COURSE

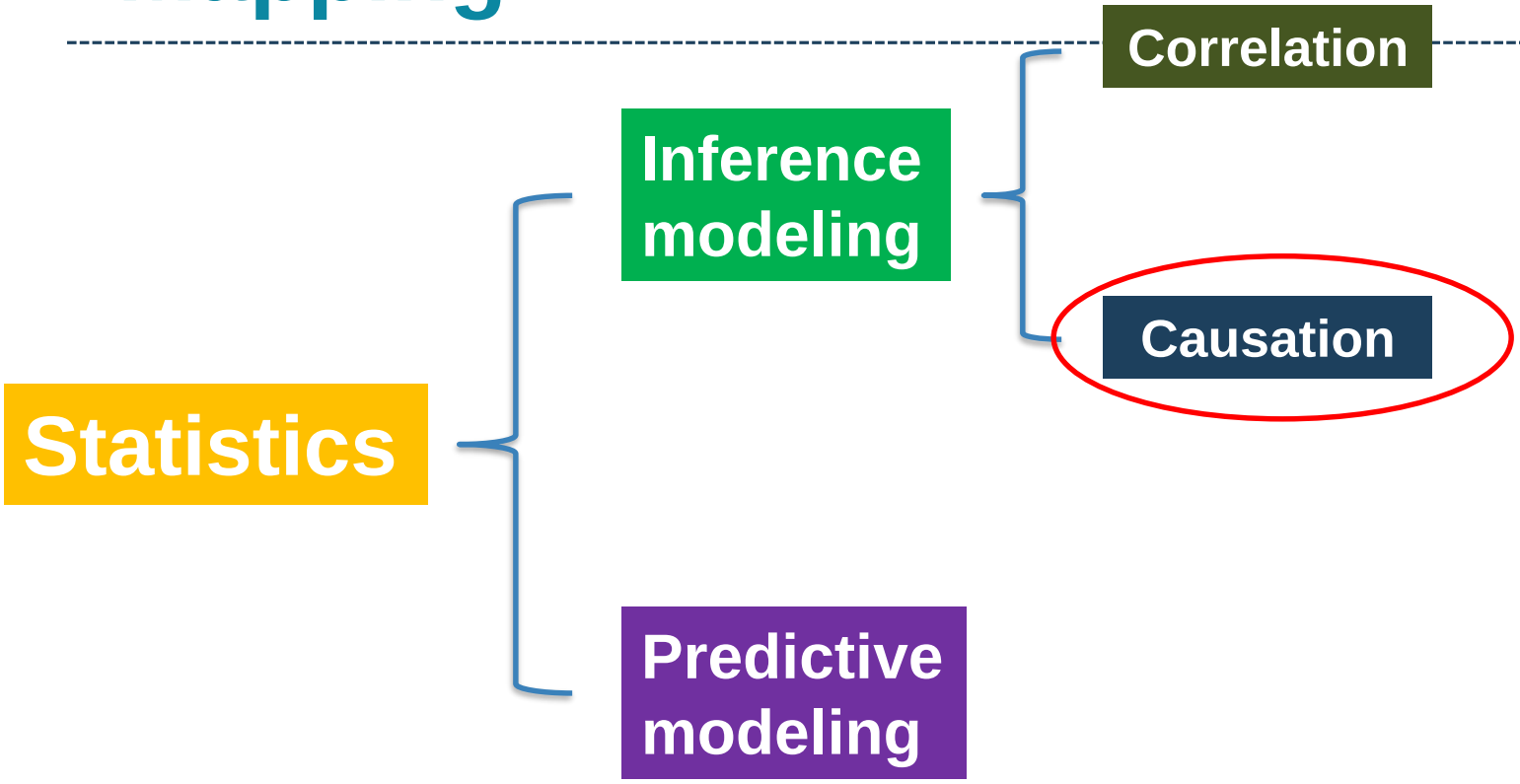
Directed Acyclic Graphs

Phần 1: Vấn đề Confounding

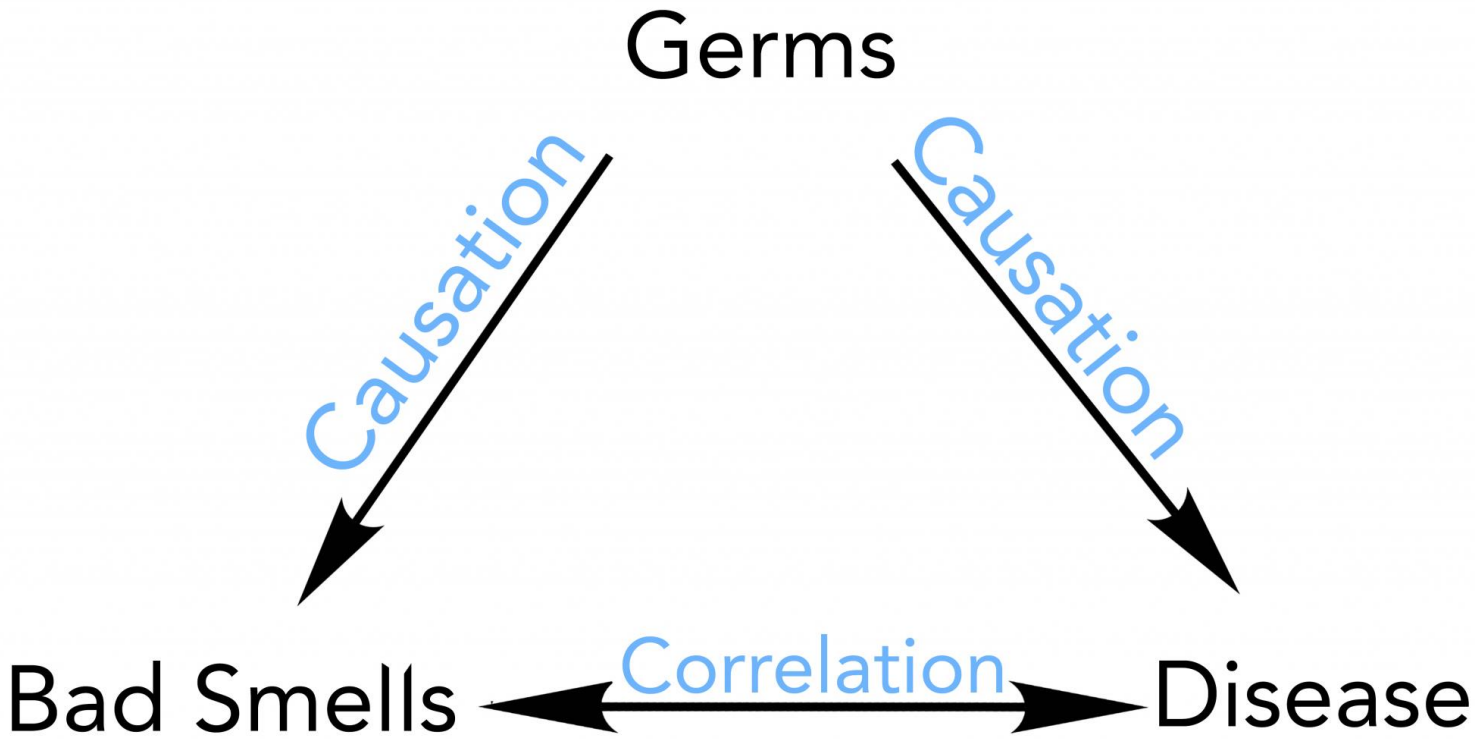
Lớp Phân tích thống kê cơ bản

Khương Quỳnh Long
Hà Nội, 06-08/06/2020

Mapping


Correlation
Inference modeling
Causation
Statistics
Predictive modeling

Correlation ≠ Causation



Motivation

- Các nhà dịch tễ học muốn tìm hiểu sự liên quan giữa phơi nhiễm và biến cố
- Nghiên cứu thực nghiệm (experimental study) cung cấp bằng chứng mạnh về mối liên hệ nhân quả. Tuy nhiên, khó/không khả thi, tốn kém, vấn đề y đức…

