Introduction à la causalité
Théorie des DAGs

Webinar QuanTIM du 18 octobre 2024

Emilie Lanoy emilie.lanoy@aphp.fr
Unité de recherche Clinique
AP-HP

Objectifs

- Rappels sur notion de causalité
- Introduire la théorie des directed acyclic graphs (DAGs)
- Représenter les indépendances conditionnelles par les DAGs
- Présenter l’approche structurelle (par DAGs) pour décrire les biais:
  - En situation observationnelle
  - En situation interventionnelle
- Introduction à la Causal discovery

background (1)

Some causal questions

	Exposure	Outcome
1	Matches in the pocket	VO2 Max
1	Sprinkling water	Grass humidity
2	Yellow fingers	Lung cancer
3	Butterbeer	Happiness in wizards at Hogwarts
4	Ice cream sales	Burglaries

1. Hernan course
2. Maathuis course
3. Petersen exercice http://www.ucbbiostat.com/#!labs/c10ce
4. Boston university Causal inference

Some causal questions

	Exposure	Outcome	Confounder(s)
1	Matches in the pocket	VO2 Max	Smoking status
1	Sprinkling water	Grass humidity	Rain
2	Yellow fingers	Lung cancer	Smoking status
3	Butterbeer	Happiness in wizards at Hogwarts	Age, house, gender, friendship with Dumbledore, enemy with Snape

1. Hernan course
2. Maathuis course
3. Petersen exercice http://www.ucbbiostat.com/#!labs/c10ce

background (1)

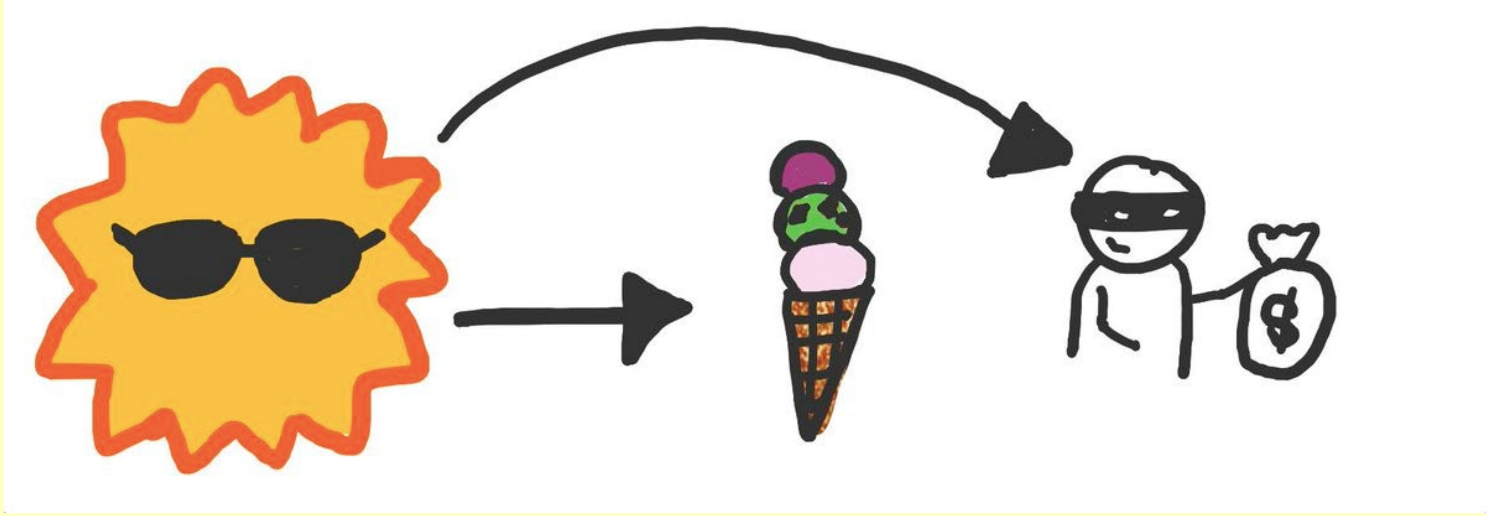
Ellie Murray twitter @EpiEllie

Epidemiology assistant prof [@BUSPH]

| Associate editor for social media [@amjepi]

| cohost [@casualinfer] [sites.bu.edu/causal]

podcast | [#epitwitter] | [http://github.com/eleanormurray]



Some causal questions in oncology

Exposure	Outcome	Confounder(s)
Red meat consumption	Occurrence of cancer	BMI, smoking status, etc.
Professional occupation : air hostess	Occurrence of breast cancer	Socio-professional status, age of first child's birth
Sedentary lifestyle	Occurrence of cancer	BMI, gender, food intakes, …
EGFR mutation	Response to treatment in NSCLC	ECOG, TNM stage, …

DAGs (1)

- DAGs theory
- Directed Acyclic Graph
  - Judea Pearl, UCLA (1988,1995,2000)

- Unknown effect of T on M

$$T \longrightarrow M$$

- No causal effect of T on M

$$T \qquad\qquad M$$

- Information is in lack of arrow
- NB : under H_0, T and M are independent then lack of arrow (conventional notation)

DAGs (2)

- Open pathway between A and B

$$A \longrightarrow C \longrightarrow B$$

$$A \longleftarrow C \longrightarrow B$$

- C collider between A and B : close pathway

$$A \longrightarrow C \longleftarrow B$$

- Conditionning on C (put a box)
  - Close the pathway

$$A \longleftarrow \boxed{C} \longrightarrow B$$

  - Open the pathway (if collider)

$$A \longrightarrow \boxed{C} \longleftarrow B$$

DAGs (3)

- Aim: to know how to treat? For example, to compare two first line treatments

- Measuring causal effect of treatment T on occurrence of event M

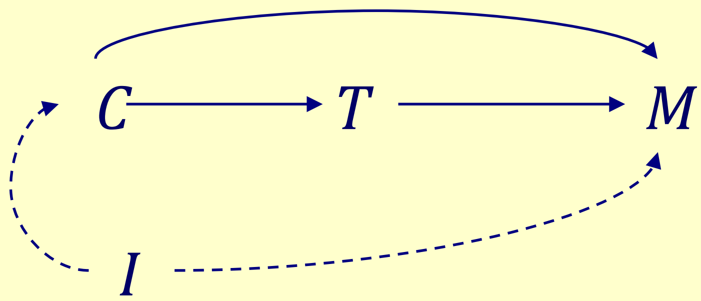
$$T \longrightarrow M$$

  - T : treatment (experimental drug vs. standard, bilateral vs. unilateral graft)
  - M : death, progression

- Statistically, association between T and M will be assessed. Does it mean a causal effect of T on M?

