The unhealthy causal salad: causal inference, DAGs and propensity scores

M. T. Liuzza L. Sità

Handzone – 27 marzo 2025

Causal salad

Say No to the Causal Salad!

I tre criteri dell'inferenza causale

- Covariazione
- Precedenza temporale
- Esclusione di cause alternative

Think before you regress: Directed Acyclic Graphs (DAG)

Introdotti da Judea Pearl, i **Directed Acyclic Graphs (DAG)** aiutano a ragionare sulla causalità

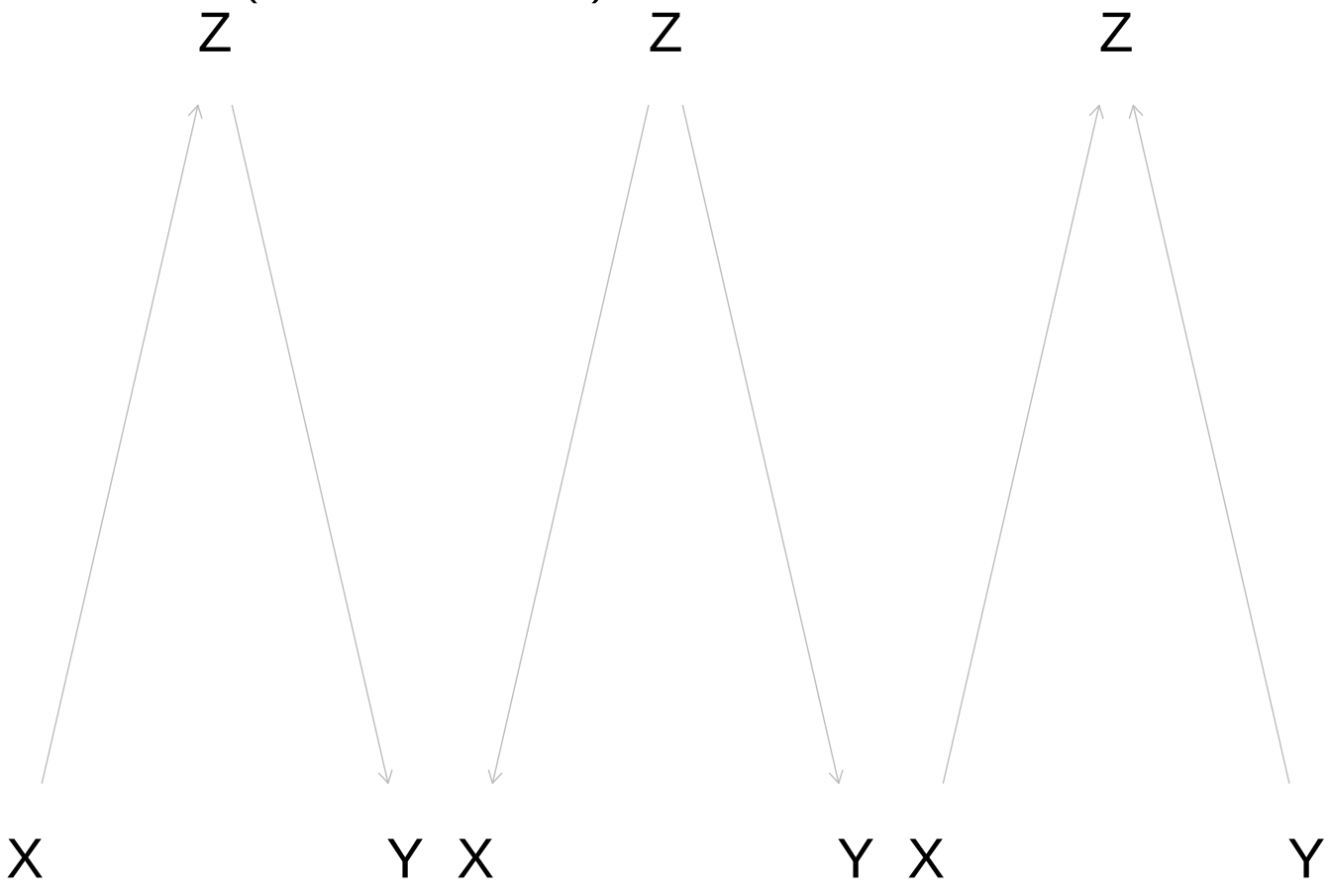
- *Directed:* le connessioni hanno frecce direzionali
- *Acyclic*: le cause non possono tornare indietro su se stesse
- *Graphs*: *nodi* e *connessioni*
  - I nodi possono essere genitori (*parents)* di un altro nodo figlio (*child*) se sono immediatamente antecedenti al nodo
  - Sono antenati (*ancestors*) se causano i genitori, oppure discendenti (*descendants*) se seguono causalmente i figli.

DAG e Structural Causal Models

- I DAG sono una rappresentazione grafica intuitiva di *Structural Causal Models* (SCM).
- Negli SCM abbiamo variabili:
  - *esogene* (U), che non possono essere discendenti perché non spieghiamo da cosa siano causate
  - *endogene* (V),
- Nei DAG la freccia è usata quando si ipotizza una relazione *causale* ad es. da X a Y (X ->Y).
  - La elazione bidirezionale (X <-> Y), equivale alla presenza du una variable non osservata - o latente - che causa entrambe (X <- U -> Y).
- Nei DAG possiamo rappresentare l'effetto di un insieme di variabili, chiamate *exposures* su altre variabili che chiameremo *outcomes*

DAG e ragionamento causale 1

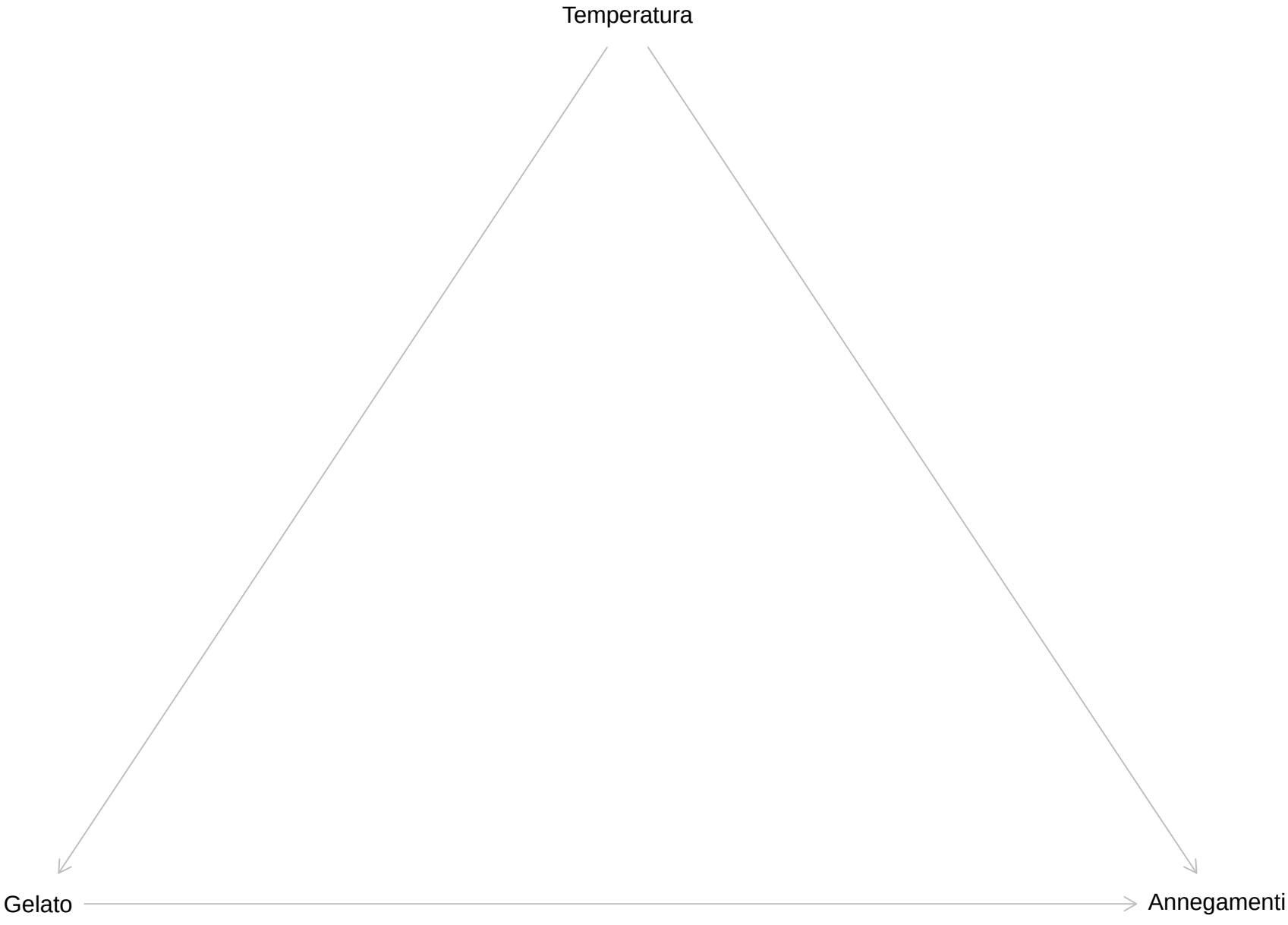
- *Chains* (X -> Z -> Y)
- *Forks* (X <- Z -> Y)
- *Colliders* (X -> Z <- Y)



