

# Identifier un ensemble de variables d'ajustement à partir du diagramme causal

Dans le cadre des activités scientifiques de l'axe SP-POS

Janie Coulombe

[janie.coulombe@umontreal.ca](mailto:janie.coulombe@umontreal.ca)

Professeure adjointe

Département de mathématiques et de statistique

Université de Montréal

18 octobre 2022

# Plan pour aujourd'hui



Objectifs

Le diagramme causal

Rappels: Confusion et stratification sur un facteur de collision

Choix d'un ensemble d'ajustement

Discussion

# Objectifs

Pour un problème donné en inférence causale,

1. Comprendre ce que sont les biais de confusion et de stratification sur un facteur de collision
2. Choisir un ensemble de variables de base pour notre diagramme causal
3. Trouver un ensemble suffisant de variables d'ajustement pour estimer (ou identifier, dans le langage causal) un effet causal (c'est-à-dire, pour estimer de façon consistante l'effet causal).

# Avant de commencer...



Différentes raisons de choisir un ensemble d'ajustement (ou de faire de la sélection de variables) (Guo et al., 2022).

Exemples:

- ▶ Trouver un ensemble qui permet d'identifier l'effet causal (aujourd'hui)
- ▶ Réduction des dimensions
- ▶ Robustesse à la mauvaise spécification des modèles
- ▶ Efficacité (réduction de la variance d'estimation)
- ▶ Éthique ou raisons économiques

On utilise la notation des issues potentielles (Neyman 1923, Rubin 1974) pour définir le paramètre qui nous intéresse.

Aujourd'hui on s'intéressera surtout à l'effet causal marginal d'un traitement  $A$  (deux catégories,  $A = 0, 1$ , ex., prise ou non de diurétiques) sur l'issue continue  $Y$  (ex., pression systolique).

Soit  $Y^1$  l'issue potentielle sous le traitement 1, et  $Y^0$  l'issue potentielle sous le traitement 0, alors on cherche à estimer

$$\mathbb{E} [Y^1 - Y^0] .$$

# Quoi faire avec l'ensemble d'ajustement?



Une fois qu'on aura identifié un ensemble d'ajustement, on peut utiliser

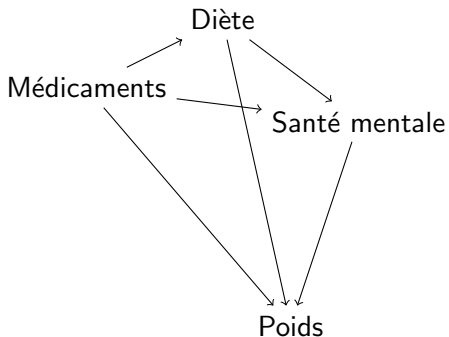
- ▶ Poids inverse à la probabilité de traitement
- ▶ Régression avec l'ensemble de variables et le traitement comme prédicteurs
- ▶ *Targeted maximum likelihood estimation* (TMLE)
- ▶ *G-estimation, G-formula, etc.*

## Le diagramme causal



# Le diagramme causal

Exemple simple:

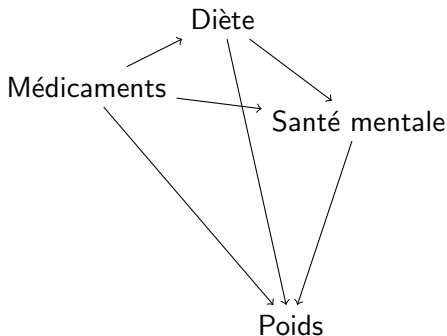


À quoi nous sert-il? Quelles sont ses caractéristiques?

1. Représente les relations (assumées) entre les variables
2. Connaissances sur le sujet (biologiques et non pas statistiques même si elles impliquent des relations de dépendance)
3. Nous permet de déterminer si l'on peut estimer de façon consistante (sans biais) l'effet d'une intervention ou d'un traitement ( $A$ ) sur un résultat ( $Y$ )

# Le diagramme causal

Ici, par exemple, on pourrait s'intéresser à l'effet causal de la diète sur le poids. On aurait besoin de considérer un sous-ensemble spécifique de variables à mettre dans le diagramme causal (à venir).



Qu'est-ce qu'un DAG?

