Link to slides

Overview of Causal Inference

Danielle Stevens
IRTA postdoctoral fellow
National Institute of Environmental Health Sciences

Link to slides

Outline & references

- What is causal inference (Smith et al)
  - Research questions
  - Causal framework
- G-methods (Naimi et al)
  - Propensity scores (IPTW)
    - R tutorial
  - G-computation
    - R tutorial

Link to slides

Outline & references

- Jennifer Ahern & Laura Balzer
  - SER Causal Inference
  - SER Long'l Causal Inference
- Maya Petersen & Laura Balzer
  - Introduction to Causal Inference
  - https://www.ucbbiostat.com/
- Ehsan Karim & Hanna Frank
  - https://ehsanx.github.io/TMLEworkshop/
- Thomas Love
  - https://github.com/thomaselove/ichps2018

Link to slides

SER Annual Meeting 2024

8:00 AM - 8:00 AM	**High-dimensional propensity score and its machine learning and double robust extensions in residual confounding control in pharmacoepidemiologic studies** Location: Virtual Speaker: Ehsan karim
8:00 AM - 8:00 AM	**Causal inference with transfer entropy: An introduction for beginners** Location: Virtual Speakers: Roni Barak Ventura, Maurizio Porfiri, James Macinko, Manuel Ruiz Marín
8:00 AM - 8:00 PM	**An overview of Difference-in-Difference and Synthetic Control Methods: Classical and Novel Approaches** Location: Virtual Speakers: Roch Nianogo, Tarik Benmarhnia
8:00 AM - 8:00 AM	**Introduction to Difference in Differences Using Stata** Location: Virtual Speaker: Chuck Huber
8:00 AM - 8:00 AM	**What would it take to change your inference? Quantifying the Discourse about Causal Inferences in Epidemiology** Location: Virtual Speaker: Kenneth Frank

Tue, Jun 18, 2024

8:30 AM - 12:30 PM	**Modern Causal Mediation Analysis** Location: Waterloo 4 Session Chair: Kara Rudolph, Ivan Diaz, Nima Hejazi
8:30 AM - 12:30 PM	**Unlocking the Mysteries of Mixed Exposures: Targeted Learning for Robust Discovery and Causal Inference in Epidemiology** Location: Waterloo 5/6 Session Chair: David McCoy

Research questions

Current paradigm

What is the expected difference in an outcome for a one-unit increase in exposure in our study population?

Current paradigm

What is the expected difference in birthweight for a one-unit increase in airborne metal exposure in Milwaukee, 2011-2013?

Keil et al. 2021, Am J Epidemiol

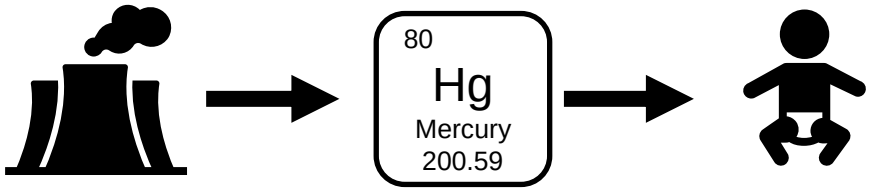
Research questions

Causal inference

If we intervened to reduce one or more exposures in a specific way in our study population, how would outcome distributions change?

Causal inference

What is the difference in mean birthweight observed if we closed 3 coal-fired plants releasing airborne metals in Milwaukee, 2011-2013?



Keil et al. 2021, Am J Epidemiol

Research questions

Current paradigm

What is the expected difference in birthweight for a one-unit increase in airborne metal exposure in Milwaukee, 2011-2013?

Causal inference

What is the difference in the birthweight distribution observed if we closed 3 coal-fired plants releasing airborne metals in Milwaukee, 2011-2013?

When research questions have to do with how things work or how best to intervene to improve health, they are often causal questions.

Keil et al. 2021, Am J Epidemiol

Pros of Causal Inference

- Not every research question is causal, but thinking causally about the research we do can have benefits:
  - Think critically about research
  - Real-world implications of research
  - Feasibility of examining research question
  - Control for time-varying confounding

What is a reasonable and meaningful unit change in exposure to examine?

Pros of causal inference

- Not every research question is causal, but thinking causally about the research we do can have benefits:
  - Think critically about research
  - Real-world implications of research
  - Feasibility of examining research question
  - Control for time-varying confounding

Thinking in terms of interventions or policies instead of associations

Pros of causal inference

- Not every research question is causal, but thinking causally about the research we do can have benefits:
  - Think critically about research
  - Real-world implications of research
  - Feasibility of examining research question
  - Control for time-varying confounding

Infeasible to have RCT for some exposures

Cons of causal inference

- … And some downsides:
  - Overconfidence in assumptions
    - Thinking that your findings apply more broadly than they do
  - Misinterpreting results
    - Thinking that your observational analysis perfectly replicates RCT
  - Sometimes more difficult to formulate and answer a causal question
    - Complexity >> common sense

Follow a formal causal framework

Current paradigm: statistical inference

- Sample individuals from underlying population and for each subject observe:

  X = Prenatal vitamin use (1=use, 0=no use)

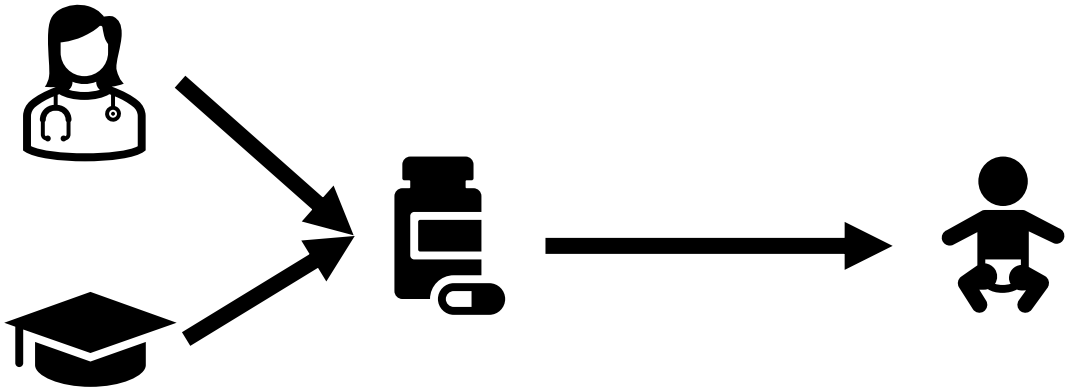
  Y = Preterm delivery (1=preterm, 0=not preterm)

- Estimate association between taking prenatal vitamin vs not on risk of preterm delivery in study population as:

  $P(Y=1 \mid X=1) - P(Y=1 \mid X=0)$

Statistical vs causal inference

- Statistical inference tells us the probability of occurrence in our data
  - Exposure doesn't occur in all participants, only self-selected group
    - e.g., take vitamins because you know you are pregnant, encouraged by medical provider, educated about benefits
- Statistical inference cannot extend into the hypothetical



Causal inference

- Causal inference tells us how a data distribution would change if we intervened to change exposure

How would preterm delivery risk change if all pregnant persons in our study had taken vitamins?

Causal inference

- Causal inference tells us how a data distribution would change if we intervened to change exposure

How would preterm delivery risk change if we assigned pregnant persons who had had a prior preterm delivery in our study to take vitamins?

Causal inference

- Causal inference tells us how a data distribution would change if we intervened to change exposure

How would preterm delivery risk change if we assigned pregnant persons who had had a prior preterm delivery in our study to take vitamins?

- Causal inference goal: draw inference about parameters for a distribution we do not (fully) observe in our data

Causal inference framework

1. State the research question & hypothetical experiment
2. Define causal model & parameter of interest
3. Link causal model to observed data & define statistical model
4. Link causal effect to parameter estimable in observed data
5. Choose & apply estimator
6. Make inferences

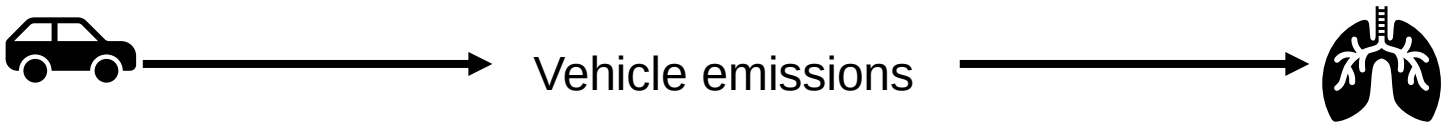
Causal inference framework

1. State the research question & hypothetical intervention
2. Define causal model & parameter of interest
3. Link causal model to observed data & define statistical model
4. Link causal effect to parameter estimable in observed data
5. Choose & apply estimator
6. Make inferences

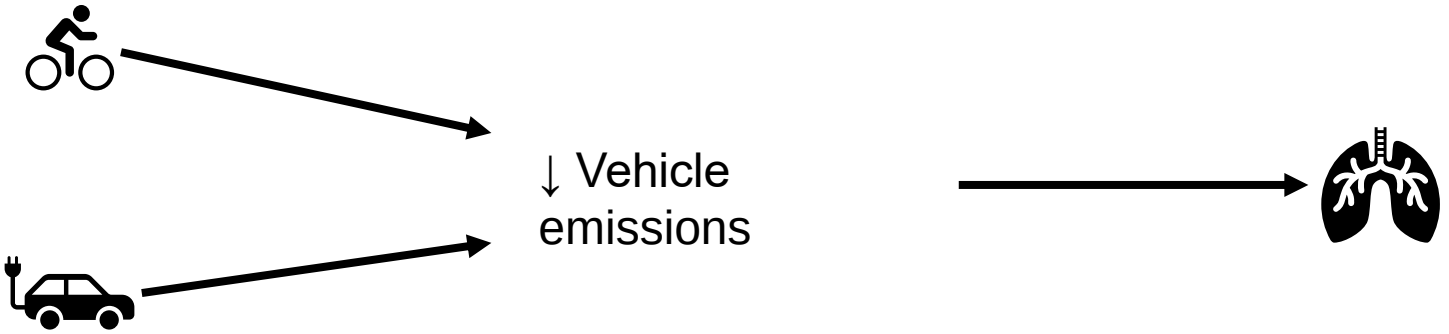
State the research question & hypothetical intervention

- Be explicit about:
  - Study (target) population
  - Exposure
  - Outcome
  - How to feasibly change the exposure
    - What is the hypothetical intervention?
    - What is the RCT?

