AUEB-NKUA-Indiana Conference in Biostatistics & Health Analytics

Multivariable Analysis: Confounding or Interaction? How Much Complexity?

Giota Touloumi and Nikos Pantazis

Department of Hygiene, Epidemiology and Medical Statistics, Medical School, National and Kapodistrian University of Athens, Athens, Greece

The components of GLMs

- **The random part:** the distribution the components of Y have
- **The systematic component:** covariates $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$

  produce a **linear predictor η** given by $$\eta = \sum_{j=1}^{p} x_{ij}\beta_j$$
- The **link** between the random and the systematic components: **g(μ)= η**

Normal	**$\eta = \mu$**	***Identity***
Poisson	**$\eta = \log(\mu) = \log(\lambda)$**	***Log of rate***
Binomial	**$\eta = \log[\pi/(1-\pi)]$**	***Logit***

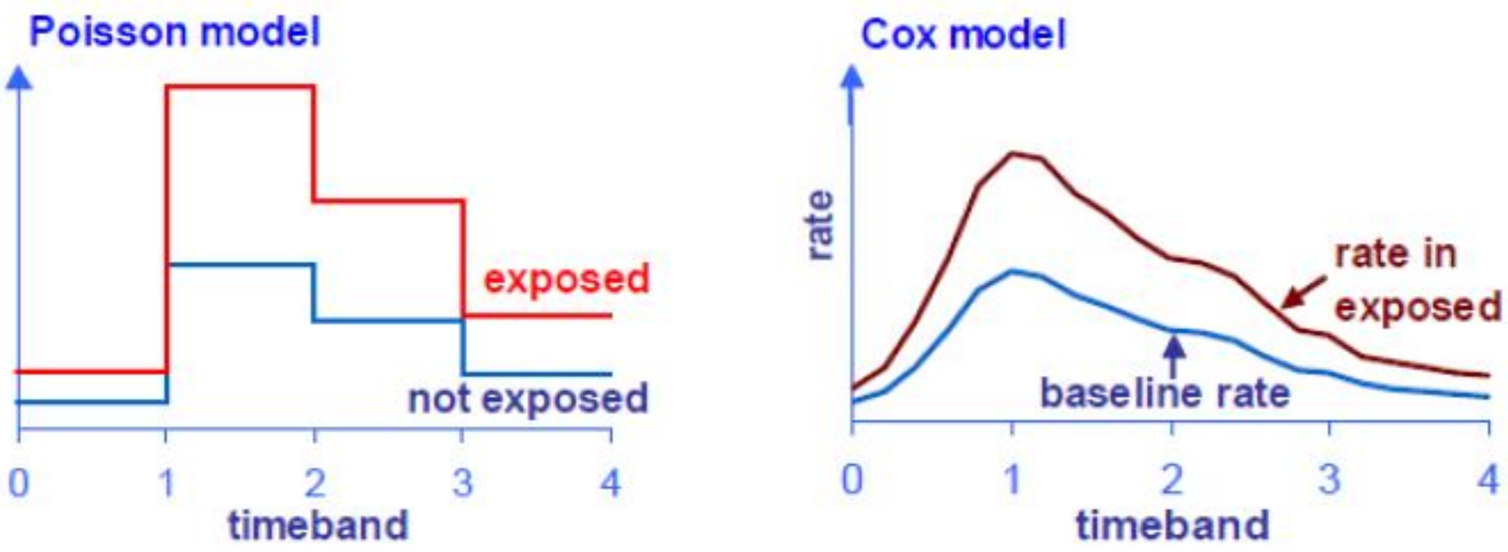
Multivariable Models

$$g(\mu) = b_0 + b_1X_1 + b_2X_2 + \dots b_pX_p = \sum_0^p b_iX_i$$

- g(μ)=μ linear regression
  - Mean change in response variable
- g(μ)=log(odds)=log(π/1-π) Logistic regression
  - Probability of an event (odds ratio)
- g(μ)=log(μ)=log(rate) Poisson regression
  - rate of a new event (rate ratio)

- $\ln(h_t) = \ln(h_{0\,(t)}) + \sum_1^p b_iX_i$ Cox proportional hazards model
  - Time to event (Hazard ratio)

Multivariable Models: Poisson vs Cox model



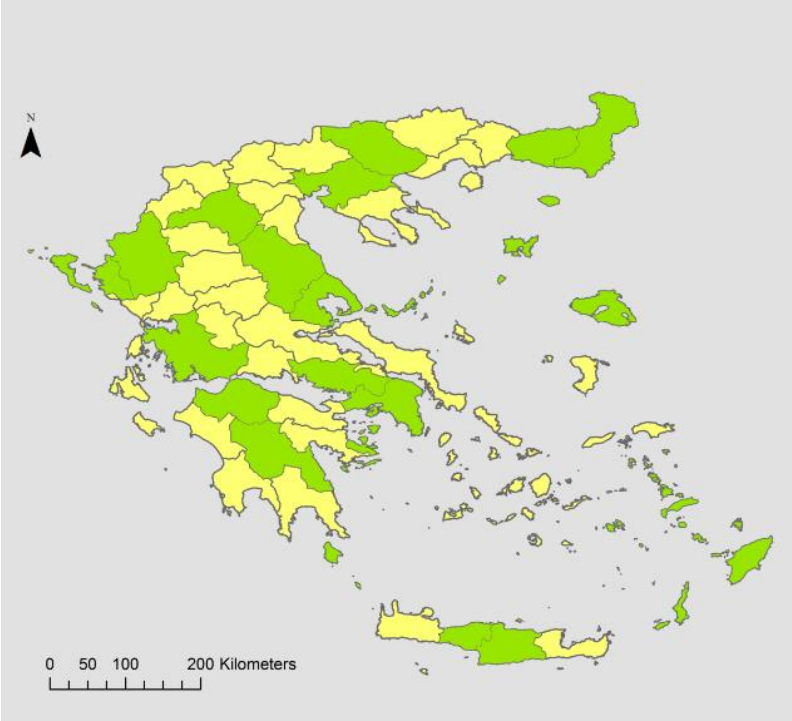
Multivariable Models: How much complexity?

- **Linearity**
  - How to check, how to present/interpret
- **Confounding**
  - What is it? How we control for?
- **Interactions**
  - What is it? How we present/interpret results?
  - Confounding or Interactions?

National Survey of Morbidity and Risk Factors (EMENO)

Εθνική ΜΕλέτη ΝΟσηρότητας και Παραγόντων Κινδύνου
E.ME.NO
Έρευνα Υγείας

Stage1:Prefectures/urbanization



Stage 2: **City blocks**

Stage 3: **Households**

Stage 4: **Individual** (with most recent birth date)

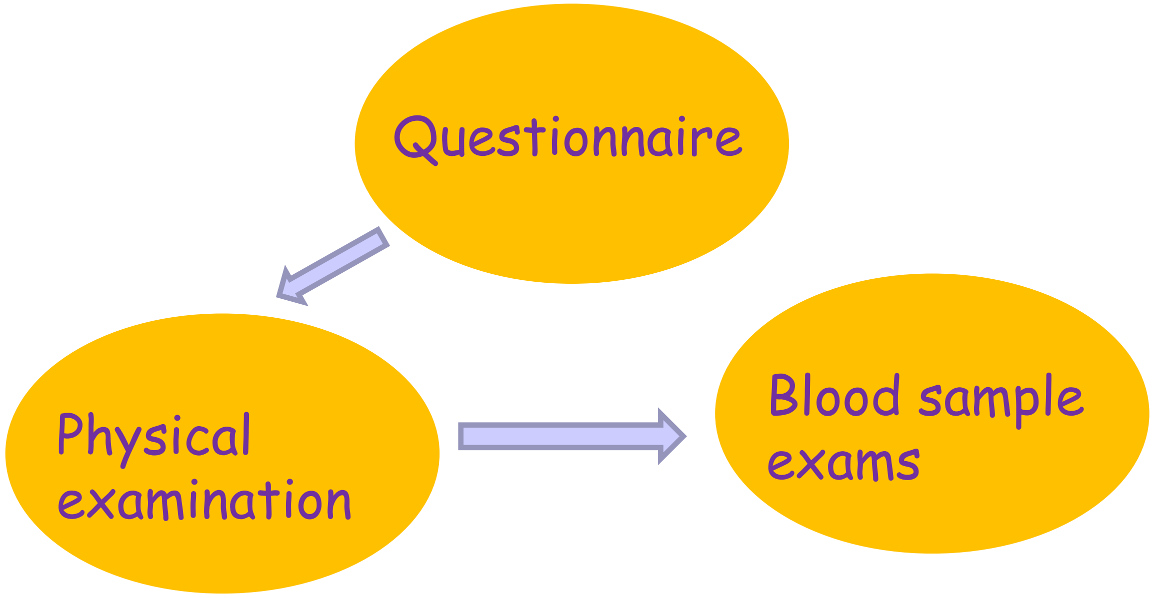
Sample selection: Multistage Stratified Random sampling

- Interviews "door-to-door"

6006 participants

Ιατρική Σχολή Πανεπιστήμιου Αθηνών
Εργ. Υγιεινής, Επιδημιολογίας & Ιατρικής Στατιστικής

EMENO: Health Examination Survey



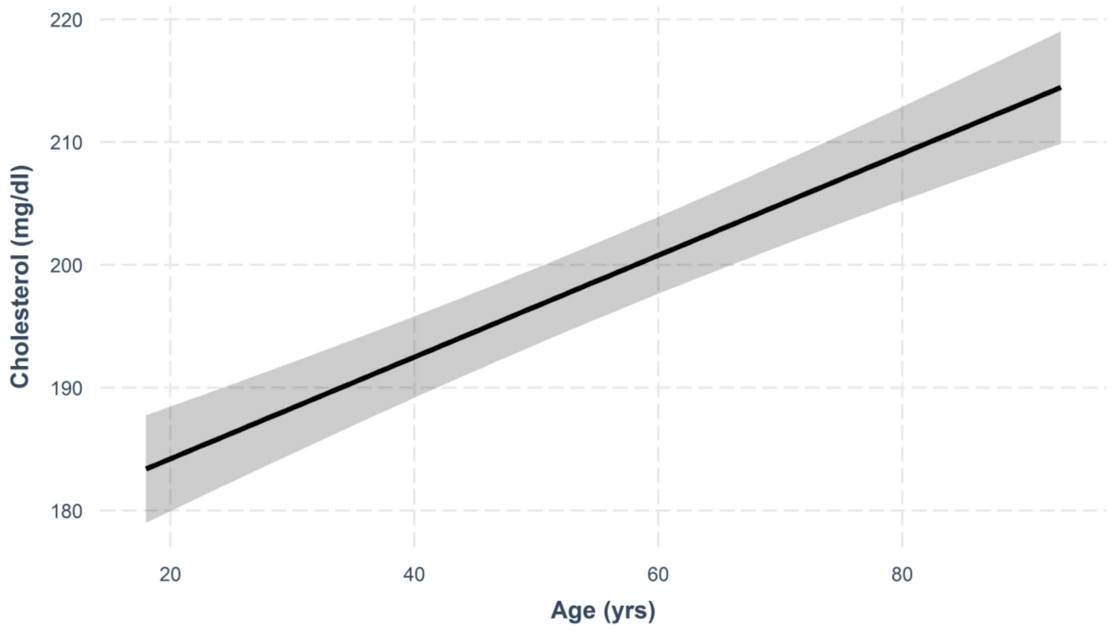
Height
Weight
Blood pressure
Total, HDL, LDL Cholesterol
Lipids
Glucose

Linearity $g(\mu) = b_0 + b_1 X_1 + b_2 X_2 + \ \dots\ b_p X_p = \sum_0^p b_i X_i$

$$\mathrm{E(Total\ Chol)} = b_0 + b_1 Age + b_2 Female + b_3 Diabetes + b_4 Alcoho + b_5 > 30Walking$$