- **Directed:** Dirigé; les relations représentées sont dirigées grâce aux têtes des flèches. Une flèche aura sa tête ( $>$ ) du côté qui subit l'effet, c'est-à-dire qui est causé par la variable à l'autre bout de la flèche. Ainsi,  $A \rightarrow B$  se lit *A cause B*.

Qu'est-ce qu'un DAG?

- ▶ **Directed:** Dirigé; les relations représentées sont dirigées grâce aux têtes des flèches. Une flèche aura sa tête ( $>$ ) du côté qui subit l'effet, c'est-à-dire qui est causé par la variable à l'autre bout de la flèche. Ainsi,  $A \rightarrow B$  se lit *A cause B*.
- ▶ **Acyclic:** Il n'y a pas de cycle; pensez à ces relations qui sont temporelles. Une variable cause une autre variable qui en cause une autre (qui ne peut pas revenir dans le temps pour causer la première). Il y a une structure temporelle inhérente au DAG.

Qu'est-ce qu'un DAG?

- ▶ **Directed:** Dirigé; les relations représentées sont dirigées grâce aux têtes des flèches. Une flèche aura sa tête ( $>$ ) du côté qui subit l'effet, c'est-à-dire qui est causé par la variable à l'autre bout de la flèche. Ainsi,  $A \rightarrow B$  se lit *A cause B*.
- ▶ **Acyclic:** Il n'y a pas de cycle; pensez à ces relations qui sont temporelles. Une variable cause une autre variable qui en cause une autre (qui ne peut pas revenir dans le temps pour causer la première). Il y a une structure temporelle inhérente au DAG.
- ▶ **Graph.**

Qu'est-ce qu'un DAG?

- ▶ **Directed:** Dirigé; les relations représentées sont dirigées grâce aux têtes des flèches. Une flèche aura sa tête ( $>$ ) du côté qui subit l'effet, c'est-à-dire qui est causé par la variable à l'autre bout de la flèche. Ainsi,  $A \rightarrow B$  se lit *A cause B*.
- ▶ **Acyclic:** Il n'y a pas de cycle; pensez à ces relations qui sont temporelles. Une variable cause une autre variable qui en cause une autre (qui ne peut pas revenir dans le temps pour causer la première). Il y a une structure temporelle inhérente au DAG.
- ▶ **Graph.**

On utilise les DAG (parfois appelés DAG causaux) comme diagrammes causaux pour inférer si un effet causal peut être identifié.

## Certaines caractéristiques importantes du DAG



- Un trajet sur lequel toutes les flèches pointent dans la même direction est appelé un trajet causal.

Ex. 1. (Causal)  $A \longrightarrow B \longrightarrow C \longrightarrow D$

Ex. 2. (Pas causal)  $A \longrightarrow B \longleftarrow C \longrightarrow D$



# Certaines caractéristiques importantes du DAG



- Un trajet sur lequel toutes les flèches pointent dans la même direction est appelé un trajet causal.

Ex. 1. (Causal)  $A \longrightarrow B \longrightarrow C \longrightarrow D$

Ex. 2. (Pas causal)  $A \longrightarrow B \longleftarrow C \longrightarrow D$

- Similairement, un trajet bloqué vs un trajet ouvert. Exemples:

Ouvert (1) entre  $A$  et  $D$ :  $A \longrightarrow B \longrightarrow C \longrightarrow D$

Ouvert (2) entre  $A$  et  $D$ :  $A \longrightarrow D$

Ouvert (3) entre  $A$  et  $D$ :  $A \longleftarrow C \longrightarrow D$

Fermé (1) entre  $A$  et  $D$ :  $A \longrightarrow C \longleftarrow D$

Fermé (2) entre  $A$  et  $D$ :  $A \longrightarrow C \longleftarrow B \longrightarrow E \longrightarrow D$

- Une variable  $A$  qui en cause une autre (ex.,  $A \rightarrow B$ ) vient nécessairement avant la variable causée  $B$  dans le temps (temporalité).

# Certaines caractéristiques importantes du DAG



- Un trajet sur lequel toutes les flèches pointent dans la même direction est appelé un trajet causal.