Envisioning hypothetical interventions in observational studies

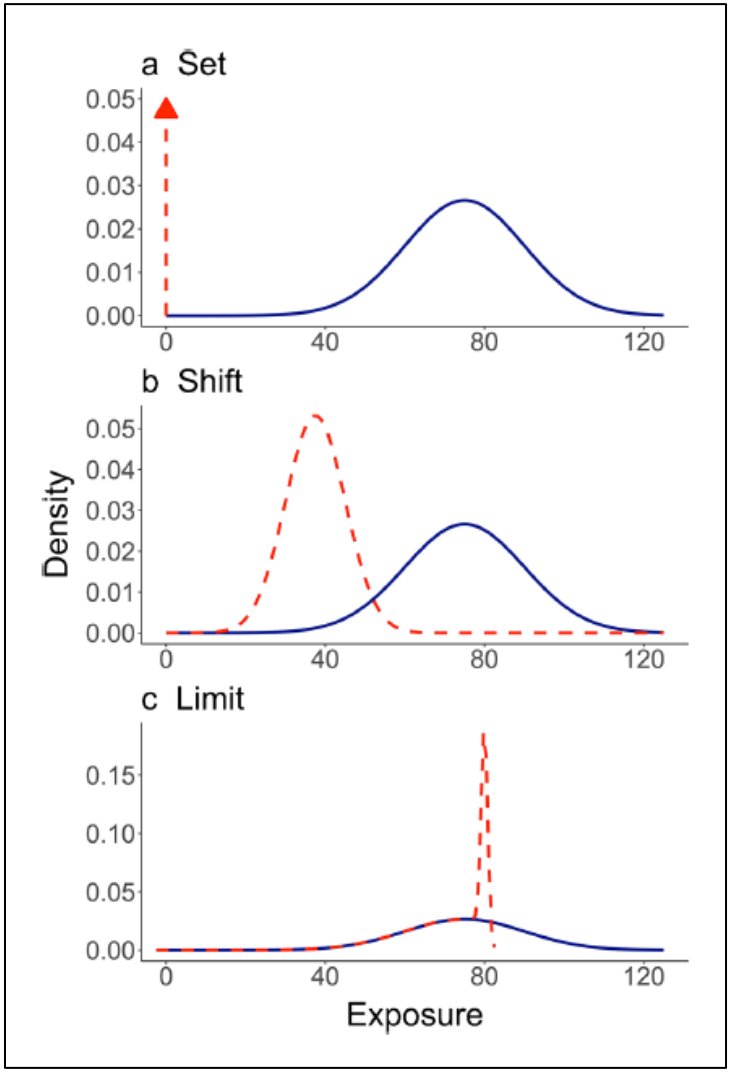


Envisioning hypothetical interventions in observational studies



While we can examine just vehicle emissions & asthma incidence, helpful to consider implications of different interventions that can be feasibly taken

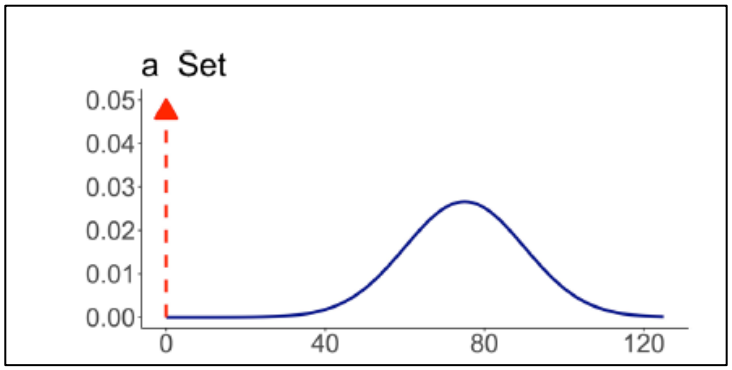
- Link study comparisons to public health actions that can be feasibly taken by decision-makers
- Operationalize intervention as a modification to measured exposure



- Set exposure to specific value
- Shift exposure distribution
- Limit exposure to threshold value

Smith et al. 2022, Curr Environ Health Rep

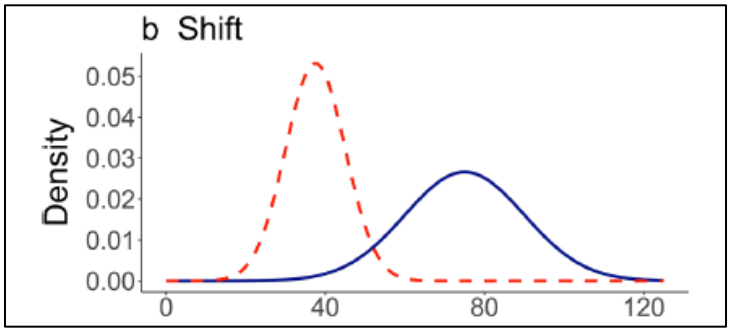
- Link study comparisons to public health actions that can be feasibly taken by decision-makers
- Operationalize intervention as a modification to measured exposure



- Set exposure to specific value
- e.g., set everyone in dataset unexposed vs exposed

Smith et al. 2022, Curr Environ Health Rep

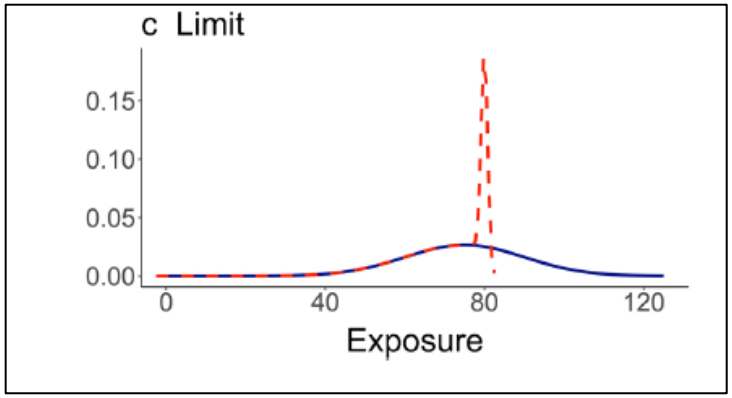
- Link study comparisons to public health actions that can be feasibly taken by decision-makers
- Operationalize intervention as a modification to measured exposure



- Shift exposure distribution
- e.g., everyone in dataset has 50% less exposure

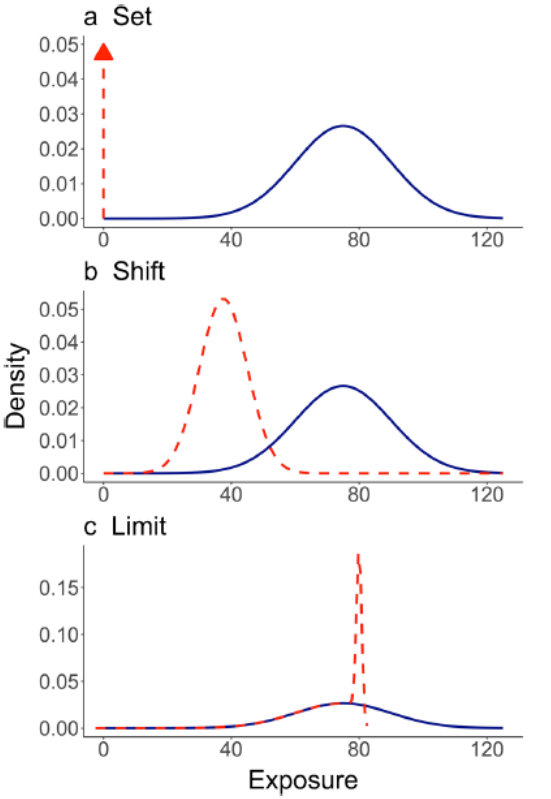
Smith et al. 2022, Curr Environ Health Rep

- Link study comparisons to public health actions that can be feasibly taken by decision-makers
- Operationalize intervention as a modification to measured exposure



- Limit exposure to threshold value
- e.g., apply regulatory thresholds to exposure

Smith et al. 2022, Curr Environ Health Rep



- Garcia et al. & Urman et al.
- PM2.5 and NO2 effects on childhood asthma incidence and lung development in CA, 1993-2014
  - **a. Set** PM2.5 or NO2 to baseline values observed in 1993
  - **b. Shift** PM2.5 or NO2 distributions downward by 10, 20, or 30%
  - **c. Limit** PM2.5 or NO2 values at hypothetical regulatory limits of 15, 12, 10 ug/m3 and 30, 20, 10 ppb, respectively
- Comparison was “natural course” or air pollution concentrations as observed over follow-up

Smith et al. 2022, Curr Environ Health Rep

State the research question & hypothetical intervention

Interested in the impact of tap water lead on neurodevelopmental outcomes in school-aged children

Statistical inference research question:

What is the association between a one-unit increase in tap water lead and risk of adverse neurodevelopmental outcomes in school-aged children?