DAGs (4)

- Given C, vector of measured factors associated with T and M



  - T : treatment by experimental or reference drug combination
  - M : occurrence of serious morbidity
  - I : real biological and clinical status
  - C : biomarkers measurement, prognosis score, Karnofsky index, body composition, …

DAGs (5)

- **Potential confounder**: variable associated with treatment allocation

$$C \longrightarrow T \longrightarrow M$$

- **True confounder**: variable associated with both treatment allocation and prognosis

$$\begin{tikzcd} C \arrow[r] \arrow[rr, bend left] & T \arrow[r] & M \end{tikzcd}$$

- **Non confounder** : variable neither associated with treatment nor with prognosis

$$C \qquad T \longrightarrow M$$

DAGs (6)

- Causal effect of T on M

$$\begin{tikzcd} C \arrow[r] \arrow[rr, bend left] & T \arrow[r] & M \end{tikzcd}$$

- Part of association between T and M are associated, partly because they are common effects of C
  - Ex: expression of molecular alteration by the tumor could be associated with higher morbidity risk and could guide treatment choice
- Exposed (treated) and unexposed (untreated) non exchangeables
- Indication bias* : C confunder
- How to measure causal effet of T on M?

*indication bias *Epidemiology* ⇔ selection bias *Statistic*

DAGs (6)

- Statistical DAG = complete DAG with all possible directed arrows

$$\begin{tikzcd} V_1 \arrow[r] \arrow[rr, bend left] \arrow[rrr, bend right] & V_2 \arrow[r] \arrow[rr, bend right] & V_3 \arrow[r] & V_4 \end{tikzcd}$$

- Causal DAG = DAG in which all the information is in missing arrows

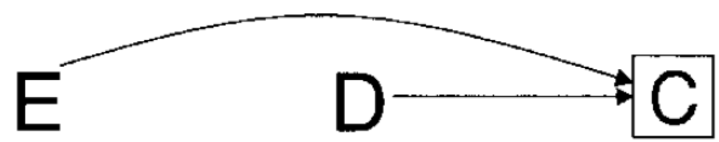
$$\begin{tikzcd} V_1 \arrow[r] \arrow[rr, bend left] & V_2 \arrow[rr, bend right] & V_3 \arrow[r] & V_4 \end{tikzcd}$$

- Causal Markov assumption: any variable not caused by a given variable V_i will be independent of V_i conditional on the direct causes of V_i

DAGs (7)

DAG as screening tool for selection bias

FIGURE 1. Common cause L of exposure E and outcome D.



Berkson's bias

FIGURE 3. Conditioning on a common effect C of exposure E and outcome D.

- Hernan, Epidemiology 2004

DAGs (8)

Berkson's bias

$$
\begin{tikzcd}
R \arrow[rr, bend left] & P \arrow[r] & D
\end{tikzcd}
$$

- R: rareness of a stamp
- P: prettiness of a stamp
- D: stamp on display

ALL COLLECTION		Prettiness		Total
		Pretty	**Not pretty**	
Rareness	**Rare**	50	50	**100**
	Not rare	450	450	**900**
Total		**500**	**500**	**1000**

$$P(R = 1) = P[(R = 1)/(P = 1)] = P[(R = 1)/(P = 0)] = 10\%$$

https://en.wikipedia.org/wiki/Berkson′s_paradox

DAGs (9)

Berkson's bias

$$\begin{tikzcd} R \arrow[rr, bend left] & P \arrow[r] & \boxed{D} \end{tikzcd}$$

- R: rareness of a stamp
- P: prettiness of a stamp
- D: stamp on display

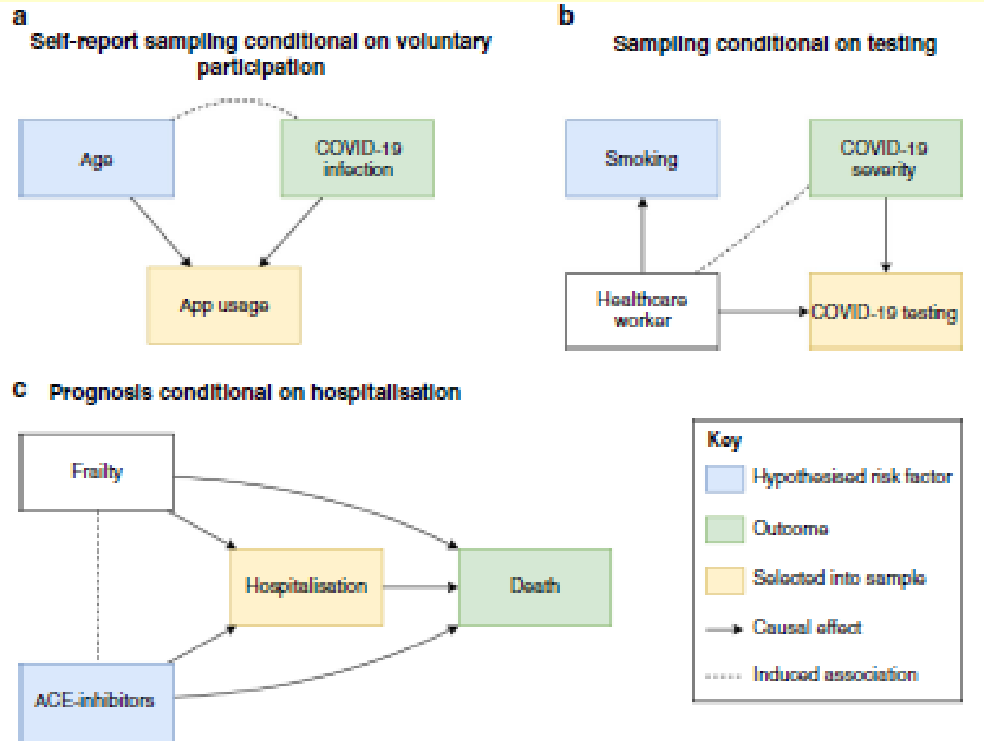
ON DISPLAY		**Prettiness**		**Total**
		Pretty	**Not pretty**	
Rareness	**Rare**	50	50	**100**
	Not rare	450	0	**450**
Total		**500**	**50**	**550**

$$P(R=1) = 18\%$$

$$P[(R=1)/(P=1)] = 10\% \neq P[(R=1)/(P=0)] = 100\%$$

https://en.wikipedia.org/wiki/Berkson's_paradox

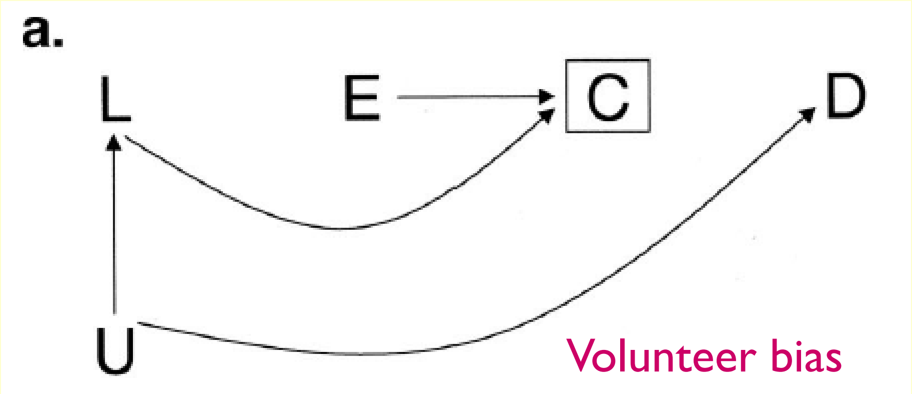
(…)en particulier le redoutable biais de collision (collider bias), par exemple en fondant ces analyses uniquement sur des données de dépistage spontané [les personnes demandant un test de dépistage ne sont pas représentatives de l'ensemble des personnes vaccinées/non vaccinées]. S Korsia-Meffre [*https://www.vidal.fr/actualites/27895-vaccins-contre-la-covid-19-la-troisieme-dose-pourquoi-pour-qui.html*](https://www.vidal.fr/actualites/27895-vaccins-contre-la-covid-19-la-troisieme-dose-pourquoi-pour-qui.html)