Ex. 1. (Causal)  $A \longrightarrow B \longrightarrow C \longrightarrow D$

Ex. 2. (Pas causal)  $A \longrightarrow B \longleftarrow C \longrightarrow D$

- Similairement, un trajet bloqué vs un trajet ouvert. Exemples:

Ouvert (1) entre  $A$  et  $D$ :  $A \longrightarrow B \longrightarrow C \longrightarrow D$

Ouvert (2) entre  $A$  et  $D$ :  $A \longrightarrow D$

Ouvert (3) entre  $A$  et  $D$ :  $A \longleftarrow C \longrightarrow D$

Fermé (1) entre  $A$  et  $D$ :  $A \longrightarrow C \longleftarrow D$

Fermé (2) entre  $A$  et  $D$ :  $A \longrightarrow C \longleftarrow B \longrightarrow E \longrightarrow D$

- Une variable  $A$  qui en cause une autre (ex.,  $A \rightarrow B$ ) vient nécessairement avant la variable causée  $B$  dans le temps (temporalité).
- Le DAG nous permet de déterminer s'il y a une dépendance entre deux variables (sous l'hypothèse qu'on a représenté toutes les variables importantes dans le DAG – détails à venir).

Rappels: Confusion et stratification sur un facteur de collision

# Supposons...



Que l'on souhaite estimer l'effet causal des habitudes alimentaires (HA) sur l'asthme (AS), exemple modifié tiré de Dumas et al. (2014).

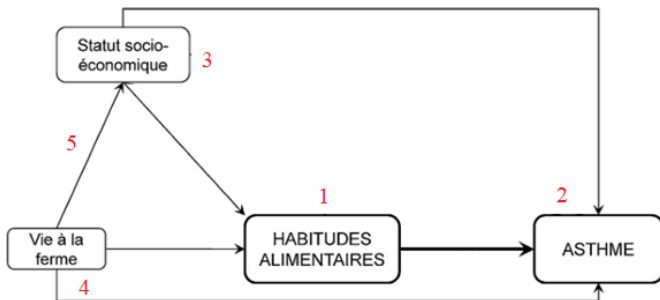
Tout d'abord, quelles variables doivent être ajoutées au DAG? Nous dénoterons par  $A$  un traitement (ex.: binaire) et par  $Y$  une issue.

Le diagramme causal doit absolument contenir les variables:

- ▶ Le traitement  $A$
- ▶ L'issue ou la réponse  $Y$
- ▶ Toute variable qui est le parent de deux variables ou plus déjà présentes dans le diagramme

## Supposons...

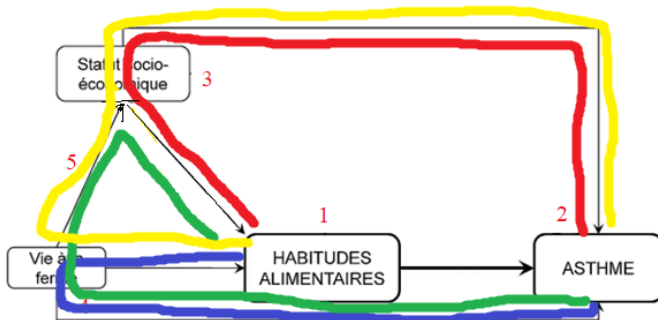
On commence donc par dessiner A (HA) et Y (AS), puis on ajoute les variables qui sont parents de A et Y simultanément. Et ainsi de suite...



Ici, on a mis l'exposition (HA) et la réponse (AS) en première étape. Ensuite, on a ajouté le statut SSÉ, puis la vie à la ferme, puis on a indiqué que la vie à la ferme affecte aussi le SSÉ et HA.

# La confusion

Causée par un trajet qu'on appelle *de porte arrière* à l'exposition qui n'est pas bloqué (pas de facteur de collision  $\rightarrow \leftarrow$ ) et se rend à la réponse:

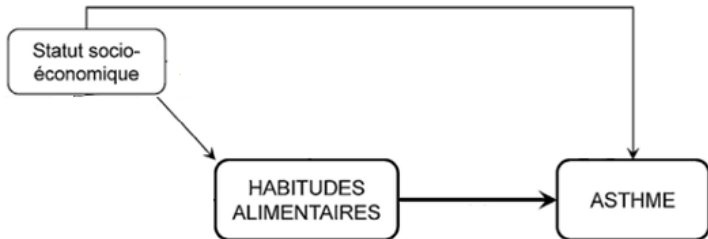


Alors que l'effet causal de *HA* sur *AS* ici est clairement celui qui passe par la flèche *HA* → *AS*, ces trajets de porte arrière créent des associations entre *HA* et *AS* qui ne sont pas causales et sont donc trompeuses.