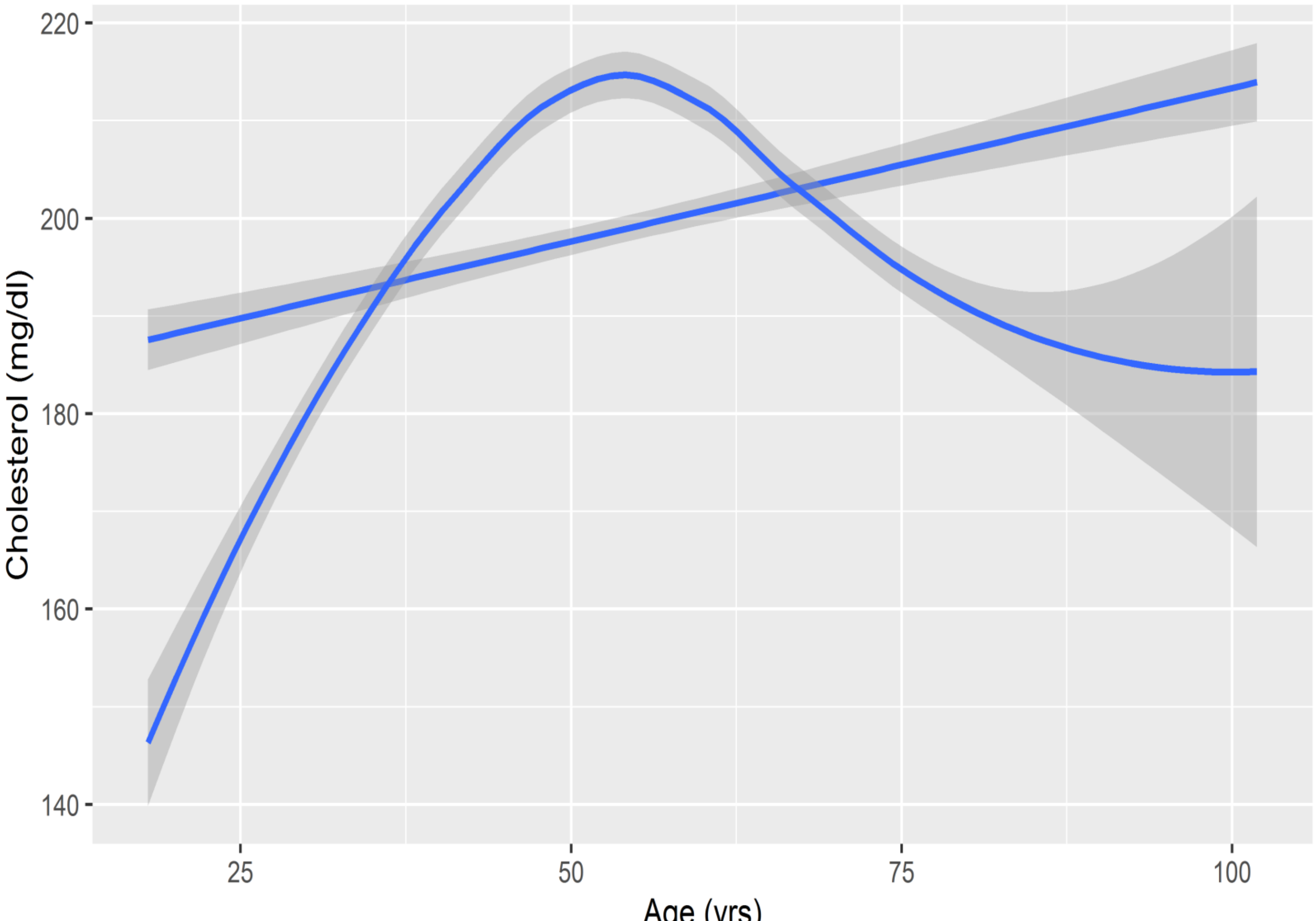
Characteristic	Beta	95% CI	p-value
Age (yrs)	0.414	0.328, 0.501	<0.001
sex			
Male			
Female	5.54	2.54, 8.54	<0.001
diab			
No			
Yes	-16.9	-21.2, -12.6	<0.001
Alcohol (categories)			
0-6			
7+	5.00	0.777, 9.23	0.020
Walking			
<30 min/day			
>=30 min/day	-2.81	-5.68, 0.069	0.056

Predicted cholesterol levels by age (cont. vars at mean, factors at base level)



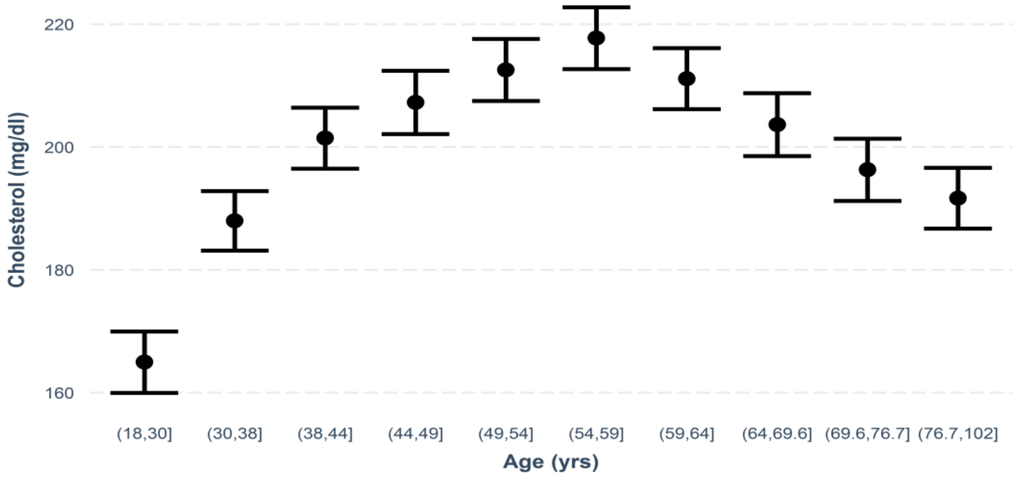
Cholesterol age relationship: initial investigation

Scatterplot with smooth (loess) line

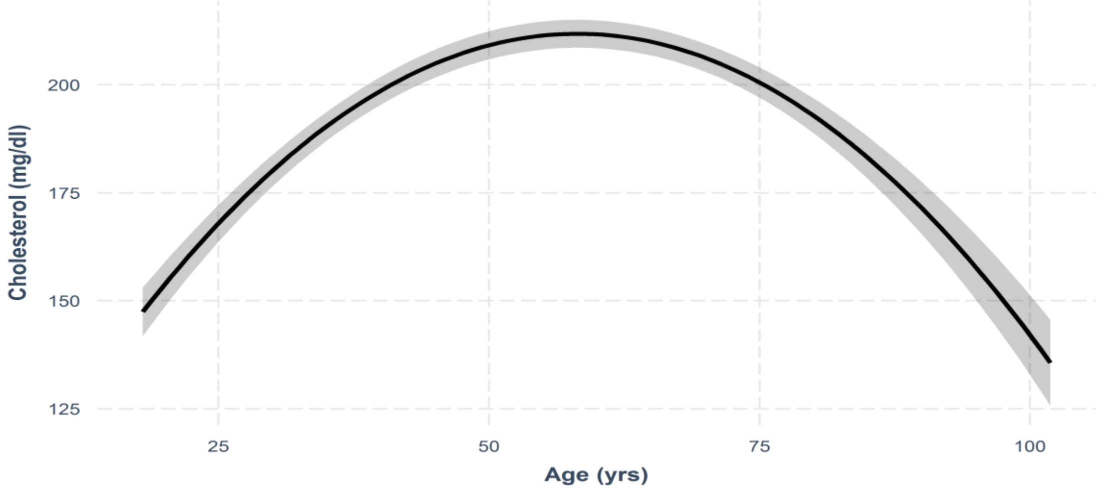


Cholesterol age relationship: model checking (1)

Characteristic	Beta	95% CI	p-value
Age (categories - yrs)			
(18,30]			
(30,38]	23.0	17.1, 28.9	<0.001
(38,44]	36.5	30.5, 42.4	<0.001
(44,49]	42.3	36.2, 48.4	<0.001
(49,54]	47.6	41.6, 53.5	<0.001
(54,59]	52.8	46.8, 58.7	<0.001
(59,64]	46.2	40.1, 52.2	<0.001
(64,69.6]	38.7	32.6, 44.7	<0.001
(69.6,76.7]	31.3	25.2, 37.4	<0.001
(76.7,102]	26.7	20.6, 32.8	<0.001

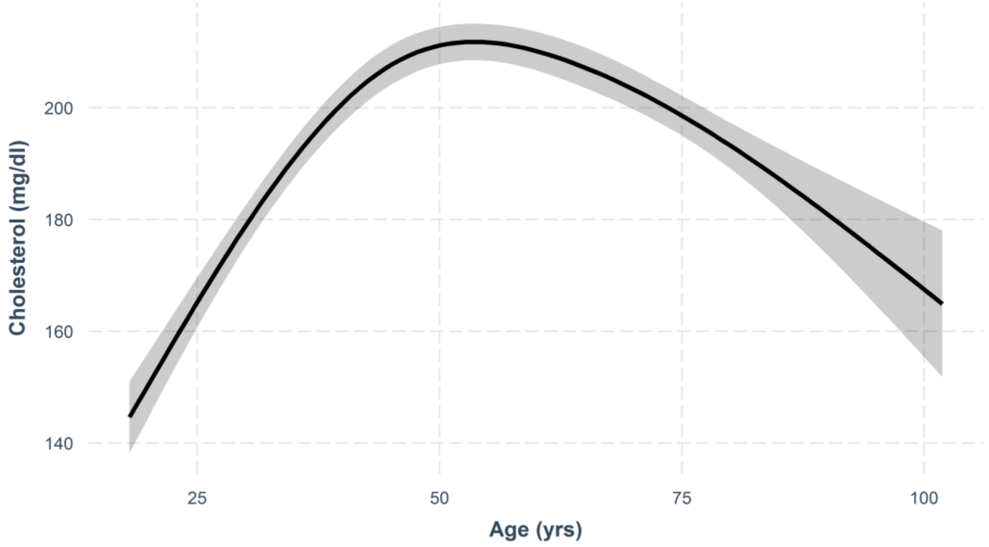


Characteristic	Beta	95% CI	p-value
Age			
Age	4.64	4.19, 5.09	<0.001
Age^2	-0.040	-0.044, -0.036	<0.001



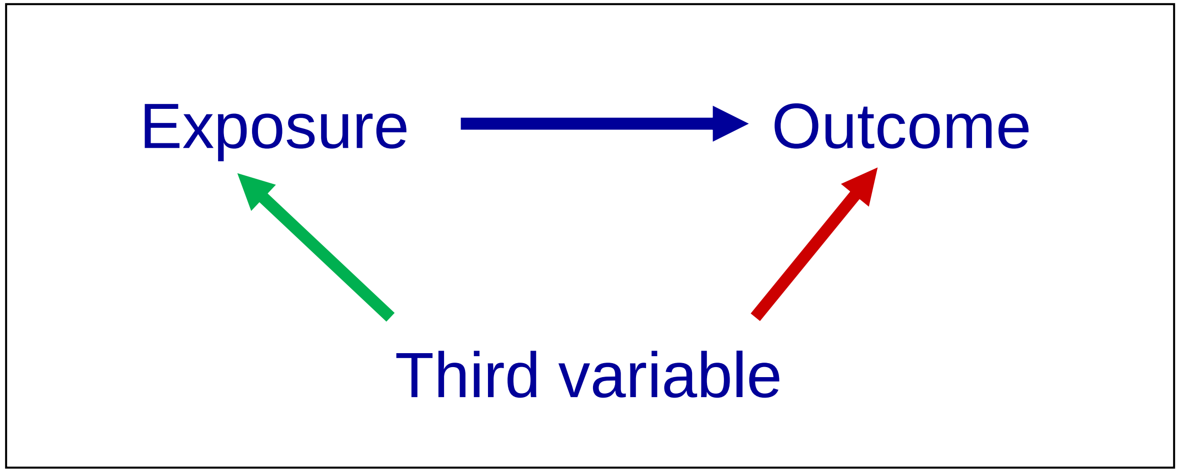
Cholesterol age relationship: model checking (2)