➔ Suy luận mối liên hệ nhân quả từ dữ liệu quan sát (observational/non-experimental data)

Motivation

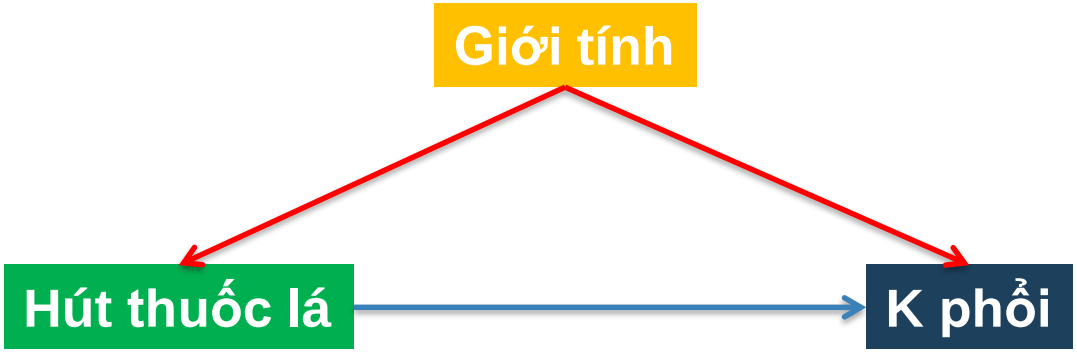
- Suy luận mối liên hệ nhân quả từ dữ liệu quan sát cần giải quyết vấn đề gây nhiễu (Confounding)
- Tất cả các phương pháp “khử nhiễu” (như propensity score, stratification, adjustment..), trước tiên cần xác định các yếu tố gây nhiễu (Confounders)
- Directed Acyclic Graphs (DAGs) là phương pháp giúp xác định một cách hệ thống các yếu tố gây nhiễu này

Nội dung

1. Giới thiệu DAGs
2. Sử dụng DAGs để xác định nhóm tối tiểu các biến gây nhiễu cần hiệu chỉnh (minimal sufficient adjustment set)
3. Giới thiệu phần mềm DAGitty
4. Ví dụ

Confounder là gì?

- Yếu tố bên ngoài, tác động đồng thời lên yếu tố phơi nhiễm và kết cuộc
- Không nằm trên đường từ phơi nhiễm → kết cuộc



Chiến lược xác định yếu tố cần hiệu chỉnh

- Sử dụng các phương pháp chọn biến tự động, như **stepwise selection**

✓ Ưu điểm: nhanh, tốn ít thời gian thực hiện, không cần suy luận

✓ Vấn đề:

- Không phải biến nào được chọn cũng là biến gây nhiễu
- Biến gây nhiễu cần hiệu chỉnh không được chọn

Chiến lược xác định yếu tố cần hiệu chỉnh

- So sánh adjusted và unadjusted effect estimates

✓ So sánh effect estimates trước và sau hiệu chỉnh, nếu thay đổi (ví dụ 10% …) thì giữ lại biến số đó

✓ Ngụ ý là tất cả các biến có ảnh hưởng tới effect estimates thì cần được hiệu chỉnh

✓ Vấn đề: không phải tất cả các yếu tố làm thay đổi effect estimates đều là yếu tố nhiễu cần hiệu chỉnh (vd overadjustment)

Chiến lược xác định yếu tố cần hiệu chỉnh

- Kiểm tra các tiêu chuẩn của biến số gây nhiễu

✓ 3 tiêu chuẩn

- Vấn đề:

  - Xem xét các biến gây nhiễu một cách rời rạc
  - Khó xác định khi nhiều biến số, các biến số có liên hệ với nhau
  - Tất cả các yếu tố nhiễu cần được hiệu chỉnh trong mô hình?

Ví dụ 1

- Ảnh hưởng của **hút thuốc lá** tới **sảy thai**
- Cần hiểu chỉnh cho **tiền sử sảy thai** hay không?

Ví dụ 2

- Xác định ảnh hưởng của **hút thuốc lá** tới **huyết áp tâm thu**
- Các biến số thu thập bao gồm: tuổi, giới, BMI, tình trạng kinh tế xã hội, uống rượu, tình trạng dinh dưỡng, Cholesterol, hoạt động thể lực
- Hiệu chỉnh cho biến số nào?

Ví dụ 3. Birth weight paradox

American Journal of Epidemiology
Copyright © 2006 by the Johns Hopkins Bloomberg School of Public Health
All rights reserved; printed in U.S.A.

Vol. 164, No. 11
DOI: 10.1093/aje/kwj275
Advance Access publication August 24, 2006

Practice of Epidemiology

The Birth Weight "Paradox" Uncovered?

Sonia Hernández-Díaz[1,2], Enrique F. Schisterman[3], and Miguel A. Hernán[1]

[1] Department of Epidemiology, Harvard School of Public Health, Boston, MA.
[2] Slone Epidemiology Center, Boston University, Boston, MA.
[3] Epidemiology Branch, National Institute of Child Health and Human Development, Bethesda, MD.

Birth weight paradox

MATERIALS AND METHODS

We identified all infants born alive ($n = 4{,}115{,}494$) in the United States in 1991 through the national linked birth/infant-death data sets assembled by the National Center for Health Statistics (11). These data contain information on dates and causes of death, birth weight, maternal smoking, and other medical and sociodemographic characteristics systematically recorded on US birth certificates. The infant mortality rate was defined as number of deaths within the first year of life per 100,000 livebirths. LBW was defined as birth weight below 2,500 g. We excluded from the analyses infants with missing information on birth weight or maternal smoking; California was excluded because of lack of smoking data. The final study population included 3,001,621 livebirths.

Birth weight paradox

As figure 1 shows, infants born to women who smoked had a lower average birth weight (mean = 3,145 g; prevalence of LBW = 11.4 percent) than infants born to nonsmokers (mean = 3,370 g; prevalence of LBW = 6.4 percent). The infant mortality rate was 1,235 per 100,000 livebirths for infants born to smokers and 805 per 100,000 livebirths for infants born to nonsmokers. Compared with nonsmokers, the infant mortality rate ratio for smokers was 1.55 (95 percent confidence interval (CI): 1.50, 1.59). This rate ratio changed to 1.09 (95 percent CI: 1.05, 1.12) upon adjustment for birth weight.

Birth weight paradox

Infant mortality increased as birth weight decreased among infants born to both smokers and nonsmokers (figure 2). However, the birth-weight-specific mortality rate curve of infants born to smokers crossed over that of infants born to nonsmokers around the interval of 2,000–2,250 g. For babies weighing less than 2,000 g at birth, mortality was higher among infants born to nonsmokers. The infant mortality rate ratio for exposed infants versus nonexposed infants was 0.79 (95 percent CI: 0.76, 0.82) among LBW infants and 1.80 (95 percent CI: 1.72, 1.88) among infants with higher birth weights.