DAG e ragionamento causale 2

- Nelle *chains* e nelle *forks*, controllare per Z **blocca** il percorso (*path*) introducendo un'*indipendenza condizionata t*ra X e Y: $X \perp Y \mid Z$
- Nei *collider*, controllare per Z **apre** il percorso (*path*) introducendo una dipendenza condizionata tra X e Y: $X \not\perp\!\!\!\perp \mid Z$
- La *d-separation* di due variabili si ha quando, attraverso una covariata, si blocca ogni percorso (*path*) tra loro. Viceversa, si ha la *d-connection*.
  Questi concetti sono importanti per DAG con più variabili. Ad esempio, controllare per un discendente di un collisore rischia di creare una *d*-connection tra variabili che prima erano *d*-separated.

Struttura dei DAG



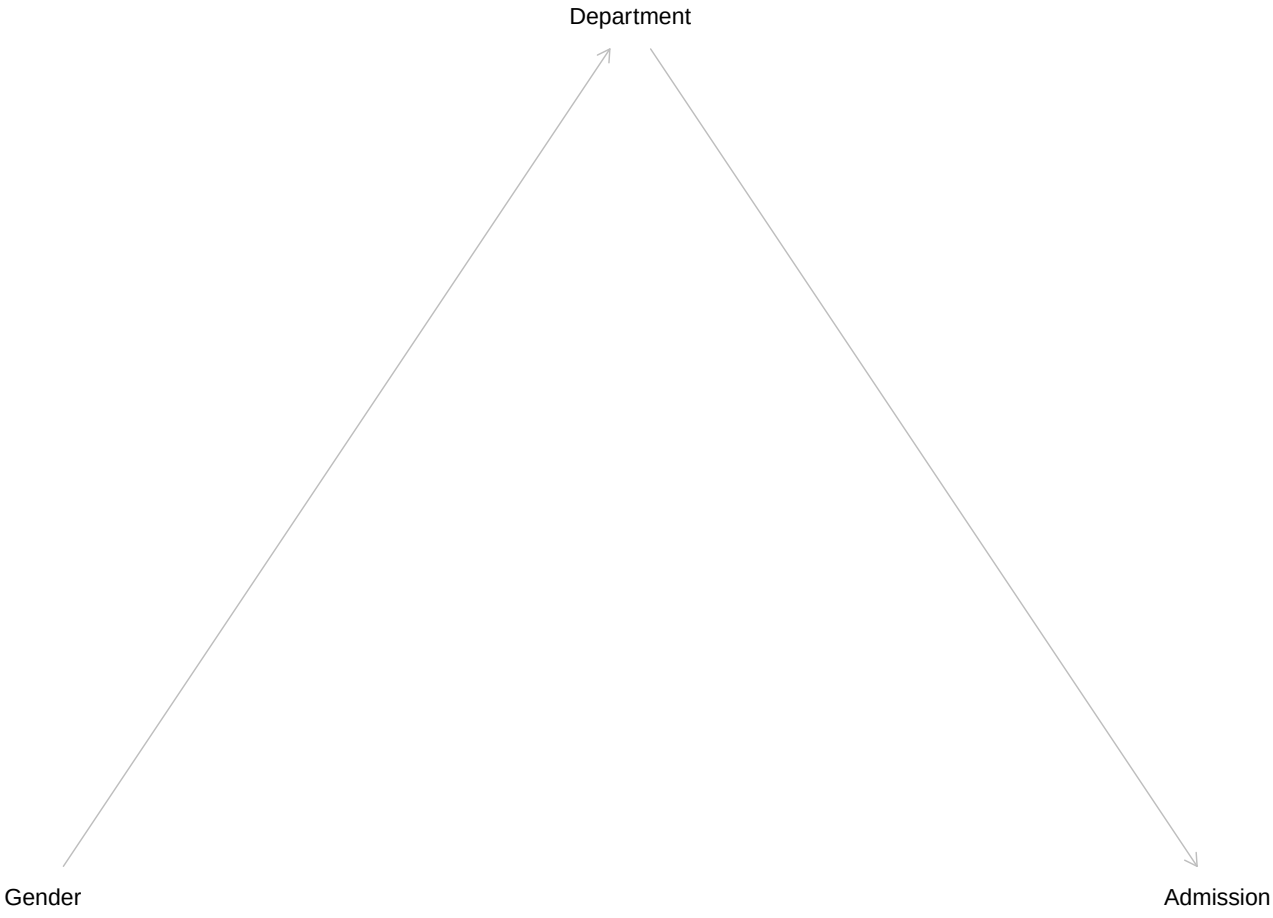
Variabili confondenti

Vanno sempre controllate

Variabili mediatrici 1

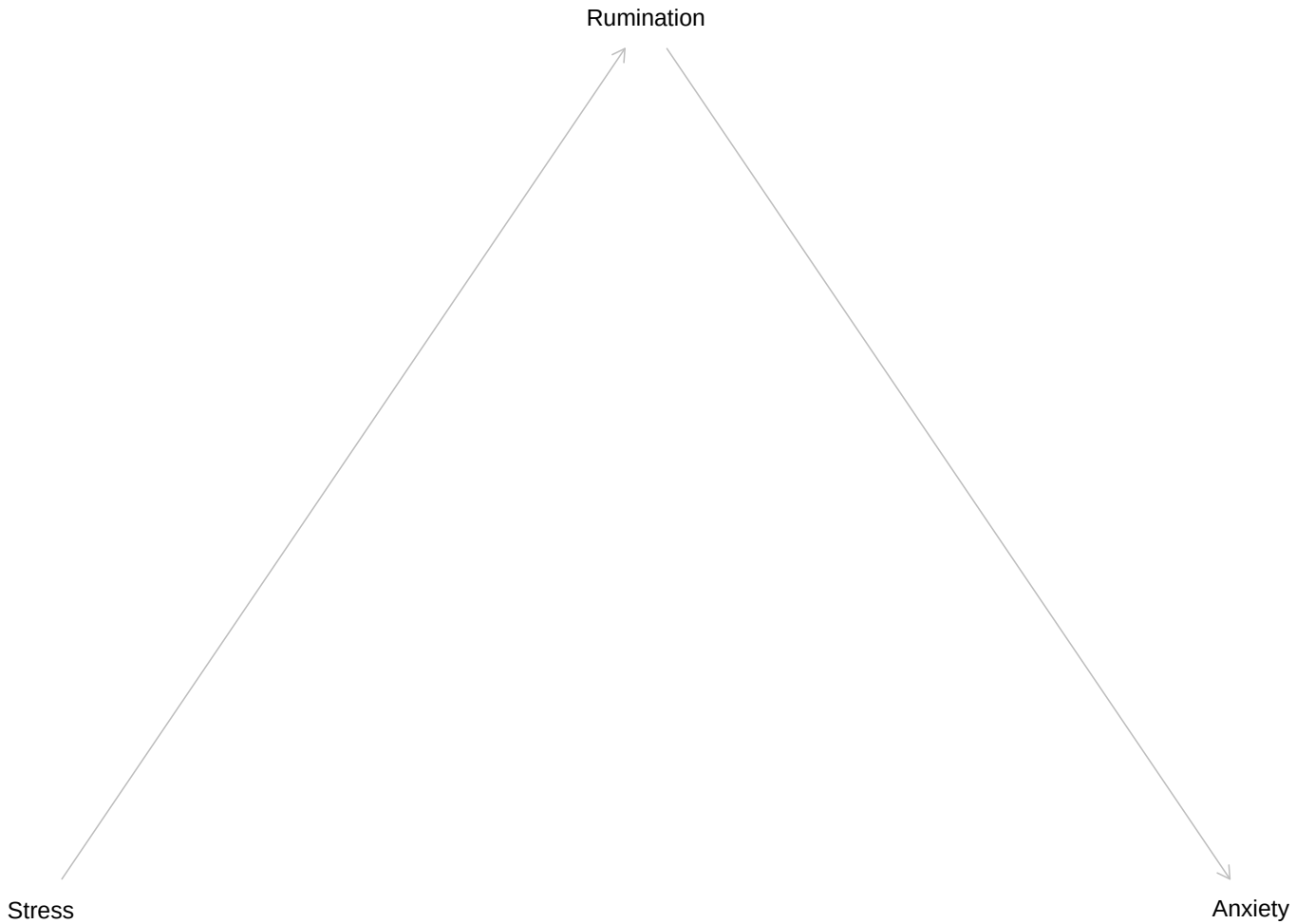
- **vanno controllate se ci interessa l'effetto diretto**
- non vanno controllate se ci interessa l'effetto totale