Griffith G, Nature communications, 2020

DAGs
(10)

DAG as screening tool for selection bias



- Hernan, Epidemiology 2004

Interventional DAGs (1)

- Comparative randomized controlled trial RCT

$$\begin{tikzcd} C \arrow[rr, bend left] & t \arrow[r] & M \end{tikzcd}$$

- Treatment t is allocated randomly and independently of patient's characteristics (ex: prognostic score)

- Gold-standard method

- Measured association between T and M corresponds to causal effect of T on M

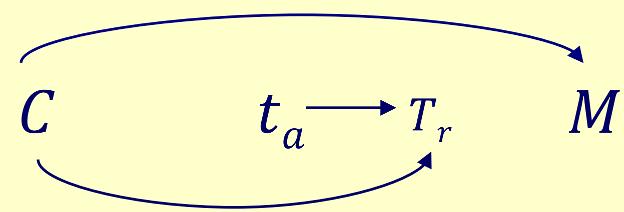
Interventional DAGs (2)

- Randomly assigned treatment
  - to avoid selection bias
  - characteristics balanced across the groups
- RCT limitations
  - Feasibility questionable when many treatment arms
  - Difficult for endpoints occurring rarely or after very long follow-up duration
  - Loss to follow-up depending on treatment arm could introduce selection bias

$$\begin{tikzcd} t \arrow[r] \arrow[rr, bend left] & M \arrow[r] & L \end{tikzcd}$$

Interventional DAGs (3)

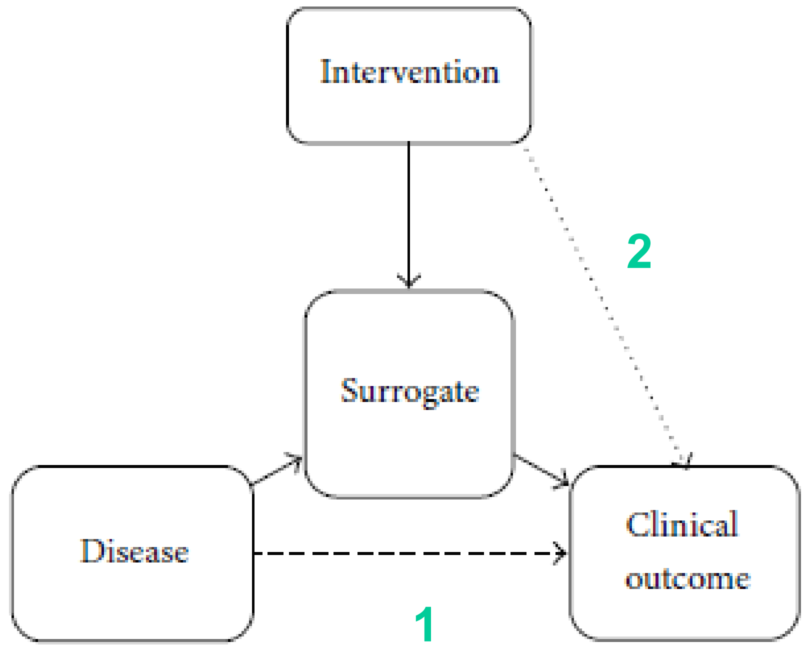
- Per protocol analyses in which protocol deviation are excluded
  - loss to follow-up, non-compliance or poor observance
  - biais risk +++ : loss of randomization and comparability of groups
  - question results credibility
  - decrease number of patients: loss of precision and power



T_r : treatment actually received depends on allocated treatment t_a and confounders C pronostic of M

Interventional DAGs (4)

Surrogate endpoints



After Lavery Mult Scler Int. 2014

- Prentice Stat Med. 1989

Validation of surrogate endpoint if:

1. All effect of disease on clinical outcome mediated by the surrogate endpoint
2. All causal effect of intervention on clinical outcome mediated by the surrogate

Interventional DAGs (5)

- Perfect surrogate endpoint represented on a DAG

$$\begin{tikzcd} C \arrow[rr, bend right] & t_a & S \arrow[r] & M \end{tikzcd}$$

- How to measure causal effect of treatment on OS when:

1. primary outcome=surrogate
2. cross-over allowed after occurrence of surrogate endpoint

- ITT analysis unbiased but dilution of treatment effect
- How to deal with informative censoring at cross-over?

Etudes observa-tionnelles

Aspirine et accident vasculaire cérébral cf. Miguel Hernan

- Traitement par aspirine en prévention de accident vasculaire cérébral

- Artères détériorées cause sous-jacente (non connue) de maladie cardio-vasculaire
- Maladie cardio-vasculaire entraîne traitement par aspirine
- Artères détériorées entraînent accident vasculaire cérébral

- L'aspirine est-elle un traitement efficace de prévention de l'accident vasculaire cérébral ?

Etudes observa-tionnelles

Aspirine et accident vasculaire cérébral cf. Miguel Hernan

- L'aspirine évite-t-elle l'accident vasculaire cérébral ? C'est-à-dire, l'aspirine a-t-elle un effet causal sur la survenue d'accident vasculaire cérébral ?
- Les traités par aspirine ont plus souvent des maladies cardiovasculaires et ne sont pas comparables aux non traités
- Biais d'indication

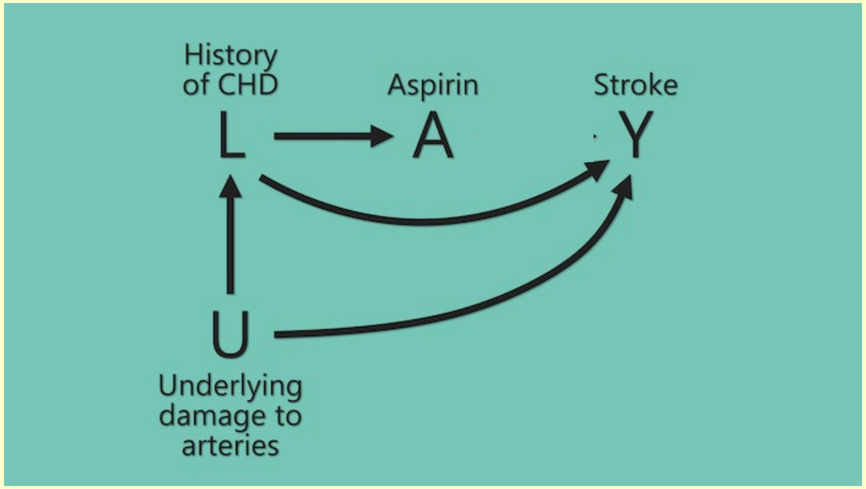
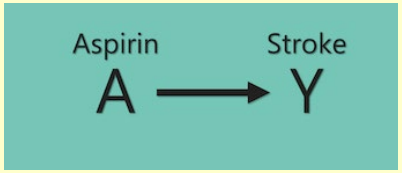