Causal inference research question:

How would the risk of adverse neurodevelopmental outcomes in school-aged children change if we intervened to set tap water lead levels to below EPA standards (15 ppb)?

State the research question & hypothetical intervention

Interested in the impact of tap water lead on neurodevelopmental outcomes in school-aged children

Statistical inference research question:

What is the association between a one-unit increase in tap water lead and risk of adverse neurodevelopmental outcomes in school-aged children?

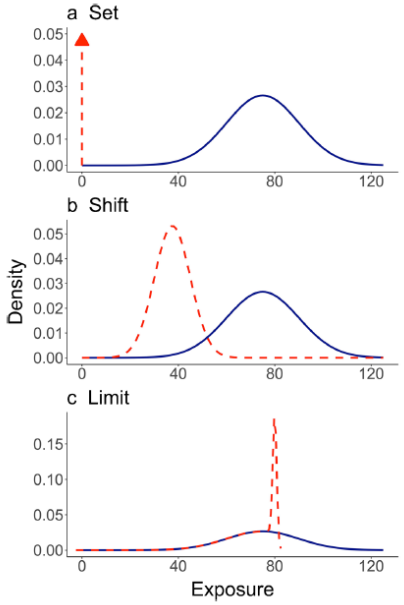
Causal inference research question:

How would the risk of adverse neurodevelopmental outcomes in school-aged children change if we had provided participants with a water filter that removes 90% of lead & other pollutants from tap water?

State the research question & hypothetical intervention

Interested in impact of noise from a local airport on sleep quality in older adults

Causal inference research questions:

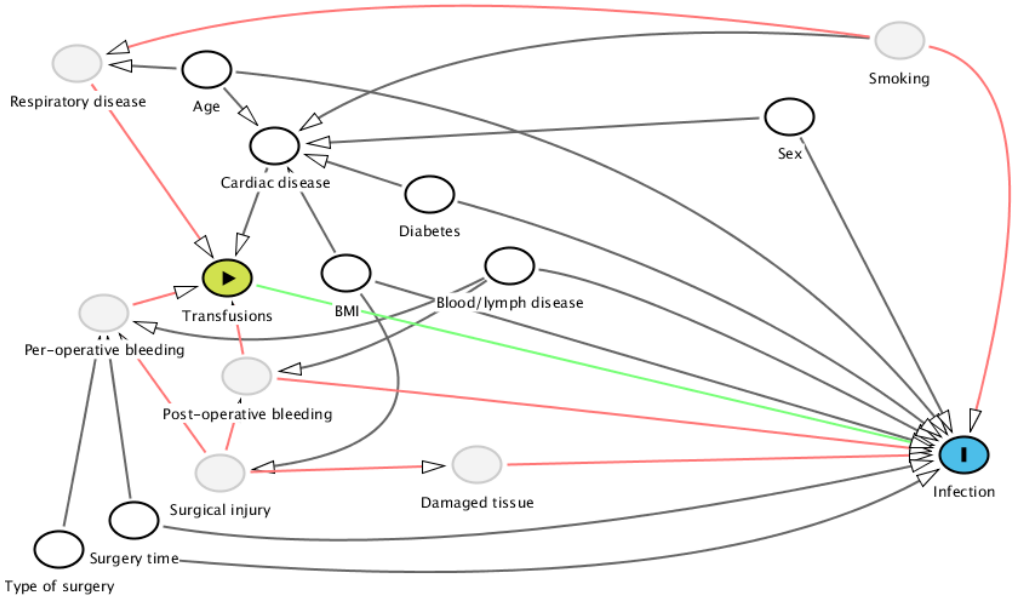


Causal inference framework

1. State the research question & hypothetical intervention
2. Define causal model & parameter of interest
3. Link causal model to observed data & define statistical model
4. Link causal effect to parameter estimable in observed data
5. Choose & apply estimator
6. Make inferences

Define causal model & parameter of interest

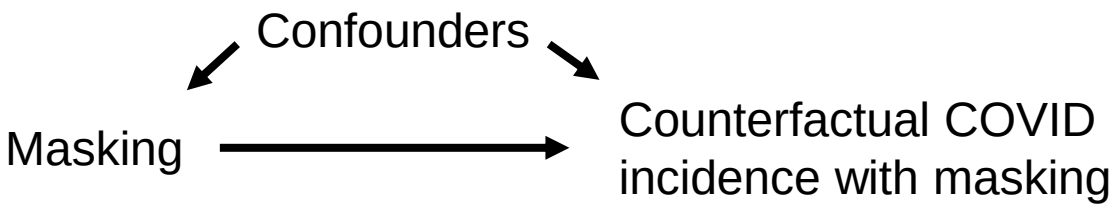
- In other words, draw a DAG!
  - Informed by prior knowledge
  - Often these are more complicated than we want them to be
- Use counterfactuals to define the causal parameter

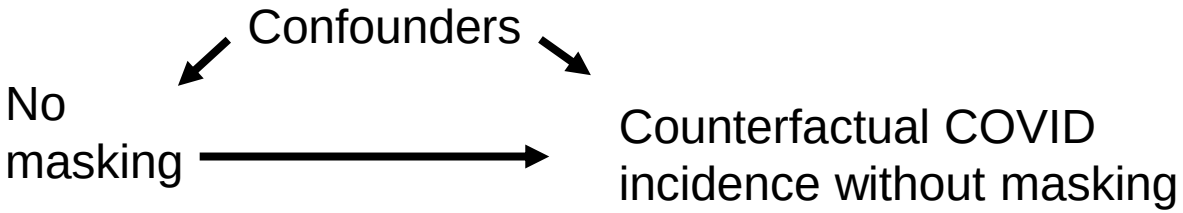


Source: https://gforge.se/2014/05/dag/

Define causal model & parameter of interest

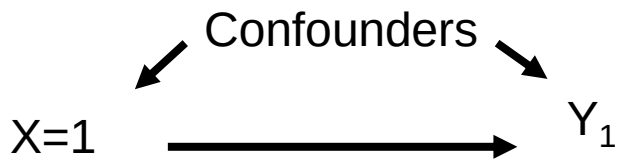
- Simple static intervention, using counterfactuals

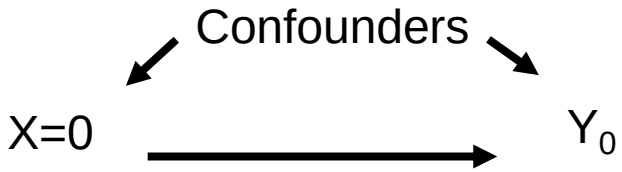




Define causal model & parameter of interest

- Simple static intervention, using counterfactuals





Average treatment effect:

$E[Y_1] - E[Y_0]$

Causal risk difference:

$P(Y_1=1) - P(Y_0=1)$

Many other causal parameters are possible!

Causal inference framework

1. State the research question & hypothetical intervention
2. Define causal model & parameter of interest
3. Link causal model to observed data & define statistical model
4. Link causal effect to parameter estimable in observed data
5. Choose & apply estimator
6. Make inferences

Link causal model to observed data & define statistical model

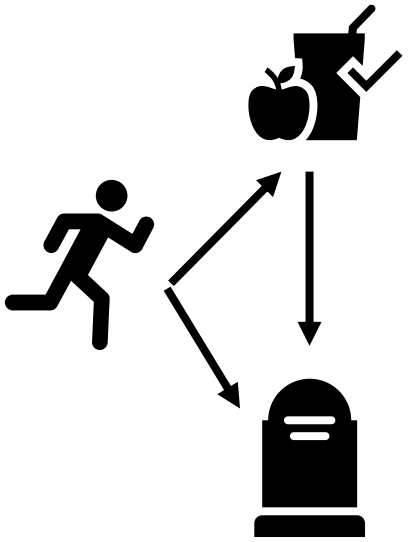
- Causal model = DAG
  - Describes the set of processes that give rise to your observed data
- Causal model implies the statistical model
  - Statistical model is set of possible distributions of observed data
  - Often no assumptions on distribution of unmeasured factors or functional form of equations
    - If we do know the form of the function between X, confounders, and Y then we should specify that
    - Statistical model often non-parametric

Causal inference framework

1. State the research question & hypothetical intervention
2. Define causal model & parameter of interest
3. Link causal model to observed data & define statistical model
4. Link causal effect to parameter estimable in observed data
5. Choose & apply estimator
6. Make inferences

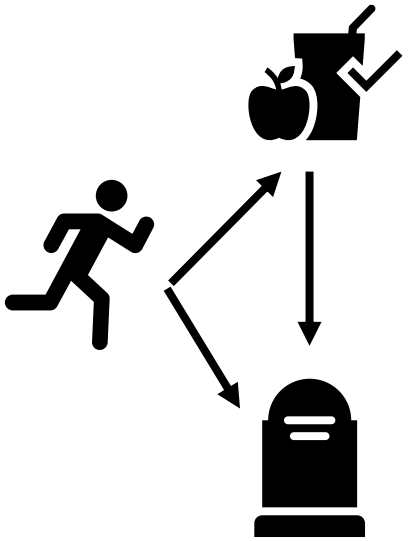
Assess identifiability

- Identifiability: link causal effect to parameter estimable in observed data
  - Are data sufficient to answer the causal question under model assumptions?