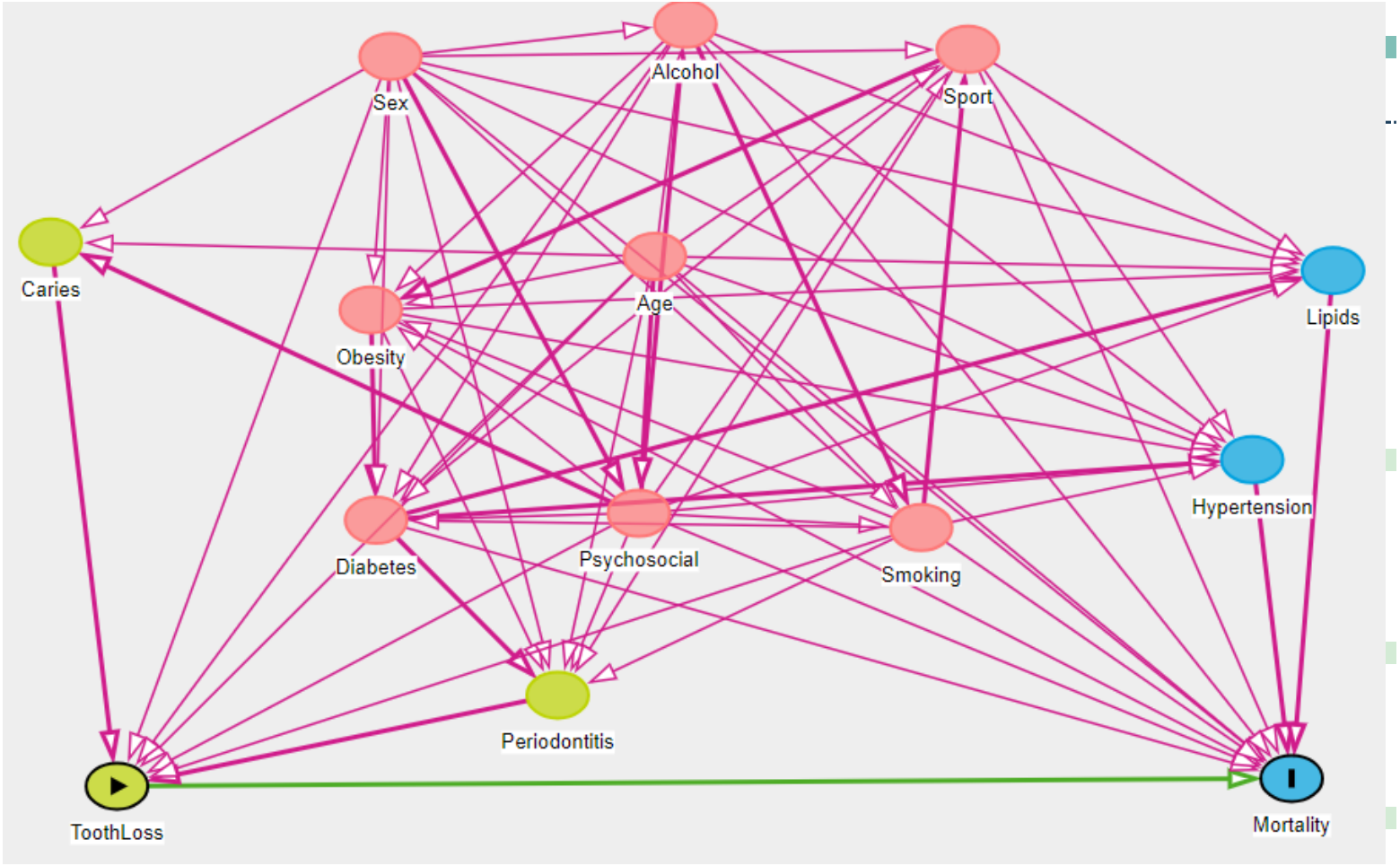
Birth weight paradox

- Không hiệu chỉnh: RR 1.55 (1.50 – 1.59)
- Hiệu chỉnh cho biến cân nặng: RR 1.09 (1.05 - 1.12)
- Cân nặng < 2000 g: RR 0.79 (0.76 - 0.82)
- Cân nặng > 2000g: RR 1.80 (1.72 - 1.88)

➔ Ở trẻ sơ sinh nhẹ cân (bw <2000g), mẹ hút thuốc lá làm giảm nguy cơ tử xong trẻ?

- Còn cách giải thích nào khác?

Directed Acyclic Graphs


Alcohol
Sex
Sport
Caries
Age
Lipids
Obesity
Hypertension
Diabetes
Psychosocial
Smoking
Periodontitis
ToothLoss
Mortality

Directed Acyclic Graphs

- Bản đồ thể hiện liên hệ nhân quả giữa các biến
- **Directed**: mũi tên 1 chiều thể hiện nguyên nhân trực tiếp
- **Acyclic**: Không có tính chu kỳ

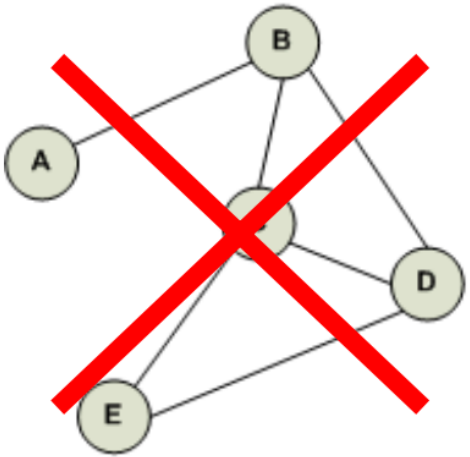
✓ A → B → A : Không cho phép

✓ A → B → C : Cho phép

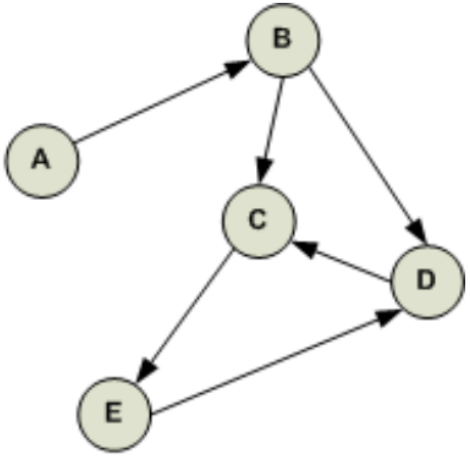
- **Graph**: Những node biểu diễn biến quan sát được (observed variables) và cả biến không quan sát được (unobserved/latent variables)