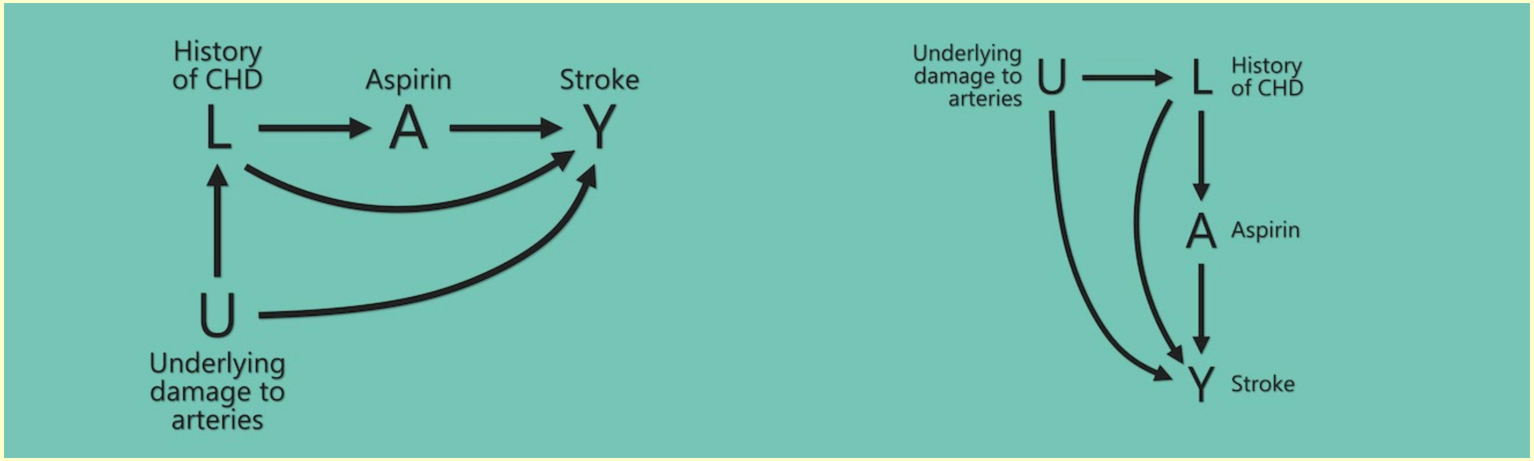
Figure by Miguel Hernan

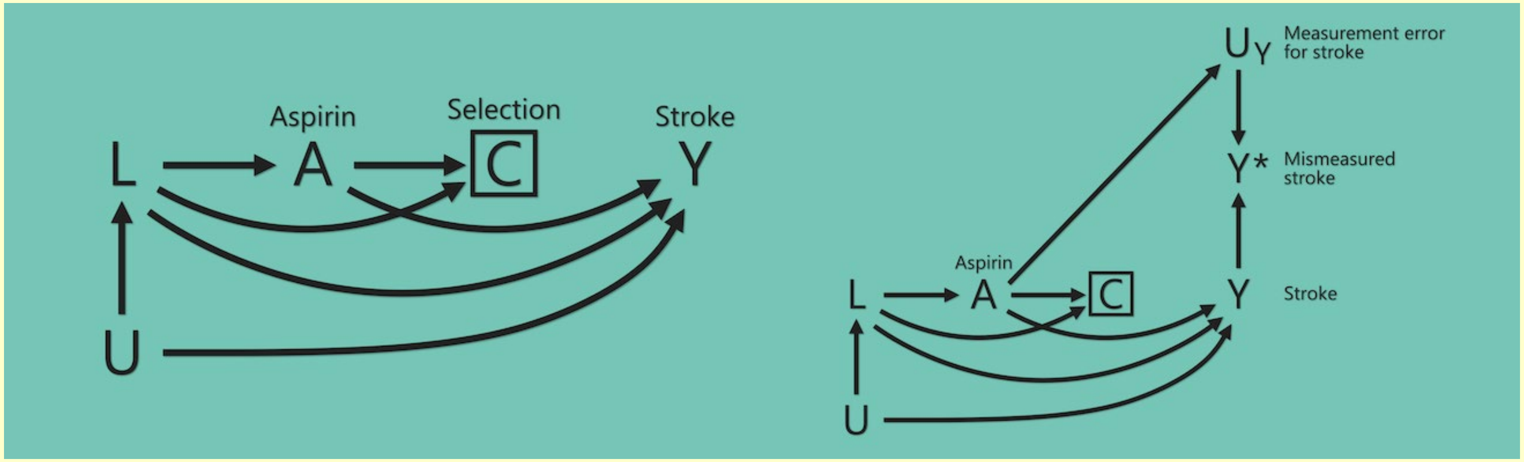
Etudes observa-tionnelles

Aspirine et accident vasculaire cérébral cf. Miguel Hernan

De + en + complexe







\+ lire thread intéressant sur survenue d'extrasystole ventriculaire après infarctus et risque de décès : https://twitter.com/FZores/status/1243459929618423808

Etudes observa-tionnelles

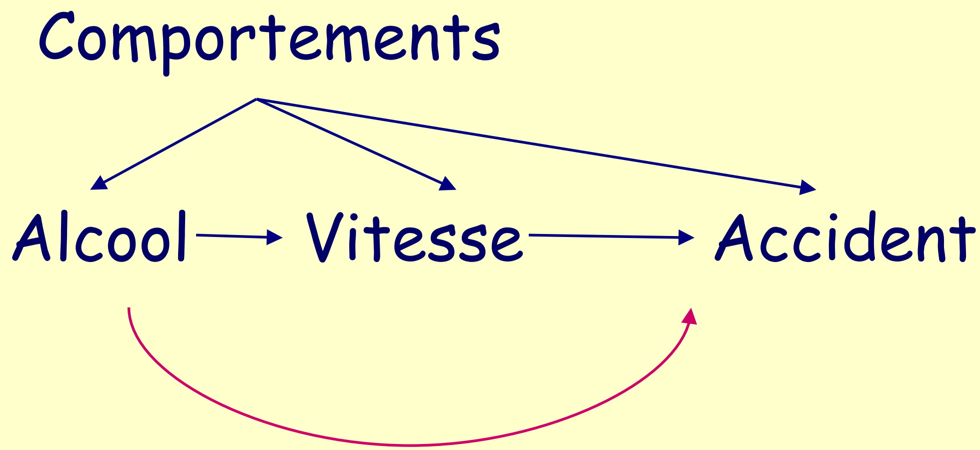
Médiation et sécurité routière Thèse de Marine Dufournet supervisée par Vivian Vialon

- Alcool augmente risque d'accident
- Vitesse excessive augmente risque d'accident
- Personne exposée à l'alcool roule à une vitesse inadaptée à la situation

