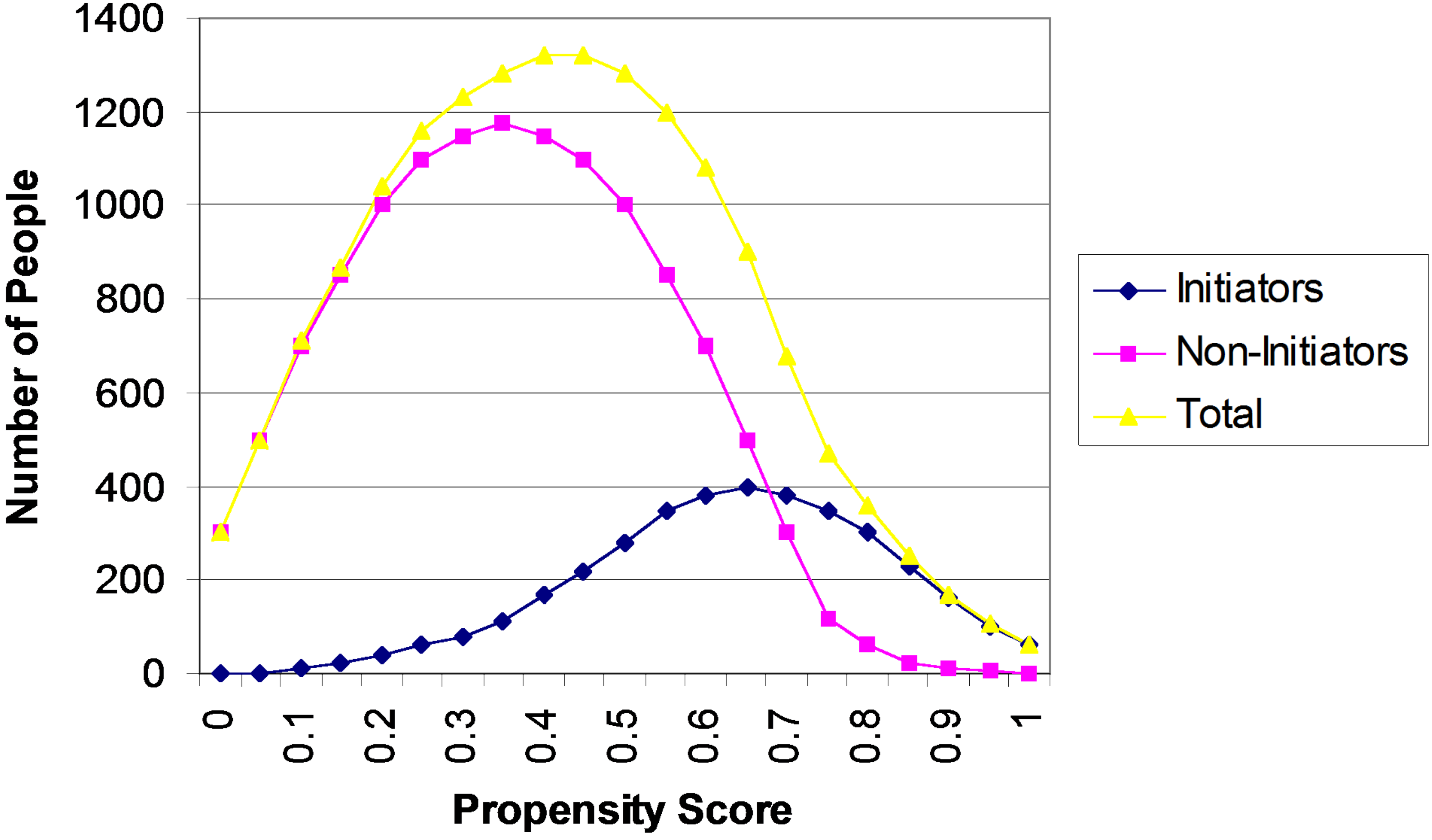
IPTW

- Weight observations by inverse probability of actual treatment, given covariates
  - Treated (exposed): 1/PS
  - Untreated (unexposed): 1/(1-PS)
- After weighting, ‘crude’ effect in the ‘pseudopopulation’ should be unconfounded
- Effects: Risk, difference, ratio
- Target Population: Total
- PS model must be specified correctly