Plus généralement, un facteur confondant est une variable qui, simultanément:

- ▶ est associée à l'exposition
- ▶ ne vient pas après l'exposition dans le temps (elle vient donc avant mais peut être un ancêtre sans être un parent de l'exposition)
- ▶ et cause la réponse

(Par ex., le SSÉ dans ce DAG, ou la vie à la ferme)



# La stratification sur un facteur de collision, un problème différent



- ▶ Plutôt que d'avoir, d'emblée, un trajet (de porte arrière) problématique, débloqué, qui crée une association entre  $A$  et  $Y$ , la stratification sur un facteur de collision crée un biais après avoir stratifié sur le dit facteur de collision



# La stratification sur un facteur de collision, un problème différent

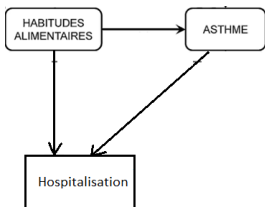


- ▶ Plutôt que d'avoir, d'emblée, un trajet (de porte arrière) problématique, débloqué, qui crée une association entre  $A$  et  $Y$ , la stratification sur un facteur de collision crée un biais après avoir stratifié sur le dit facteur de collision
- ▶ Ainsi, à prime abord, il n'y a pas d'association trompeuse due au facteur de collision

# La stratification sur un facteur de collision, un problème différent



- ▶ Plutôt que d'avoir, d'emblée, un trajet (de porte arrière) problématique, débloqué, qui crée une association entre  $A$  et  $Y$ , la stratification sur un facteur de collision crée un biais après avoir stratifié sur le dit facteur de collision
- ▶ Ainsi, à prime abord, il n'y a pas d'association trompeuse due au facteur de collision
- ▶ L'association trompeuse se crée lorsqu'on sélectionne les données sur un sous-groupe qui correspond à une des catégories du facteur de collision



Par ex.:

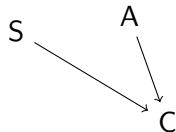
hospitalisation est un facteur de collision.

# Un autre exemple de stratification sur un facteur de collision (intuition du biais)



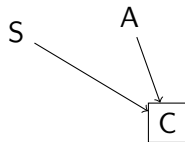
Soit une population où:

- ▶ l'âge (A) et le sexe (S) sont marginalement indépendants
- ▶ les risques de maladie du coeur (C) augmentent avec l'âge et sont plus élevés chez les personnes de sexe masculin

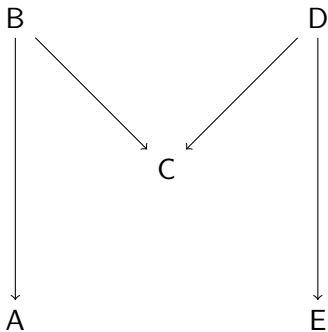


Que pouvons-nous inférer

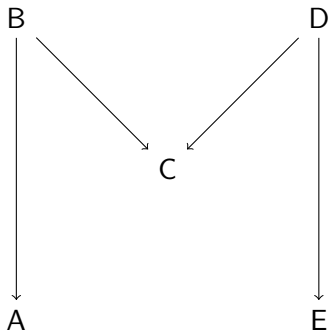
- ▶ à propos d'un individu avec maladie du coeur ( $C=1$ ) de sexe féminin?
- ▶ à propos d'un individu sans maladie du coeur ( $C=0$ ) très âgé?



Un exemple particulier où on combine trajet de porte arrière et collision..



Un exemple particulier où on combine trajet de porte arrière et collision..



Ici, n'ajuster pour aucune variable nous permet d'évaluer l'effet causal de  $A$  sur  $E$ . Par contre, ajuster pour  $C$  seulement en pensant que c'est le seul facteur confondant ( $A \leftarrow C \rightarrow E$ ) serait problématique.

Choix d'un ensemble d'ajustement

Rappel: Les différentes variables pour lesquelles on pourrait ajuster (Austin, 2011)

1. Toutes les variables mesurées à l'entrée dans la cohorte
2. Toutes les variables mesurées à l'entrée associées au traitement
3. Toutes les variables qui affectent le résultat
4. Toutes les variables qui affectent le traitement et le résultat simultanément

Différents 'conflits' dans la littérature. En général, on ira avec 4).

# Consolider les informations vues précédemment



- Une fois qu'on a un ensemble de variables qui peuvent créer des trajets de porte arrière (celles discutées au 3e item de la p. 13), on ajoutera aussi toute variable sur laquelle on conditionne qui est reliée avec au moins une des variables dans le diagramme.