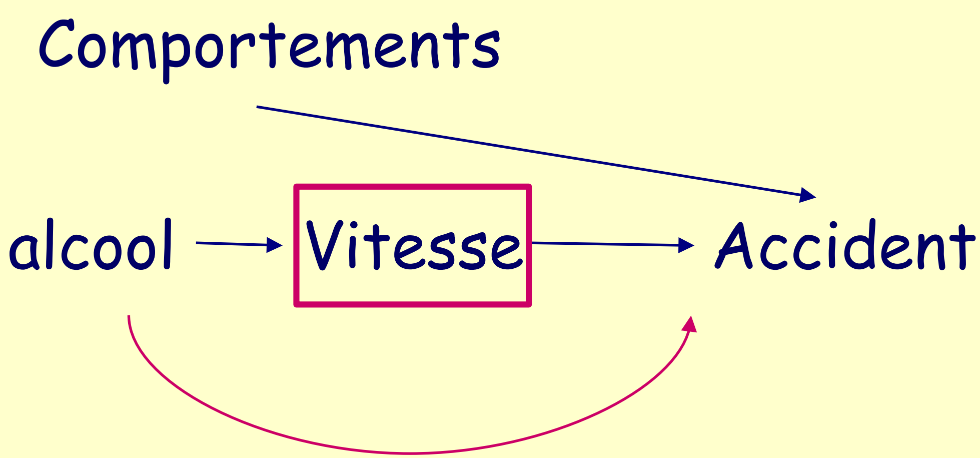
- Effet direct de l'alcool et effet médié par la vitesse en présence de facteur de confusion (comportements)

Etudes observa-tionnelles

Médiation et sécurité routière
Thèse de Marine Dufournet supervisée par Vivian Vialon



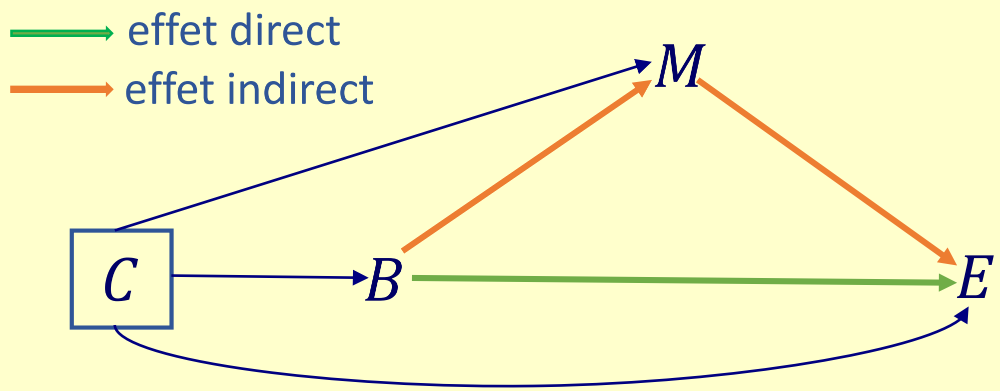
- Design d'essai interventionnel pour identifier l'effet non médié



Etudes observationnelles

Effet du bruit des avions sur la santé
Thèse d'Antoine Kodji

Médiation

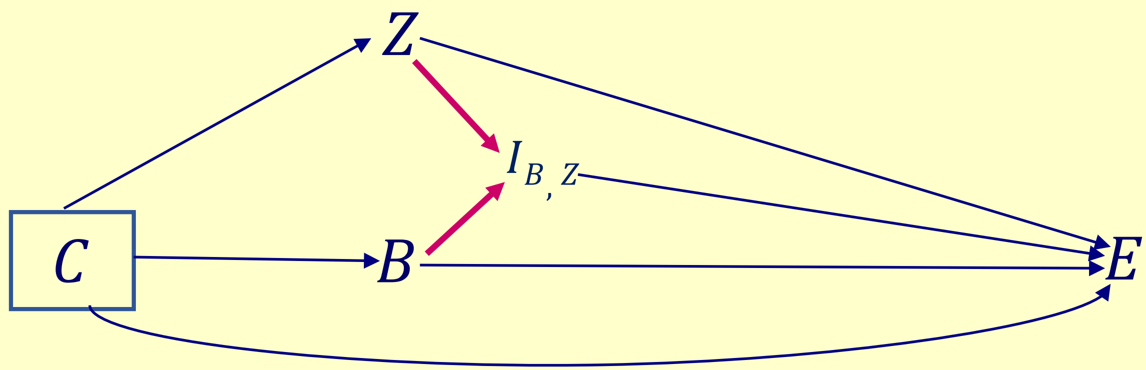


B : exposition au bruit des avions
M : médiateur : gène occasionnée par le bruit
E : événement de santé (état de santé perçu)
C : facteurs de confusion (facteurs socio-démographiques, lieu d'habitation, etc.)