> All common causes must be shown on a DAG or it is not considered causal

Directed Acyclic Graphs

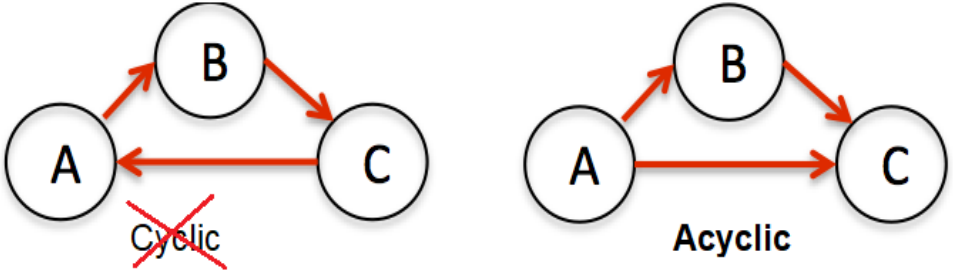


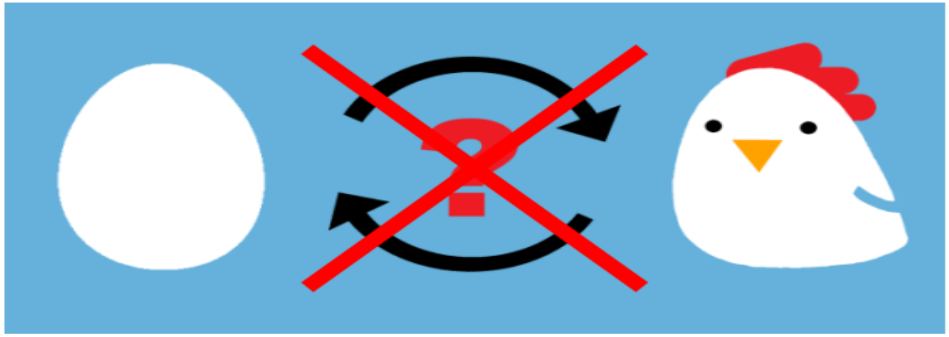
Undirected Graph



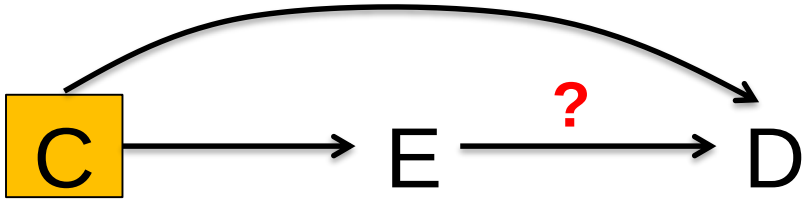
Directed Graph

Directed Acyclic Graphs

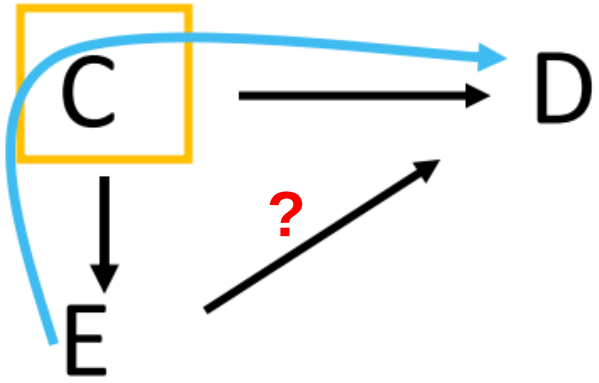




DAG - confounding



Tác động của E lên D là bao nhiêu ?



- $D = \alpha + \beta * E$
- ✓ $D = \alpha + \beta_1 * E + \beta_2 * C$

Thuật ngữ

Parents, Children

- Parents: một biến trực tiếp gây ra biến khác mà không thông qua biến trung gian (mediator)
- Child/children: Biến chịu ảnh hưởng trực tiếp từ một biến khác mà không thông qua trung gian
- A → B → C : A là parent của B, B là child của A

Ancestor, Descendant

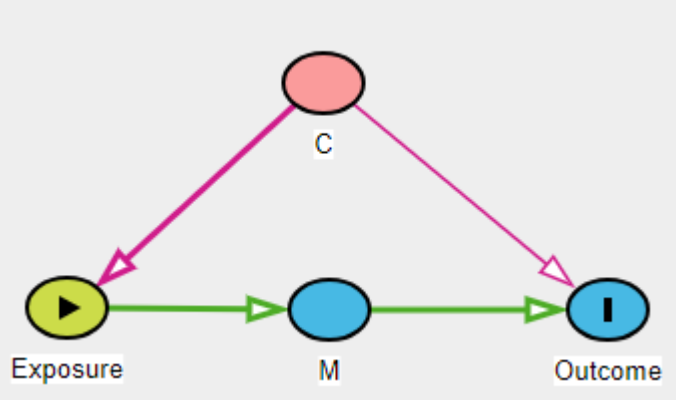
- Ancestor: một biến gây ra gián tiếp biến khác thông qua mediator
- Descendant: một biến chịu ảnh hưởng gián tiếp của biến khác thông qua mediator
- A → B → C:

✓ A là ancestor của C (B: mediator)

✓ C là descendant của A

Path

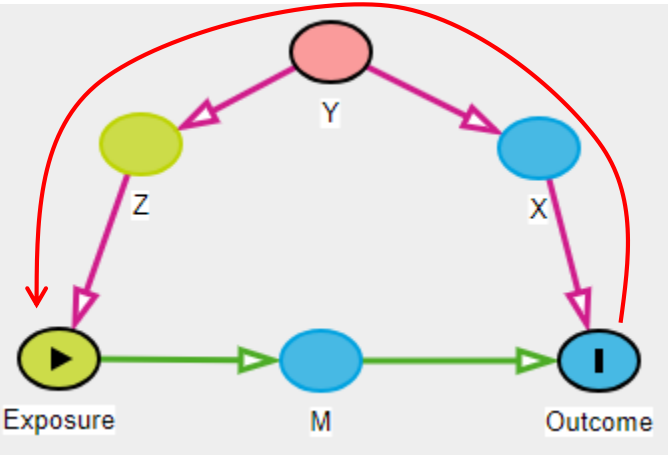
- Path: một chuỗi mũi tên nối 2 hay nhiều “node” (biến)
- Directed path: đường mũi tên cùng hướng theo chiều nguyên nhân → kết quả



Directed path:
Exposure → M → Outcome

Backdoor Path

- Khái niệm quan trọng của DAGs
- Đường đi ngược lại từ **kết cuộc** và hướng đến **nguyên nhân**
- Chính là đường thể hiện confounding

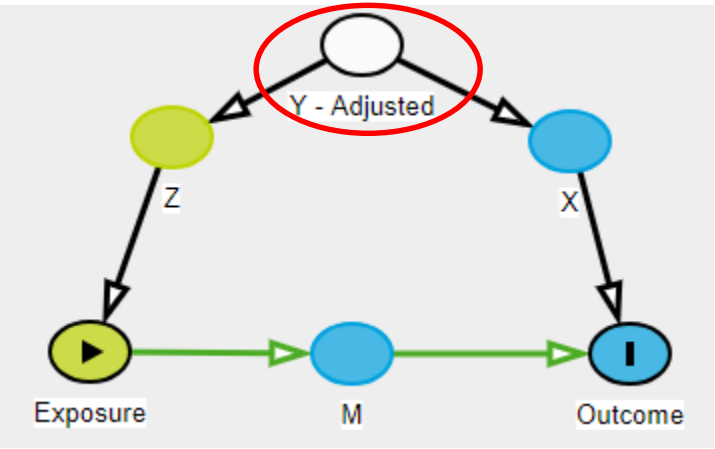


Outcome ←X← Y
→Z→Exposure

Outcome ← M ← Exposure
có phải backdoor?

