An introduction to statistical, causal and epistatic interaction

Stijn Vansteelandt
Ghent University, Belgium
London School of Hygiene and Tropical Medicine, U.K.

COST, Antwerp 2016

Acknowledgement

Slides partly build on work and course materials by

Tyler VanderWeele, Harvard University

as well as joint work with him and

Christophe Lange, Eric Tchetgen Tchetgen and James Robins, Harvard University

Outline

- Statistical interactions
- Causal interactions
- Robustness

Part I. Statistical interaction

Interaction

- The effect of one exposure often depends in some way on the presence or absence of another exposure.
- We then say that there is interaction between the two exposures.
- E.g. gene-environment interaction, gene-gene interaction, ...
- The process giving rise to illness and health is often inherently complex.
- Interaction is one manifestation of this complexity.

Example: breast cancer, alcohol and XRC3-T241M

- Odds ratios for breast cancer:

	No Alcohol	Alcohol
T/T or T/M	1	1.12 (0.81 to 1.54)
M/M	1.21 (0.70 to 2.09)	2.09 (1.16 to 3.78)

(Figueiredo et al., 2004)

- Gene-environment interaction:
  XRC3-T241M appear to be associated with breast cancer only when accompanied by alcohol consumption.
- Other examples are
  - APOE and dietary cholesterol on serum cholesterol;
  - PPARG2 and dietary fat on obesity;

(Hunter, 2005)

Additive interaction

- Consider the following disease risks in a cohort study.

	Unexposed ($E = 0$)	Exposed ($E = 1$)
a/a ($G = 0$)	$p_{00} = 0.02$	$p_{01} = 0.05$
a/A or A/A ($G = 1$)	$p_{10} = 0.04$	$p_{11} = 0.15$

- We then say there is additive interaction because the effect of the G and E together

$$p_{11} - p_{00} = 0.13$$

differs from the sum of the effects of each separately:

$$p_{10} - p_{00} + p_{01} - p_{00} = 0.02 + 0.03 = 0.05$$

- The additive interaction contrasts these numbers:

$$p_{11} + p_{00} - p_{01} - p_{10} = 0.13 - 0.05 = 0.08.$$

- We say there is a positive or superadditive interaction (as opposed to negative or subadditive).

Multiplicative interaction

- Consider the following disease risks in a cohort study.

	Unexposed ($E = 0$)	Exposed ($E = 1$)
a/a ($G = 0$)	$p_{00} = 0.02$	$p_{01} = 0.05$
a/A or A/A ($G = 1$)	$p_{10} = 0.04$	$p_{11} = 0.15$

- We then say there is multiplicative interaction because the effect of the G and E together

$$R_{11} = p_{11}/p_{00} = 7.5$$

differs from the product of the effects of each separately:

$$R_{10} \times R_{01} = (p_{10}/p_{00}) \times (p_{01}/p_{00}) = 2 \times 2.5 = 5$$

- The multiplicative interaction contrasts these numbers:

$$\frac{R_{11}}{R_{10} \times R_{01}} = 1.5.$$

Interactions are scale-dependent

- Consider the following disease risks in a cohort study.

	Unexposed ($E = 0$)	Exposed ($E = 1$)
a/a ($G = 0$)	$p_{00} = 0.02$	$p_{01} = 0.05$
a/A or A/A ($G = 1$)	$p_{10} = 0.04$	$p_{11} = 0.10$

- Is there additive and/or multiplicative interaction?
- Is the interaction positive or negative?

Interactions are scale-dependent

- Consider the following disease risks in a cohort study.

	Unexposed ($E = 0$)	Exposed ($E = 1$)
a/a ($G = 0$)	$p_{00} = 0.02$	$p_{01} = 0.05$
a/A or A/A ($G = 1$)	$p_{10} = 0.07$	$p_{11} = 0.10$

- Is there additive and/or multiplicative interaction?
- Is the interaction positive or negative?

Interactions are scale-dependent

- Consider the following disease risks in a cohort study.

	Unexposed ($E = 0$)	Exposed ($E = 1$)
a/a ($G = 0$)	$p_{00} = 0.01$	$p_{01} = 0.05$
a/A or A/A ($G = 1$)	$p_{10} = 0.04$	$p_{11} = 0.10$

- Is there additive and/or multiplicative interaction?
- Is the interaction positive or negative?
- Suppose we have 100 doses of a drug (E) and the outcome means 'cured'.
- There are 100 patients in the $G = 0$ group and 100 in the $G = 1$ group.
- Who should we treat?

Additive versus multiplicative interactions

- In most published epidemiologic studies, interactions are evaluated and reported on the multiplicative scale.
- Interaction on the additive scale are often not; perhaps only about 1 in 50 in epidemiology reported.

  (Knol et al., 2009)
- Previous example shows that additive interactions are more interesting from a public health perspective.
- Recommendation: report estimates and confidence intervals on both scales.

Case-control designs

- The main reasons why there is so much focus on multiplicative interactions have to do with
  - the fact that the multiplicative scale sometimes naturally corresponds to the biological mechanisms;

    (Siemiatycki and Thomas, 1981)
  - confounding control (later);
  - case-control designs.
- Disease risks cannot be estimated from case-control designs.
- Odds ratios can, and approximate relative risks when disease risks are small.

Multiplicative interaction in case-control designs

- Consider the following numbers of cases/controls in a case-control study.

	Unexposed ($E = 0$)	Exposed ($E = 1$)
a/a ($G = 0$)	10/52	50/50
a/A or A/A ($G = 1$)	40/51	100/47

- Is there (approximately) multiplicative interaction?

$$
\begin{aligned}
R_{11} &\approx \frac{100 \times 52}{47 \times 10} = 11 \\
R_{10} &\approx \frac{40 \times 52}{51 \times 10} = 4 \\
R_{01} &\approx \frac{50 \times 52}{50 \times 10} = 5
\end{aligned}
$$

- Is there (approximately) additive interaction?