Etudes observationnelles

Effet du bruit des avions sur la santé Thèse d' Antoine Kodji

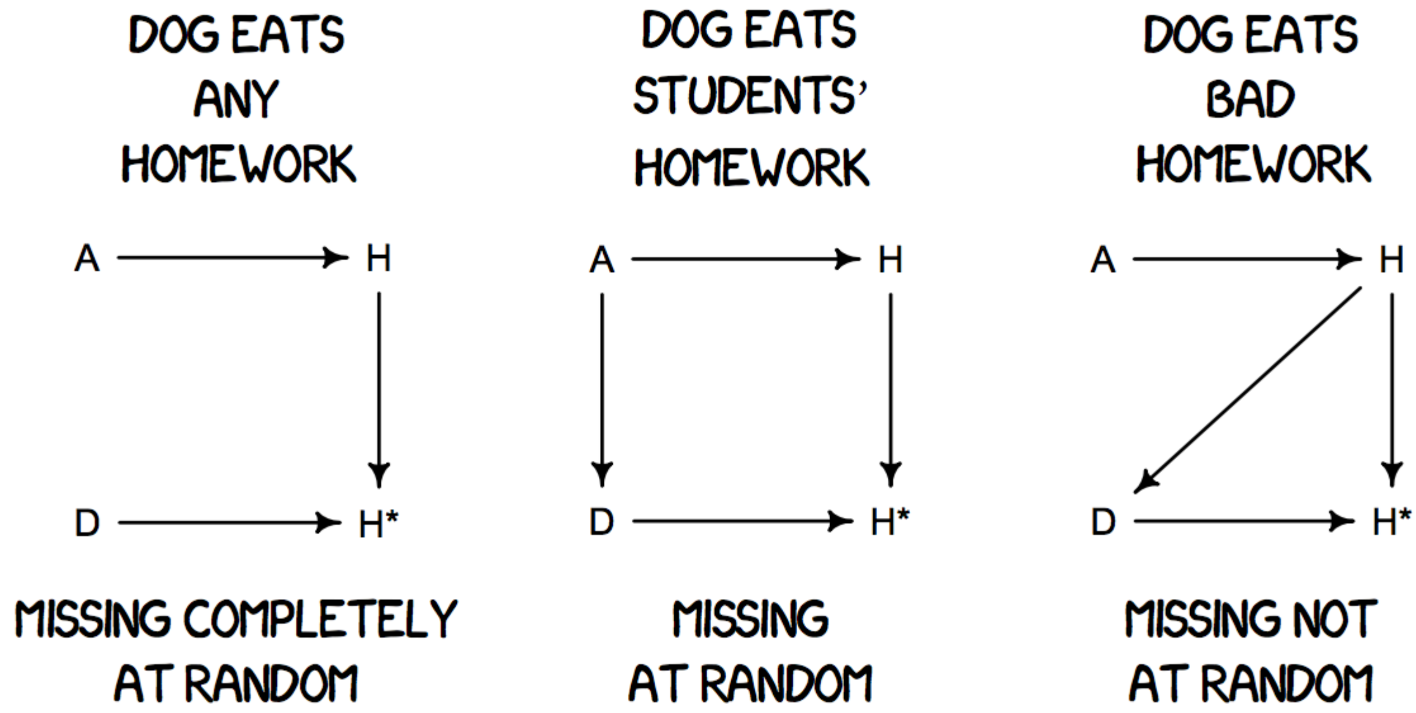
Modération



B : exposition au bruit des avions
Z : modérateur : sensibilité au bruit
$I_{B,Z}$: interaction entre le facteur d'exposition et le modérateur
E : événement de santé (état de santé perçu)
C : facteurs de confusion (facteurs socio-démographiques, lieu d'habitation, etc.)

Typologie MCAR, MAR, MNAR

H: Homework
H*: Homework with missing values
A: Attribute of student
D: Dog (missingness mechanism)



From Richard McElreath twitter

Etudes observa-tionnelles

œstrogènes et cancer endométrial cf. Miguel Hernan

- Traitement hormonal substitutif de la ménopause par œstrogènes et survenue d'un cancer utérin
- Association trouvée dès le milieu des années 90

- Cancer utérin entraîne saignements
- Œstrogènes entraîne saignements qu'il y ait cancer utérin ou non

- Beaucoup de cancers non diagnostiqués
- Saignements entraîne dépistage et diagnostic de cancer utérin

- Œstrogènes sont-ils un facteur de risque de cancer utérin ?

Etudes observa-tionnelles

œstrogènes et cancer endométrial cf. Miguel Hernan

- Quel est l'effet de l'exposition aux œstrogènes sur la survenue de cancer utérin ?
- On dispose des données sur les traitements et les **diagnostics** de cancer utérin
- Biais de diagnostic, de surveillance

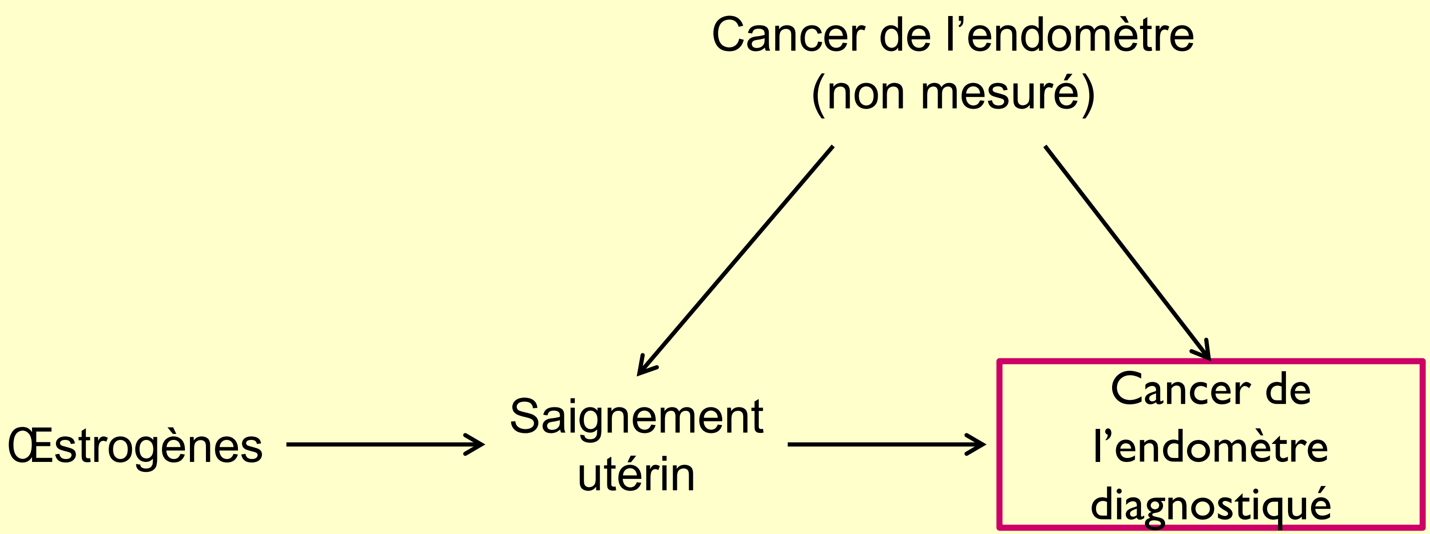
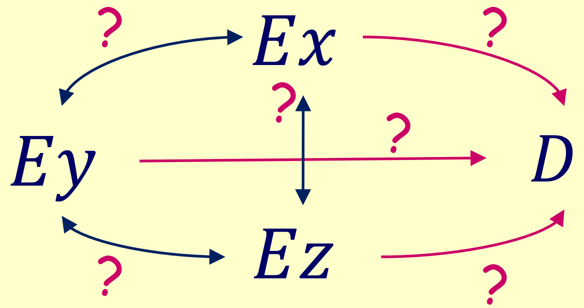


Figure d'après Leslie Myint

Quand le DAG est inconnu (1)

Modélisation des effets causaux de biomarqueurs multidimensionnels

- Données de grande dimension : Intégration de biomarqueurs (E) par approche causale



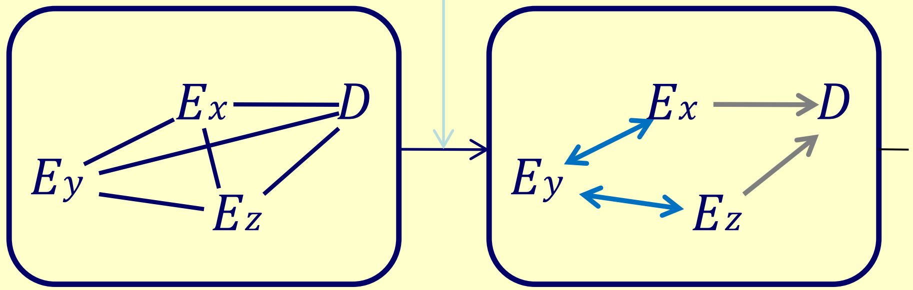
1. Relations d’ordre inconnues
2. Intervention impossible

Quand le DAG est inconnu (2)