Esempio: *Simpson's paradox* sui dati di Berkley del 1979



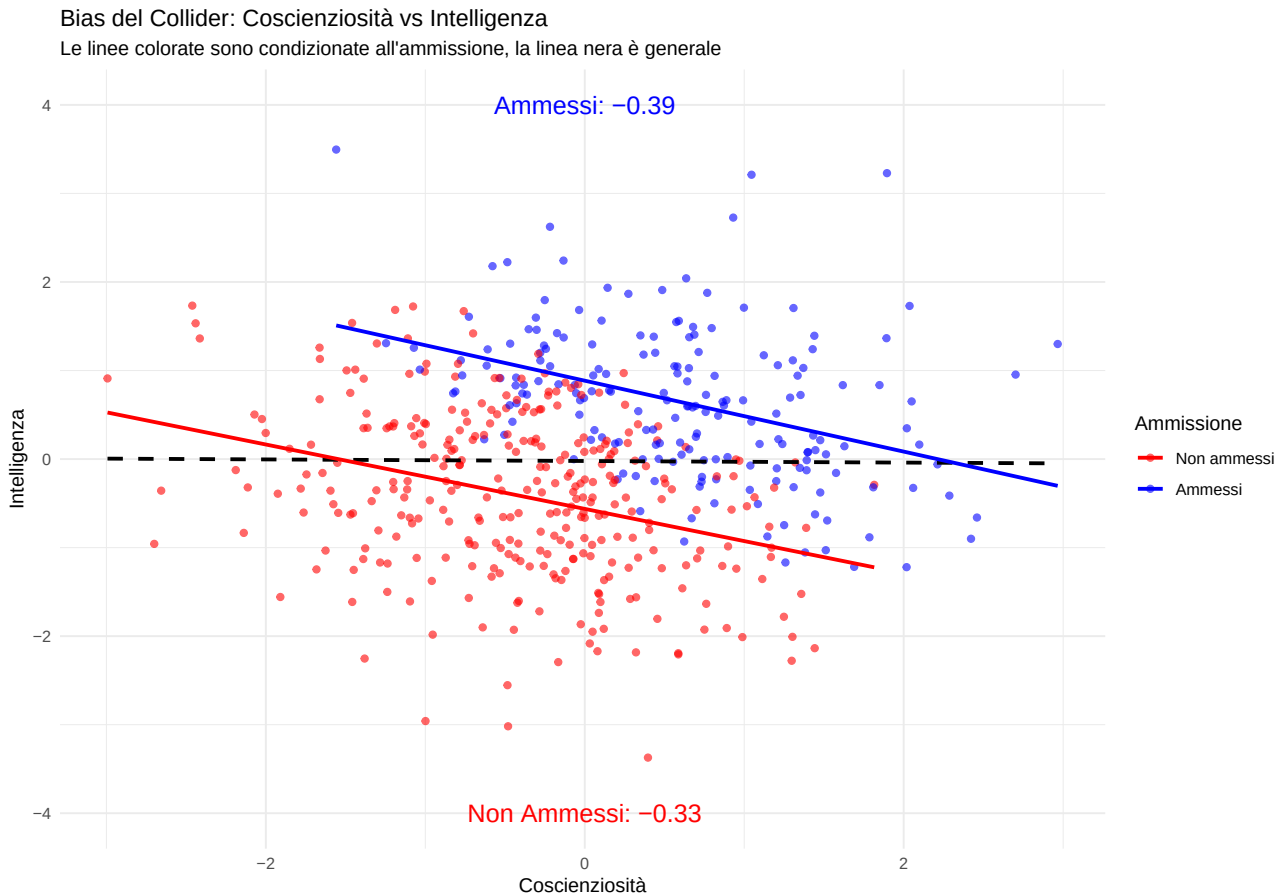
Variabili mediatrici 2

- vanno controllate se ci interessa l'effetto diretto
- **non vanno controllate se ci interessa l'effetto totale**

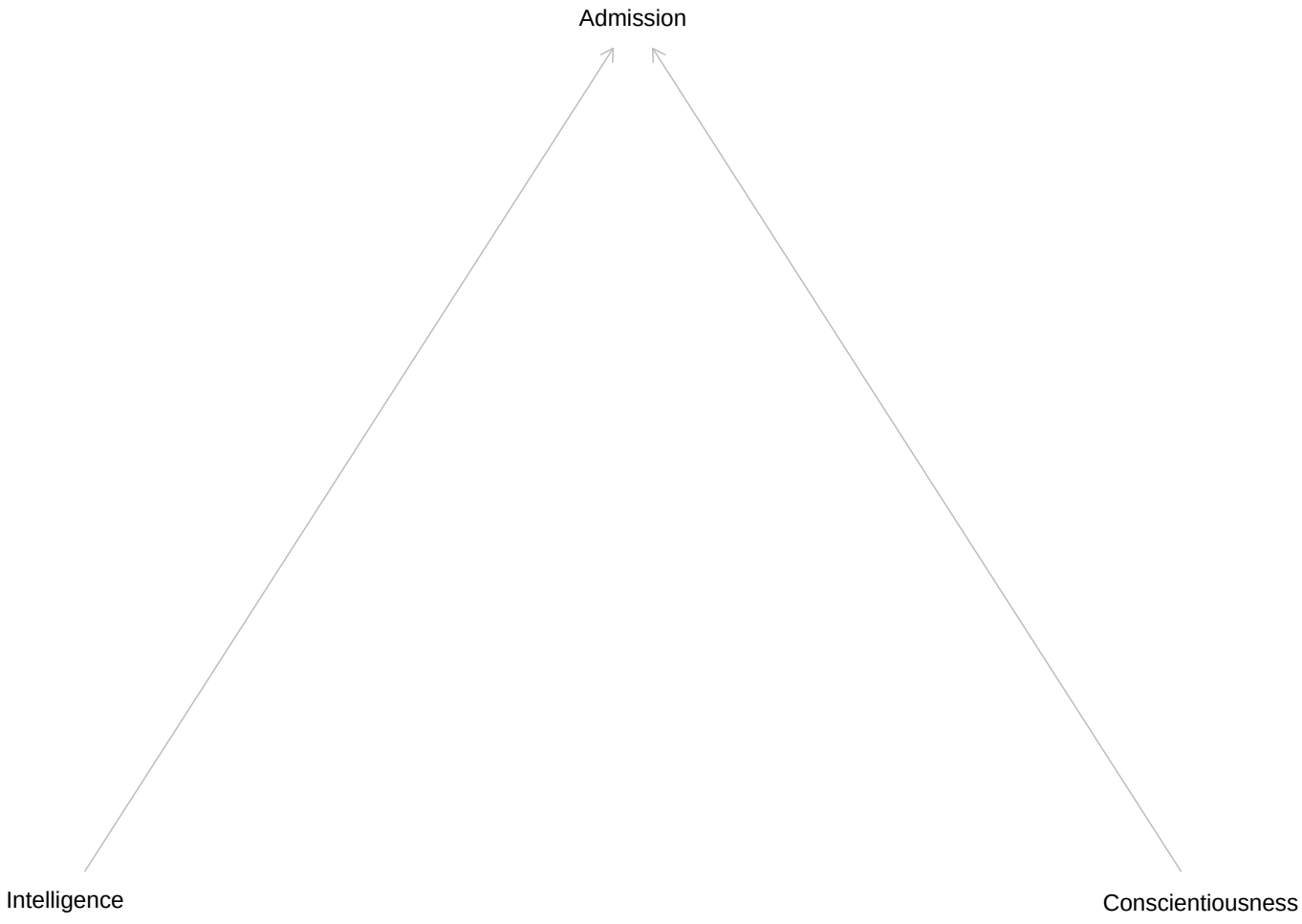


Collider

- Quando una variabile è causata da due variabili tra loro non correlate , si crea un *collider bias* che crea una dipendenza condizionata tra le variabili.
- Detto anche *Berkson's paradox*, si osserva spesso quando si introducono bias di selezione