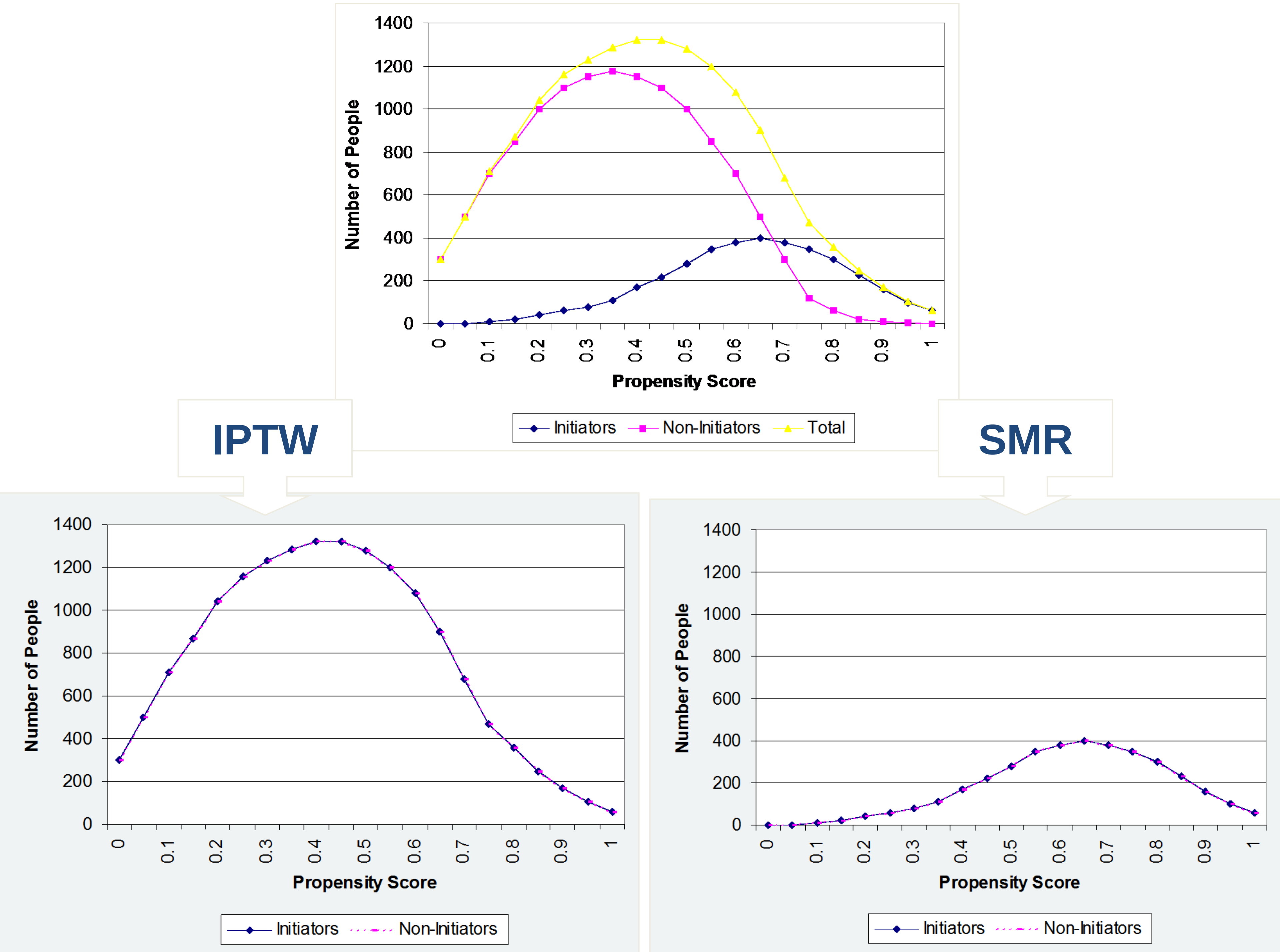
SMR weights

- Standardized mortality/morbidity ratio (SMR)
- Weight observations by
  - 1 in the treated
  - Propensity odds in the untreated, PS/(1-PS)
- Target Population: Treated
- Effects: Risk (observed), difference, ratio
- Assumes PS model correctly specified