Assess identifiability

- Identifiability: link causal effect to parameter estimable in observed data
  - Are data sufficient to answer the causal question under model assumptions?
- What are some assumptions?
  - Counterfactual consistency
    - Potential outcome under observed exposure is indeed observed outcome
  - Conditional exchangeability
    - Adequate control for confounding and selection bias
  - Positivity
    - Sufficient variability in exposure within confounder strata
    - Empirically assess by examining exposure distributions within strata of confounders



Assess identifiability

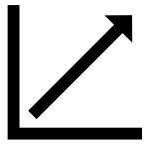
- Likely that data and model are not sufficient
- What then?
  - Get more (better) data
  - Do the best job with what you have, understand limitations, & make convenience assumptions
- Convenience assumptions, e.g.,
  - “No unmeasured confounding” when we suspect there may be some confounding that we did not measure

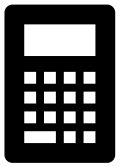
Causal inference framework

1. State the research question & hypothetical intervention
2. Define causal model & parameter of interest
3. Link causal model to observed data & define statistical model
4. Link causal effect to parameter estimable in observed data
5. Choose & apply estimator
6. Make inferences

Choose & apply estimator

- This is when we get into the statistics!
- Several options available:
  - Substitution estimators (e.g., g-computation)
  - Propensity score based (e.g., IPTW)
  - Doubly robust (e.g., TMLE)





Estimators

- Parametric
  - Assume that we know the relationship between your exposure, covariates, & outcome
  - Specify this relationship with parameters e.g., regression with main terms & some interactions or squared terms

Estimators

- Non-parametric
  - We know nothing about form of relationship between exposure, covariates, & outcome
  - Divide data into all possible combinations of exposure-covariate relationships and average stratum-specific exposure-outcome relations
    - This would be a lot of work!

Estimators

- Semi-parametric
  - Smooth over data with weak support during estimation using data-adaptive estimation or machine learning

- Statistical parameter

$$E\{E(Y|A=1,W) - E(Y|A=0,W)\}$$

Y = outcome

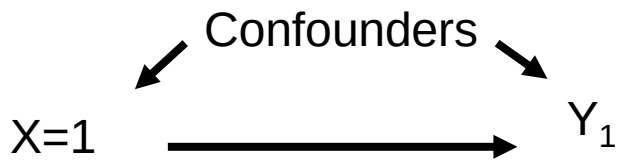
A = exposure

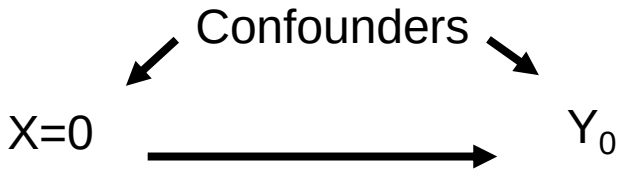
W = covariates

  - Equals the ATE if identifiability assumptions hold

Causal model & parameter of interest

Simple static intervention, using counterfactuals





Average treatment effect:

$E[Y_1] - E[Y_0]$

Causal risk difference:

$P(Y_1=1) - P(Y_0=1)$

Simple substitution estimator

1. Estimate outcomes for all participants under exposed and unexposed conditions, controlling for confounders

- Estimate mean outcome as function of exposure and confounders

- Use estimate to predict outcomes while “setting” exposure to different hypothetical values
  - e.g., values of exposed vs unexposed

Simple substitution estimator

2. Average and compare predicted outcomes

- Average predictions to estimate marginal risks under exposed vs unexposed conditions
- Take difference in means

Causal inference framework

1. State the research question & hypothetical intervention
2. Define causal model & parameter of interest
3. Link causal model to observed data & define statistical model
4. Link causal effect to parameter estimable in observed data
5. Choose & apply estimator
6. Make inferences

Make inferences

- Interpret findings
  - Consider assumptions that were and were not met
- If major violations of identifiability assumptions, maybe you just estimated an association, not a cause
  - As close as we can get!
- If no major violations of identifiability assumptions, can interpret parameter as ATE

Reminder that identifiability is just, "What can be estimated from the data"

Link to slides

Outline & References

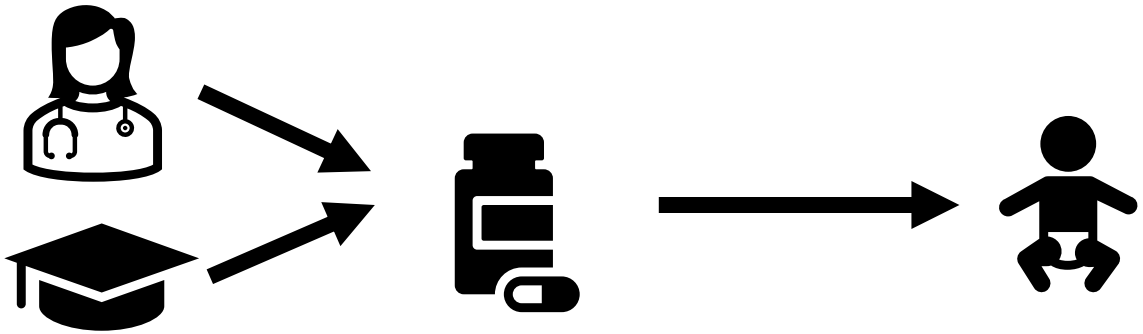
- What is causal inference (Smith et al)
  - Research questions
  - Causal framework
- G-methods (Naimi et al)
  - Data for R tutorial
  - Propensity scores (IPTW)
    - R tutorial
  - G-computation
    - R tutorial

Data for R tutorial

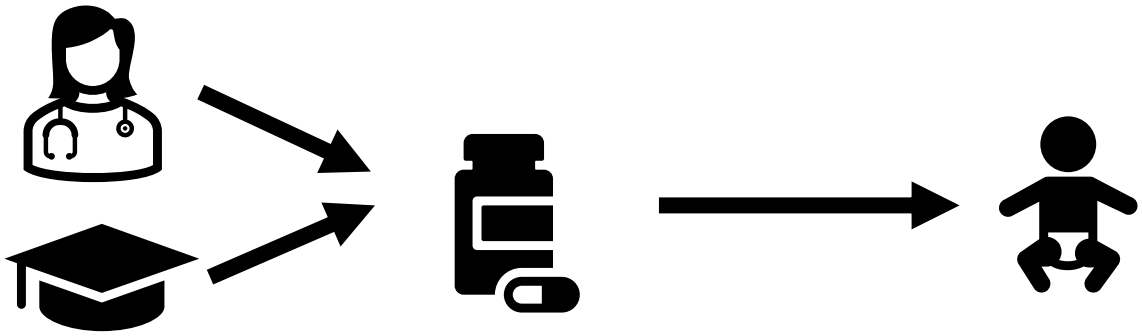
- Data at https://ehsanx.github.io/TMLEworkshop/rhc-data-description.html
- There is a widespread belief among cardiologists that the right heart catheterization (RHC) is helpful in managing critically ill patients in the intensive care unit, and can decrease the length of stay in the hospital
  - Connors et al, 1996
- RHC dataset
  - A = exposure = RHC
  - Y = outcome = Length of hospital stay
  - L = 49 covariates

Propensity scores (IPTW)

- Think of confounding as problem of biased sampling
  - Some groups under- or over-represented in observational study relative to RCT



Propensity scores (IPTW)



- Use weights to correct for biased sampling
  - Up-weight under-represented
  - Down-weight over-represented
- Other types of propensity scores methods, e.g., matching

Propensity scores (IPTW)

1. Fit a model of exposure given confounders to obtain conditional probabilities of exposure given confounders
2. Fit weighted regression of outcome on exposure using inverse of the conditional probabilities as weights

Note: Difficult to estimate step 1 with continuously measured exposures

Propensity scores (IPTW)