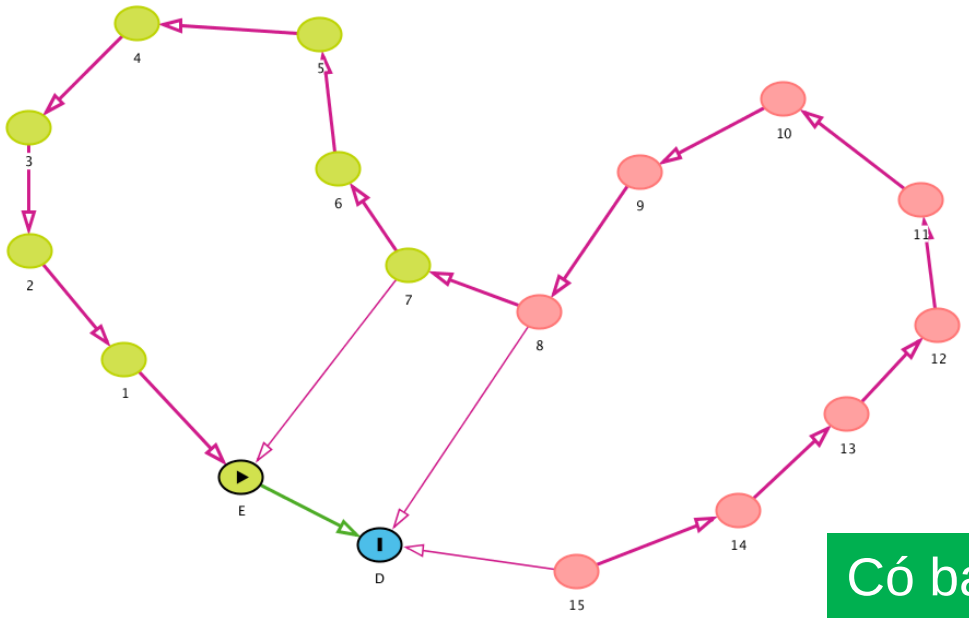
Backdoor Path

- Là đường thể hiện confounding → cần "block"
- Chỉ cần adjust 1 node trên đường backdoor thì toàn bộ đường backdoor đó bị block



$$\text{Outcome} = \alpha + \beta_1 * \text{Exposure} + \beta_2 * (\text{X or Y or Z})$$

Backdoor Path

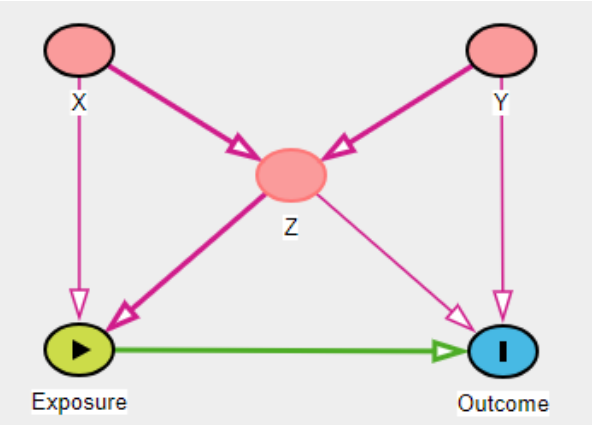


Có bao nhiêu đường backdoor?

Hiệu chỉnh biến nào để “tiết kiệm” nhất?

Collider

- Một node trở thành Collider nếu trên path đó có 2 mũi tên hướng đến nó (children chung của 2 nodes)
- Một node có thể là collider của path này nhưng không phải collider của path khác

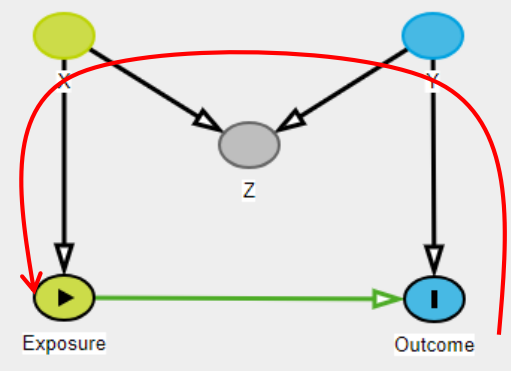


Outcome ← Y → Z ← X → Exposure
Z là collider

Outcome ← Z ← X → Exposure
Z không phải collider

Collider

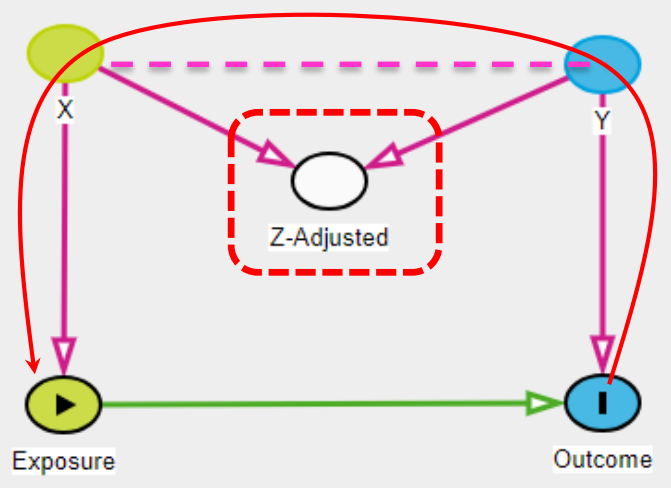
- Trên đường backdoor, nếu xuất hiện collider thì đường backdoor tự động block (i.e., không cần hiệu chỉnh)
- Nếu hiệu chỉnh cho collider: Backdoor đang blocked sẽ unblocked → xuất hiện confounding



Backdoor:
Outcome ← Y → Z ← X → Exposure

Tuy nhiên backdoor đang bị block → Không có confounding giữa **Exposure → Outcome**

Collider

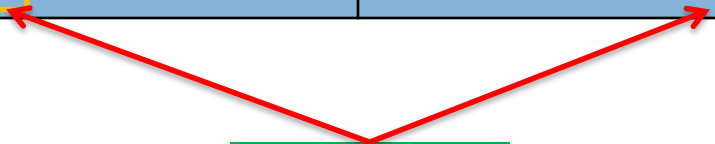


- Khi hiệu chỉnh cho Z (collider), backdoor từ **Outcome** ← **Y** → **Z** ← **X** → **Exposure** đang blocked → unblocked

➔ xuất hiện confounding cho Exposure → Outcome

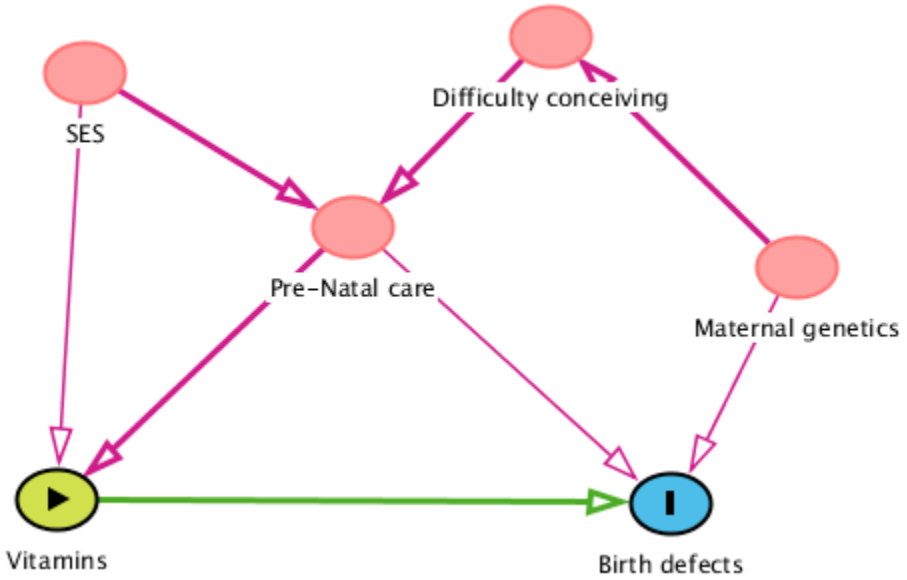
Tóm tắt

Backdoor Open	Backdoor Blocked
→ A →	→ A →
← A →	← A →
← A ←	← A ←
→ A ←	→ A ←



Ví dụ

- Có tất cả bao nhiêu đường backdoor?
- Hiểu chỉnh cho biến nào?