Modélisation des effets causaux de biomarqueurs multidimensionnels

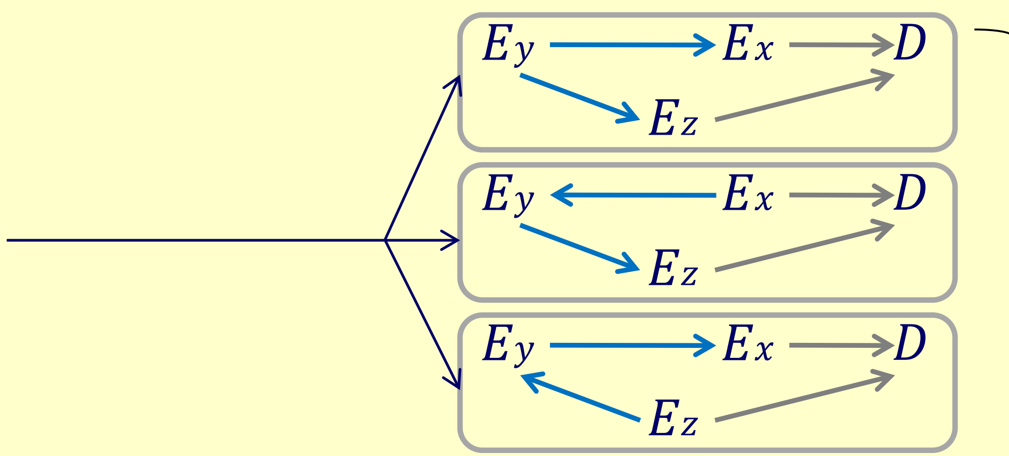
- Intervention calculus when the DAG is absent (IDA Maathuis 2009)

① PC-algorithm Peter Spirtes and Clark Glymour ; 2000



Données completed partially DAG

② Do-calculus



$\theta_i = \min(\theta_{i,j})$

DAGs

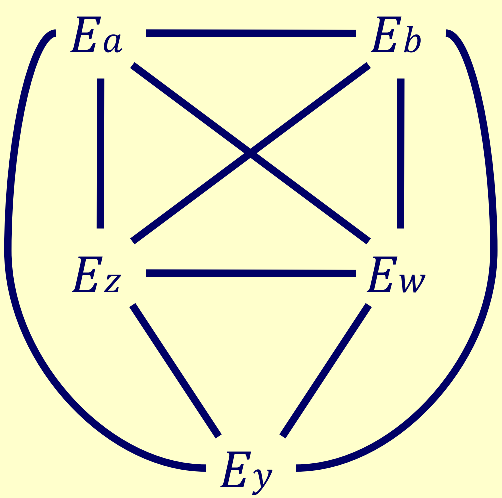
Quand le DAG est inconnu (3)

Modélisation des effets causaux de biomarqueurs multidimensionnels

IDA : ① *PC-algorithm*

Objectif :

A partir d'un graphe non orienté, supprimer un lien si indépendance au seuil **α**



Quand le DAG est inconnu (4)

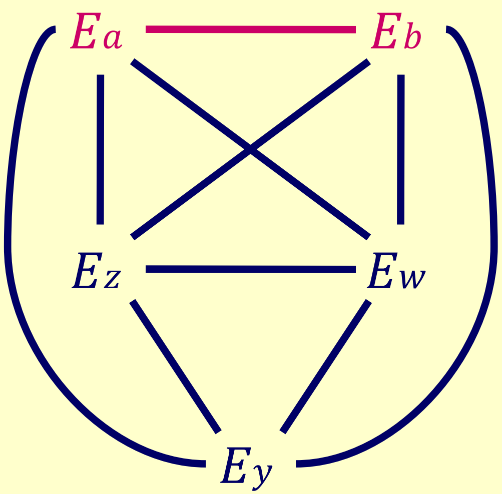
Modélisation des effets causaux de biomarqueurs multidimensionnels

IDA : ① *PC-algorithm*

Objectif :

A partir d'un graphe non orienté, supprimer un lien si indépendance au seuil **α**

$$E_a \perp\!\!\!\perp E_b \,|\, \varnothing ?$$



Quand le DAG est inconnu (5)

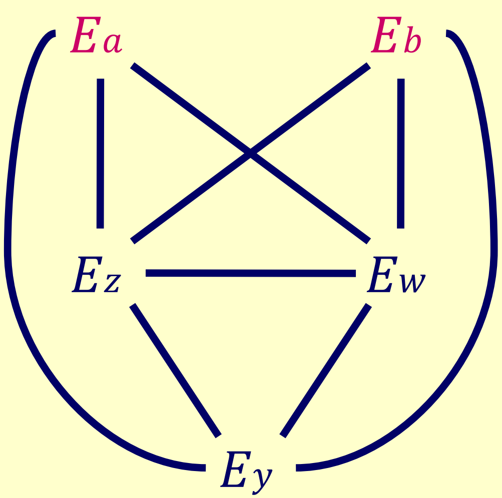
Modélisation des effets causaux de biomarqueurs multidimensionnels

IDA : ① *PC-algorithm*

Objectif :

A partir d'un graphe non orienté, supprimer un lien si indépendance au seuil **α**

$E_a \perp\!\!\!\perp E_b \mid \varnothing$? Oui



Quand le DAG est inconnu (6)

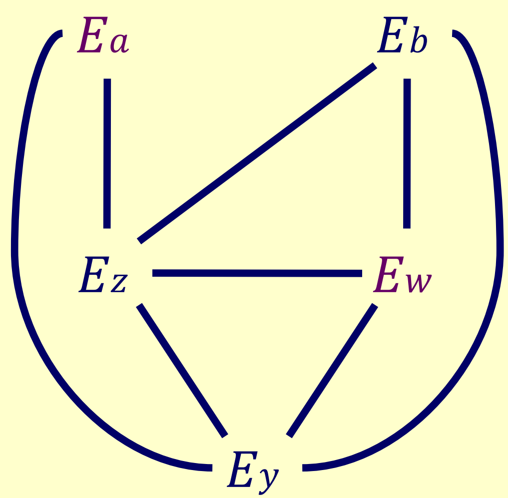
Modélisation des effets causaux de biomarqueurs multidimensionnels

IDA : ① *PC-algorithm*

Objectif :

A partir d'un graphe non orienté, supprimer un lien si indépendance au seuil **α**

$E_a \perp\!\!\!\perp E_b \,|\varnothing$? Oui $E_a \perp\!\!\!\perp E_w \,|\varnothing$? Oui



Quand le DAG est inconnu (7)

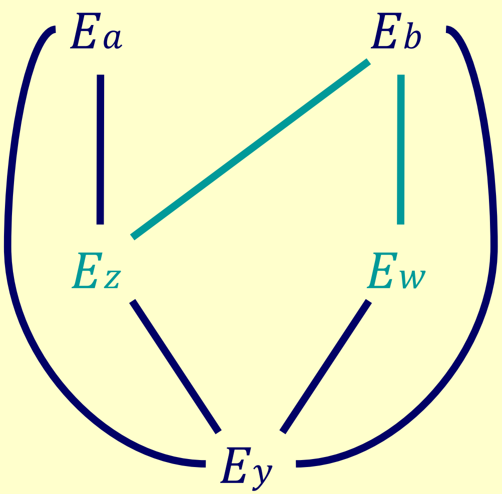
Modélisation des effets causaux de biomarqueurs multidimensionnels