# Consolider les informations vues précédemment



- ▶ Une fois qu'on a un ensemble de variables qui peuvent créer des trajets de porte arrière (celles discutées au 3e item de la p. 13), on ajoutera aussi toute variable sur laquelle on conditionne qui est reliée avec au moins une des variables dans le diagramme.
- ▶ Ainsi, si on sélectionne les gens à l'hôpital, et que ceux-ci ont plus de chance d'être asthmatiques, alors on ajouterait la variable hôpital dans le diagramme de l'effet de HA sur AS.

# Consolider les informations vues précédemment

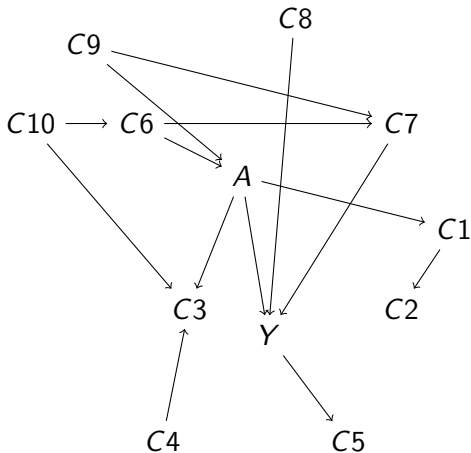


- ▶ Une fois qu'on a un ensemble de variables qui peuvent créer des trajets de porte arrière (celles discutées au 3e item de la p. 13), on ajoutera aussi toute variable sur laquelle on conditionne qui est reliée avec au moins une des variables dans le diagramme.
- ▶ Ainsi, si on sélectionne les gens à l'hôpital, et que ceux-ci ont plus de chance d'être asthmatiques, alors on ajouterait la variable hôpital dans le diagramme de l'effet de HA sur AS.
- ▶ L'ensemble final de toutes les variables ajoutées dans notre DAG est un ensemble possible d'ajustement (pas nécessairement un bon!). Comment choisir les variables à garder dans l'ajustement?

La méthode discutée en p. 14 (ajout des variables importantes, parents de deux variables dans le DAG, etc.) mène généralement à un ensemble d'ajustement qui permet d'identifier l'effet. C'est-à-dire qu'il n'y a pas besoin de le modifier, s'il ne contient pas de variables sur lesquelles il y a sélection ou qui viennent après le traitement, dans le temps.

La méthode dont on discute maintenant nous permettra plutôt de s'assurer qu'un ensemble de variables est adéquat pour identifier l'effet causal lorsqu'on débute avec un diagramme déjà construit.

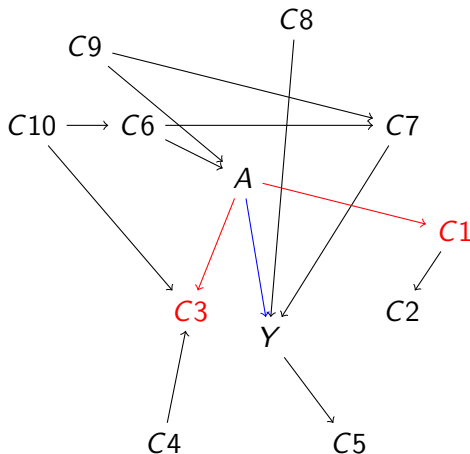
## Un DAG un peu plus complexe...



Et l'ensemble d'ajustement possible  
 $\{C1, C2, C3, C4, C5, C6, C7, C8, C9, C10\}$ .

# Approche en 6 étapes ( Pearl, 2000; discuté dans Shrier et Platt, 2008)

**Étape 1.** Un facteur confondant ne peut pas être un descendant du traitement. Enlever tout **descendant** de l'ensemble possible.



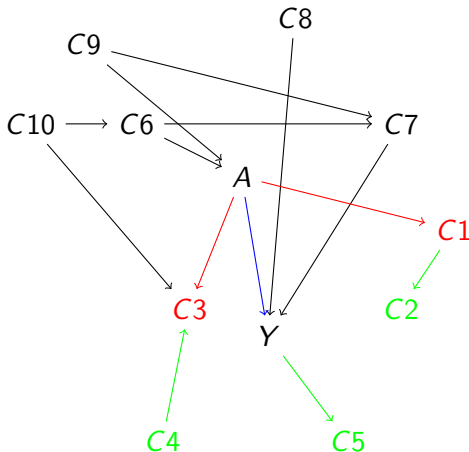
Il nous reste donc:  $\{C2, C4, C5, C6, C7, C8, C9, C10\}$ . Bien sûr,  $Y$  ne fait pas partie de l'ensemble.