Minimal sufficient adjustment set (MSAS)

- Thứ cần tìm trong thực hành!
- Tập hợp những biến số tối thiểu để loại bỏ ảnh hưởng của nhiễu
- Tối thiểu vì:

✓ Nếu hiệu chỉnh thiếu 1 biến → sai lệch có thể xảy ra

✓ Nếu thêm 1 biến → không ảnh hưởng kết quả

Xác định MSAS

- Nguyên tắc: hiệu chỉnh ít biến nhất nhưng block được tất cả các backdoor

1. Xác định tất cả các đường backdoor
2. Hiệu chỉnh cho biến có khả năng block nhiều đường backdoor nhất (thường là giao điểm của nhiều backdoor)
3. Xác định lại các backdoor nếu biến hiệu chỉnh là collider
4. Tiếp tục hiệu chỉnh cho đến khi tất cả các backdoor bị block

Chú ý

- Có thể có nhiều MSAS khác nhau cho 1 model
- Lựa chọn MSAS phù hợp và khả thi

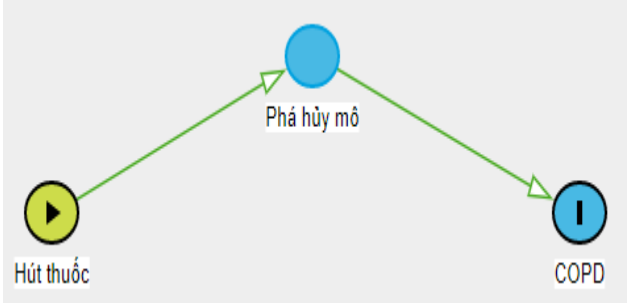
Các bước xây dựng DAGs

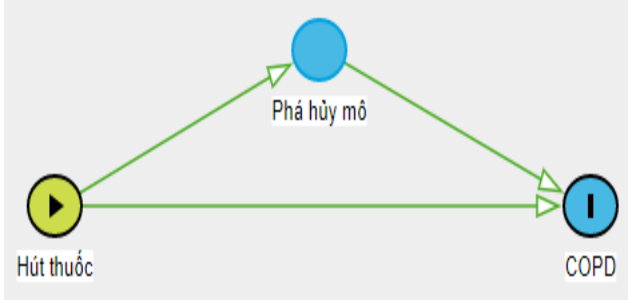
Các bước

1. Xác định biến **Exposure** & **Outcome**
2. Thêm chiều(s) tác động **Exposure** → **Outcome**
3. Thêm tất cả các nguyên nhân của **Outcome**
4. Xác định thêm nếu biến nguyên nhân của **Outcome** có là nguyên nhân của **Exposure**
5. Xác định mối liên hệ nhân quả giữa các biến trong DAG
6. Kiểm tra nếu thiếu các ancestors phổ biến cho từng cặp biến số (dù có số liệu hay không)
7. Xác định các backdoor path, collider, MSAS…

Chú ý

1. Trong DAGs, thiếu chiều mũi tên đồng nghĩa chắc chắn không có mỗi liên hệ nhân quả, nếu nghi ngờ → cân nhắc thêm mũi tên
2. Trong chuỗi các biến số, kiểm tra nếu có thể còn đường tác động nào khác





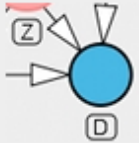
Chú ý

3. Tất cả những nguyên nhân lên **Outcome** và **Exposure** cần được cân nhắc, cho dù có số liệu hay không
4. Tất cả các đường backdoor cần phải được block
5. Cẩn thận khi Adjust cho collider
6. **Có thể có nhiều DAGs cho một vấn đề, không có DAGs "đúng"**

Phần mềm DAGitty

DAGitty: http://www.dagitty.net/

Welcome to DAGitty!

Launch	Download	Learn	Code
 Launch DAGitty online in your browser	 Download DAGitty's source for offline use	 Learn more about DAGs and DAGitty	 The R package "dagitty" is available on CRAN or github

What is this?

DAGitty is a browser-based environment for creating, editing, and analyzing causal models (also known as directed acyclic graphs or causal Bayesian networks). The focus is on the use of causal diagrams for minimizing bias in empirical studies in epidemiology and other disciplines. For background information, see the "learn" page.