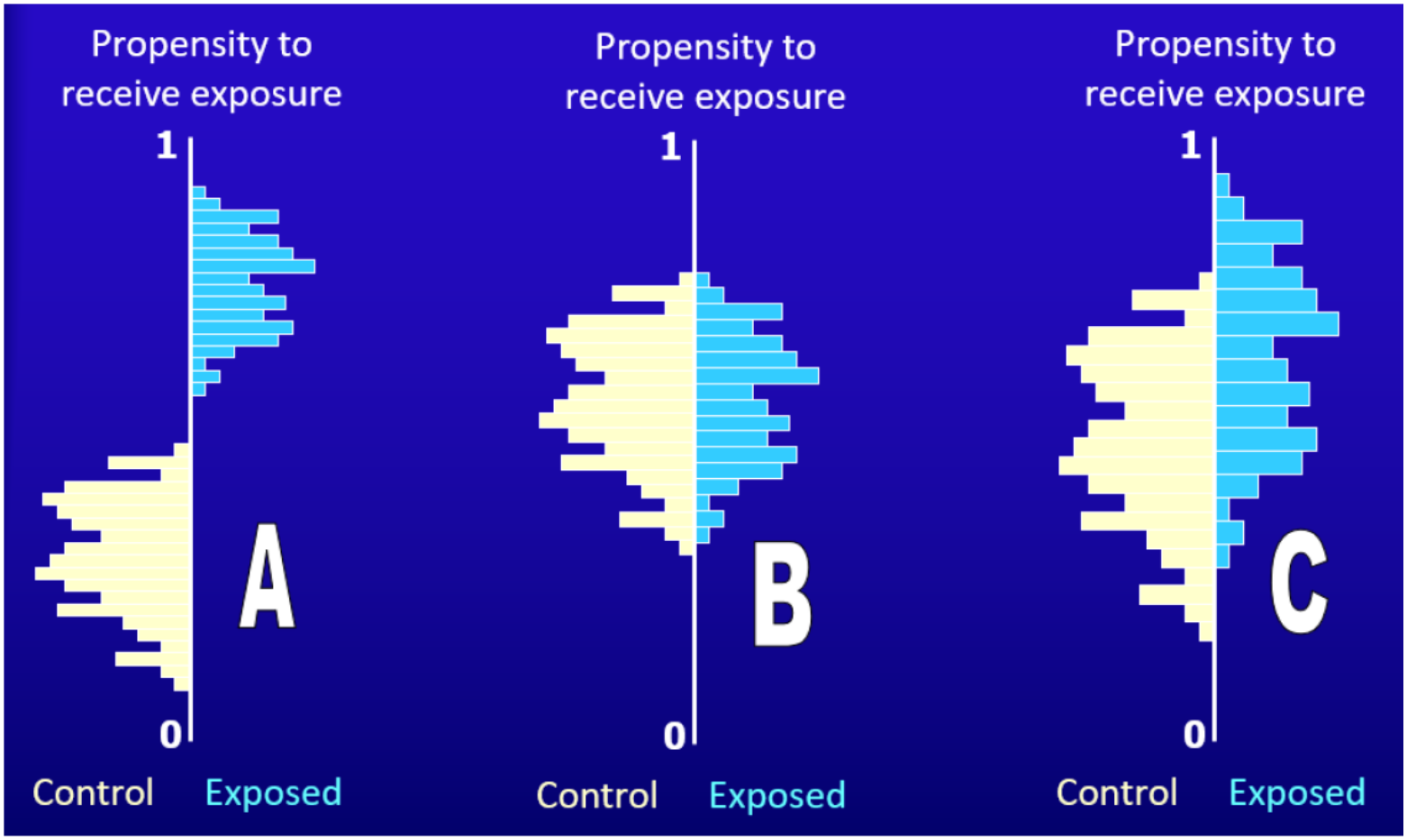
1. Fit a model of exposure given confounders to obtain conditional probabilities of exposure given confounders

1. Estimate probability of being exposed as a function of measured confounders, e.g., via logistic regression

   $P(X=1 \mid W)$

2. Calculate IPTW
   For exposed: $1 / P(X=1 \mid W)$
   For unexposed: $1 / P(X=0 \mid W)$

Diagnostics: plot weights



Diagnostics: standardized differences

- In observational study, expect covariates to differ between exposed and unexposed
  - e.g., those taking prenatal vitamin will be multiparous, have higher education, initiate prenatal care earlier in pregnancy
- Quantify covariate imbalance
  - Assess differences in group distributions by covariates

How different are characteristics by gestational weight gain status?

Bodnar & Hutcheon. 2022, Epidem

Diagnostics: standardized differences

TABLE 2. Balance in Standard and Detailed Confounders Before and After Matching on Propensity Score for Low Pregnancy Weight Gain in the Nulliparous Pregnancy Outcomes Study: Monitoring Mothers-to-be Cohort (2010–2013), n = 8978

	Before Propensity Score Matching			After Propensity Score Matching		
	Mean			Mean		
Variable	Low Weight Gain [a]	Adequate Weight Gain	% bias (Standardized Differences)	Low Weight Gain	Adequate Weight Gain	% bias (Standardized Differences)
Standard confounders						
Maternal age, years	26	27	−24[b]	26	26	−1.2
Prepregnancy body mass index, kg/m^2	25	24	19[b]	25	25	2.8
Smoker, %	16	13	8.4	16	15	−2.3
Married, %	56	67	−25[b]	56	57	−2.3
Public insurance, %	41	26	33[b]	41	40	2.4
Some college education, %	30	26	9.6	30	29	2.0
Bachelor's degree, %	23	31	−18[b]	23	24	−1.4
Graduate degree, %	21	27	−14[b]	21	22	−2.3

Goal is to minimize standardize differences (achieve balance in covariates) before moving to next step

Bodnar & Hutcheon. 2022, Epidem

Note on weights

- In cases where some participants have extreme weights after PS calculation, can “trim” dataset to exclude those participants
  - Could trim all weights above some criteria (e.g., 95th percentiles)
- If you have other weights (e.g., survey weights), you multiply the two weights

Propensity scores (IPTW)

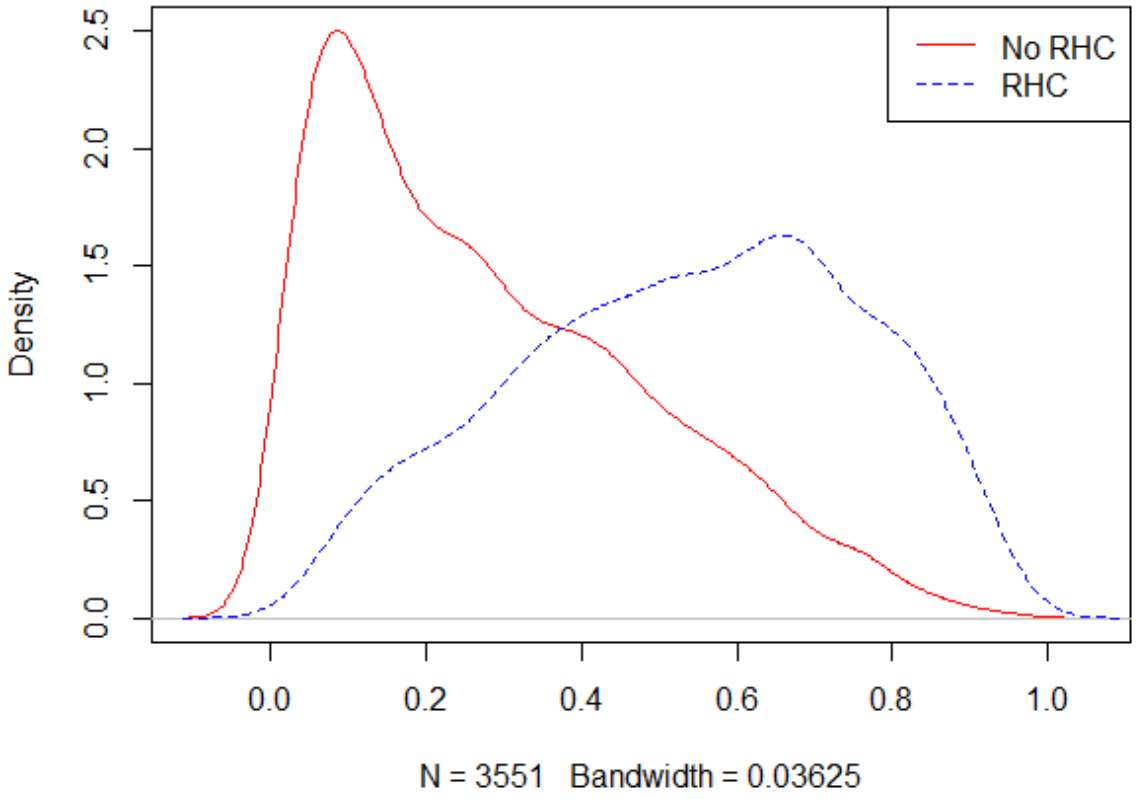
2. Fit weighted regression of outcome on exposure using inverse of the conditional probabilities as weights