Notes en lien avec l'étape 1:

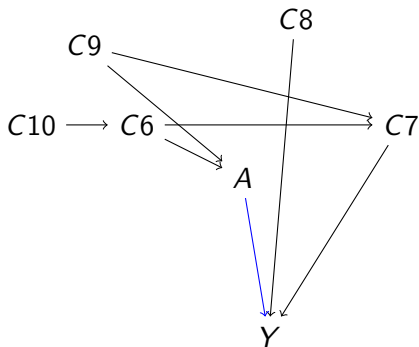
- ▶ On sait qu'ajuster sur des variables qui viennent après le traitement est dangereux: On pourrait conditionner sur des facteurs de collision
- ▶ Il est aussi inapproprié d'ajuster pour une variable associée assez fortement avec une variable descendante de  $A$  (exemple:  $C2$ )

**Étape 2.** Enlever toute variable qui **satisfait toutes les conditions suivantes** de l'ensemble possible:

- ▶ N'est pas un ancêtre du traitement  $A$
- ▶ N'est pas un ancêtre de l'issue  $Y$
- ▶ N'est pas un ancêtre des variables dans l'ensemble possible



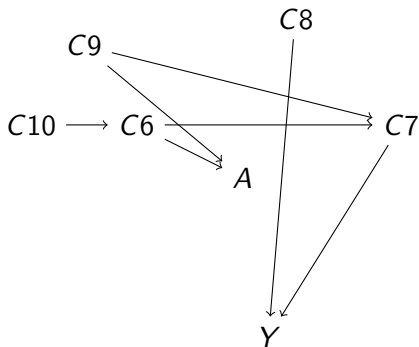
**Étape 2.** Après avoir enlevé les variables des étapes 1 et 2:



Il nous reste donc:  $\{C6, C7, C8, C9, C10\}$ .

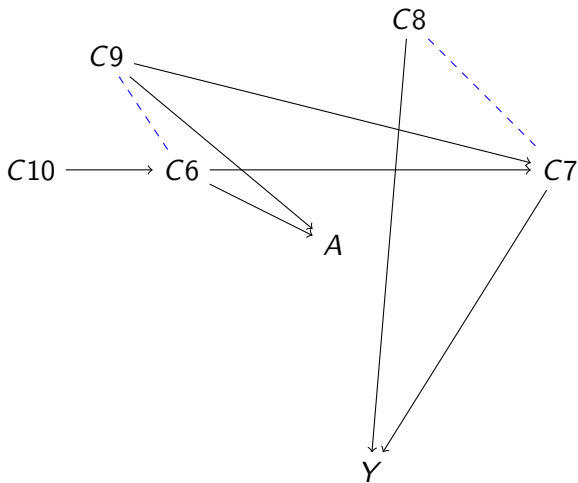


**Étape 3.** Enlever toutes les flèches qui sortent du traitement A



Il nous reste toujours  $\{C6, C7, C8, C9, C10\}$ .

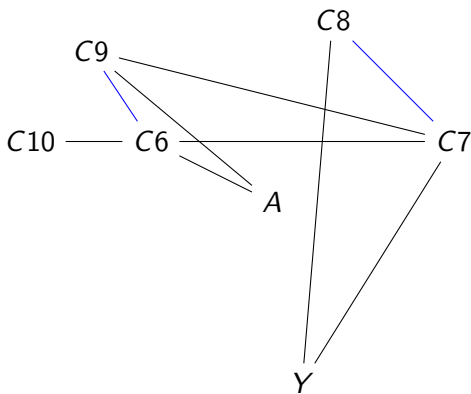
**Étape 4.** Connecter toutes les paires de parents dans le DAG qui ont un enfant en commun (celles qui n'étaient pas déjà connectées).



Note par rapport à l'étape 4: Cette étape permet de tenir compte des cas où deux variables causent une 3e variable et où l'ajustement/la condition sur la 3e variable crée donc une association conditionnelle des deux premières.