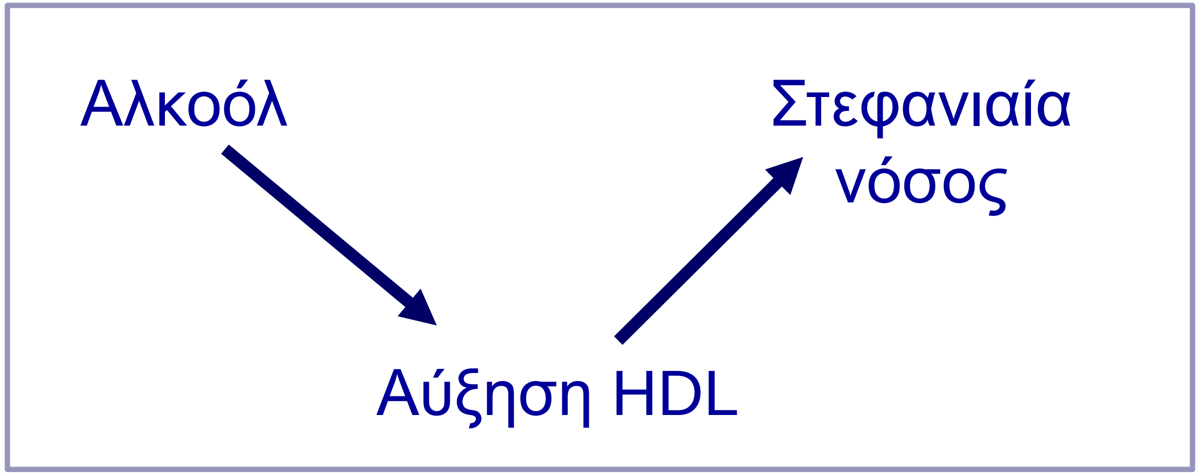
Characteristic	Beta	95% CI	p-value
Natural Spline of Age			
ns(Age, df = 3)1	38.9	32.3, 45.4	<0.001
ns(Age, df = 3)2	100	84.4, 116	<0.001
ns(Age, df = 3)3	-16.2	-29.1, -3.37	0.013



	df	AIC	Δ(AIC)	LogLikelihood
Linear	7	39341.32	351.94	-19663.64
Quadratic	8	39006.60	17.22	-19495.28
Spline	**9**	**38989.38**	**0.00**	**-19485.67**
Categorical	15	39001.58	12.20	-19485.73

Confounding

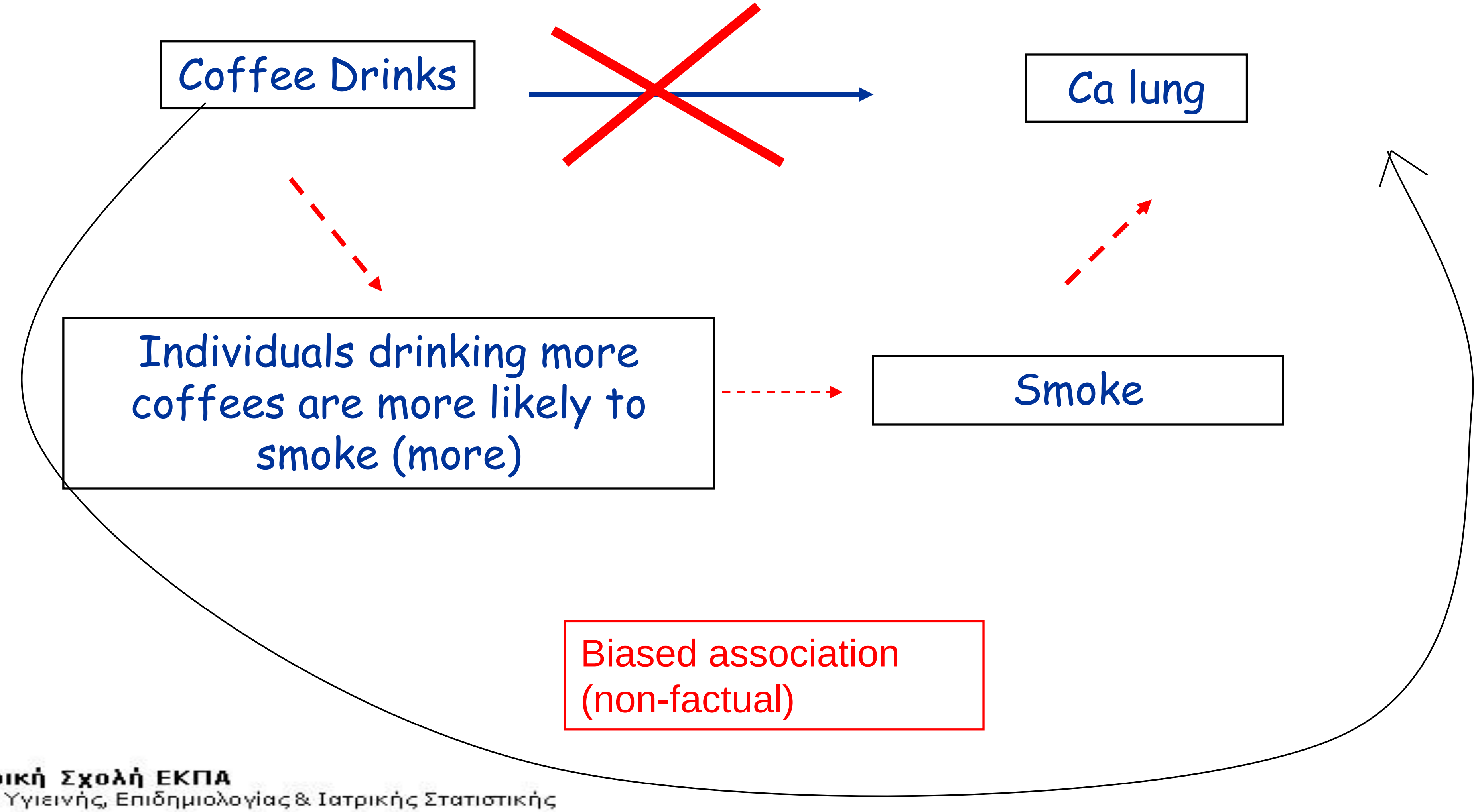




1. Prognostic factor of outcome
2. Associated with the third variable
3. It is not in the causal pathway

HDL in the causal pathway
HDL: Not a confounder;
HDL: No need for adjustment

Confounding: example



Direction of bias: Confounding

- Compare treatment A with treatment B (exposure)
- Event: 5-years survival
- Confounder: age (lower survival at older ages; treatment A tended to be given to older people)

True association:
A better than B

Trt A Mean age 60 years	Trt B Mean age 45 years
survival 70%	survival 70%

Conclusion:
A equivalent B
(false negative finding)

True association:
A equivalent to B

Trt A Mean age 60 years	Trt B Mean age 45 years
survival 50%	survival 70%

Conclusion :
B better than A
(false positive finding)

True association:
B better than A

Trt A Mean age 60 years	Trt B Mean age 45 years
survival 30%	survival 80%

Conclusion:
B better than A
(correct direction but underestimation of the effectiveness of B)

Channeling effect

Cohort studies:

Hormone Replacement Therapy

↓

Lower rates of Coronary Heart Disease

Association:

Lower rates of CHD in women who used HRT

Channeling effect

Randomized Controlled Trials:

Hormone Replacement Therapy

showed HIGHER rates of Cardiovascular Disease

Lower rates of Coronary Heart Disease

Channeling effect

Explanation:

Health-Conscious Behavior → Hormone Replacement Therapy

Health-Conscious Behavior → **Lower** rates of Cardiovascular Disease

Women who were more health-conscious tended to use HRT
&
tended to have lower rates of cardiovascular disease

Channeling effect

Explanation:

Hormone Replacement Therapy

Health-Conscious Behavior

Women who were more health-conscious tended to use HRT and tended to have lower rates of cardiovascular disease

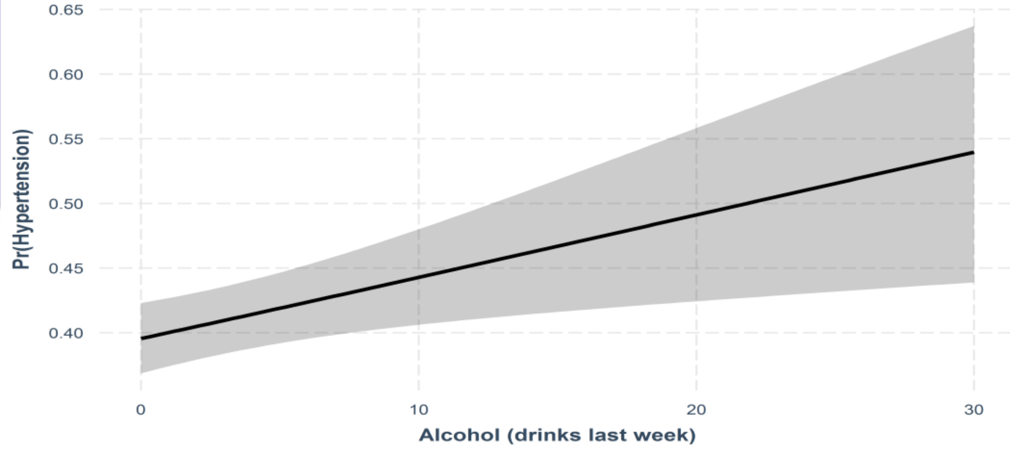
Higher rates of Cardiovascular Disease

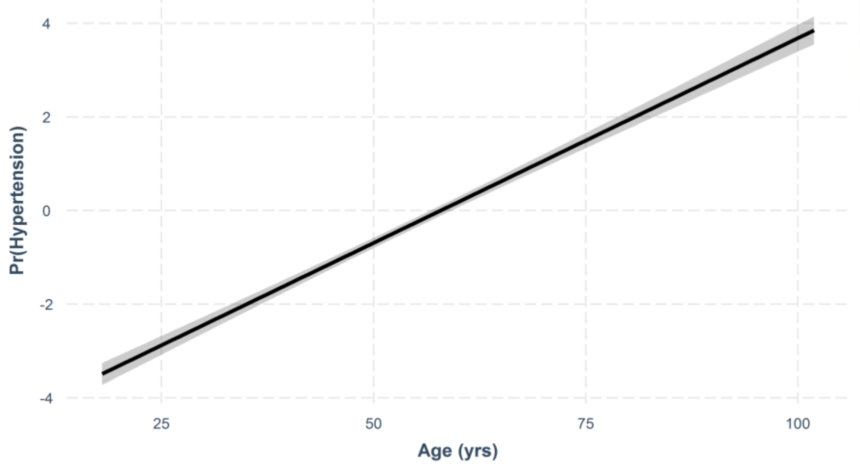
When the protective effect of healthy behavior was removed, HRT actually led to higher rates of Cardiovascular disease.