Additive interaction in case-control designs

- Upon dividing

$$p_{11} + p_{00} - p_{01} - p_{10}$$

by p_{00}, we obtain

$$R_{11} + 1 - R_{01} - R_{10},$$

which is the Relative Excess Risk due to Interaction (RERI).

(Rothman, 1986)

- In the previous example, we conclude that

$$\mathrm{RERI} = 11 + 1 - 4 - 5 > 0,$$

suggesting a positive additive interaction.

Example: breast cancer, alcohol and XRC3-T241M

- Odds ratios for breast cancer:

	No Alcohol	Alcohol
T/T or T/M	1	1.12 (0.81 to 1.54)
M/M	1.21 (0.70 to 2.09)	2.09 (1.16 to 3.78)

(Figueiredo et al., 2004)

- There is positive interaction on the multiplicative scale:

$$\frac{2.09}{1.21 \times 1.12} = 1.54 > 1$$

- There is positive interaction on the additive scale:

$$\mathrm{RERI} \approx 2.09 - 1.21 - 1.12 + 1 = 0.76 > 0$$

Part II. Causal interaction

Interaction

- The interest in gene-environment or gene-gene interactions is typically motivated by a desire to understand better the disease etiology.
- Previous interaction measures have shortcomings in such case.

- They are based on contrasts between groups that may not be entirely comparable, due to confounding.

Controlling for confounding

- To understand how to best control for confounding, note that different causal questions can be phrased that all capture the notion of interaction.
- *Is the effect of smoking on lung cancer different for people with different variants of rs8034191?*
- To investigate this question, one must control for common causes of smoking and lung cancer.

Controlling for confounding

- *Is the effect of SNP rs8034191 different for smokers versus non-smokers?*
- To investigate this question,
  one must control for common causes of the SNP and lung cancer.
- This may be easier to attain because of random inheritance (although population admixture; see later).

Controlling for confounding

- *Does smoking modify the effect of SNP rs8034191 on lung cancer?*
- Here, we consider the effect of the exposure and genotype jointly.
- This requires control for confounding of both associations.

Controlling for confounding

- *Are there individuals who would develop lung cancer if they, at the same time, smoke and carry the risk allele on SNP rs8034191, but not if they have only one of these exposures?*
- This is a question about causal mechanism, synergism.
- We will address this later.

Additive interaction

- Additive interaction can be estimated as the term β in the linear regression model

$$E(Y|G,E) = \gamma_0 + \gamma_1 G + \gamma_2 E + \beta GE$$

- Indeed, one can verify that

$$\begin{aligned}\beta \;=\; & \underbrace{E(Y|G=1,E=1)}_{\gamma_0+\gamma_1+\gamma_2+\beta} + \underbrace{E(Y|G=0,E=0)}_{\gamma_0} \\ & - \underbrace{E(Y|G=0,E=1)}_{\gamma_0+\gamma_2} - \underbrace{E(Y|G=0,E=1)}_{\gamma_0+\gamma_1}\end{aligned}$$

Additive interaction

- This easily enables control for confounding
  by including measured confounders X
  (e.g. substructure-informative loci, principal components, ...)
  in the model

$$E(Y|G,E,X) = \gamma_0 + \gamma_1 G + \gamma_2 E + \gamma_x X + \beta GE$$

- One can now verify that

$$\begin{aligned} \beta &= E(Y|G=1,E=1,X) + E(Y|G=0,E=0,X) \\ &\quad -E(Y|G=0,E=1,X) - E(Y|G=1,E=0,X) \end{aligned}$$

- This can also be used for non-dichotomous outcomes,
  genotypes or exposures.

Multiplicative interaction

- Multiplicative interaction can be estimated as the term β in the loglinear regression model

$$\log\ E(Y|G,E,X) = \gamma_0 + \gamma_1 G + \gamma_2 E + \gamma_x X + \beta GE$$

- Related interactions can be estimated as the term β in the logistic regression model

$$\text{logit}\ E(Y|G,E,X) = \gamma_0 + \gamma_1 G + \gamma_2 E + \gamma_x X + \beta GE$$

- The latter, unlike the other models, gives valid results in case-control designs.

Case-only designs

- Gene and environment are often independent, when the environmental exposure is not genetically affected.
- If that is the case, then interaction term β in model

$$\text{logit } E(Y|G,E,X) = \gamma_0 + \gamma_1 G + \gamma_2 E + \gamma_x X + \beta GE$$

  can be estimated more precisely from less information.
- The multiplicative interaction β_3 then equals the odds ratio between G and E in the subgroup of cases (with given X).

  (Yang et al., 1999; Piegorsch et al., 1994; Moerkerke et al., 2013)
- It is thus obtained by fitting the logistic regression model

$$\text{logit } E(G|Y=1,E,X) = \alpha_0 + \beta E + \alpha_x X$$

- To assess multiplicative interaction, only data for cases are then needed.
- This is referred to as the case-only estimator of interaction.

Statistical versus mechanistic interaction

- Ultimately, geneticists are interested in mechanistic interdependencies between genes or genes and environmental exposures.
- Standard model-based tests for interaction do not signal such mechanistic interaction processes.
- Tests for sufficient cause interactions have been developed for this purpose.

  (Rothman, 1976; VanderWeele and Robins, 2007)
- These aim to signal the presence of individuals for whom the outcome (e.g., disease) would occur if both exposures were 'present', but not if only one of the two were present.

Sufficient causes

- How can we formalize this notion?
- Rothman (1976) defined a 'sufficient cause'
  as minimal set of events, conditions or characteristics
  that gave rise to a process that inevitably produced disease.