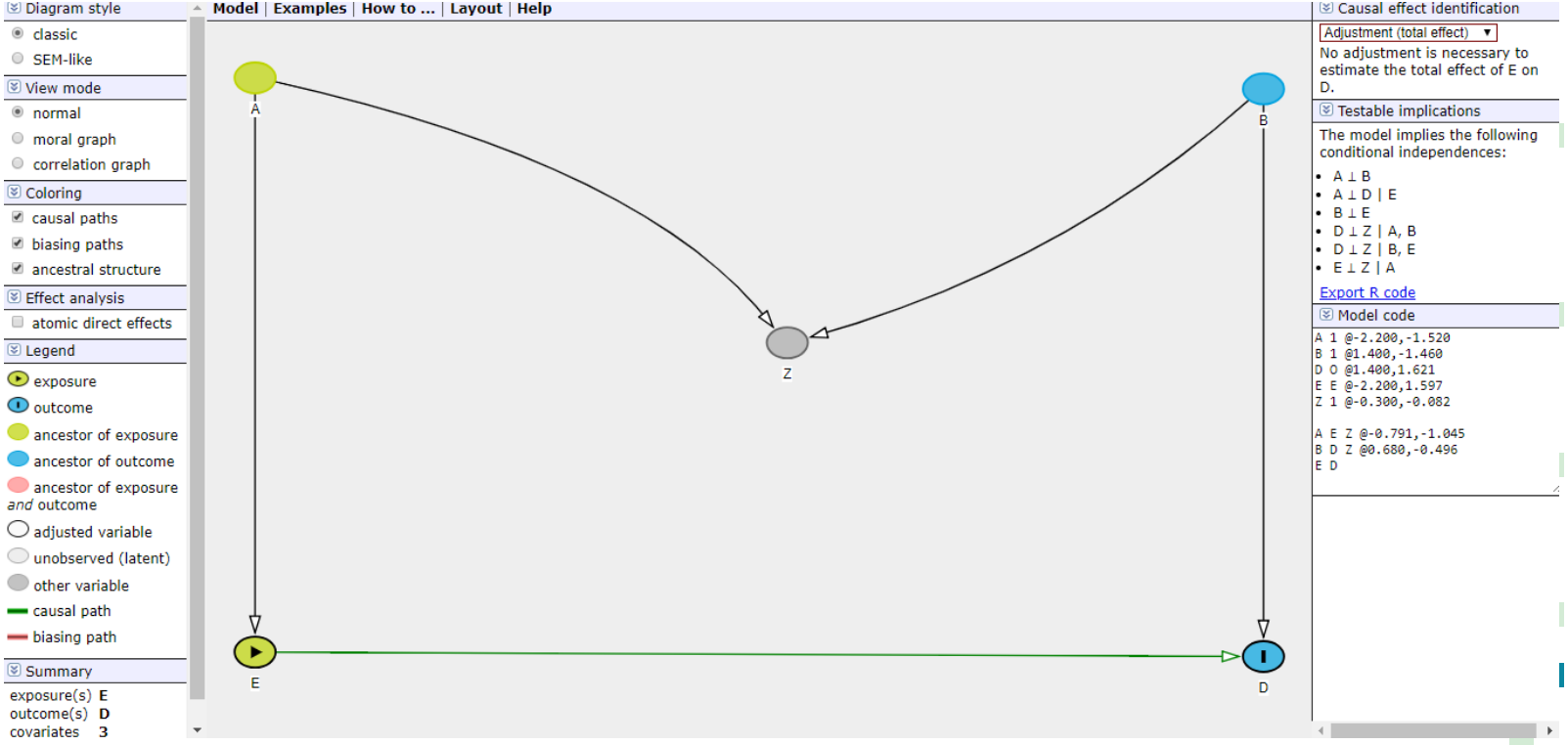


Diagram style
classic
SEM-like
View mode
normal
moral graph
correlation graph
Coloring
causal paths
biasing paths
ancestral structure
Effect analysis
atomic direct effects
Legend
exposure
outcome
ancestor of exposure
ancestor of outcome
ancestor of exposure and outcome
adjusted variable
unobserved (latent)
other variable
causal path
biasing path
Summary
exposure(s) E
outcome(s) D
covariates 3
Model | Examples | How to ... | Layout | Help
A
B
Z
E
D
Causal effect identification
Adjustment (total effect)
No adjustment is necessary to estimate the total effect of E on D.
Testable implications
The model implies the following conditional independences:
A ⊥ B
A ⊥ D | E
B ⊥ E
D ⊥ Z | A, B
D ⊥ Z | B, E
E ⊥ Z | A
Export R code
Model code
A 1 @-2.200,-1.520
B 1 @1.400,-1.460
D O @1.400,1.621
E E @-2.200,1.597
Z 1 @-0.300,-0.082
A E Z @-0.791,-1.045
B D Z @0.680,-0.496
E D

Model | **Examples** | **How to ...** | **Layout** | **Help**

Diagram style
- classic
- SEM-like

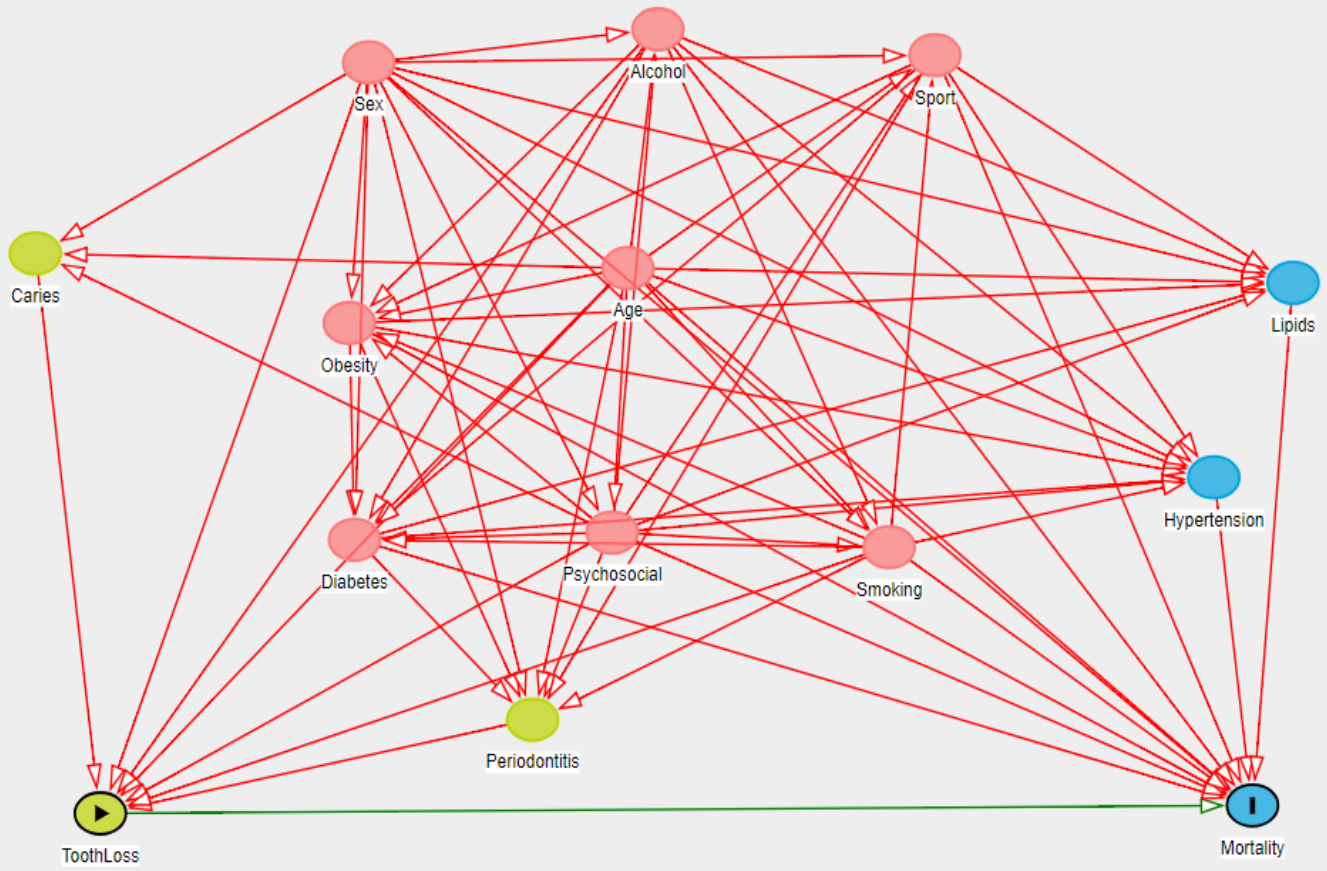
View mode
- normal
- moral graph
- correlation graph

Coloring
- causal paths
- biasing paths
- ancestral structure

Effect analysis
- atomic direct effects

Legend
- exposure
- outcome
- ancestor of exposure
- ancestor of outcome
- ancestor of exposure *and* outcome
- adjusted variable
- unobserved (latent)
- other variable
- causal path
- biasing path

Summary
exposure(s) **ToothLoss**
outcome(s) **Mortality**
covariates **12**



Causal effect identification
Adjustment (total effect)

Minimal sufficient adjustment sets for estimating the total effect of ToothLoss on Mortality:

- Age, Alcohol, Diabetes, Obesity, Psychosocial, Sex, Smoking, Sport
- Age, Alcohol, Periodontitis, Psychosocial, Sex, Smoking

Testable implications
The model implies the following conditional independences:

- ToothLoss ⊥ Diabetes | Age, Alcohol, Periodontitis, Psychosocial, Sex, Smoking
- ToothLoss ⊥ Obesity | Age, Alcohol, Periodontitis, Psychosocial, Sex, Smoking
- ToothLoss ⊥ Hypertension | Age, Alcohol, Diabetes, Obesity, Psychosocial, Sex, Smoking, Sport
- ToothLoss ⊥ Hypertension | Age, Alcohol, Periodontitis, Psychosocial, Sex, Smoking
- ToothLoss ⊥ Sport | Age, Alcohol, Periodontitis, Psychosocial, Sex, Smoking
- ToothLoss ⊥ Lipids | Age, Alcohol, Diabetes, Obesity, Psychosocial, Sex, Sport
- ToothLoss ⊥ Lipids | Age, Alcohol, Periodontitis, Psychosocial, Sex, Smoking

MSAS

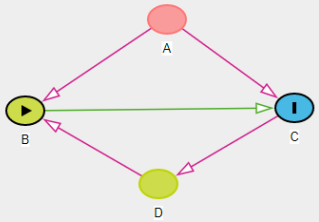
Causal effect identification

Adjustment (total effect) ▾

Minimal sufficient adjustment sets for estimating the total effect of ToothLoss on Mortality:

- Age, Alcohol, Diabetes, Obesity, Psychosocial, Sex, Smoking, Sport
- Age, Alcohol, Periodontitis, Psychosocial, Sex, Smoking

Testable implications



Adjustment (total effect) ▾

The total effect cannot be estimated by covariate adjustment.