Collider con i DAG



Riicapitoliamo: in possibili bias nella stima

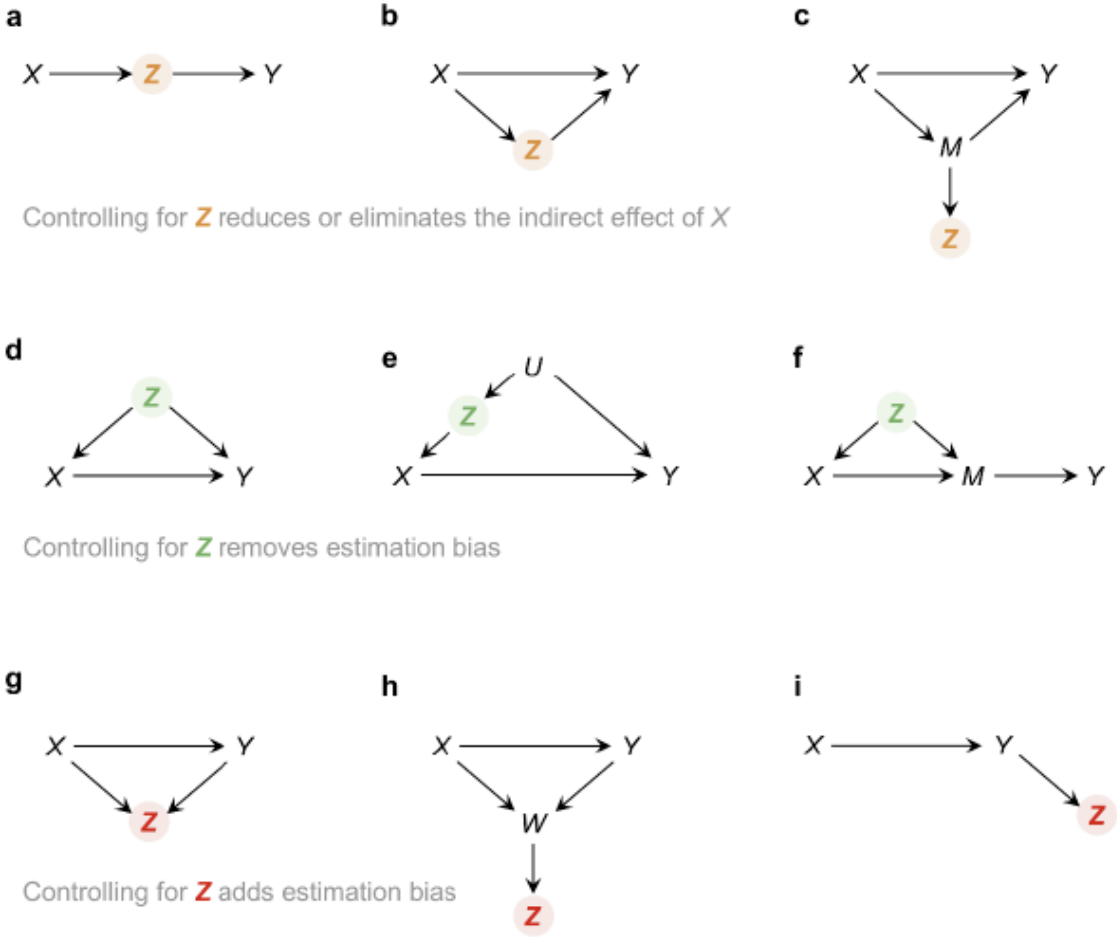
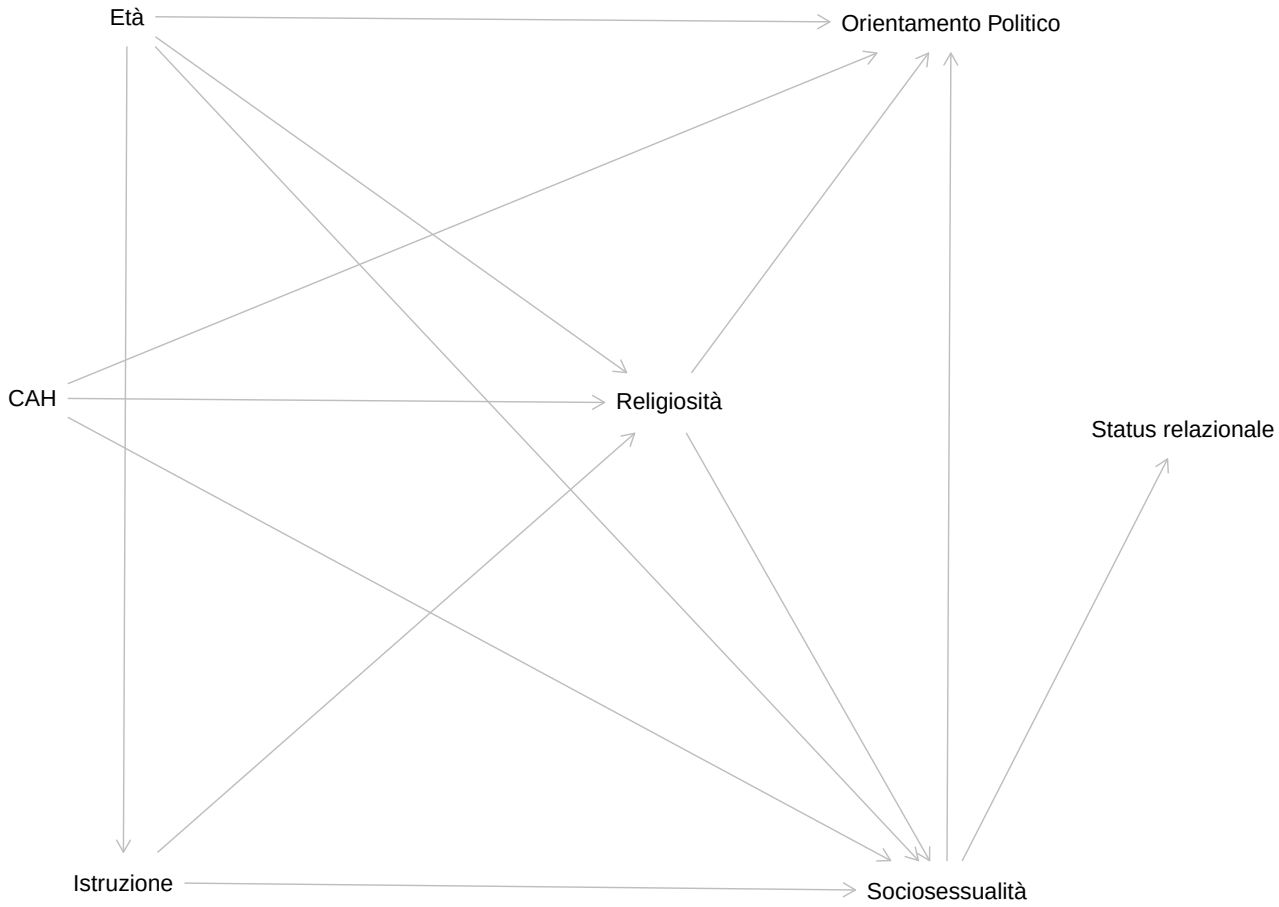


Figure 1. Simple causal models that illustrate the effects of covariate selection on the estimation of the effect of interest ($X \rightarrow Y$). In (a), (b), and (c), controlling for Z reduces or eliminates the indirect (mediated) effect of X on Y. In (d), (e), and (f), controlling for Z removes estimation bias by de-confounding the $X \rightarrow Y$ effect. In (g), (h), and (i), controlling for Z adds estimation bias to the $X \rightarrow Y$ effect.