Example

example

- Suppose we have 3 genes of interest: G_1, G_2, G_3.
- Let $Y_{g_1 g_2 g_3}$ be the counterfactual disease status (1: disease; 0: no disease) that would be observed if genotype (g_1, g_2, g_3) were present.
- Suppose the following sufficient cause representation holds:

  (VanderWeele and Robins, 2007)

$$Y_{g_1 g_2 g_3} = B_1 g_1 g_2 \bigvee B_2 g_1 g_3 \bigvee B_3 g_3$$

  where B_1, B_2, B_3 are background causes.
- Then $B_1 g_1 g_2, B_2 g_1 g_3$ and $B_3 g_3$ are sufficient causes for disease.

Example

example

$$Y_{g_1 g_2 g_3} = B_1 g_1 g_2 \bigvee B_2 g_1 g_3 \bigvee B_3 g_3$$

- There will be individuals (those with $B_1 = 1, B_2 = B_3 = 0$) for whom the disease can only occur if both genotypes $g_1 = 1$ and $g_2 = 1$ are 'present'.
- We then say that sufficient cause interaction is present between G_1 and G_2 because both genotypes are present in the same sufficient cause.

Subtleties

Generally, more than one set of sufficient causes will replicate a set of potential outcomes.

example

- Suppose the outcome is death and we are interested in the effects of 2 life-threatening 'killer' genes G_1 and G_2.
- Suppose that $Y_{g_1 g_2} = 1$ if and only if $G_1 = 1$ or $G_2 = 1$.
- Then both sufficient cause representations

$$Y_{g_1 g_2} = B_1 g_1 \bigvee B_2 g_2$$

and

$$Y_{g_1 g_2} = B_1 g_1 \overline{g}_2 \bigvee B_2 \overline{g}_1 g_2 \bigvee B_3 g_1 g_2$$

are compatible with this.

Sufficient cause interaction

- We will say that there is a sufficient cause interaction between G_1 and G_2 if in every sufficient cause representation for Y, there is a sufficient cause in which G_1 and G_2 are both present.
- If there is a sufficient cause interaction there must be a causal mechanism which requires both G_1 and G_2 to operate.
- The previous example suggests that no consistent tests for sufficient cause interaction can be developed.
- By using the relation between sufficient cause interactions and counterfactual outcomes, empirical tests for sufficient cause interactions can however be constructed.

Empirical conditions for sufficient cause interaction

- Define $p_{ge|x} = E(Y|G = g, E = e, X = x)$.
- A sufficient cause interaction is present if

(VanderWeele et al., 2007)

$$p_{11|x} - p_{10|x} - p_{01|x} > 0 \quad \text{or} \quad \text{RERI} > 1$$

- When the additive risk model:

$$E(Y|G, E, X) = \gamma_0 + \gamma_1 G + \gamma_2 E + \beta GE + \gamma_x X$$

holds, where X suffices to control for confounding of the effect of G and E on Y, then this amounts to

$$\beta > \gamma_0$$

Empirical conditions for sufficient cause interaction

- The previous result can be strengthened under monotonicity.
- Monotonicity requires the effect always operates in the same direction for all individuals.
- This might be plausible sometimes (e.g. the effect of smoking on lung cancer), but not always (e.g. alcohol on stroke).
- When both exposures have monotonic effects on the outcome, a sufficient cause interaction is present if

  (VanderWeele et al., 2007)

$$p_{11|x} - p_{10|x} - p_{01|x} + p_{00|x} > 0 \quad \text{or} \quad \text{RERI} > 0$$

which amounts to

$$\beta > 0$$

Example: breast cancer

- Odds ratios for breast cancer:

	No Alcohol	Alcohol
T/T or T/M	1	1.12 (0.81 to 1.54)
M/M	1.21 (0.70 to 2.09)	2.09 (1.16 to 3.78)

(Figueiredo et al., 2004)

- Here

$$\mathrm{RERI} \approx 2.09 - 1.21 - 1.12 + 1.00 = 0.76 > 0$$

- Ignoring sampling variability,
  this suggests evidence for sufficient cause interaction with the assumption that both alcohol and the M/M polymorphism have monotonic effects on the outcome.

Example: diarrheal disease

- Case-control study in Northwestern Ecuador (2003-2008) indicates that Giardia, rotavirus and E. coli/Shigella, are all associated with increased risk of diarrheal disease.

  (Bhavnani et al., 2012)

- Giardia rotavirus: $\text{RERI} = 10.7 - 2.6 - 1.1 + 1 = 7.9$ (95% CI: 3.1 to 18.9)
- E. coli/Shigella rotavirus: $\text{RERI} = 13.2 - 2.6 - 1.6 + 1 = 9.9$ (95% CI: 2.6 to 28.4)
- E. coli/Shigella giardia: $\text{RERI} = 3.0 - 1.1 - 1.6 + 1 = 1.2$ (95% CI: -1.4, 3.1)
- For Giardia rotavirus and for E. coli/Shigella and rotavirus, there is strong evidence of mechanistic interaction even without making any monotonicity assumptions.

  (VanderWeele, 2012)

Epistasis

- The original notion of epistasis is similar.

  (Bateson, 1909; Cordell, 2002; VanderWeele, 2010)
- For two genetic factors G_1 and G_2,
  we say that G_1 is epistatic to G_2 if it masks the effect of G_2.
- That is, there are individuals who can only get disease
  if both 'exposures' are 'present'.
- VanderWeele (2010) refers to this as compositional epistasis.