Testable implications

Either the model does not imply any conditional independencies or the implied ones are untestable due to unobserved variables.

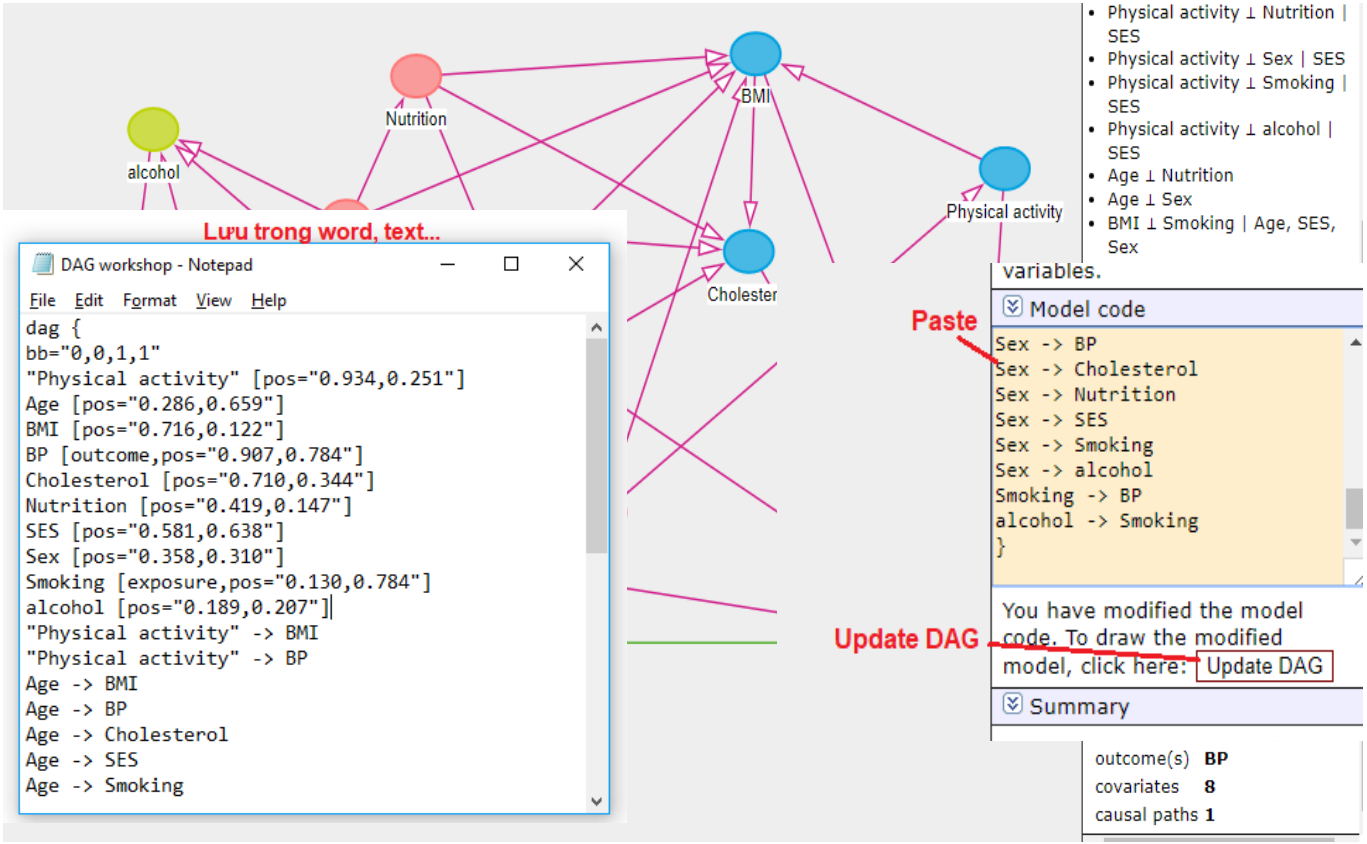
Model code

```
dag {
bb="0,0,1,1"
A [pos="0.491,0.362"]
B [exposure,pos="0.330,0.513"]
C [outcome,pos="0.637,0.509"]
D [pos="0.482,0.635"]
A -> B
A -> C
B -> C
C -> D
```

Summary

Model contains cycle:
C→D→B→C

Save DAGs


Nutrition
BMI
alcohol
Physical activity
Cholester
Lưu trong word, text...
DAG workshop - Notepad
File Edit Format View Help
dag {
bb="0,0,1,1"
"Physical activity" [pos="0.934,0.251"]
Age [pos="0.286,0.659"]
BMI [pos="0.716,0.122"]
BP [outcome,pos="0.907,0.784"]
Cholesterol [pos="0.710,0.344"]
Nutrition [pos="0.419,0.147"]
SES [pos="0.581,0.638"]
Sex [pos="0.358,0.310"]
Smoking [exposure,pos="0.130,0.784"]
alcohol [pos="0.189,0.207"]
"Physical activity" -> BMI
"Physical activity" -> BP
Age -> BMI
Age -> BP
Age -> Cholesterol
Age -> SES
Age -> Smoking
Physical activity ⊥ Nutrition | SES
Physical activity ⊥ Sex | SES
Physical activity ⊥ Smoking | SES
Physical activity ⊥ alcohol | SES
Age ⊥ Nutrition
Age ⊥ Sex
BMI ⊥ Smoking | Age, SES, Sex
variables.
Model code
Paste
Sex -> BP
Sex -> Cholesterol
Sex -> Nutrition
Sex -> SES
Sex -> Smoking
Sex -> alcohol
Smoking -> BP
alcohol -> Smoking
}
Update DAG
You have modified the model code. To draw the modified model, click here: Update DAG
Summary
outcome(s) BP
covariates 8
causal paths 1

Hotkeys

- **E = Exposure**
- **O = Outcome**
- A = Adjusted
- U = Unobserved
- D = Delete
- R = Rename

Thực hành: (Ví dụ 2)

- Xác định ảnh hưởng của **hút thuốc lá** tới **huyết áp tâm thu**
- Các biến số cân nhắc: tuổi, giới, BMI, tình trạng kinh tế xã hội, uống rượu, tình trạng dinh dưỡng, Cholesterol, hoạt động thể lực
- Xác định MSAS?

Tóm tắt DAGs

- Bản đồ thể hiện mối liên hệ nhân quả (dựa vào kiến thức và y văn)
- Giúp xác định confounders
- Giúp xác định MSAS cần hiệu chỉnh
- Giúp xây dựng khung lý thuyết trong quá trình xây dựng đề cương

Tóm tắt DAGs

- DAGs là cách tiếp cận định tính, do đó không cho biết mức độ tác động
- Kết hợp với SEM và Baysian Network

THANK YOU!