Figure 1: Del Giudice e Gangestead, 2021, SM

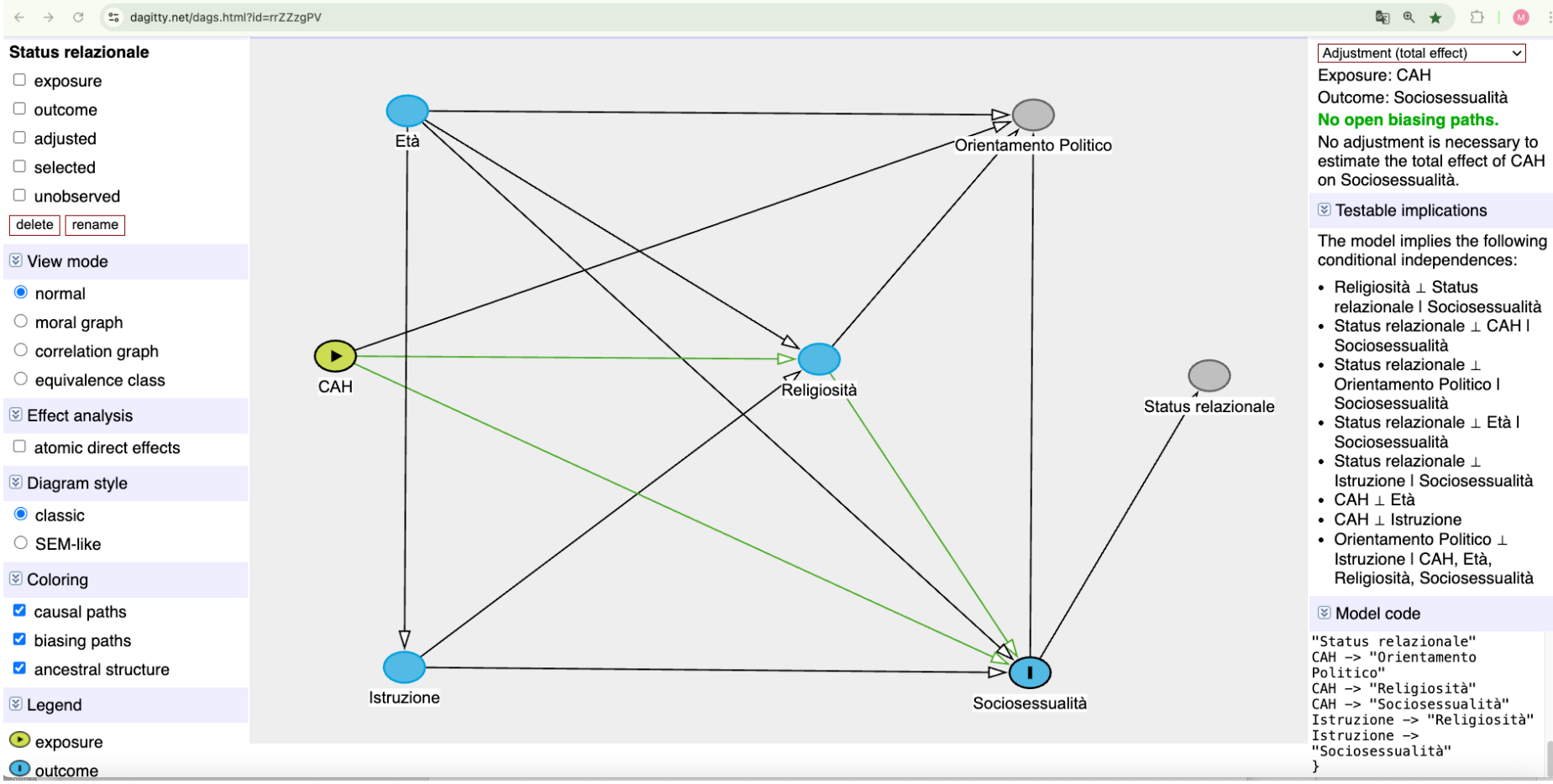
Un esempio: scegliere un gruppo di controllo

- Studio su donne con iperplasia surrenale congenita (CAH) e sociosessualità promiscua
  - Per qulai variabili appaiare i controlli? Primi candidati: età, livello di istruzione, orientamento politico, religiosità, status relazionale...ma ha senso questa insalata causale?



Andare su DAGitty.net

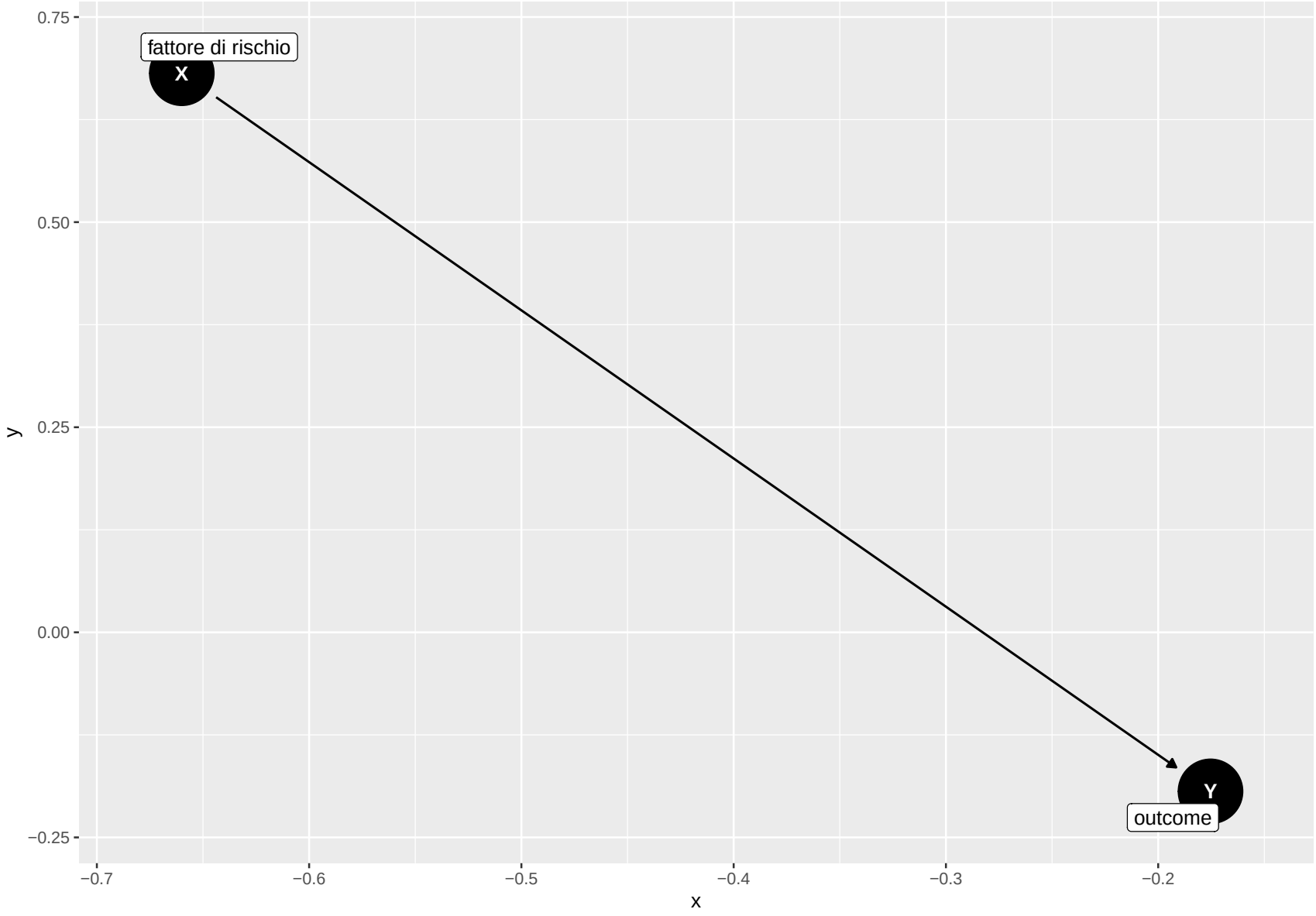
https://dagitty.net/dags.html?id=EP9fXebg