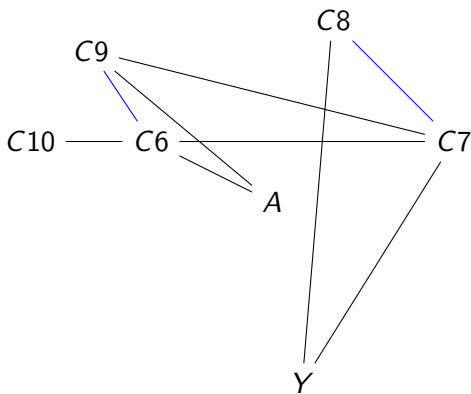
Truc pour faciliter la tâche: Pour chaque variable (une à une), vérifier quels sont ses parents et les relier ensemble (pas les ancêtres mais bien les parents).

**Étape 5.** Enlever toutes les têtes des flèches.



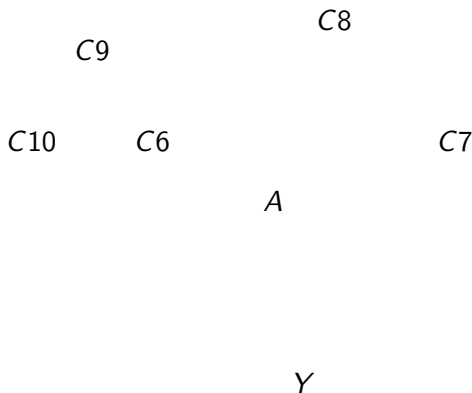
En fait, les têtes ne nous servaient qu'à pouvoir identifier les variables qui sont des facteurs de collision. Une fois qu'on a rejoint les parents d'un facteur de collision par une ligne (pour représenter l'association conditionnelle), on peut enlever les têtes.

**Étape 6A.** Ensemble final à confirmer.



Supposons qu'on finit la procédure avec cet ensemble de variables. L'ensemble d'ajustement est donc  $\{C6, C7, C8, C9, C10\}$ . Vérifions d'abord en enlevant les lignes émanant de ces variables (diapositive suivante).

**Étape 6B.** Après avoir enlevé les lignes émanantes:



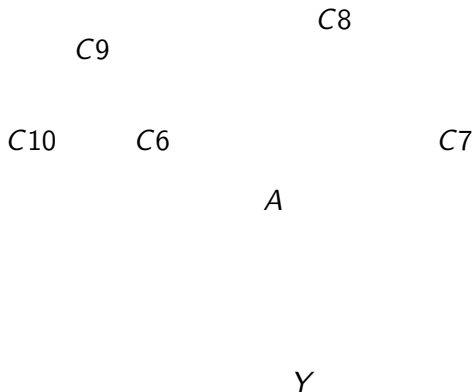
L'ensemble d'ajustement  $\{C6, C7, C8, C9, C10\}$  est approprié! En effet, il n'y a plus de ligne reliant  $A$  et  $Y$ .

Cela ne veut pas dire qu'il s'agit de l'ensemble minimal. En particulier, les variables  $C8$  et  $C10$  ne sont clairement pas essentielles (revoir le DAG du début). On pourrait refaire la procédure en les enlevant dès le départ.

## Ensemble alternatif



Si je choisissais l'ensemble d'ajustement  $\{C_6, C_9\}$  dès le départ, j'obtiendrais en étape 6:



La même chose!

## Discussion



## Notes (1)

- ▶ Cette procédure ne nous permet pas de trouver le meilleur ensemble (ou le plus petit ensemble) d'ajustement. Elle nous permet de trouver un ensemble qui, s'il est bien spécifié, brise les trajets de porte arrière qui biaisent l'estimation de l'effet causal.
- ▶ On peut reproduire ces 6 étapes sur tous les ensembles de variables possibles et trouver soi-même le plus petit ensemble d'ajustement, bien qu'il ne soit pas nécessaire d'avoir le plus petit ensemble.

## Notes (2)

- ▶ Les facteurs de collision ne doivent tout simplement pas faire partie de l'ensemble d'ajustement (sauf dans certains cas très rares où ils sont sur des trajets de porte arrière et où on pourrait bénéficier d'ouvrir ces trajets où il y a des facteurs de collision).
- ▶ Par contre, si les données constituent déjà en un sous-groupe de la population sélectionné sur un facteur de collision, il faudra jumeler l'ajustement pour la confusion et l'ajustement pour la sélection (poids IIV pour les visites, poids inverses à la probabilité d'être manquant, etc.).