1. Apply weights within a second model that estimates effect of exposure on outcome

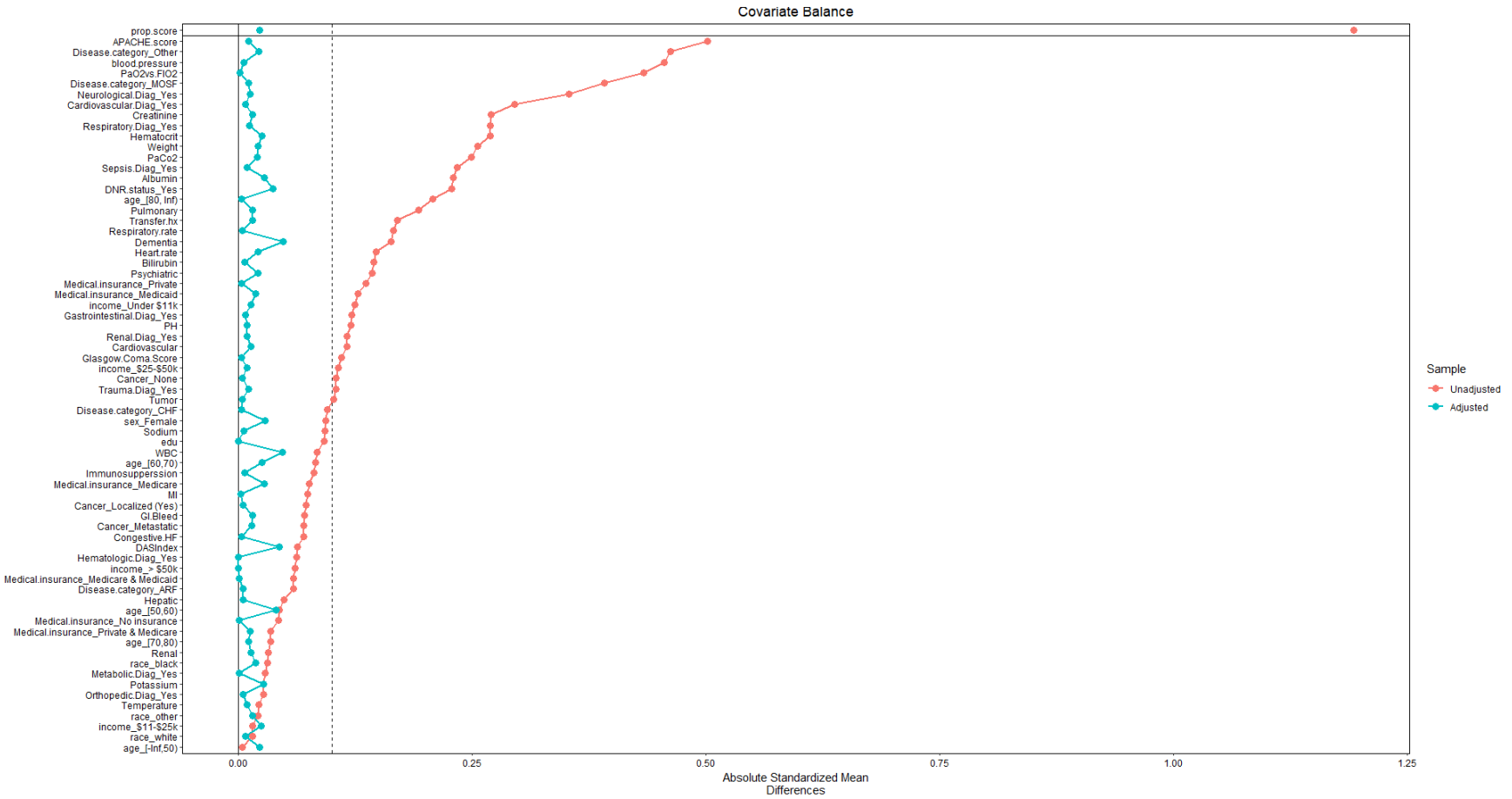
   $P(Y=1 \mid W, X)$
   $P(Y=0 \mid W, X)$

2. Calculate differences
   $P(Y=1 \mid W, X) – P(Y=0 \mid W,X)$

Run in R: PS weights



Run in R: Standardized differences



Run in R: Interpret

- What was the ATE?
- How is this interpreted?
- How does the ATE differ from unadjusted estimates?

G-computation: Dataset with L=sex

sex	A	Y
Male	0	9
Female	1	45
Female	1	60
Female	0	37
Male	1	2
Female	0	7

G-computation: Restructure data

- Outcomes under different exposure scenarios are in different columns

Subject ID	Sex	RHC status (A)	Y when A=1 (RHC)	Y when A=0 (no RHC)
John	Male	0		9
Emma	Female	1	45	
Isabella	Female	1	60	
Sophia	Female	0		37
Luke	Male	1	2	
Mia	Female	0		7
			36	**18**

Think of as a missing data problem

G-computation: Restructure Data

- Outcomes under different exposure scenarios are in different columns

Subject ID	Sex	RHC status (A)	Y when A=1 (RHC)	Treatment Effect
John	Male	0		
Emma	Female	1	45	
Isabella	Female	1	60	
Sophia	Female	0		
Luke	Male	1	2	
Mia	Female	0		
			36	18	18

Predict Ys when A=1 for those where A = 0 in the observed data

G-computation: Restructure Data

- Outcomes under different exposure scenarios are in different columns

Subject ID	Sex	RHC status (A)	Y when A=0 (no RHC)	Treatment Effect
John	Male	0	9	
Emma	Female	1		
Isabella	Female	1		
Sophia	Female	0	37	
Luke	Male	1		
Mia	Female	0	7	
			36	18	18

Predict Ys when A=0 for those where A = 1 in the observed data

G-computation: Estimate Individual Differences

Subject ID	Sex	RHC status (A)	Y when A=1 (RHC)		Y when A=0 (no RHC)		Treatment Effect
John	Male	0	36	-	9	=	27

G-computation: Average Individual Differences

Subject ID	Sex	RHC status (A)	Y when A=1 (RHC)	Y when A=0 (no RHC)	Treatment Effect
John	Male	0	36	9	27
Emma	Female	1	45	18	27
Isabella	Female	1	60	18	42
Sophia	Female	0	36	37	-1
Luke	Male	1	2	18	-16
Mia	Female	0	36	7	29
			36	**18**	**18**

G-computation

- Switch over to a regression approach, taking into account covariates

Steps:

1. Fit model of outcome given exposures and confounders
2. From fitted model, predict outcome under exposure levels corresponding to intervention of interest
3. Calculate difference of the mean predictions for exposure contrasts of interest

G-computation

1. Fit model of outcome given exposures and confounders
2. From fitted model, predict outcome under exposure levels corresponding to intervention of interest
   - P(Y=1 | A, L)
   - P(Y=0 | A, L)
3. Calculate difference of the mean predictions for exposure contrasts of interest
   - Individual differences
     - P(Y=1 | A, L) - P(Y=0 | A, L) for each participant
   - Average over individual differences
     - Average and then calculate CIs using bootstrapping

Run in R: Interpret

- What was the ATE?
- How is this interpreted?
- How does the ATE differ from unadjusted and adjusted estimates?

A couple of notes…

- We just walked through simple examples
  - Other treatment contrasts possible
  - Longitudinal settings possible
  - Complex exposure scenarios (e.g., mixtures)
- There are packages that can do this as well!
  - lptw, twang
  - gfoRmula

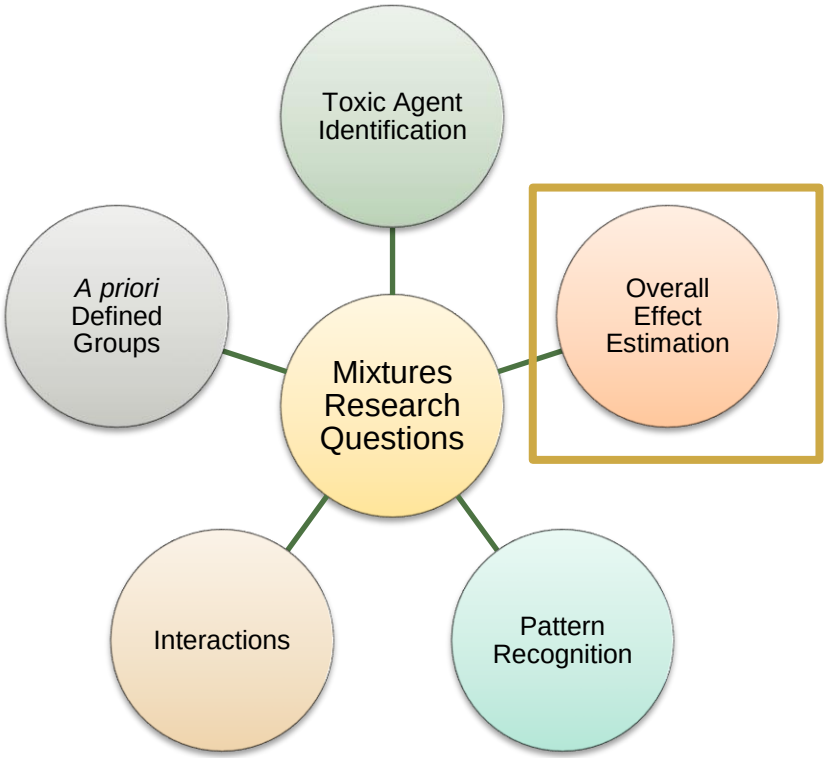
A couple of notes…

- Non- or semi-parametric methods
  - Machine learning
- Doubly robust: TMLE, LTMLE
- Remember to check your assumptions (if possible)
  - Counterfactual consistency
    - Not verifiable
  - Conditional exchangeability
    - Not verifiable
  - Positivity
    - Empirically assess by examining exposure distributions within strata of confounders