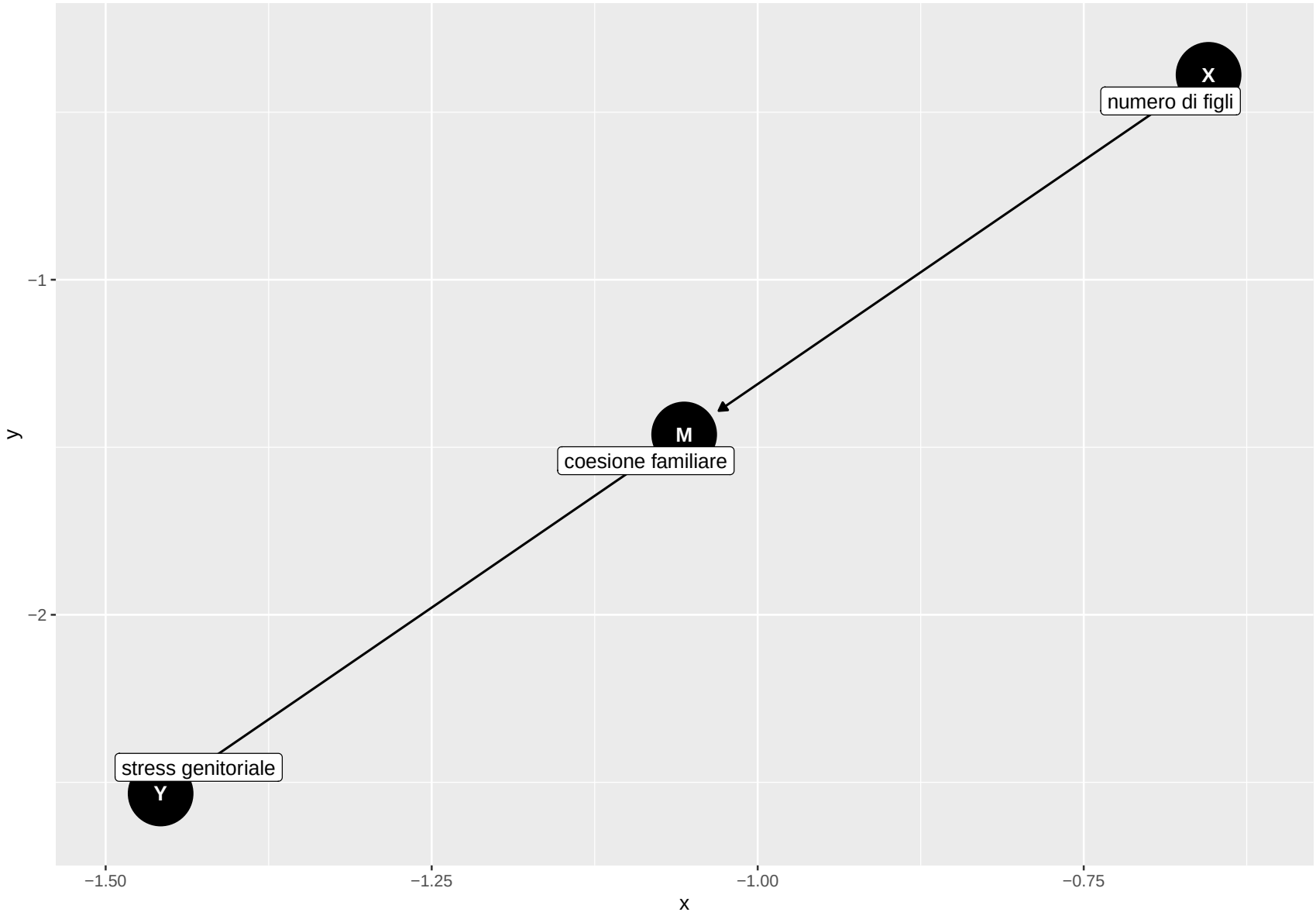
DAG con ggdag() 1

```
dag1<-dagify(Y~X,

             exposure = "X",

             outcome = "Y",

             labels = c("X"="fattore di rischio",
                        "Y"="outcome"))

dag1<-ggdag(dag1, use_labels = "label", text = TRUE)
```

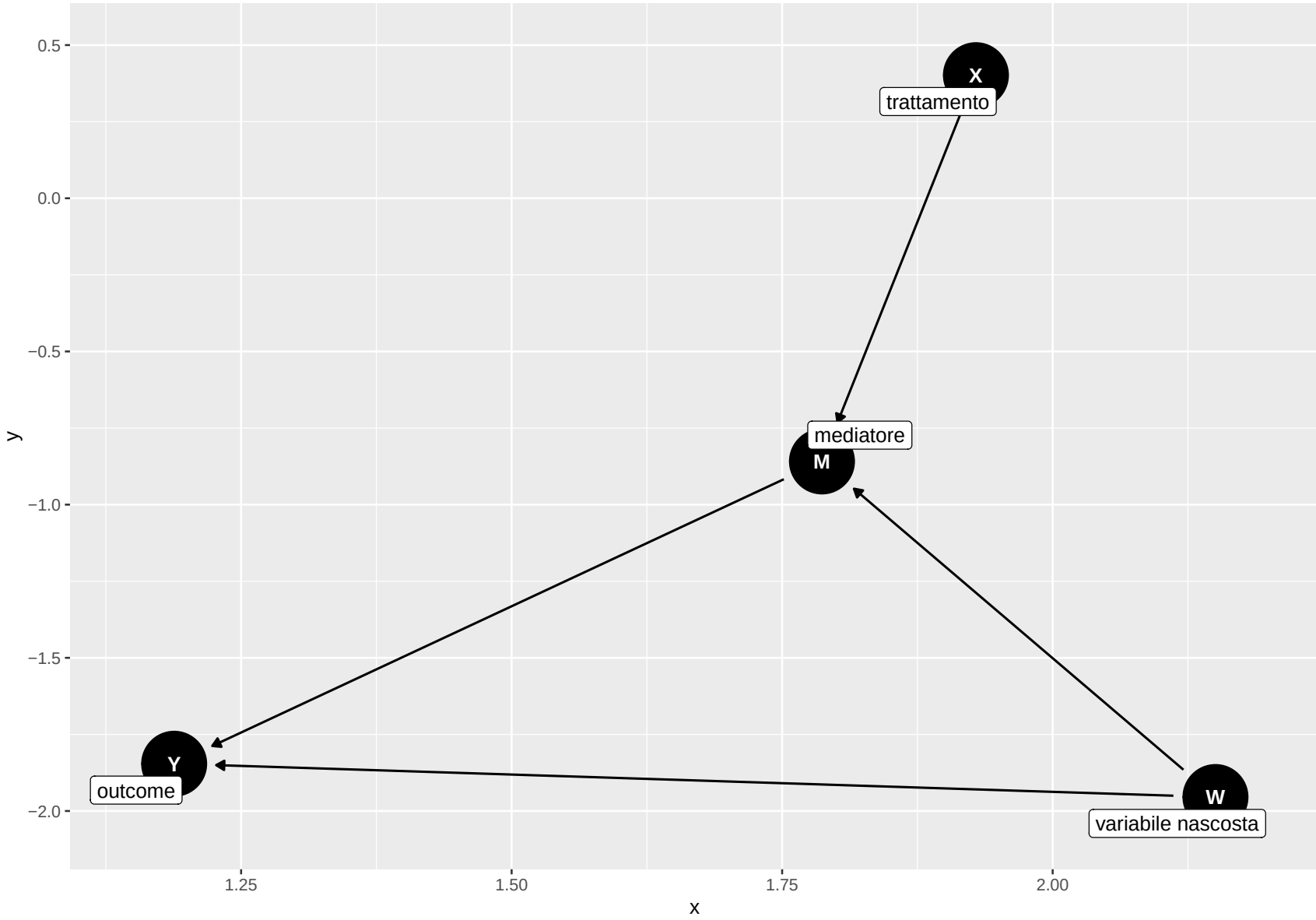
DAG con ggdag (1)