## Notes (3)

- ▶ Ajouter des variables à un ensemble d'ajustement approprié peut mener à un nouvel ensemble d'ajustement inapproprié! Plus de variables  $\neq$  mieux.
- ▶ Il faut aussi penser à comment on spécifiera nos modèles (traitement, ou réponse) pour l'ajustement de la confusion.
  - ▶ Cas 1: Si je suis incapable de spécifier le format de l'âge dans le modèle pour le traitement, et que c'est un facteur confondant, alors l'estimation ne sera pas consistante même si j'ai gardé âge dans mon ensemble d'ajustement.
  - ▶ Cas 2. Si je suis incapable de trouver la forme de l'âge dans mon modèle traitement mais que je peux trouver un autre ensemble d'ajustement approprié qui ne contienne pas l'âge, ça peut être intéressant...

# Retour (1)



Guo et al., 2022. Raisons de faire de la sélection de variables:

- ▶ Trouver un ensemble qui permet d'identifier l'effet causal (aujourd'hui)
- ▶ Réduction des dimensions
- ▶ Robustesse à la mauvaise spécification des modèles
- ▶ Efficacité (réduction de la variance d'estimation)
- ▶ Éthique ou raisons économiques

Guo et al., 2022. Raisons de faire de la sélection de variables:

- ▶ Trouver un ensemble qui permet d'identifier l'effet causal (aujourd'hui)
- ▶ Réduction des dimensions
- ▶ Robustesse à la mauvaise spécification des modèles
- ▶ Efficacité (réduction de la variance d'estimation)
- ▶ Éthique ou raisons économiques

Pour la raison 1 discutée aujourd'hui, l'inférence dépendra aussi de si on a la connaissance structurelle complète (tout le diagramme), partielle, ou aucune connaissance structurelle. Il existe des méthodes pour estimer un effet causal lorsqu'on a de la connaissance partielle (voir par ex., VanderWeele et Shpitser, 2011).

Nos objectifs aujourd'hui:

1. Comprendre ce que sont les biais de confusion et de stratification sur un facteur de collision
2. Choisir un ensemble de variables de base pour notre diagramme causal
3. Trouver un ensemble suffisant de variables d'ajustement pour estimer (ou identifier, dans le langage causal) un effet causal (c'est-à-dire, pour estimer de façon consistante l'effet causal).

# Conclusion



Je vous invite à aller voir les articles en références pour plus de détails.

Merci beaucoup pour votre attention. Merci à Mr. Denis Talbot pour l'invitation.

Les questions sont bienvenues!

[janie.coulombe@umontreal.ca](mailto:janie.coulombe@umontreal.ca)

Diapo. disponibles au <https://janiecoulombestat.github.io/>

# Références I



Peter C Austin, An introduction to propensity score methods for reducing the effects of confounding in observational studies, *Multivariate Behavioral Research* **46** (2011), no. 3, 399–424.



Hailey R Banack and Jay S Kaufman, From bad to worse: Collider stratification amplifies confounding bias in the “obesity paradox”, *European Journal of Epidemiology* **30** (2015), no. 10, 1111–1114.



O Dumas, V Siroux, N Le Moual, and R Varraso, Approches d'analyse causale en épidémiologie, *Revue d'épidémiologie et de santé publique* **62** (2014), no. 1, 53–63.



F Richard Guo, Anton Rask Lundborg, and Qingyuan Zhao, Confounder selection: Objectives and approaches, *arXiv preprint arXiv:13871* (2022).








Miguel A Hernán, Sonia Hernández-Díaz, and James M Robins, A structural approach to selection bias, *Epidemiology* **15** (2004), no. 5, 615–625.



# Références II



-  Jerzy S Neyman, On the application of probability theory to agricultural experiments. Essay on principles. Section 9 (translation published in 1990), Statistical Science **5** (1923), no. 4, 472–480.
-  Judea Pearl, Models, reasoning and inference, Cambridge, UK: Cambridge University Press **19** (2000), no. 2.
-  Donald B Rubin, Estimating causal effects of treatments in randomized and nonrandomized studies, Journal of Educational Psychology **66** (1974), no. 5, 688–701.
-  Ian Shrier and Robert W Platt, Reducing bias through directed acyclic graphs, BMC Medical Research Methodology **8** (2008), no. 1, 1–15.
-  Tyler J VanderWeele and Ilya Shpitser, A new criterion for confounder selection, Biometrics **67** (2011), no. 4, 1406–1413.