Confounding: Pr (Hypertension)

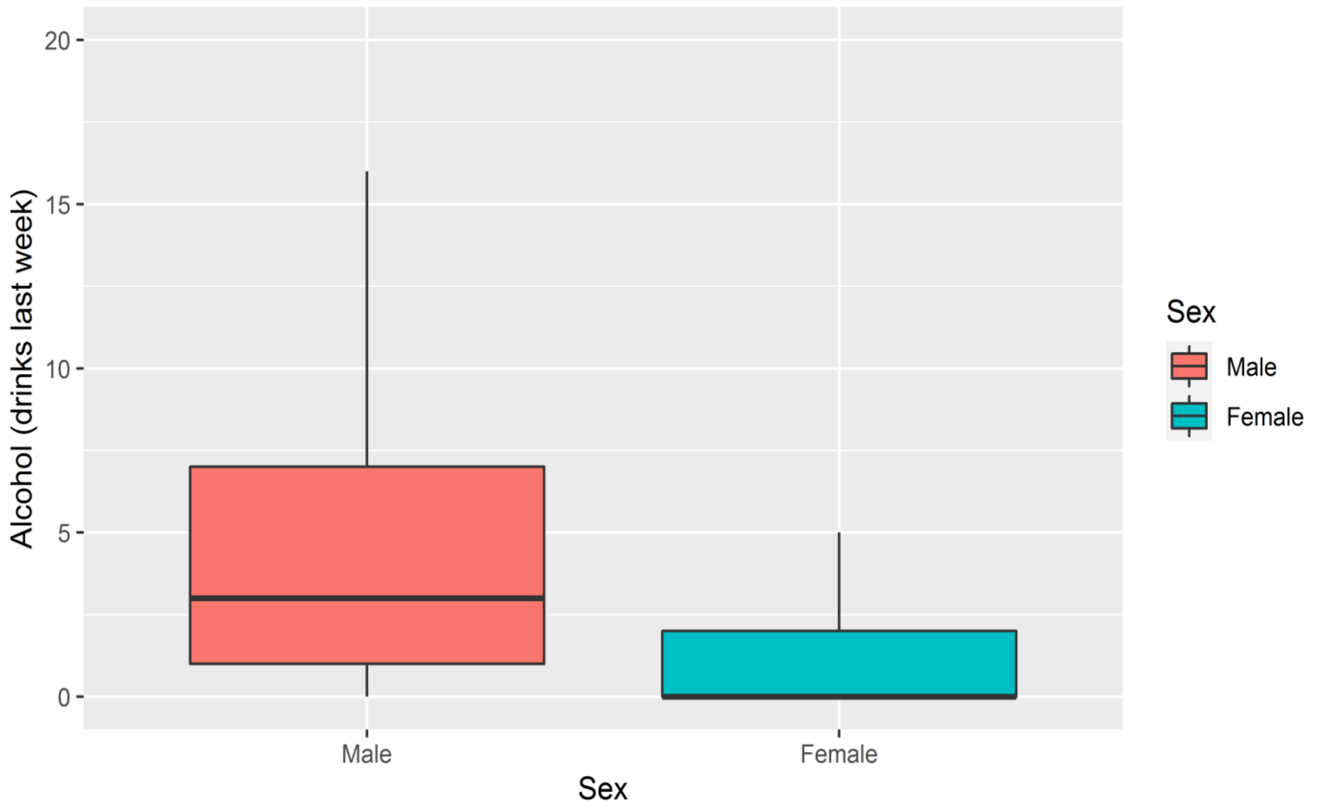
$$\ln\left(\frac{\pi}{1-\pi}\right) = \log(odds) = b_0 + b_1 Alcohol + b_2 Age + b_3 Semi - urban + b_4 Rural$$

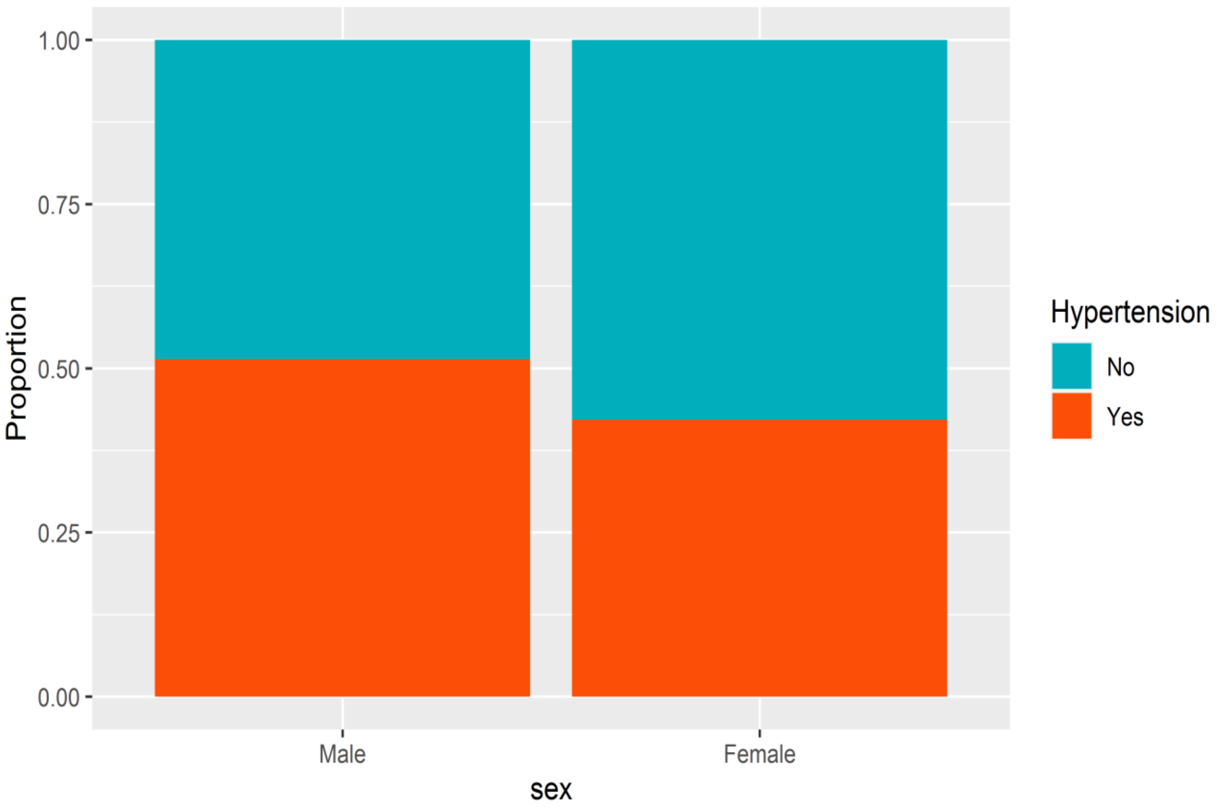
Characteristic	OR	95% CI	p-value
Alcohol (drinks last week)	1.02	1.01, 1.03	0.008
Age (yrs)	1.09	1.09, 1.10	<0.001
Urbanity			
Urban	—	—	
Semi-urban	1.20	0.98, 1.47	0.077
Rural	1.32	1.11, 1.57	0.002





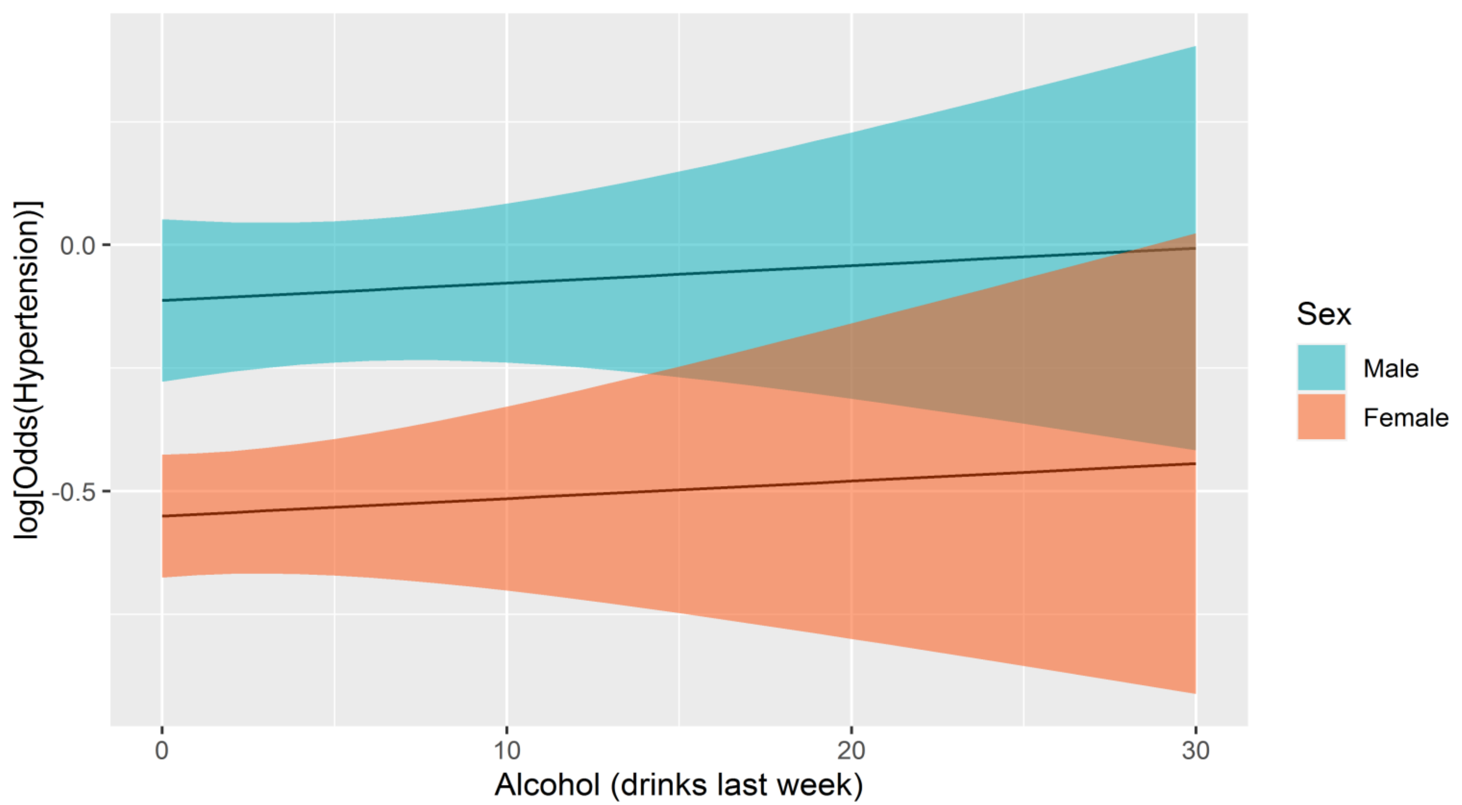
Sex: Confounder?





Adjusting for sex

Characteristic	OR	95% CI	p-value
Alcohol (drinks last week)	**1.00**	**0.99, 1.02**	**0.655**
Sex			
Male			
Female	**0.646**	**0.545, 0.764**	**<0.001**
Age (yrs)	1.09	1.0[illegible]	[illegible]
Urbanity			
Urban			
Semi-urban	1.20	0.9[illegible]	[illegible]
Rural	1.31	1.1[illegible]	[illegible]



Effect modifier (Interactions)

Birth Weight → **Leukemia**

sex

OR = 1.5

Is the weight effect differentiated by sex?