Empirical conditions for compositional epistasis

- Compositional epistasis is present if

  (VanderWeele et al., 2010)

$$p_{11|x} - p_{10|x} - p_{01|x} - p_{00|x} > 0 \quad \text{or} \quad \text{RERI} > 2$$

- Under the additive risk model:

$$E(Y|G, E, X) = \gamma_0 + \gamma_1 G + \gamma_2 E + \beta GE + \gamma_x X$$

  this amounts to

$$\beta > 2\gamma_0$$

Empirical conditions for compositional epistasis

- When one of the exposures is monotonic, this can be strengthened to

$$p_{11|x} - p_{10|x} - p_{01|x} > 0 \quad \text{or} \quad \text{RERI} > 1$$

which amounts to

$$\beta > \gamma_0$$

- When both exposures are monotonic, this can be strengthened to

$$p_{11|x} - p_{10|x} - p_{01|x} + p_{00|x} > 0 \quad \text{or} \quad \text{RERI} > 0$$

which amounts to

$$\beta > 0$$

Example: esophageal cancer

- Yang et al. (2005) examine interaction in the effects of Arg variants on ADH2 (chr 4) and Glu/Glu versus Glu/Lys on ALDH2 (chr 12) on esophageal cancer.
- Based on the odds ratios (OR), we find
  RERI = OR11 - OR10 - OR01 + 1
  $= 7.20 - 1.40 - 3.52 + 1 = 3.28$ (95% CI: 0.4 to 6.16)
- The estimate RERI $= 3.28 > 2$ would suggest compositional epistasis without any assumptions about monotonicity.
- The confidence interval implies compositional epistasis only if both variants had monotonic effects on esophageal cancer.

Part III. Robustness

Model misspecification

- Misspecification of the main effects may seriously bias standard interaction estimates.

  (Greenland, 1993)

- This can form a major concern,
  especially in genome-wide screening
  and because of the possibility of high-dimensional confounders.

Model misspecification

- The concern about misspecification gets exacerbated when there are strong correlations between covariates.
  - Regression methods are then prone to extrapolate.
  - Important confounders may be dismissed in the model building process.
- Standard model building procedures may prioritize interactions over other higher order terms, thus leading to an inflation of the Type I error rate.
- The concern about model misspecification is especially pertinent when studying additive interaction or sufficient cause interaction.

Concerns about tests for additive interaction

- Conditions like

$$p_{11|x} - p_{10|x} - p_{01|x} > 0$$

  could alternatively be tested on the basis of contrasts between predicted values $p_{ge|x}$ from logistic regression models.
- However, such models are highly non-additive.
- By thus imposing non-additivity under parsimonious, but misspecified logistic regression models, one would induce a bias towards non-additivity on the risk difference scale, leading to an inflation of the Type I error.

  (Vansteelandt, VanderWeele and Robins, 2012)
- In addition,
  the contrast $p_{11|x} - p_{10|x} - p_{01|x}$ would then depend on x, thus requiring separate tests at each confounder level.

  (Skrondal, 2003)

Robust assessment of interactions

question

Can we estimate/test for statistical interaction
in a way that is robust against model misspecification?

We will consider 2 approaches:

- marginal structural models;

  (VanderWeele, Vansteelandt and Robins, 2010; Vansteelandt, VanderWeele and Robins, 2012)

- semi-parametric interaction models;

  (Vansteelandt et al., 2008)

Marginal structural additive risk models

- One strategy is to use marginal structural additive risk models

$$p_{ge} \equiv P(Y_{ge} = 1) = \beta_0 + \beta_1 g + \beta_2 e + \beta_3 ge$$

- These models are saturated (when G and E are dichotomous) and thus guaranteed not to be misspecified:

$$\begin{aligned} p_{00} &= \beta_0 \\ p_{01} &= \beta_0 + \beta_2 \\ p_{10} &= \beta_0 + \beta_1 \\ p_{11} &= \beta_0 + \beta_1 + \beta_2 + \beta_3 \end{aligned}$$

Empirical conditions for marginal synergism

- The condition

$$p_{11} - p_{10} - p_{01} > 0$$

amounts to

$$\beta_3 > \beta_0;$$

it entails the presence of a sufficient cause interaction between G and E.

- When the effects of G and E are monotonic, the condition

$$p_{11} - p_{10} - p_{01} + p_{00} > 0$$

amounts to

$$\beta_3 > 0;$$

it entails the presence of a sufficient cause interaction between G and E.

Fitting marginal structural additive risk models

- The marginal structural additive risk model

$$P(Y_{ge} = 1) = \beta_0 + \beta_1 g + \beta_2 e + \beta_3 ge$$

  cannot be directly fitted
  because the counterfactuals Y_{ge} are not all observed.
- It can instead be fitted via weighted least squares regression of the corresponding additive risk model

$$P(Y = 1|G, E) = \beta_0 + \beta_1 G + \beta_2 E + \beta_3 GE$$

  with weights

$$\frac{P(G)}{P(G|X)} \frac{P(E|G)}{P(E|G, X)}$$

  whose components can be obtained via logistic regression.

  (VanderWeele, Vansteelandt and Robins, 2010)
- Related approach for the RERI is based on marginal structural linear odds models.

  (VanderWeele and Vansteelandt, 2011)

Application

- We apply the methods described above to data from a cross-sectional study of 11 062 individuals in Bangladesh.
- The study concerns interactions between the effects of exposure to high levels of arsenic in drinking water ($> 100\mu$g/L in well water) and current or past tobacco smoking on premalignant skin lesions.
- These were defined as the presence of melanosis or hyperkeratosis, which are precursor lesions of basal and squamous cell skin cancers in an arsenic-exposed population.

Weights

- Weights are constructed using logistic regressions of high levels of arsenic exposure (G) and smoking (E).
- Confounding variables (X) are sex, age, education, body mass index, land and TV ownership (markers of socioeconomic status in Bangladesh), fertilizer use, and pesticide use.

1. We thus fit a logistic model for $P(G)$ and $P(G|X)$.
   - Subjects with $G = 1$ receive weight

$$w_1 = \frac{P(G = 1)}{P(G = 1|X)}$$

   - Subjects with $G = 0$ receive weight

$$w_1 = \frac{1 - P(G = 1)}{1 - P(G = 1|X)}$$

Fitting marginal structural additive risk models

2. Next, we fit a logistic model for $P(E|G)$ and $P(E|G,X)$.
   - Subjects with $E = 1$ receive weight

$$w_2 = \frac{P(E=1|G)}{P(E=1|G,X)}$$

   - Subjects with $E = 0$ receive weight

$$w_2 = \frac{1-P(E=1|G)}{1-P(E=1|G,X)}$$

3. We then constructed a weight for each subject as

$$w_1 w_2$$

Following suggestions from Cole and Hernán (2008), weights are truncated at the 1st and 99th percentiles.