IDA : ① *PC-algorithm*

Objectif :

A partir d'un graphe non orienté, supprimer un lien si indépendance au seuil **α**

$E_z \perp\!\!\!\perp E_w \mid E_b$? Oui



Quand le DAG est inconnu (8)

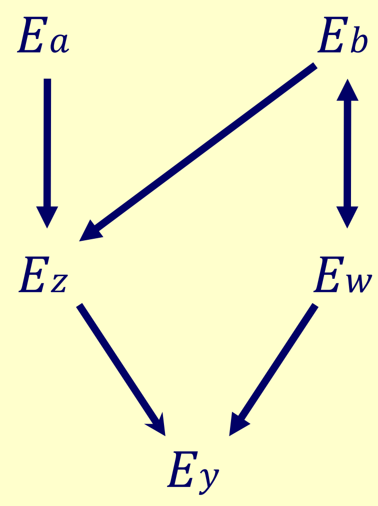
Modélisation des effets causaux de biomarqueurs multidimensionnels

IDA : ① *PC-algorithm*

Objectif :

A partir d'un graphe non orienté, supprimer un lien si indépendance au seuil **α**

Completed partially DAG (CPDAG)



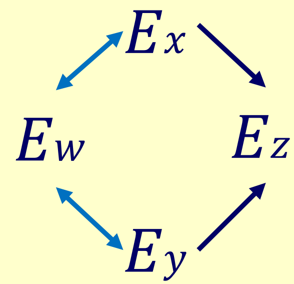
Quand le DAG est inconnu (9)

Modélisation des effets causaux de biomarqueurs multidimensionnels

② *Do-calculus : orienter les flèches bidirigées du CPDAG*

⇨ *Classe de DAGs équivalents (même squelette, même V-structure)*

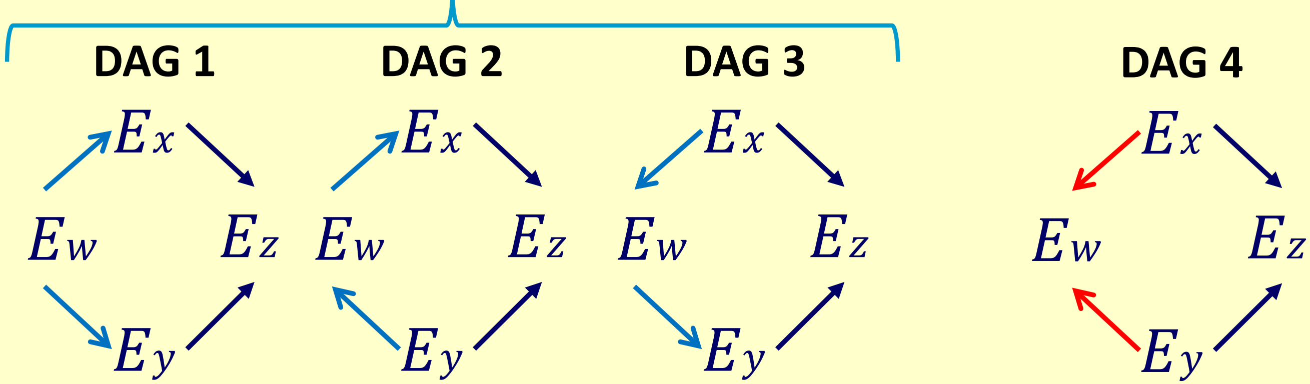
CPDAG



$E_x \perp\!\!\!\perp E_y$? Non

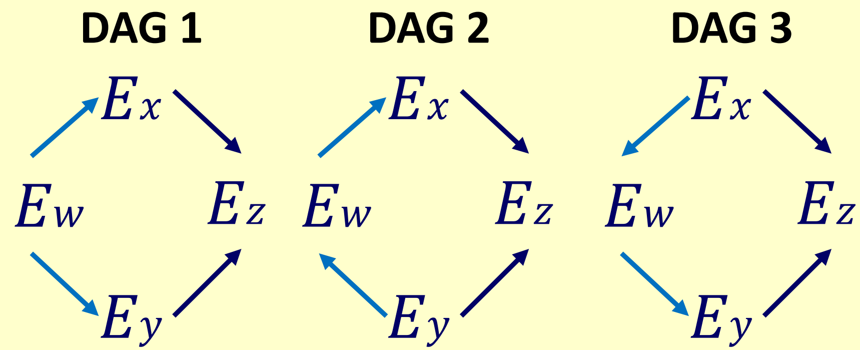
$E_x \perp\!\!\!\perp E_y \mid E_w$? Oui

DAGs de classe Markov équivalente encodant les mêmes indépendances conditionnelles



Quand le DAG est inconnu (10)

Modélisation des effets causaux de biomarqueurs multidimensionnels



Objectif : dans chaque DAG, estimer l'effet causal de chaque prédicteur sur l'événement

Estimateur unique : effet médian, effet minimal, etc..

Comparaison de méthodes statistiques d'analyse de données omics dans le traitement néoadjuvant dans le cancer du sein (Ternès et al. 2014*)

encadrement Nils Ternès (Master 2)

Quand le DAG est inconnu (11)

Intégration de mesures répétées de biomarqueurs multidimensionnels

- Identifier des marqueurs immunologiques précoces des effets des traitements
  - Inférence causale dans l'analyse de biomarqueurs (laboratoire d'immunomonitoring en oncologie UMS CNRS 3655, Inserm US23)
  - Marqueurs mesurés avant et pendant le traitement
  - Critères de jugement : toxicité/réponse au traitement / survie
- Thèse de Vahé Asvatourian « Développer une approche causale dans l'évaluation des immunothérapies »

Take home message

- In DAGs, all information is in missing arrows
- Classical statistical inference could be represented with DAGs
- DAGs = screening tools to identify selection bias:
  - especially in observational settings
  - but also in interventional trials
- DAG have been developped in omics settings
- Expert knowledge is required +++ to draw DAGs

References

- Harvard School Public Health causal inference program
  http://www.hsph.harvard.edu/causal/
- University of Berkeley M Petersen and L Balzer course
  http://www.ucbbiostat.com/#!lectures/cpk
- Pearl, J., 1995. Causal diagrams for empirical research. *Biometrika*, 82(4), pp.669–688.
- Greenland, S., Pearl, J. & Robins3, J.M., 1999. Causal Diagrams for Epidemiologic Research. *Epidemiology*, 10(1), pp.37–48.
- Hernan, M.A., Hernandez-Diaz, S. & Robins, J.M., 2004. A structural approach to selection bias. *Epidemiology*, 15(5), pp.615–625.
- Maathuis, M.H., 2009. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A), pp.3133–3164.
- Maathuis, M.H. et al., 2010. Predicting causal effects in large-scale systems from observational data. *Nature Methods*, 7(4), pp.247–248.
- Stekhoven, D.J. et al., 2012. Causal stability ranking. *Bioinformatics*, 28(21), pp.2819–23.