Hypothetical Distribution of Propensity Scores



Adapted from MA Brookhart, Counterfactuals, 2012


Number of People
Propensity Score
Initiators
Non-Initiators
Total
IPTW
SMR
Number of People
Propensity Score
Initiators
Non-Initiators
Number of People
Propensity Score
Initiators
Non-Initiators

G-computation

- Usual generalized linear outcome model
- Marginalizes the treatment effect by estimating each individual’s expected response (counter factual) under both treatment conditions
- Effects: Risk, difference, ratio
- Target Population: Total, treated, untreated
- Assumes outcome model is correctly specified

G-computation: Implementation

- Fit outcome regression model(s) to obtain parameter estimates
- Using the individual's characteristics, calculate predicted outcomes for each patient with and without treatment
- Calculate average response across all patients under each treatment condition

DR Estimator: Conceptual description

- Doubly robust (DR) estimation uses two models:
  - Propensity score model for the confounder - exposure (or treatment) relationship
  - Outcome regression model for the confounder – outcome relationship, under each exposure condition
- These two stages can use:
  - different subsets of covariates, and
  - different parametric forms.
- If either model is correct, then the DR estimate of treatment effect is unbiased.

Doubly robust estimator

$$\hat{\Delta}_{DR} = n^{-1}\sum_{i=1}^{n}\left[\frac{X_iY_i}{e(Z_i,\hat{\beta})} - \frac{\left\{X_i - e(Z_i,\hat{\beta})\right\}}{e(Z_i,\hat{\beta})} m_1(Z_i,\hat{\alpha}_1)\right]$$

$$-n^{-1}\sum_{i=1}^{n}\left[\frac{(1-X_i)Y_i}{1-e(Z_i,\hat{\beta})} + \frac{\left\{X_i - e(Z_i,\hat{\beta})\right\}}{1-e(Z_i,\hat{\beta})} m_0(Z_i,\hat{\alpha}_0)\right]$$

$$\hat{\Delta}_{DR} = [\boldsymbol{E(Y_1)} + augmentation] - [\boldsymbol{E(Y_0)} + augmentation]$$

$$\hat{\Delta}_{DR} = [E(Y_1)] - [E(Y_0)]$$

Adapted from Davidian M, [DR Presentation, 2007]

DR estimator translated

	DR_{1i}	DR_{0i}
General Form	$\frac{Y_{X=1} \times X}{PS} - \frac{\hat{Y}_1 \left(X - PS\right)}{PS}$	$\frac{Y_{X=0}(1 - X)}{1 - PS} + \frac{\hat{Y}_0 \left(X - PS\right)}{1 - PS}$
Among exposed (X=1)		
Among unexposed (X=0)		

X: exposure (0,1)
Y: outcome

Z: covariates
PS: p(X=1|Z)

$\widehat{Y}_1$: predicted Y setting X to 1
$\widehat{Y}_0$: predicted Y setting X to 0

$Y_{X=1}$: observed Y given X=1
$Y_{X=0}$: observed Y given X=0

DR estimator translated

	DR_{1i}	DR_{0i}
General Form	$\frac{Y_{X=1} \times X}{PS} - \frac{\hat{Y}_1 \left(X - PS\right)}{PS}$	$\frac{Y_{X=0}(1-X)}{1-PS} + \frac{\hat{Y}_0 \left(X - PS\right)}{1-PS}$
Among exposed (X=1)	$\frac{Y_{X=1}}{PS} - \frac{\hat{Y}_1 \left(1 - PS\right)}{PS}$	$\hat{Y}_0$
Among unexposed (X=0)	$\hat{Y}_1$	$\frac{Y_{X=0}}{1-PS} - \frac{\hat{Y}_0 \times PS}{1-PS}$

X: exposure (0,1)
Y: outcome

Z: covariates
PS: p(X=1|Z)

$\widehat{Y}_1$: predicted Y setting X to 1
$\widehat{Y}_0$: predicted Y setting X to 0