Fitting marginal structural additive risk models

4. Next we fit the additive risk model

$$P(Y = 1|G, E) = \beta_0 + \beta_1 G + \beta_2 E + \beta_3 GE$$

using standard linear regression methods,
giving each subject the constructed weight.

- Standard error for the estimate requires a sandwich estimator, e.g. via GEE-routines.
- Standard error outputed by such routines are conservative.
- Estimating an 'unbiased' standard error for the estimate is more tricky, as this should include the variability in the weights. One option is to use the bootstrap.

Results

- The estimates (reported in %) from the marginal structural model are

$$\begin{aligned}
\beta_0 &= 4.72 \text{ (95\% confidence interval: } 3.40, 6.04) \\
\beta_1 &= 3.03 \text{ (95\% confidence interval: } 0.66, 5.40) \\
\beta_2 &= 1.72 \text{ (95\% confidence interval: } -0.13, 3.57) \\
\beta_3 &= 3.64 \text{ (95\% confidence interval: } 0.13, 7.14)
\end{aligned}$$

- Thus, under the assumption of monotonicity - which is biologically plausible - and of no unmeasured confounding, the estimate suggests a sufficient cause interaction.
- It confirms the presence of individuals who would have a skin lesion if they were exposed to high levels of well arsenic and tobacco smoking, but would not have had a skin lesion if only one of these 2 exposures were present.

Results

- The estimates (reported in %) from the marginal structural model are

$$\begin{aligned}
\beta_0 &= 4.72 \text{ (95\% confidence interval: } 3.40, 6.04) \\
\beta_1 &= 3.03 \text{ (95\% confidence interval: } 0.66, 5.40) \\
\beta_2 &= 1.72 \text{ (95\% confidence interval: } -0.13, 3.57) \\
\beta_3 &= 3.64 \text{ (95\% confidence interval: } 0.13, 7.14)
\end{aligned}$$

- An estimate of the lower bound on the prevalence of such sufficient cause interactions is 3.64%.
- But the confidence interval is quite wide: 0.13, 7.14.
- Because $\beta_3 < \beta_0$, it is not possible to draw conclusions about sufficient cause interactions without the monotonicity assumption.

Semi-parametric interaction models

- An alternative strategy is to estimate statistical interactions without modeling main effects

$$E(Y|G, E, X) = \beta GE + ???$$

- We call this a semi-parametric interaction model.

  (Vansteelandt et al., 2008; Vansteelandt, VanderWeele and Robins, 2012)

- Smoothing methods would not work well when X is high-dimensional.
- Progress by using information on the joint distribution of the exposures conditional on the extraneous factors.

A priori knowledge on the exposure distribution

Such information is sometimes available.

example: randomized clinical trial

experiment with G and E both randomly assigned.

example: gene-gene interaction in family-based genetic association studies

genotype distribution is known, conditional on parental genotypes, by Mendel's law of segregation.

Estimation

- Let us start from a traditional regression model

$$E(Y|G,E,X) = \beta GE + \gamma_2 E + \gamma_1 G + \gamma_x X + \gamma_0$$

- Then we propose estimating β as the solution to an estimating equation of the form

$$\sum_{i=1}^{n} d(G_i, E_i, X_i)\epsilon_i(\beta,\gamma) = 0$$

with $d(G,E,X)$ a function of (G,E,X) satisfying

$$E\{d(G,E,X)|G,X\} = E\{d(G,E,X)|E,X\} = 0$$

and γ substituted with e.g. the OLS estimator.

Why this constraint?

- The constraint

$$E\{d(G,E,X)|G,X\} = E\{d(G,E,X)|E,X\} = 0$$

ensures that misspecification of the main effects does not harmfully affect the estimator, which is the solution to

$$\begin{aligned} 0 \quad = \quad & \sum_{i=1}^{n} d(G_i,E_i,X_i)(Y_i - \beta G_i E_i - \gamma_2 E_i \\ & -\gamma_1 G_i - \gamma_x X_i - \gamma_0) \end{aligned}$$

- How do we find a function $d(G,E,X)$ satisfying

$$E\{d(G,E,X)|G,X\} = E\{d(G,E,X)|E,X\} = 0?$$

Conditionally independent exposures

- Suppose first that both exposures are known to be conditionally independent:

$$G \perp\!\!\!\perp E | X$$

example: gene-environment interaction

gene and environment are independent when the environmental exposure is not genetically affected.

example: gene-gene interaction

unlinked genes can be assumed independent (conditional on parental mating types).

See Vansteelandt et al. (2008) for conditionally dependent exposures.

Conditionally independent exposures

- Then the constraint is met for choices of the form

$$d(G_i, E_i, X_i) = d(X_i)\Delta(G_i|X_i)\Delta(E_i|X_i)$$

with $d(X_i)$ arbitrary and

$$\Delta(G_i|X_i)\Delta(E_i|X_i) = \{G_i - E(G_i|X_i)\}\{E_i - E(E_i|X_i)\}$$

- The closed-form estimator $\hat{\beta}$ for β:

$$\frac{\sum_{i=1}^{n} \Delta(G_i|X_i)\Delta(E_i|X_i)(Y_i - \gamma_2 E_i - \gamma_1 G_i - \gamma_x X_i - \gamma_0)}{\sum_{i=1}^{n} \Delta(G_i|X_i)\Delta(E_i|X_i) G_i E_i}$$

is semi-parametric efficient.

Robustness

- When the exposure distribution is unknown, we specify a model and fit it using maximum likelihood estimation.
- Then our approach delivers multiply robust estimators.

Test and standard error

- When the models for $E(G_i|X_i)$ and $E(E_i|X_i)$ are correct, a conservative test of the *GxE* interaction null hypothesis amounts to a one-sample t-test that

$$\Delta(G_i|X_i)\Delta(E_i|X_i)(Y_i - \gamma_2 E_i - \gamma_1 G_i - \gamma_x X_i - \gamma_0)$$

  has mean zero.