A couple of notes…

- We just walked through simple examples
  - Other treatment contrasts possible
  - Longitudinal settings possible
  - **Complex exposure scenarios (e.g., mixtures)**
- There are packages that can do this as well!
  - lptw, twang
  - gfoRmula

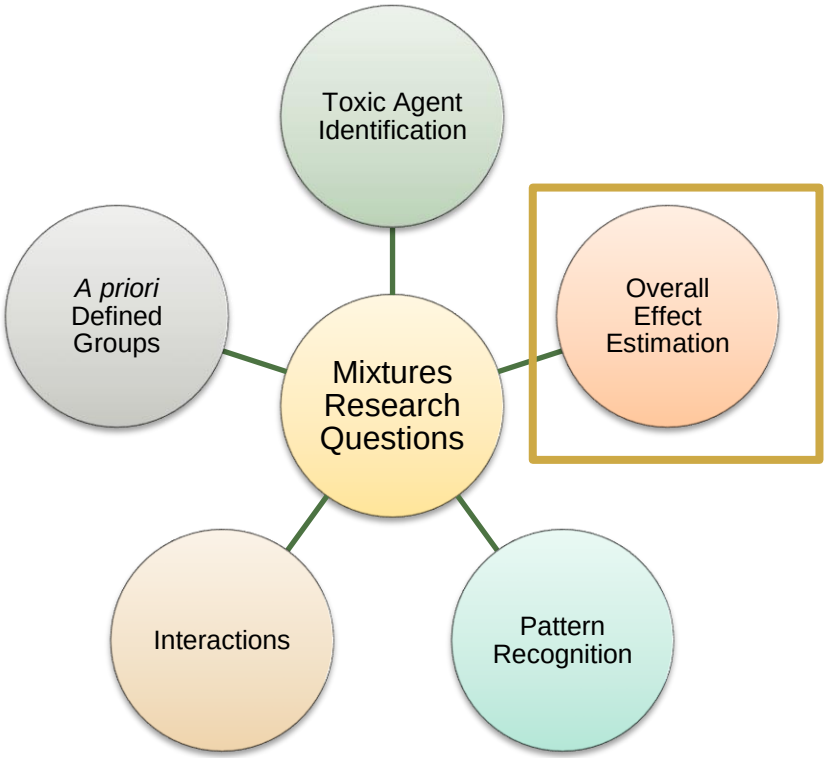
Exciting extensions: quantile g-computation



Addresses the question:

How can the mixture as a whole, rather than individual components, influence the health of the populations exposed to the multitude of components in the mixture?

Exciting extensions: quantile g-computation



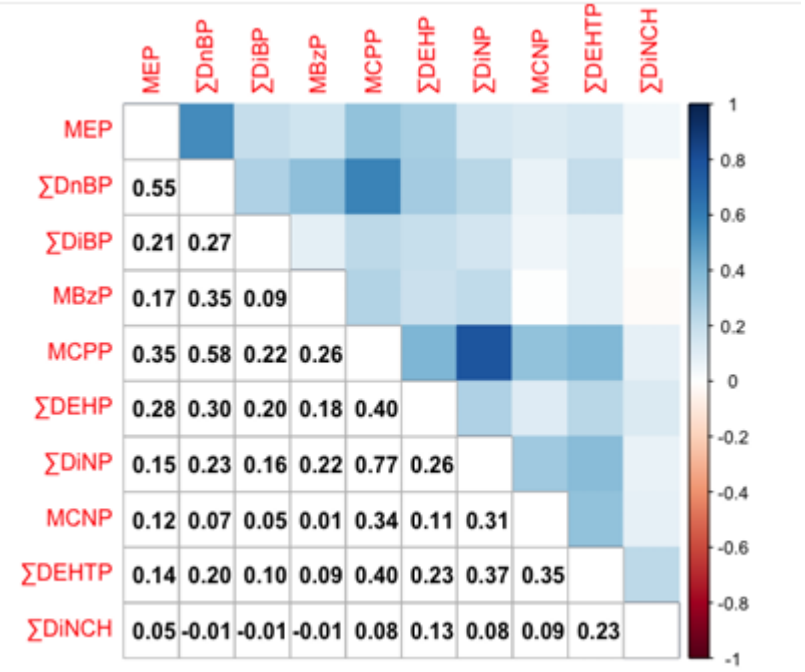
- Simple, interpretable, comparable to single-pollutant estimates
- Accommodate missing data, weights, longitudinal exposures & outcomes

Exciting extensions: quantile g-computation

- Step 1: Transform exposures into quantized versions
- Step 2: Fit a linear model: $Y_i = \beta_0 + \sum_{j=1}^{d} \beta_j X_{ji}^{q} + \varepsilon_i.$
- Step 3: Estimate the mixture effect via standard g-computation algorithms as described in Snowden et al. 2011. Briefly:
  - Fit underlying model allowing individual effects of exposures on the outcome, including interactions & nonlinear terms
  - Make predictions at set levels of the exposures
  - Fit a marginal structural model to these predictions

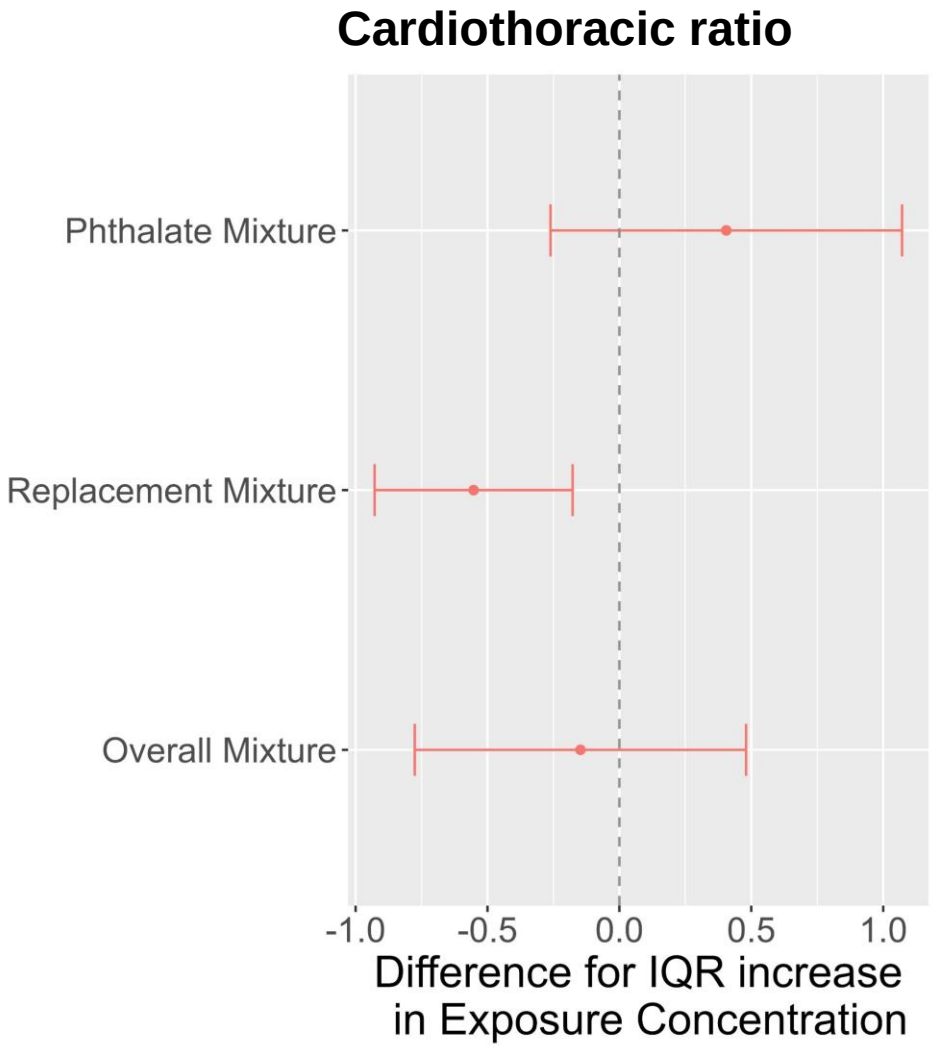
Exciting extensions: quantile g-computation

In the HPP-3D Study, we assessed exposure to 10 endocrine disrupting chemicals during pregnancy…