$Y_{X=1}$: observed Y given X=1
$Y_{X=0}$: observed Y given X=0

IPTW estimator

	DR_{1i}	DR_{0i}
General Form	$\frac{Y_{X=1} \times X}{PS} - \frac{\hat{Y}_1 \left(X - PS\right)}{PS}$	$\frac{Y_{X=0}(1-X)}{1-PS} + \frac{\hat{Y}_0 \left(X - PS\right)}{1-PS}$
Among exposed (X=1)	$\frac{Y_{X=1}}{PS} - \frac{\hat{Y}_1 \left(1 - PS\right)}{PS}$	$\hat{Y}_0$
Among unexposed (X=0)	$\hat{Y}_1$	$\frac{Y_{X=0}}{1-PS} - \frac{\hat{Y}_0 \times PS}{1-PS}$

X: exposure (0,1)
Y: outcome

Z: covariates
PS: p(X=1|Z)

$\widehat{Y}_1$: predicted Y setting X to 1
$\hat{Y}_0$: predicted Y setting X to 0

$Y_{X=1}$: observed Y given X=1
$Y_{X=0}$: observed Y given X=0

G-computation

	DR_{1i}	DR_{0i}
General Form	$\frac{Y_{X=1} \times X}{PS} - \frac{\hat{Y}_1 (X - PS)}{PS}$	$\frac{Y_{X=0}(1 - X)}{1 - PS} + \frac{\hat{Y}_0 (X - PS)}{1 - PS}$
Among exposed (X=1)	$\frac{Y_{X=1}}{PS} - \frac{\hat{Y}_1 (1 - PS)}{PS}$	$\hat{Y}_0$
Among unexposed (X=0)	$\hat{Y}_1$	$\frac{Y_{X=0}}{1 - PS} - \frac{\hat{Y}_0 \times PS}{1 - PS}$

X: exposure (0,1)
Y: outcome

Z: covariates
PS: p(X=1|Z)

$\widehat{Y_1}$: predicted Y setting X to 1
$\hat{Y}_0$: predicted Y setting X to 0

$Y_{X=1}$: observed Y given X=1
$Y_{X=0}$: observed Y given X=0

Counterfactual outcomes

	DR_{1i}	DR_{0i}
General Form	$\dfrac{Y_{X=1} \times X}{PS} - \dfrac{\hat{Y}_1 \left(X - PS\right)}{PS}$	$\dfrac{Y_{X=0}(1-X)}{1-PS} + \dfrac{\hat{Y}_0 \left(X - PS\right)}{1-PS}$
Among exposed (X=1)	$\dfrac{Y_{X=1}}{PS} - \dfrac{\hat{Y}_1 \left(1 - PS\right)}{PS}$	$\hat{Y}_0$
Among unexposed (X=0)	$\hat{Y}_1$	$\dfrac{Y_{X=0}}{1-PS} - \dfrac{\hat{Y}_0 \times PS}{1-PS}$

X: exposure (0,1)
Y: outcome

Z: covariates
PS: p(X=1|Z)

$\widehat{Y}_1$: predicted Y setting X to 1
$\hat{Y}_0$: predicted Y setting X to 0

$Y_{X=1}$: observed Y given X=1
$Y_{X=0}$: observed Y given X=0

Weighting relevant observed events

	DR_{1i}	DR_{0i}
General Form	$\frac{Y_{X=1}\times X}{PS}-\frac{\hat{Y}_1(X-PS)}{PS}$	$\frac{Y_{X=0}(1-X)}{1-PS}+\frac{\hat{Y}_0(X-PS)}{1-PS}$
Among exposed (X=1)	$\frac{Y_{X=1}}{PS}-\frac{\hat{Y}_1(1-PS)}{PS}$	$\hat{Y}_0$
Among unexposed (X=0)	$\hat{Y}_1$	$\frac{Y_{X=0}}{1-PS}-\frac{\hat{Y}_0\times PS}{1-PS}$

X: exposure (0,1)
Y: outcome

Z: covariates
PS: p(X=1|Z)

$\widehat{Y}_1$: predicted Y setting X to 1
$\widehat{Y}_0$: predicted Y setting X to 0

$Y_{X=1}$: observed Y given X=1
$Y_{X=0}$: observed Y given X=0

Subtracting?