- A conservative standard error of $\hat{\beta}$ is given by 1 over root n times the standard deviation of

$$\frac{\Delta(G_i|X_i)\Delta(E_i|X_i)(Y_i - \gamma_2 E_i - \gamma_1 G_i - \gamma_x X_i - \gamma_0)}{n^{-1}\sum_{i=1}^{n}\Delta(G_i|X_i)\Delta(E_i|X_i)G_i E_i}$$

- Vansteelandt et al. (2008) give more exact results.

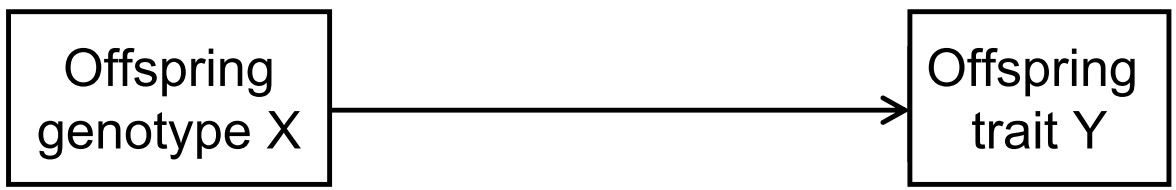
Boston Early-Onset COPD Study

- 128 extended pedigrees
  from the Boston Early-Onset COPD Study.

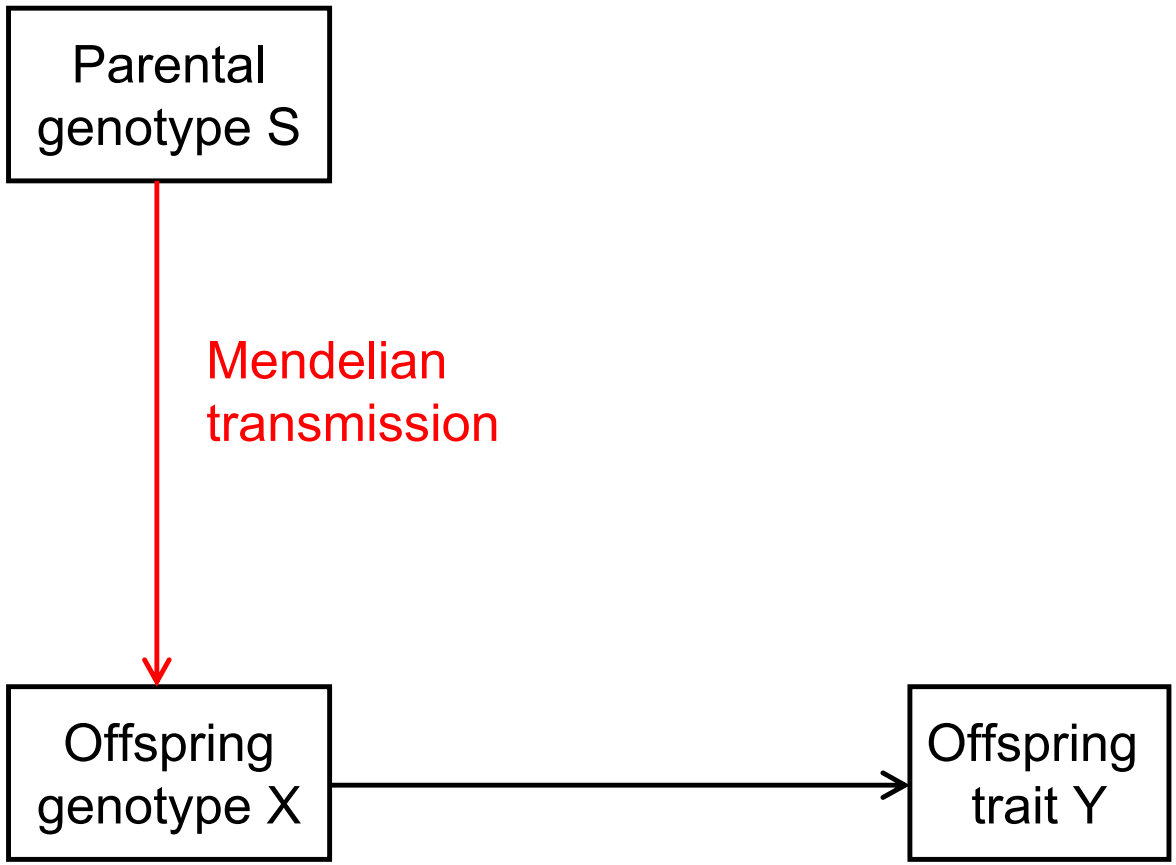
  (Silverman et al., 1998)
- We test 6 SNPs located in the SERPINE2 gene found to be in a linkage peak for COPD.

  (DeMeo et al., 2006)
- Interest in interactions of these SNPs X with pack years of smoking Z on post-bronchodilator measurements of forced expiratory volume in 1 second (FEV_1 in liters) Y.
- Note the change of notation!