Boys **Birth Weight** → **Leukemia** **OR = 1.8**

Girls **Birth Weight** —/ / → **Leukemia** **OR = 0.9**

Effect Modifier

- If the estimated effect of a risk factor on the outcome differs across the levels of a third variable
- Statistically: Interaction
- There is no point/intention to adjust for an effect modifier
- The effect of the risk factor should be presented separately at the different levels of the effect modifier

Effect Modifier vs confounder

Effect Modifier

Belong to «nature»!

Different effect at different levels of the effect modifier

Useful information…

It improves our knowledge for the underlying mechanisms

Application to public health interventions and to personalized medicine

Confounder

Belong to the study!

Same effect across the levels of the third variable

We need to adjust for the third variable (crude and adjusted estimates)

It caused confusion in the data and the results

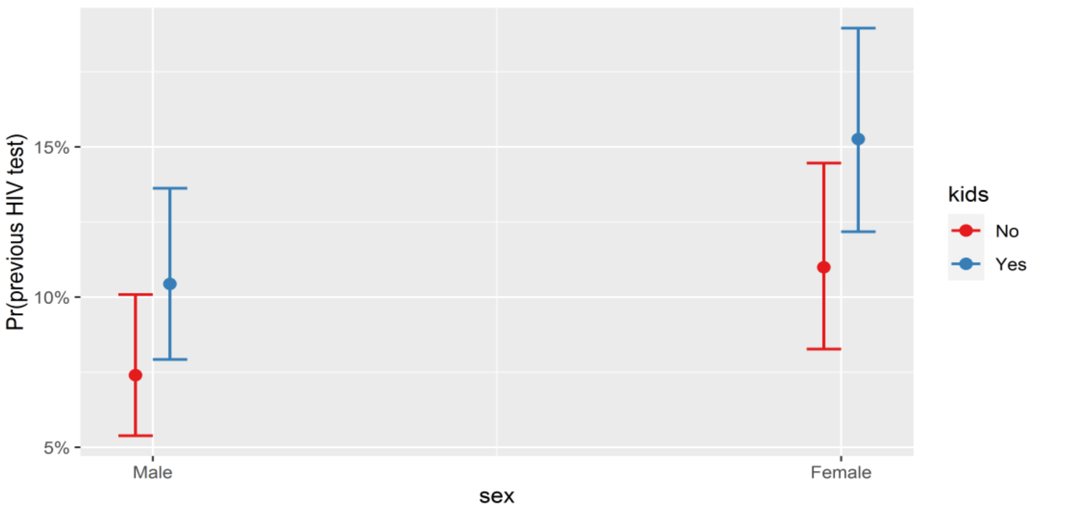
We can deal with a confounder during the data analysis

Interactions: Binary x Binary

$$Log(Odds) = b_0 + b_1 Fem + b_2 Kids + b_3 age + b_4 Ins. + b_5 Sex(6-10) + b_6 sex(11+)$$

π=Probability of past testing for HIV

Characteristic	OR	95% CI	p-value
Sex			
Male			
Female	1.55	1.28, 1.87	<0.001
Kids			
No			
Yes	1.46	1.18, 1.81	<0.001
Age (yrs)	0.97	0.96, 0.97	<0.001
Insurance			
No			
Yes	1.33	1.03, 1.72	0.033
Number of Sexual partners			
0-5			
6-10	2.02	1.61, 2.52	<0.001
11+	3.84	3.04, 4.86	<0.001



Interactions: Binary x Binary

$$Log(Odds) = b_0 + b_1 Fem + b_2 Kids + b_3 age + b_4 Ins. + b_5 Sex(6-10) + b_6 sex(11+) + \boldsymbol{b_7 (Fem * Kids)}$$

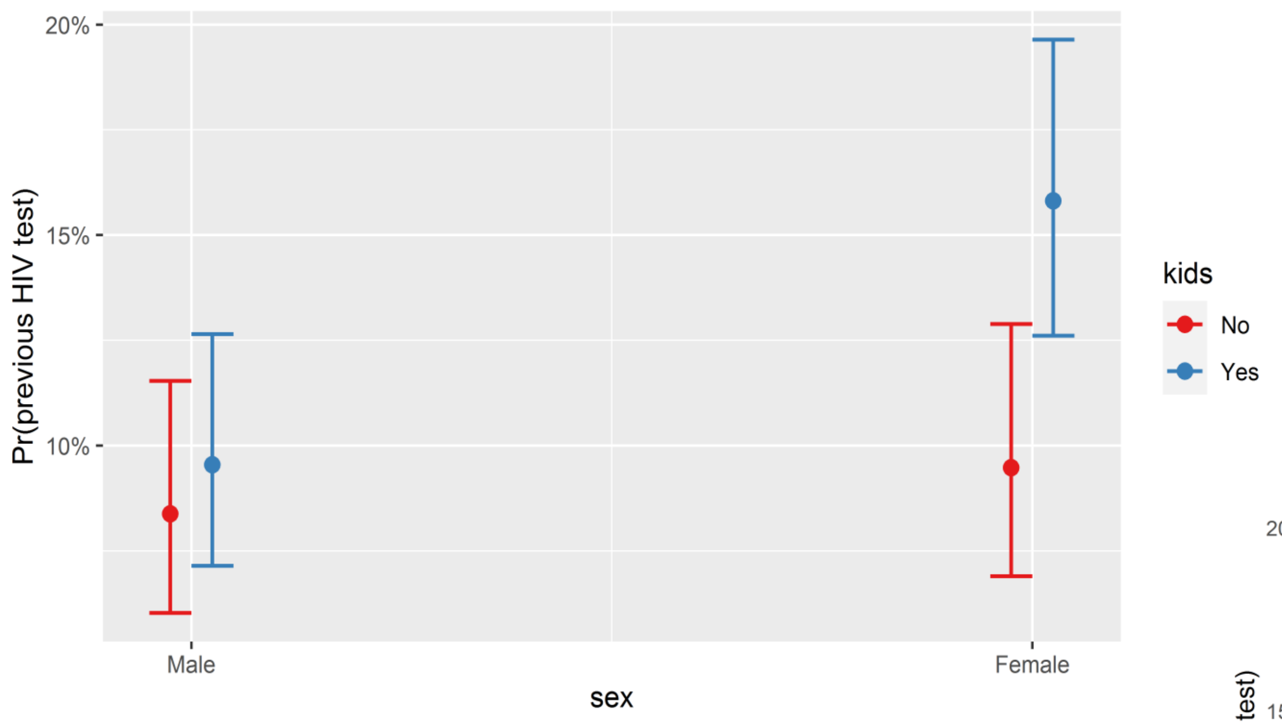
Characteristic	log(OR)	p-value
Sex		
Male		
Female	0.135	0.391
Kids		
No		
Yes	0.143	0.326
Sex * Kids		
Female * Yes	0.441	**0.016**

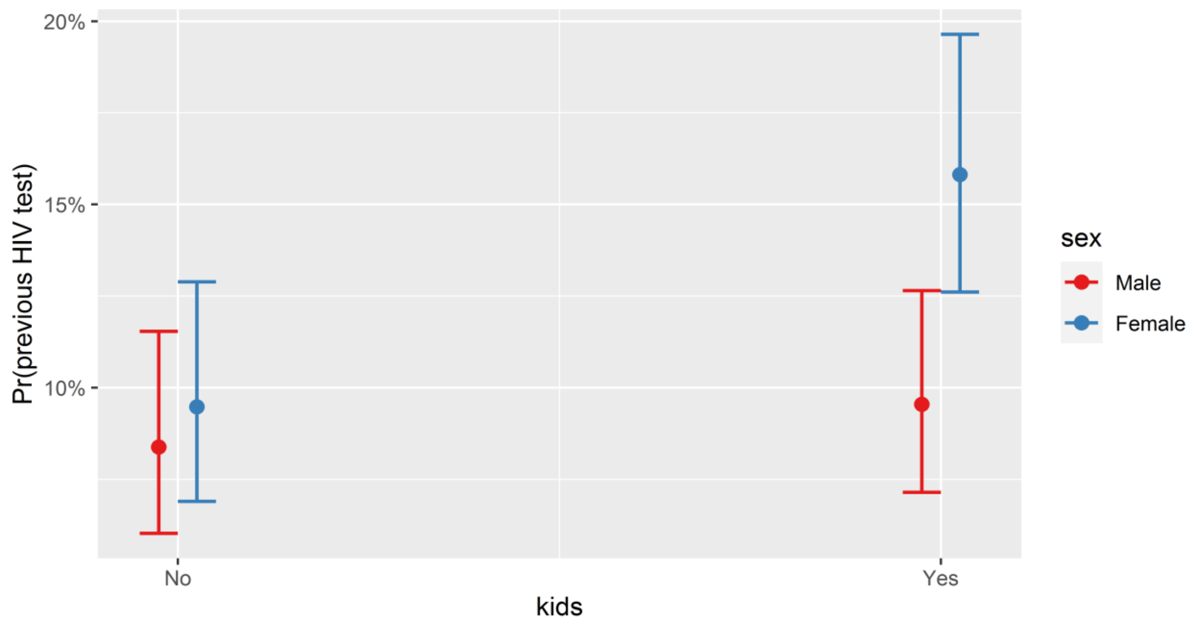
Characteristic	OR	95% CI	p-value
Sex			
Male			
Female	1.14	0.840, 1.56	0.391
Kids			
No			
Yes	1.15	0.868, 1.54	0.326
Female * Yes	1.55	1.09, 2.23	0.016

	Female/Male
No Kids	$e^{0.135} = 1.14$ (0.84-1.56)
Yes Kids	$e^{0.135+0.441} = 1.78\ (1.43 - 2.22)$
	Yes Kids/No Kids
Male	$e^{0.143} = 1.15$ (0.87-1.54)
Female	$e^{0.143+0.441} = 1.79\ (1.36 - 2.36)$

	No Kids	Yes Kids
Male	1	$e^{0.143} = 1.15$
Female	$e^{0.135} = 1.14$	$e^{0.135+0.143+0.441} = 2.05$

Effect Modifier: Graphical representation





Interactions: How to estimate combinations of bs

$$Log(Odds) = b_0 + b_1 Fem + b_2 Kids + b_3 age + b_4 Ins. + b_5 Sex(6-10) + b_6 sex(11+) + b_7(Fem * Kids)$$

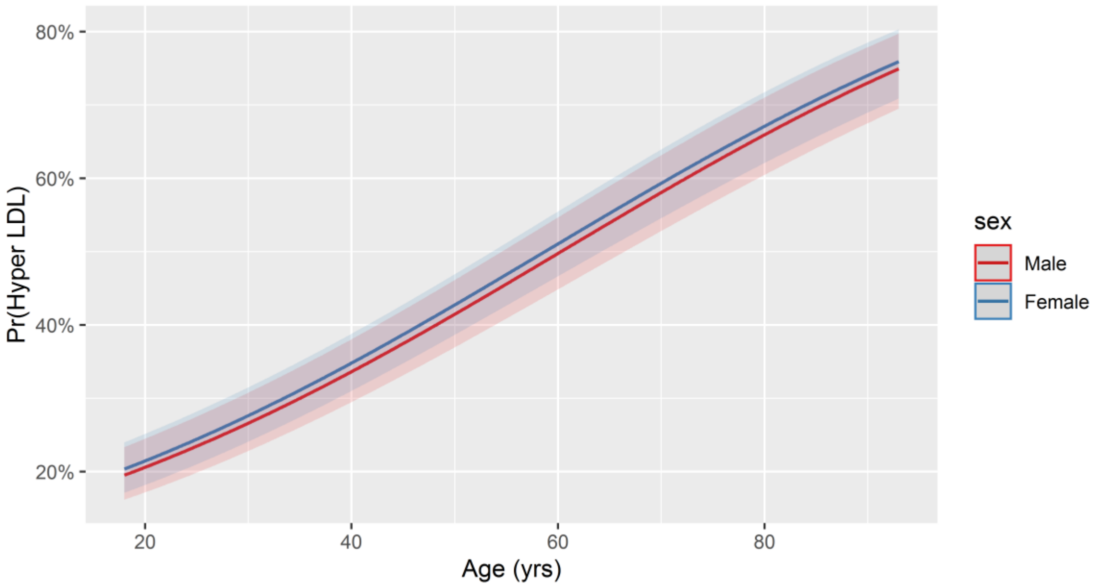
1. Define Constrain $C = (0 \quad 1 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1)'$
2. Compute C***b'**, **b**: 1xp vector of bs
3. Compute C*V(b)*C '