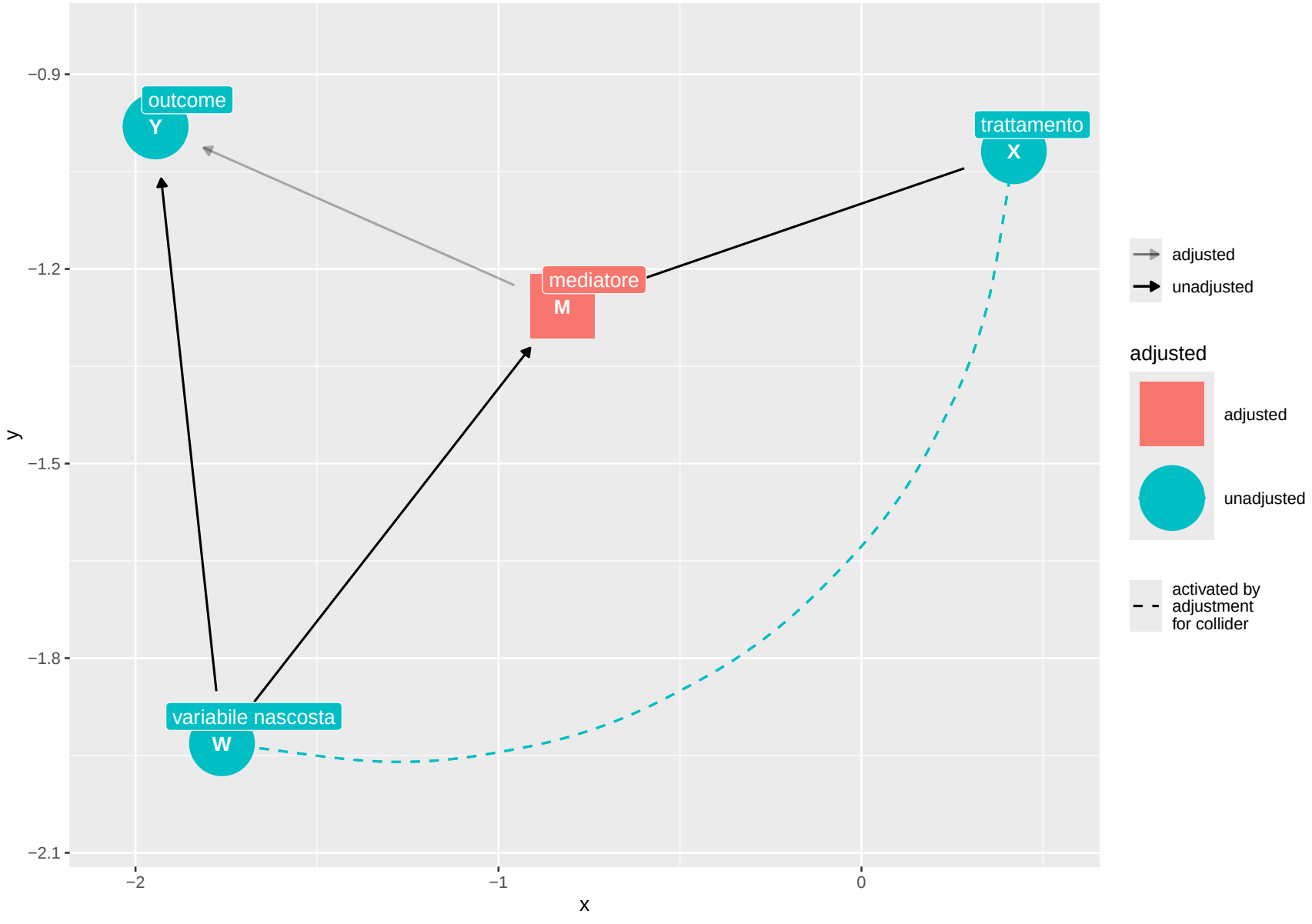
DAG con ggdag (2)



DAG con ggdag (3)

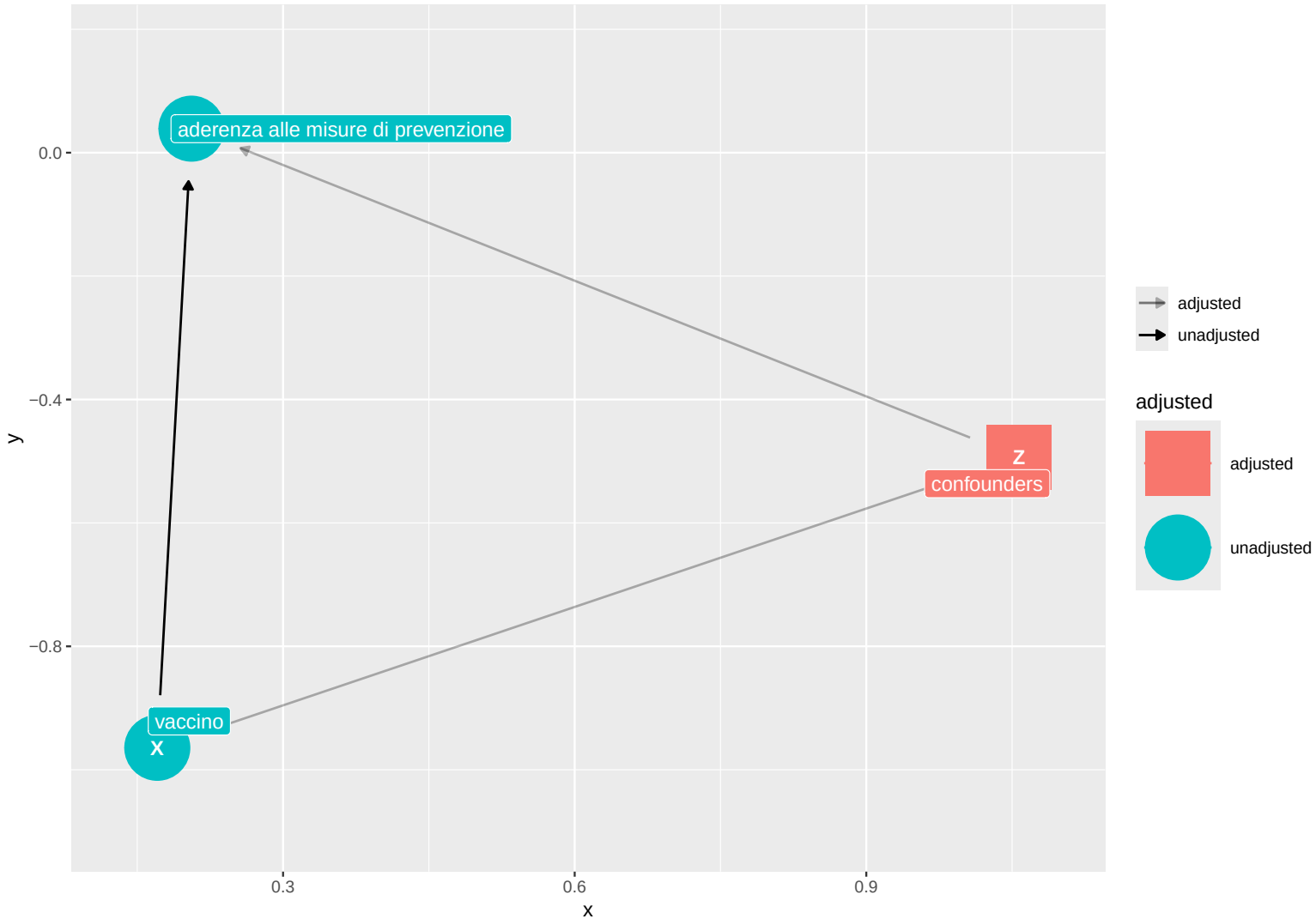


DAG con ggdag (3)