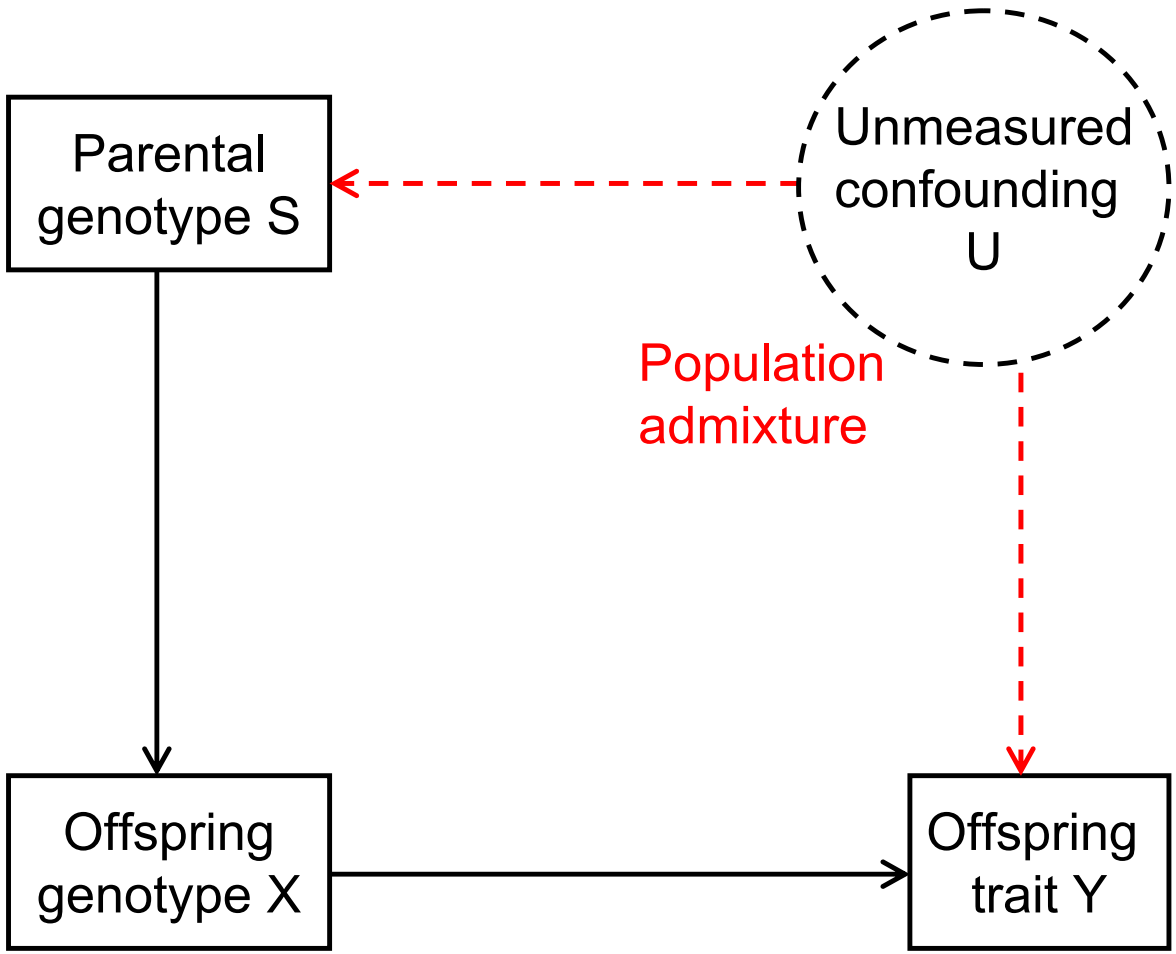
The target genetic effect



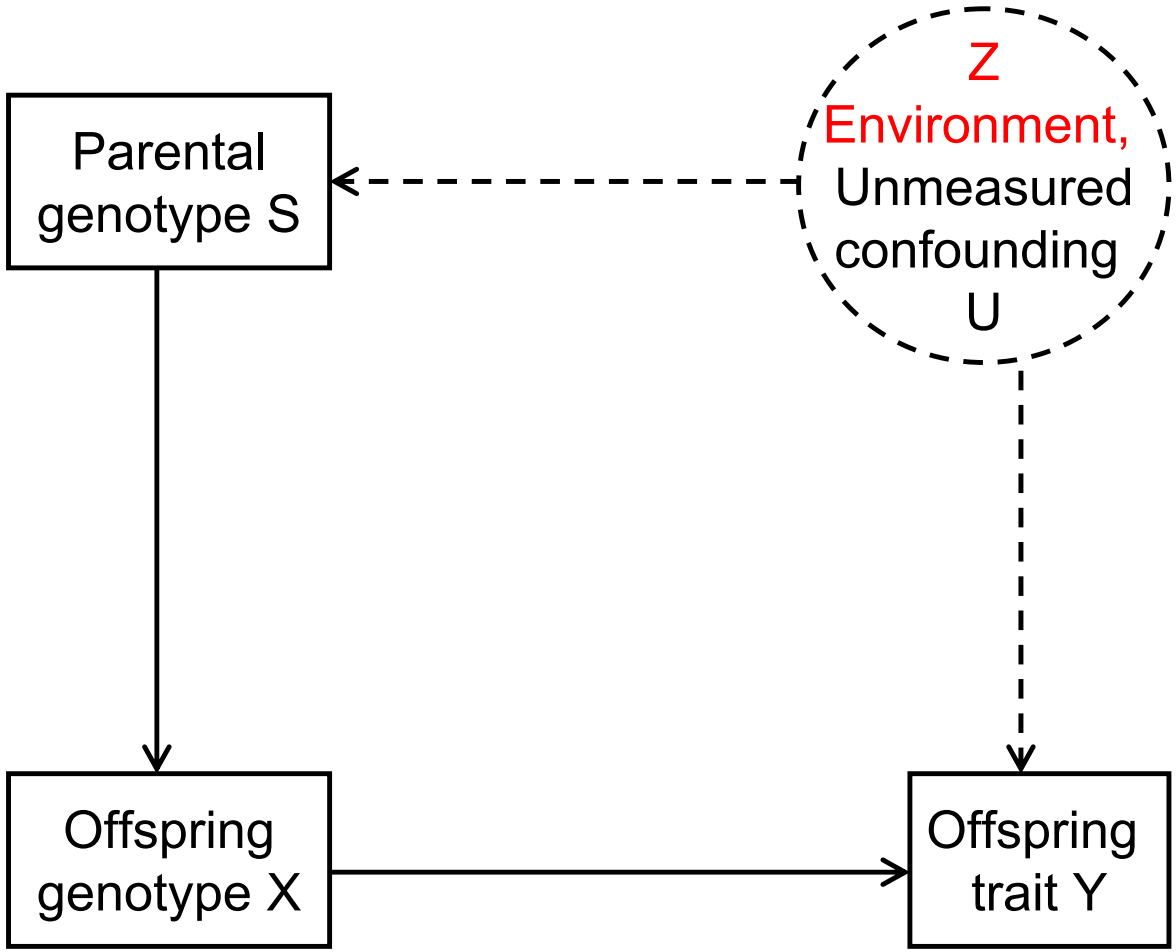
Mendelian transmission



Population admixture



Stratifying on parental mating types removes bias



Model

- The QBAT-I test essentially follows the previous principle.