Interactions: Binary x Continuous

$$Log(Odds) = b_0 + b_1 Fem + b_2 age + b_3 overw + b_4 obese + b_5 walking30 \ + b_6 walking30 +$$

π=Probability of elevated LDL

Characteristic	OR	95% CI	p-value
Sex			
Male			
Female	1.05	0.918, 1.21	0.457
Age (yrs)	1.03	1.03, 1.04	<0.001
BMI Categories			
Normal			
Overweight	1.94	1.62, 2.31	<0.001
Obese	1.85	1.55, 2.22	<0.001
Walking			
<30 min/day			
>=30 min/day	0.811	0.706, 0.931	0.003



Interactions: Binary x Continuous

$$Log(Odds) = b_0 + b_1 Fem + b_2 age + b_3 overw + b_4 obese + b_5 walking30 + b_6 walking30 + \boldsymbol{b_7 Fem * Age}$$

Characteristic	log(OR)	p-value
Sex		
Male		
Female	-0.667	0.005
Age (yrs)	0.026	<0.001
Sex * Age (yrs)		
Female * Age (yrs)	0.013	0.002

Characteristic	OR	95% CI	p-value
Sex			
Male			
Female	0.513	0.321, 0.820	0.005
Age (yrs)	1.03	1.02, 1.03	<0.001
Sex * Age (yrs)			
Female * Age (yrs)	1.01	1.01, 1.02	0.002

$Male: 0.026 * age$ OR (per 10 years):1.30 (1.23-1.38)

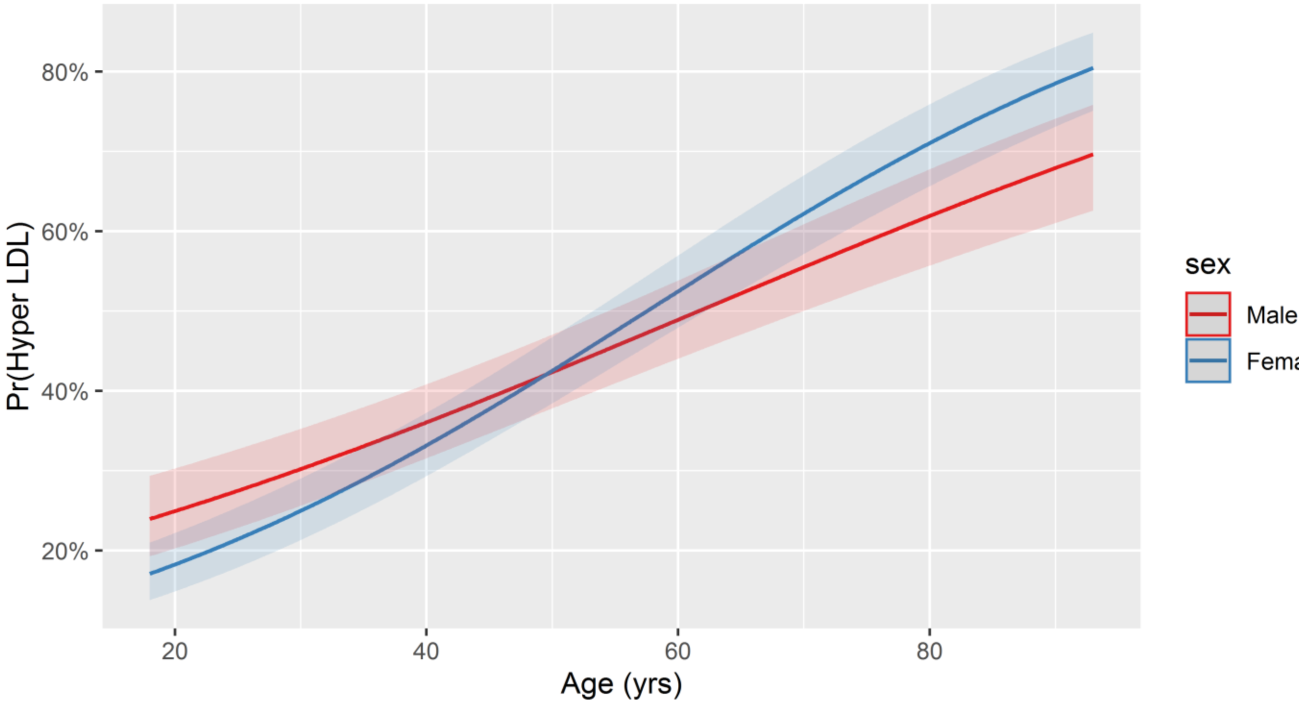
Female: (0.026+0.013)* age OR (per 10 years): 1.49 (1.40-1.58)

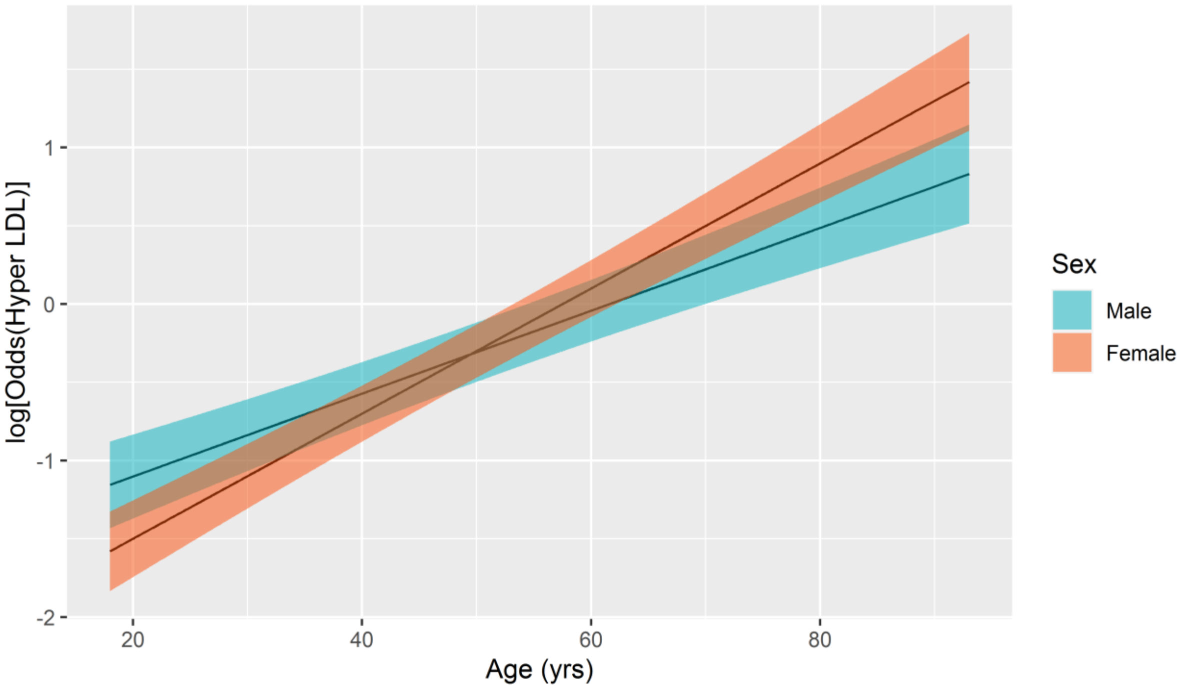
$$\frac{Female}{Male}: b_1 + b_6 * age$$

Age (years)	Female/Male: OR (95%CI)
20	0.67 (0.49-0.91)
50	1.01 (0.87-1.15)
80	1.51 (1.16-1.96)

Interactions: Binary x Continuous; Graphically

$$Log(Odds) = b_0 + b_1 Fem + b_2 age + b_3 overw + b_4 obese + b_5 walking30 \ + b_6 walking30 + \boldsymbol{b_7 Fem * Age}$$



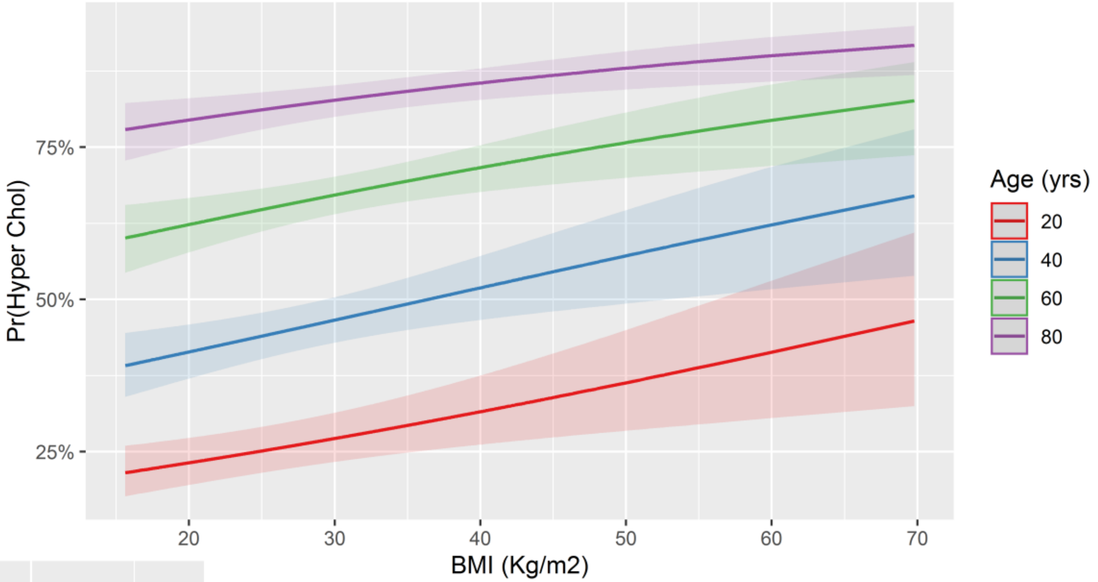


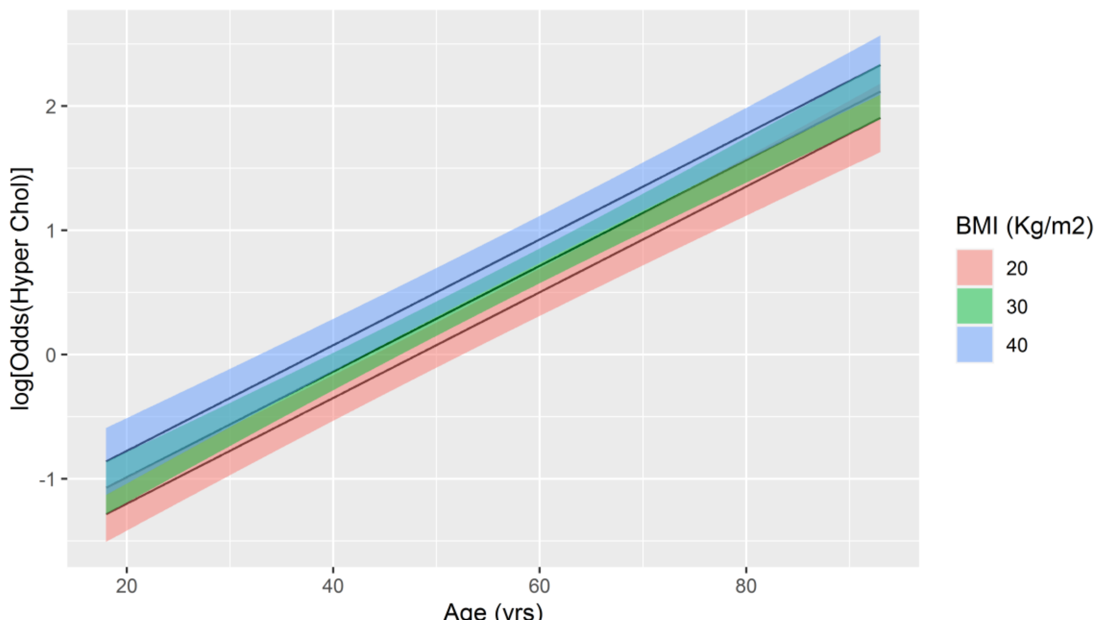
Interactions: Continuous × Continuous

$$Log(Odds) = b_0 + b_1 BMI + b_2 age + b_3 Fem + b_4 Walking30+$$

π=Probability of elevated Total Cholesterol

Characteristic	OR	95% CI	p-value
BMI (Kg/m²)	1.02	1.01, 1.03	0.001
Age (yrs)	1.04	1.04, 1.05	<0.001
Sex			
Male			
Female	1.17	1.02, 1.35	0.026
Walking			
<30 min/day			
>=30 min/day	0.848		