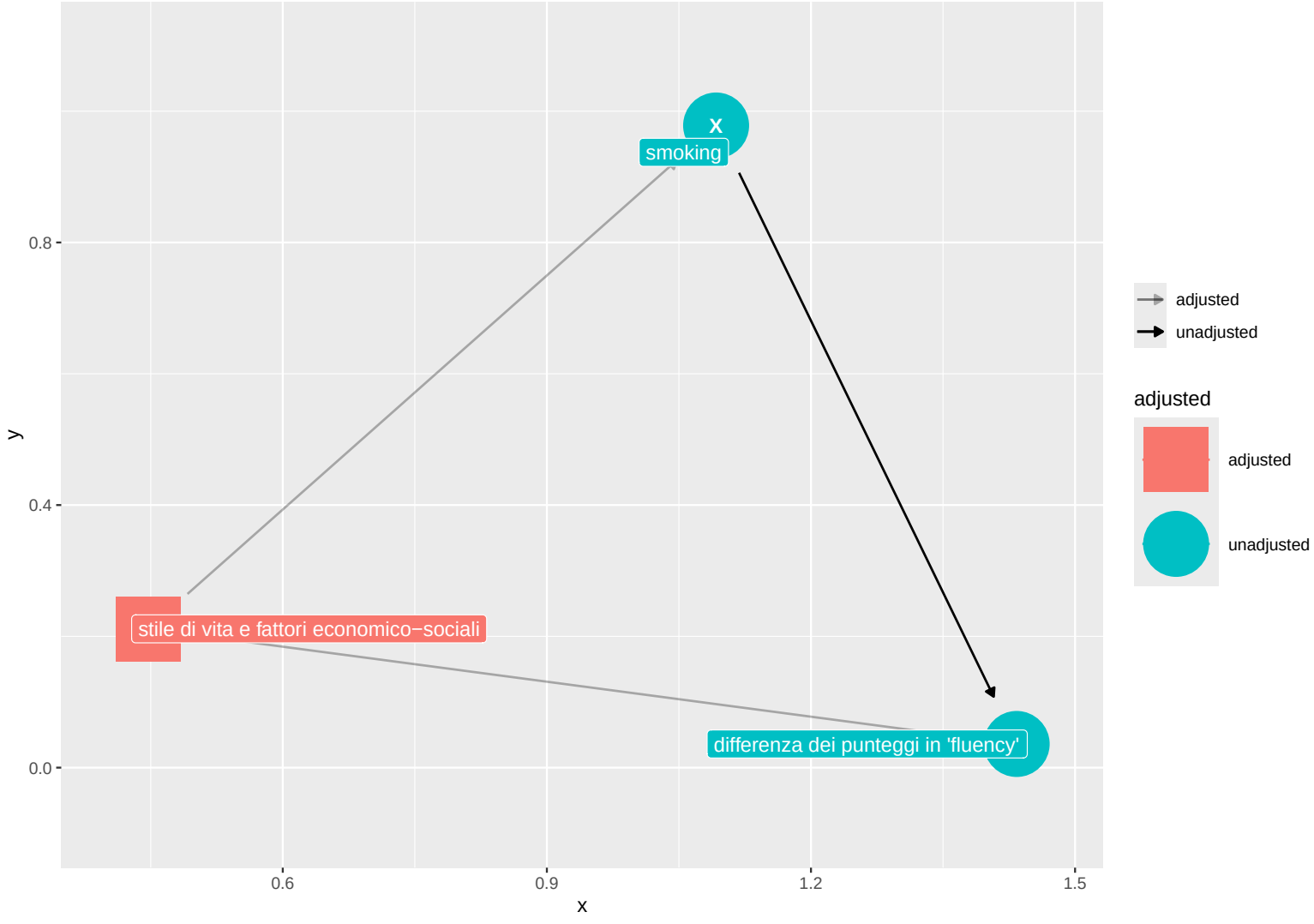
DAG con ggdag (4)

Influenza dello **status vaccinale** sull'**adesione alle misure di prevenzione contro la diffusione del COVID-19** controllando l'effetto di **possibili confounder**



DAG con ggdag (5)

Effetto del **fumo** sul **decadimento cognitivo** controllando per **stile di vita e fattori economico-sociali**



Studi quasi sperimentali (1)

Disegno quasi sperimentale: ricreare una condizione che si avvicini il più possibile alla randomizzazione

Possibile tramite **metodi di aggiustamento delle variabili confondenti**

1. aggiustamento additivo
2. tecniche di bilanciamento (es. basate sul propensity score)

Studi quasi sperimentali (2)

1) Aggiustamento additivo

Addizione delle covariate all'interno, ad esempio, di un modello di regressione lineare

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} ... + \beta_k X_{ki} + \epsilon_i$$

Studi quasi sperimentali (3)

2) Tecniche basate sul propensity score

Il **propensity score**

- esprime la probabilità che ogni individuo ha di ricevere il trattamento, sulla base del profilo di covariate che presenta
- si basa sull'**assunto di ignorabilità forte**
- viene prima stimato e poi applicato al modello dello studio

Studi quasi sperimentali (4)

2) Tecniche basate sul propensity score: stima

Stima del propensity score può avvenire in più modi (parametrici e non parametrici)

Un esempio di metodo parametrico è la **regressione logistica multipla**

$$\pi_i = Pr(Y = 1|X = x_i) = \lambda(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} ... + \beta_k X_{ki})$$

Studi quasi sperimentali (5)

2) Tecniche basate sul propensity score: applicazione

Un possibile metodo di applicazione è l'***inverse probability of treatment weights*** (**IPTW**)

Ad ogni individuo si associa un peso:

- peso dato agli individi trattati $IPTW = 1/\rho_i$
- peso dato agli individui di controllo $IPTW = 1/(1-\rho_i)$

Studio d'esempio (1)

Studio riguardo l'effetto del fumo di sigaretta sul decadimento cognitivo

- X = essere fumatore alla baseline
- Y = differenza nei punteggi di fluenza verbale nel corso di 10 anni
- confounder = variabili suggerite dall'APA + ... ?

Dati ottenuti dal database europeo SHARE relativi a soggetti italiani

- campione N = 33'525
- individui maschili N = 14'675
- individui femminili N= 18'850

Studio d'esempio (2)

I modelli di regressione che intendiamo confrontare sono corretti

1. con **aggiustamento additivo** delle covariate
2. con **aggiustamento additivo** delle covariate + aggiustamento tramite **propensity score** (applicando il metodo IPTW)

Studio d'esempio (3)

Rispettivamente, i risultati dei due modelli riportano che

1. l'effetto del fumo sul decadimento cognitivo **non è significativo**
2. l'effetto del fumo sul decadimento cognitivo **è significativo**

Possibile spiegazione della differenza nel **bias di selezione**: i fumatori tendono ad evitare le survey

Studio d'esempio (4)

Applicazione della correzione tramite propensity score:

- miglior riduzione di bias rispetto all'aggiustamento additivo da solo
- risultati in linea con la letteratura

Limiti dello studio (e dell'utilizzo di tecniche di aggiustamento in generale):

- assunto di ignorabilità forte

Take-home messages

Evitare causal salads

Thinking before regressing: riportare anche i DAG !

Per una miglior causal inference

- aggiustamento additivo con campioni grandi
- tecniche di bilanciamento con campioni piccoli e tante covariate

Bibliografia

Del Giudice, M., & Gangestad, S. W. (2021). A traveler's guide to the multiverse: Promises, pitfalls, and a framework for the evaluation of analytic decisions. *Advances in Methods and Practices in Psychological Science, 4,* 1-15.

Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological methods*, *15*(3), 234.

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in methods and practices in psychological science*, *1*(1), 27-42.

Sità, L., Caserotti, M., Zamparini, M., Lotto, L., de Girolamo, G., & Girardi, P. (2024). Impact of COVID-19 vaccination on preventive behavior: The importance of confounder adjustment in observational studies. *PloS one*, *19*(11), e0313117.

Appendice (1)

Confronto dell'aggiustamento additivo con e senza correzione tramite propensity score

```
## Analisi binomiale senza propensity score
fit1_full<-glm(difference_in_fluency_bin~smoking+drinking
            +physical_inactivity+age+gender+isced+marital_status
            +number_of_children+economic_status+home+job_status
            +health_status+gali+number_of_chronic_diseases+bmi
            +mobility+depression_scale, data=db_sel1,
            family=binomial) # modello di regressione logistica

# stepAIC backward: dal modello con tutte le variabili indicate,
# deseleziono quelle che non hanno un effetto sul decadimento della fluency
fit1_final<-stepAIC(fit1_full,direction = "backward")
t1<-tbl_regression(fit1_final,exponentiate = TRUE)

## Analisi binomiale con propensity score
# Stima del PS tramite metodo IPTW
mod_ps_smoking<-glm(I(smoking=="Yes")~physical_inactivity+drinking+age
                    +gender+isced+marital_status+number_of_children
                    +economic_status+home+job_status+health_status
                    +gali+number_of_chronic_diseases+bmi+mobility
                    +depression_scale,data=db_sel1,family=binomial)
mod_ps_smoking %>% tbl_regression(exponentiate = TRUE)
```

Appendice (2)

Confronto dell'aggiustamento additivo con e senza correzione tramite propensity score

```
# Stima delle probabilità
ps_smo<-predict(mod_ps_smoking,type="response")
# Calcolo dei pesi
db_sel1$pesi_smo<-0
db_sel1$pesi_smo[db_sel1$smoking=="Yes"]<-1/ps_smo[db_sel1$smoking=="Yes"]
db_sel1$pesi_smo[db_sel1$smoking=="No"]<-1/(1-ps_smo[db_sel1$smoking=="No"])

# Modello binomiale corretto con i pesi del PS
mod_fluency_ps_smo<-glm(difference_in_fluency_bin~smoking+drinking
            +physical_inactivity+age+gender+isced+marital_status
            +number_of_children+economic_status+home+job_status+health_status
            +gali+number_of_chronic_diseases+bmi+mobility+depression_scale,
            data=db_sel1,family=binomial,weights = pesi_smo)

t2<-tbl_regression(mod_fluency_ps_smo,exponentiate = TRUE)
```