  (Vansteelandt et al., 2008)

- It uses estimating equations of the form

$$\sum_{i,j} \begin{pmatrix} 1 \\ Z_{ij} \end{pmatrix} \{X_{ij} - \underbrace{E(X_{ij}|Z_{ij}, S_i)}_{E(X_{ij}|S_i)}\}(Y_{ij} - \beta X_{ij} - \gamma X_{ij} Z_{ij}) = 0.$$

Results of Boston Early-Onset COPD Study

- Each family was ascertained from a single proband satisfying $FEV_1 < 40\%$ of predicted.
- S_{GxE}-p extends this idea to adopt the ascertainment condition.

(Fardo et al., 2012)

	Pack Years		Ever Smoker	
Markers	S_{GxE}-p	QBAT-I-p	S_{GxE}-p	QBAT-I-p
ser37	0.118	0.126	0.152	0.163
ser8	0.662	0.243	0.951	0.958
ser51	**0.047**	**0.082**	**0.455**	**0.816**
ser55	0.209	0.244	0.703	0.774
ser50	0.213	0.271	0.745	0.825
ser6	0.232	0.249	0.870	0.941

Approach stays valid when parental genotypes are incomplete and S is replaced by their sufficient statistic (Rabinowitz and Laird, 2000).

References

Cordell, H.J. (2002) Epistasis: what it means, what it doesnÕt mean, and statistical methods to detect it in humans. Human Molecular Genetics, 11:2463-2468.

Fardo, D.W., Liu, J., Demeo, D.L., Silverman, E. and Vansteelandt, S. (2011). Gene-environment interaction testing in family-based association studies with phenotypically ascertained samples: A causal inference approach. Biostatistics, 13, 468-481.

Figueiredo JC, Knight JA, Briollais L, Andrulis IL, Ozcelik H. (2004). Polymorphisms XRCC1-R399Q and XRCC3-T241M and the risk of breast cancer at the Ontario Site of the Breast Cancer Family Registry. Cancer Epidemiology, Biomarkers and Prevention 13:583-591.

Hunter DJ. (2005). Gene-environment interactions in human diseases. Nature Reviews Genetics, 6:287-298.

Knol MJ, Egger M, Scott P, Geerlings MI, Vandenbroucke JP. When One Depends on the Other: Reporting of Interaction in Case-Control and Cohort Studies. Epidemiology. 2009; 20:161-166.

Knol, M.J. and VanderWeele, T.J. (2012). Guidelines for presenting analyses of effect modification and interaction. International Journal of Epidemiology, 41:514-520.

Moerkerke, B., Vansteelandt, S. and Lange, C. (2010). A doubly-robust test for gene-environment interaction in family-based studies of affected offspring. Biostatistics, 11, 213-225.

Rothman KJ. (1976). Causes. Am J of Epidemiol 104:587-592.

Siemiatycki J, Thomas DC (1981). Biological models and statistical interactions: an example from multistage carcinogenesis. Int. J. Epidemiol. 10:383-387.

VanderWeele, T.J. (2009). On the distinction between interaction and effect modification. Epidemiology, 20:863-871.

VanderWeele, T.J. and Knol, M.J. (2011). The interpretation of subgroup analyses in randomized trials: heterogeneity versus secondary interventions. Annals of Internal Medicine, 154:680-683.

References

VanderWeele TJ, Robins JM. (2007) The identification of synergism in the SCC framework. Epidemiol, 18:329-339.

VanderWeele, T.J. and Robins, J.M. (2007). Four types of effect modification Ð a classification based on directed acyclic graphs. Epidemiology 18:561-568.

VanderWeele, T.J. and Robins, J.M. (2008). Empirical and counterfactual conditions for sufficient cause interactions. Biometrika, 95:49-61.

VanderWeele, T.J. (2009). Sufficient cause interactions and statistical interactions. Epidemiology, 20:6-13.

VanderWeele, T.J. (2010). Empirical tests for compositional epistasis. Nature Reviews Genetics, 11:166.

VanderWeele, T.J., Vansteelandt, S. and Robins, J.M. (2010). Marginal structural models for sufficient cause interactions. American Journal of Epidemiology, 171:506-514.

VanderWeele, T.J. and Vansteelandt, S. (2011). A weighting approach to causal effects and additive interaction in case-control studies: marginal structural linear odds models. American Journal of Epidemiology, 174, 1197-1203.

Vansteelandt, S., VanderWeele, T.J., Tchetgen, E.J., Robins, J.M., (2008). Multiply robust inference for statistical interactions. Journal of the American Statistical Association, 103:1693-1704.

Vansteelandt, S., DeMeo, D., Su, J., Smoller, J., Murphy, A.J., McQueen, M., Celedon, J., Weiss, S.T., Silverman, E.K. and Lange, C. (2008). Testing and estimating gene-environment interactions in family-based association studies. Biometrics, 64, 458-467.

Vansteelandt, S., VanderWeele, T. and Robins, J.M. (2012). Semiparametric tests for sufficient cause interaction. Journal of the Royal Statistical Society - Series B, 74, 223-244.

Vansteelandt, S. and Lange, C. (2012). Causation and Causal Inference for Genetic Effects. Human Genetics, 131, 1665-1676.