Interactions: Continuous x Continuous

$$Log(Odds) = b_0 + b_1 BMI + b_2 age + b_3 Fem + b_4 Walking30 + \boldsymbol{b_5 BMI * age}$$

Characteristic	log(OR)	p-value
BMI (Kg/m2)	0.170	<0.001
Age (yrs)	0.125	<0.001
Sex		
Male		
Female	0.197	0.006
Walking		
<30 min/day		
>=30 min/day	-0.177	0.015
BMI (Kg/m2) * Age (yrs)	-0.003	<0.001

Characteristic	OR	95% CI	p-value
BMI (Kg/m2)	1.19	1.13, 1.24	<0.001
Age (yrs)	1.13	1.11, 1.16	<0.001
Sex			
Male			
Female	1.22	1.06, 1.40	0.006
Walking			
<30 min/day			
>=30 min/day	0.837	0.726, 0.97	0.015
BMI (Kg/m2) * Age (yrs)	1.00	1.00, 1.00	<0.001

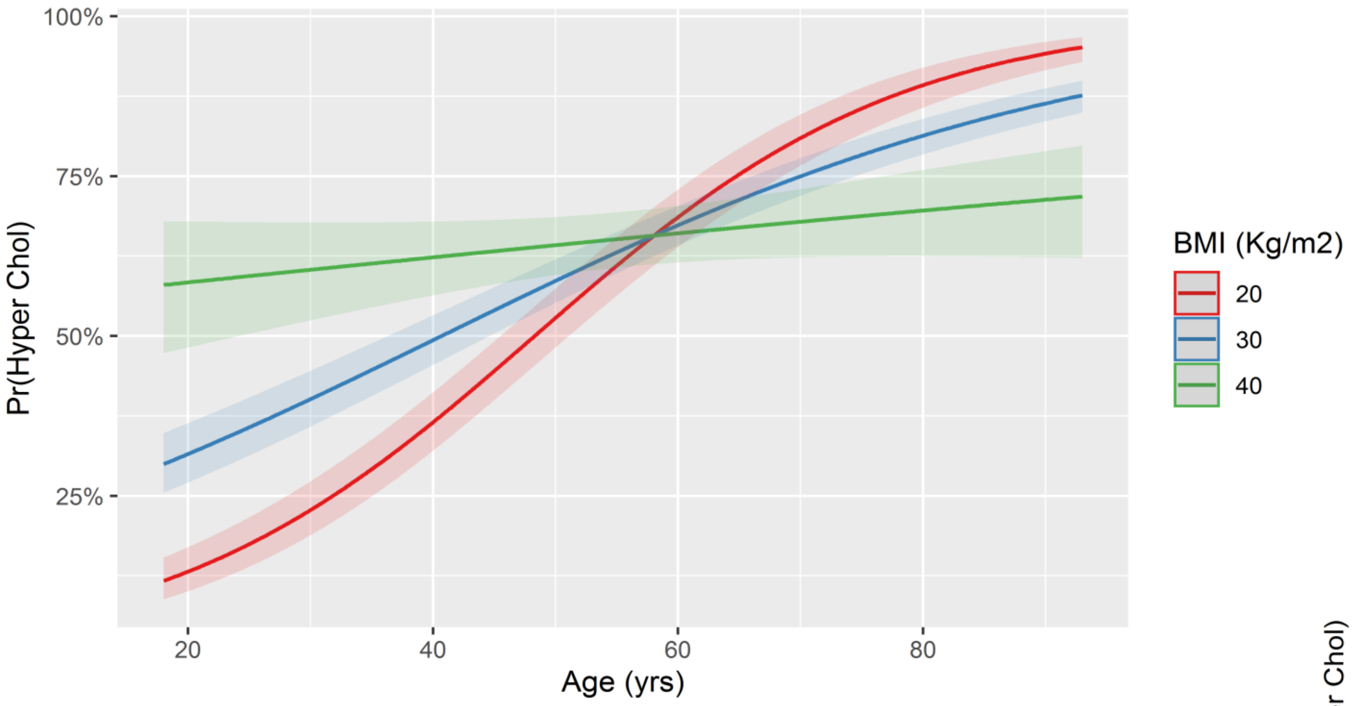
BMI (kg/m^2)	Age (per 10yrs)
20	1.94 (1.79-2.12)
30	1.45 (1.39-1.52)
Age (yrs)	**BMI (per 5 kg/m^2)**
30	1.51 (1.35-1.68)
60	0.73 (0.64-0.83)

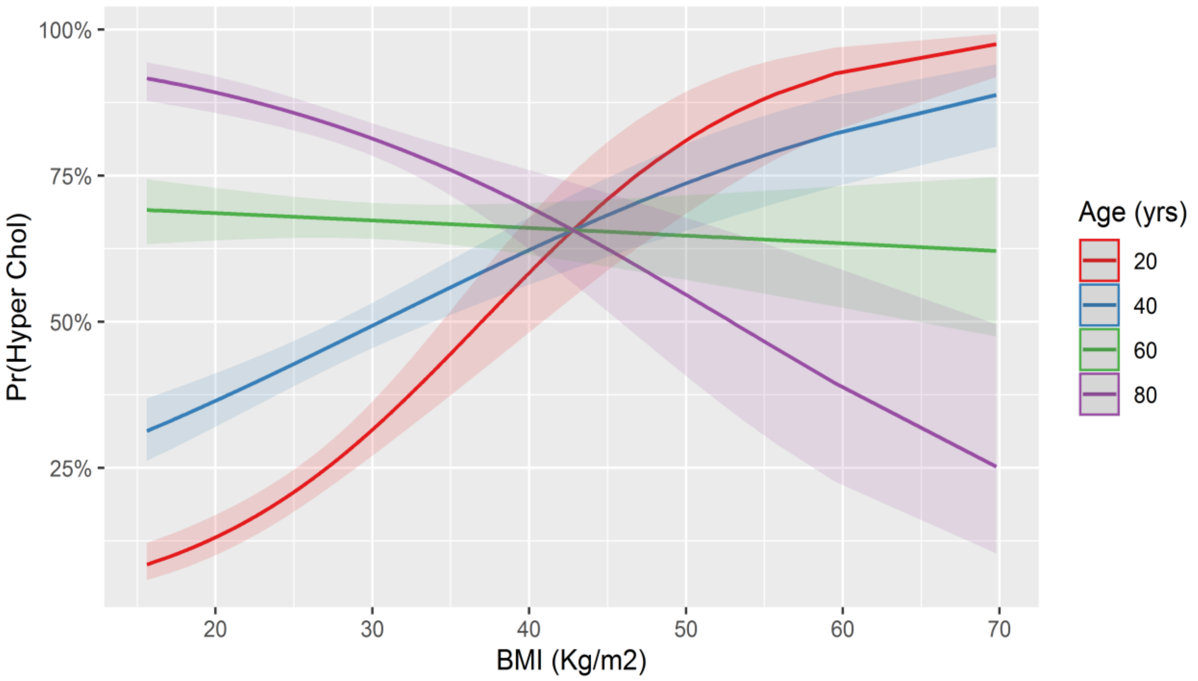
BMI (per unit): $b_1 + b_5 * age$

Age (per unit): $b_2 + b_5 * BMI$

Interactions: Continuous x Continuous graphically

$$Log(Odds) = b_0 + b_1 BMI + b_2 age + b_3 Fem + b_4 Walking30 + \boldsymbol{b_5 BMI * age}$$





Interactions: Multiple Interaction terms

Research Paper — Tobacco Prevention & Cessation

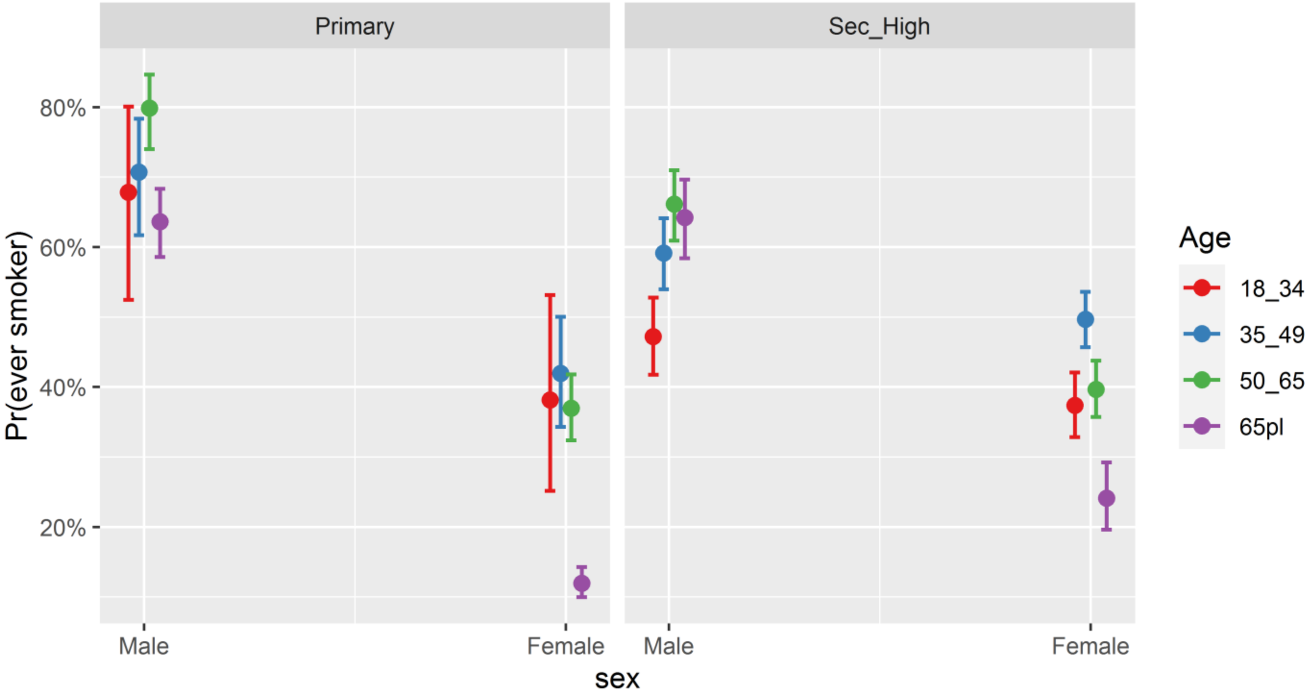
Prevalence of tobacco smoking and association with other unhealthy lifestyle risk factors in the general population of Greece: Results from the EMENO study