	DR_{1i}	DR_{0i}
General Form	$\frac{Y_{X=1} \times X}{PS} - \frac{\hat{Y}_1 \left(X - PS\right)}{PS}$	$\frac{Y_{X=0}(1-X)}{1-PS} + \frac{\hat{Y}_0 \left(X - PS\right)}{1-PS}$
Among exposed (X=1)	$\frac{Y_{X=1}}{PS} - \frac{\hat{Y}_1 \left(1 - PS\right)}{PS}$	$\hat{Y}_0$
Among unexposed (X=0)	$\hat{Y}_1$	$\frac{Y_{X=0}}{1-PS} - \frac{\hat{Y}_0 \times PS}{1-PS}$

X: exposure (0,1)
Y: outcome

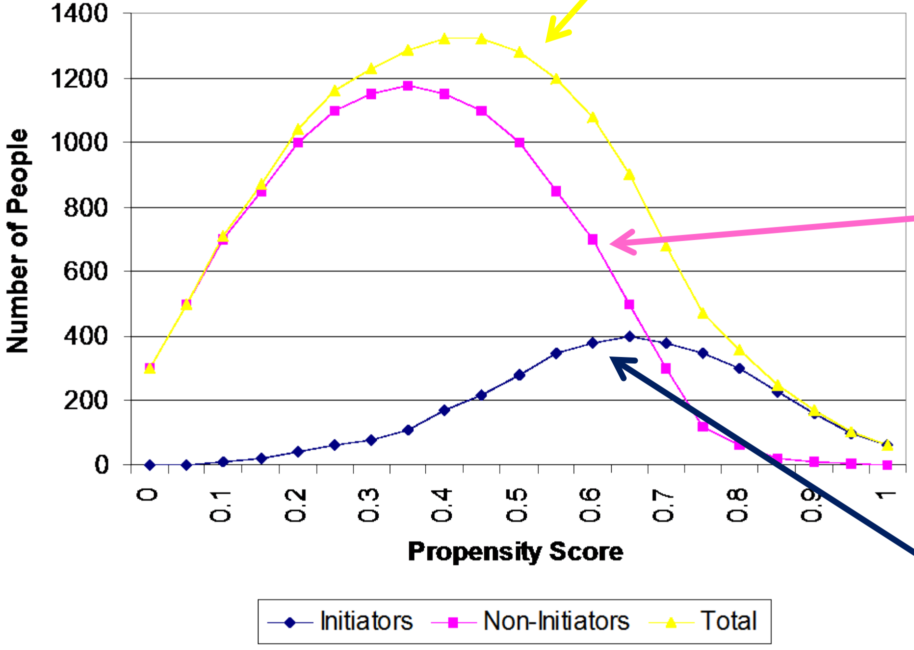
Z: covariates
PS: p(X=1|Z)

$\widehat{Y_1}$: predicted Y setting X to 1
$\widehat{Y_0}$: predicted Y setting X to 0

$Y_{X=1}$: observed Y given X=1
$Y_{X=0}$: observed Y given X=0

Net effect of combining weights

#1: IPTW * **observed** outcomes =
Response under exposure
standardized to the
total population



#2: 1/SMR * **predicted** outcomes =
Response under exposure
standardized to the
unexposed population

Subtract #2 from #1
For the net result:
Response under exposure
standardized to the
exposed population

Combining weights for relevant observed outcomes

	DR_{1i}	DR_{0i}
General Form	$\frac{Y_{X=1} \times X}{PS} - \frac{\hat{Y}_1 \left(X - PS\right)}{PS}$	$\frac{Y_{X=0}(1-X)}{1-PS} + \frac{\hat{Y}_0 \left(X - PS\right)}{1-PS}$
Among exposed (X=1)	$\frac{Y_{X=1}}{PS} - \frac{\hat{Y}_1 \left(1 - PS\right)}{PS}$	$\hat{Y}_0$
Among unexposed (X=0)	$\hat{Y}_1$	$\frac{Y_{X=0}}{1-PS} - \frac{\hat{Y}_0 \times PS}{1-PS}$