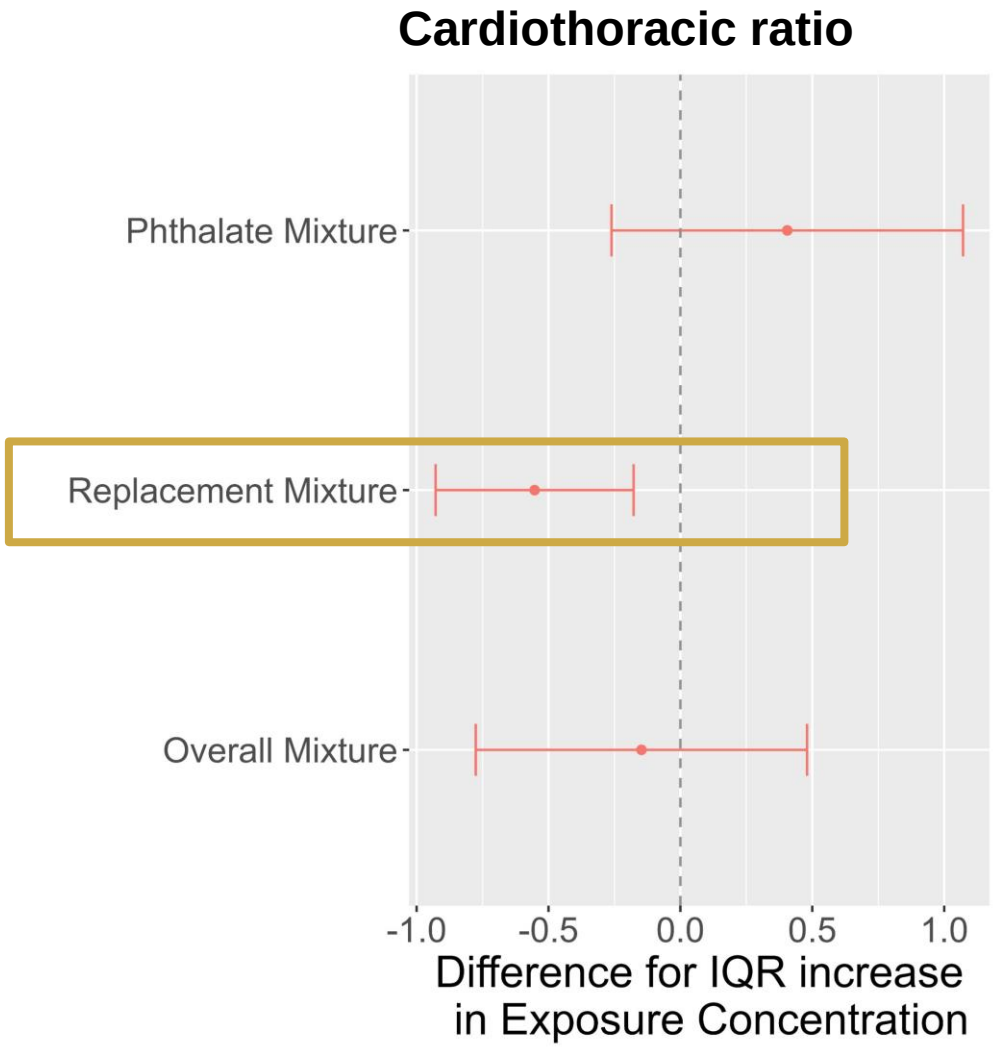
Exciting extensions: quantile g-computation

…and looked at associations with fetal cardiometabolic development



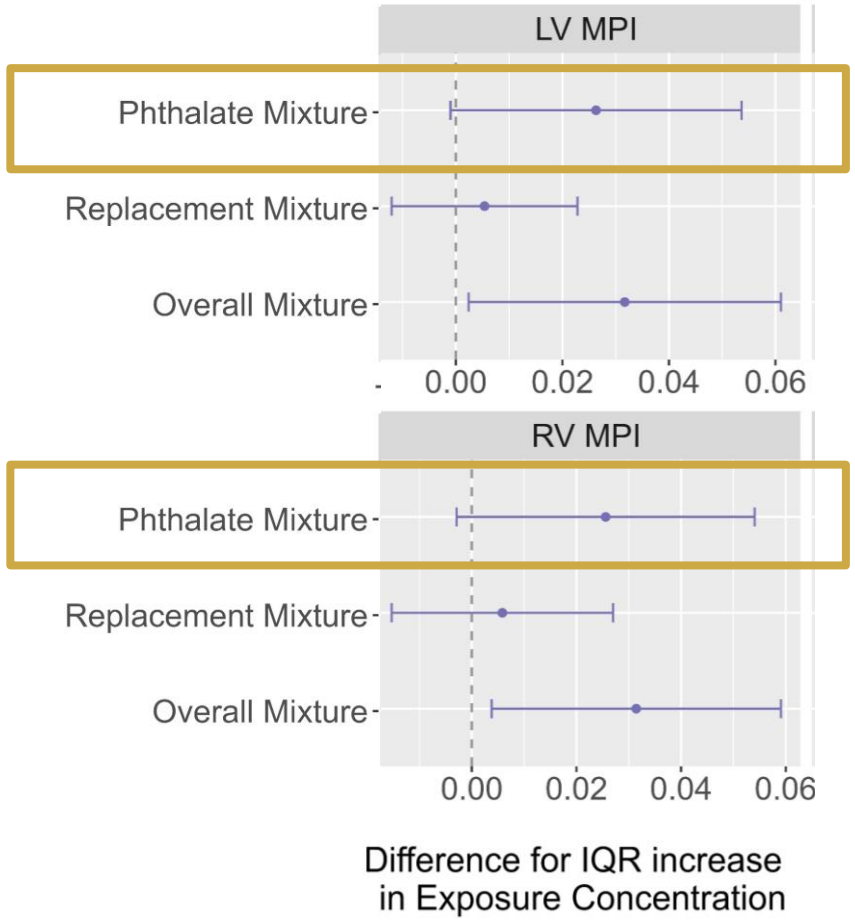
Exciting extensions: quantile g-computation

An IQR increase in gestational replacement chemical concentrations was associated with lower fetal heart size

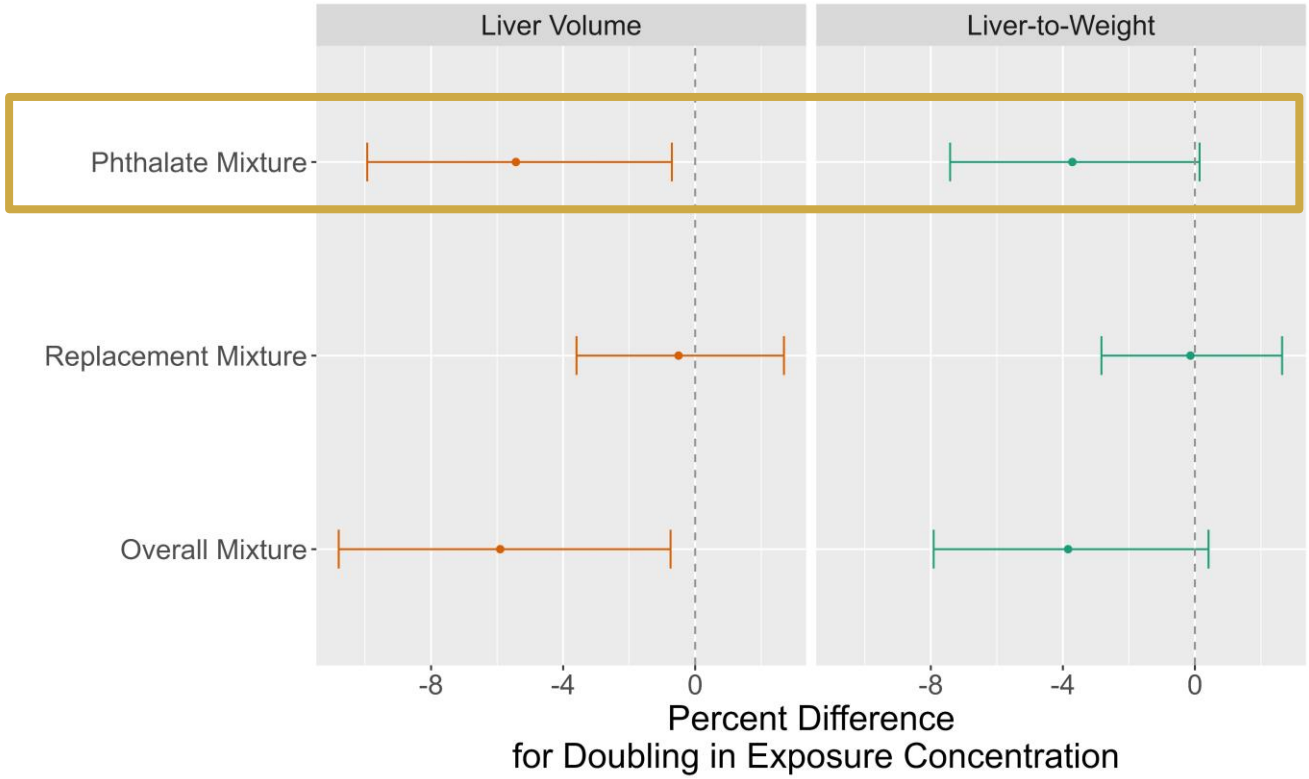


Exciting extensions: quantile g-computation

An IQR increase in gestational chemical (mostly phthalate) concentrations was associated with poorer fetal global cardiac function



Exciting extensions: quantile g-computation



A doubling in gestational phthalate exposure concentrations was associated with a lower fetal liver volume

Exciting extensions: quantile g-computation

- We concluded that gestational phthalate exposure was associated with impairments in fetal cardiometabolic development & demonstrated:
  - Qgcomp useful for estimating total mixture effects
  - Also able to isolate effects of specific chemicals from the mixture
    - Phthalates from Replacement Chemicals
    - Could do single-chemicals as well
- https://cran.r-project.org/web/packages/qgcomp/vignettes/qgcomp-vignette.html

Link to slides

Danielle Stevens

Danielle.stevens@nih.gov

https://simplydani99.github.io/