Maria Gangadi[1], Natasa Kalpourtzi[2], Magda Gavana[3], Apostolos Vantarakis[4], Gregory Chlouverakis[5], Christos Hadjichristodoulou[6], Gregory Trypsianis[7], Paraskevi V. Voulgari[8], Yannis Alamanos[9], Argiro Karakosta[2], Giota Touloumi[2], Anna Karakatsani[1*]*

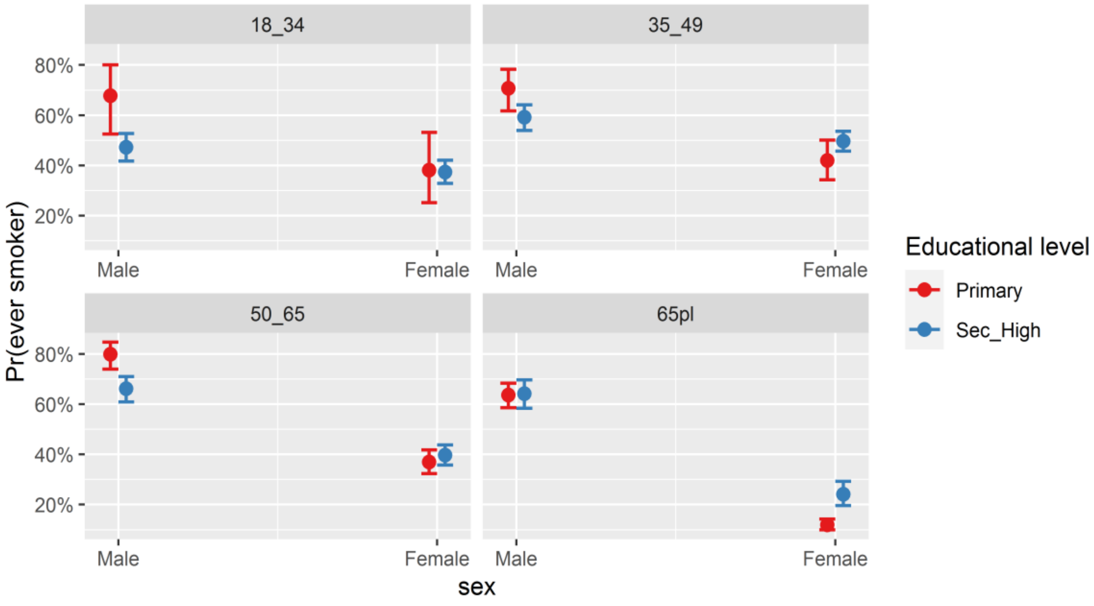
Tob. Prev. Cessation, 2021

$$
\begin{aligned}
&log[Pr(EverSmoker)]\\
&= \beta_0 + \beta_1 * EduSecHigh + \beta_2 * Age3549 + \beta_3 * Age5065 + \beta_4 * Age65pl + \beta_5 * SexFemale + \beta_6\\
&* Alcohol17 + \beta_7 * Alcohol7pl + \beta_8 * UnemployedYes + \boldsymbol{\beta_9 * EduSecHigh{:}\,Age3549} + \boldsymbol{\beta_{10}}\\
&* \boldsymbol{EduSecHigh{:}\,Age5065} + \boldsymbol{\beta_{11} * EduSecHigh{:}\,Age65pl} + \boldsymbol{\beta_{12} * EduSecHigh{:}\,SexFemale} + \boldsymbol{\beta_{13}}\\
&* \boldsymbol{Age3549{:}\,SexFemale} + \boldsymbol{\beta_{14} * Age5065{:}\,SexFemale} + \boldsymbol{\beta_{15} * Age65pl{:}\,SexFemale}
\end{aligned}
$$

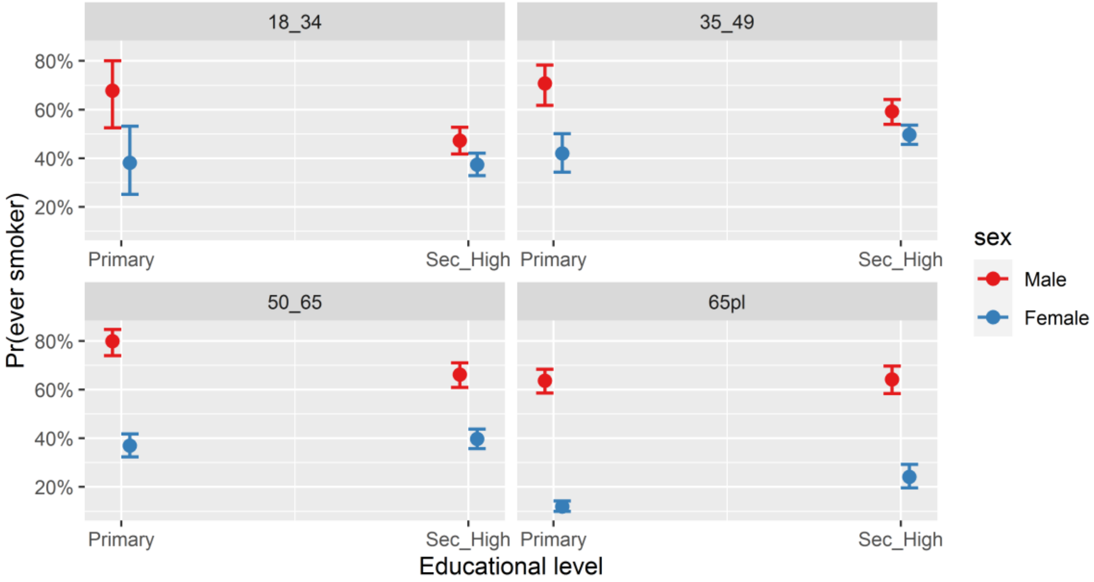
Characteristics	*OR (95% CI)*	*p*
Alcohol consumption (glasses/week)		
1–7	1.41 (1.22–1.64)	**<0.001**
>7	2.52 (1.97–3.23)	**0.001**
Unemployed/employed	1.42 (1.16–1.73)	**0.001**
Age (18–34.9 years)		
Women		
Secondary/Higher vs Primary	0.84 (0.43–1.66)	0.614
Men		
Secondary/Higher vs Primary	0.44 (0.23–0.84)	0.013
Primary education		
Women/Men	0.37 (0.23–0.59)	<0.001
Secondary/Higher education		
Women/Men	0.71 (0.54–0.94)	0.015
Age (35–49.9 years)		
Women		
Secondary/Higher vs Primary	1.28 (0.85–1.92)	0.24
Men		
Secondary/Higher vs Primary	0.67 (0.44–1.03)	0.06
Primary education		
Women/Men	0.35 (0.23–0.53)	**<0.00**
Secondary/Higher education		
Women/Men	0.67 (0.52–0.87)	0.00
Age (50–64.9 years)		
Women		
Secondary/Higher vs Primary	1.00 (0.75–1.34)	0.99
Men		
Secondary/Higher vs Primary	0.52 (0.35–0.79)	0.00
Primary education		
Women/Men	0.17 (0.12–0.26)	**<0.00**
Secondary/Higher education		
Women/Men	0.33 (0.25–0.43)	**<0.00**



Primary
Sec_High
80%
60%
40%
20%
Pr(ever smoker)
Male
Female
Male
Female
sex
Age
18_34
35_49
50_65
65pl

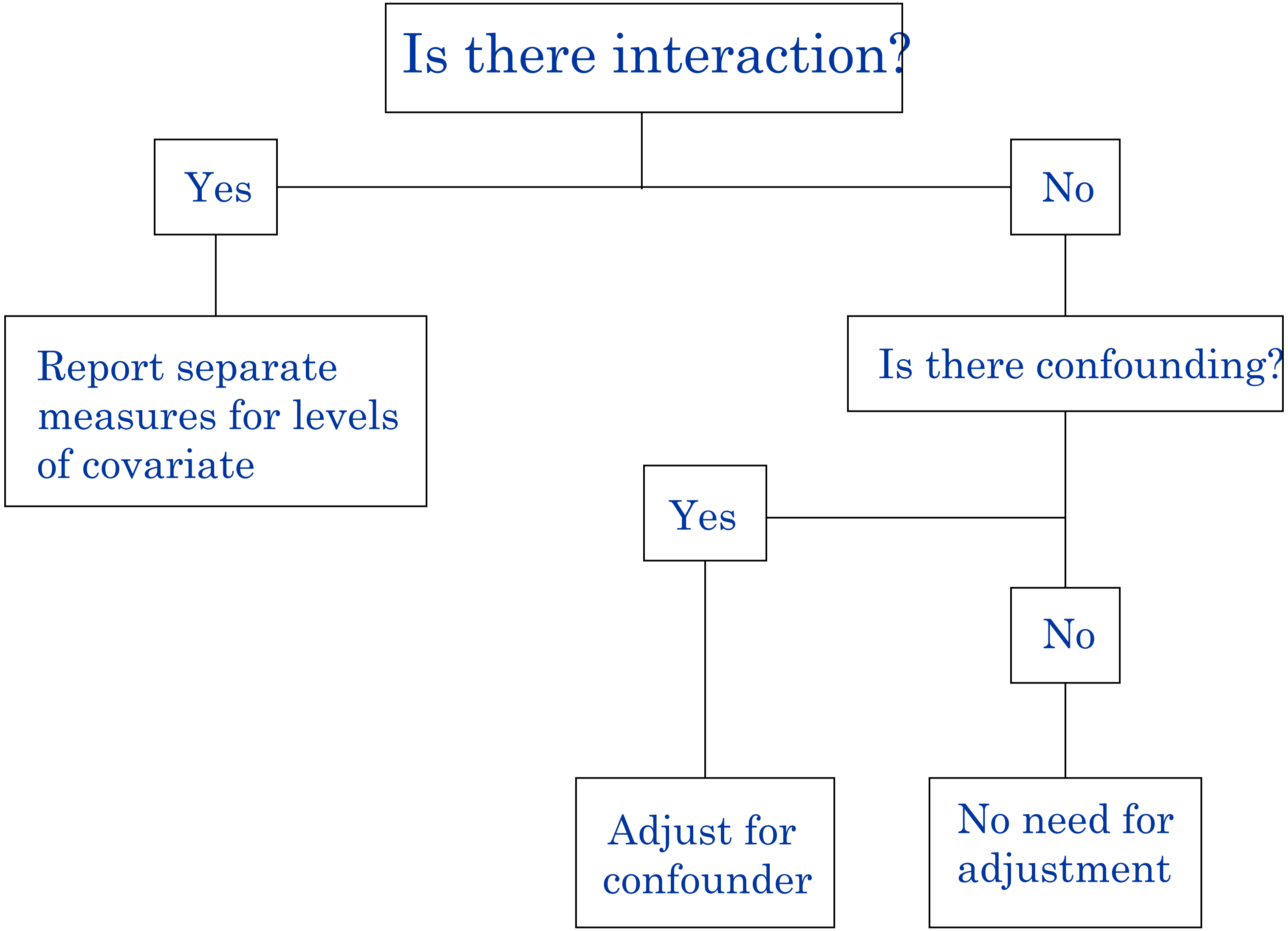


18_34
35_49
50_65
65pl
80%
60%
40%
20%
Pr(ever smoker)
Male
Female
sex
Educational level
Primary
Sec_High



18_34
35_49
50_65
65pl
80%
60%
40%
20%
Pr(ever smoker)
Primary
Sec_High
Educational level
sex
Male
Female

Approach to Interaction and Confounding



Thank you for your attention