X: exposure (0,1)
Y: outcome

Z: covariates
PS: p(X=1|Z)

$\widehat{Y}_1$: predicted Y setting X to 1
$\widehat{Y}_0$: predicted Y setting X to 0

$Y_{X=1}$: observed Y given X=1
$Y_{X=0}$: observed Y given X=0

Effect measures

- Scale
  - Risk, mean response
  - Risk difference, difference in means
  - Relative risk
  - Odds ratio
- Target Populations
  - Total
  - Treated
  - Untreated

Assumptions

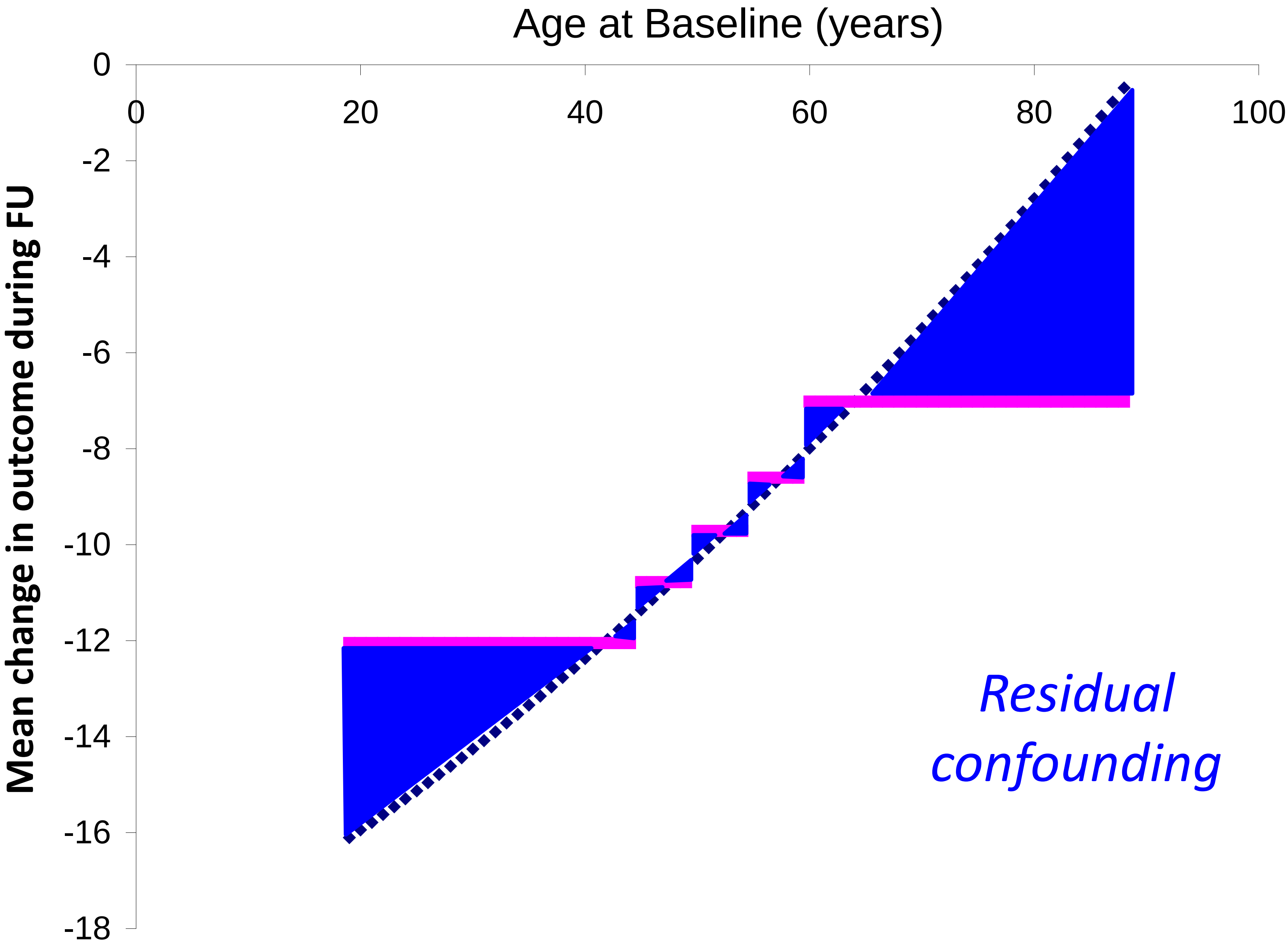
- Positivity
- Consistency
- No interference (aka independence)
- Exchangeability (aka ignorability)
  - Correct model specification for
    PS model or outcome regression models
  - No unmeasured confounding

See Cole & Hernan, AJE, 2008

Misspecified covariates

- Categorize continuous covariates (realistic scenario)
  - Simulated to mirror the distribution of common confounders
    - Age, BMI, LDL cholesterol, physical activity
  - Categories reflect ‘meaningful’ cutpoints

- True relationships known (simulated)
  - Linear or slightly quadratic



Age at Baseline (years)
0
20
40
60
80
100
0
-2
-4
-6
-8
-10
-12
-14
-16
-18
Mean change in outcome during FU
Residual
confounding

Monte Carlo simulation

- Draw a random sample (n=5000)
- Fit a model (OLS or DR)
- Save the parameter estimate & standard error
- Repeat 1000 times

- *For each of 11 scenarios x 4 treatment effects*

Root Mean Squared Error

Scenario		True TX effect			
		0	-0.41	-1.10	-1.61
Unadjusted		2.963	2.963	2.963	2.963
True	OLS	0.032	0.032	0.032	0.032
	DR	0.036	0.036	0.036	0.036
Misspecified Outcome Model					
Categorize linear covariates	OLS	0.113	0.111	0.113	0.116
	DR	0.047	0.048	0.049	0.048
Categorize nonlinear covariates	OLS	0.144	0.142	0.144	0.147
	DR	0.054	0.055	0.054	0.054
Categorize linear & nonlinear covariates	OLS	0.250	0.248	0.250	0.253
	DR	0.064	0.064	0.063	0.064

95% CI coverage

Scenario		True TX effect			
		0	-0.41	-1.10	-1.61
Unadjusted		0	0	0	0
True	OLS	94.4	95.7	95.7	94.7
	DR	94.5	95.7	95.2	95.1
Misspecified Outcome Model					
Categorize linear covariates	OLS	19.9	18.5	19.2	15.5
	DR	95.9	95.3	94.8	96.2
Categorize nonlinear covariates	OLS	17.5	17.8	16.2	14.5
	DR	96.2	96.0	94.2	96.4
Categorize linear & nonlinear covariates	OLS	0.2	0.1	0.1	0.0
	DR	96.4	94.3	95.5	95.2

Limitations

- Two poorly specified models can be worse than a single wrong maximum likelihood regression
- Standard errors tend to be slightly larger compared to a single correctly specified regression model
- Residual confounding is modest in magnitude relative to bias of crude estimate
- DR estimation is not a panacea for unmeasured confounding
- Standard errors/confidence intervals require bootstrapping
- Best practices & diagnostics still under development

Conclusions

- Observational (non-experimental) studies depend on statistical models to disentangle causal effects from confounding
- We can never be certain that the statistical model we have chosen is correct
- DR estimator is unbiased if at least one of the two component models is right and therefore provides some protection against residual confounding
- Attractive properties of marginalized effect estimates with improved efficiency relative to IPTW

Resources

- Funk MJ, Westreich D, Wiesen C, Sturmer T, Brookhart MA, Davidian M. Doubly robust estimation of causal effects. *Am J Epidemiol.* Apr 1 2011;173(7):761-767.
- **Recommended further reading**
  - Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. Biometrics. 2005;61(4):962–973.
  - Tsiatis AA. Semiparametric Theory and Missing Data. New York: Springer; 2006.
  - Van der Laan M, Robins JM. Unified Methods for Censored Longitudinal Data and Causality. New York: Springer; 2003.

SAS macro: www.unc.edu/~mfunk/dr

UNC
GILLINGS SCHOOL OF
GLOBAL PUBLIC HEALTH

Michele Jonsson Funk

www.unc.edu/~mfunk

mfunk@unc.edu