UNIVERSITÉ DU QUÉBEC À MONTRÉAL

# UTILISATION DE MODÈLES GRAPHIQUES ET DU SCORE DE PROPENSION POUR L'ESTIMATION DE L'EFFET CAUSAL : APPLICATION CHEZ LES PATIENTS ATTEINTS D'UNE HÉMORRAGIE INTRACÉRÉBRALE SOUS WARFARINE

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN MATHÉMATIQUES

CONCENTRATION STATISTIQUE

PAR

ESSAÏD OUSSAÏD

MAI 2015

UNIVERSITÉ DU QUÉBEC À MONTRÉAL
Service des bibliothèques

*Avertissement*

La diffusion de ce mémoire se fait dans le respect des droits de son auteur, qui a signé le formulaire *Autorisation de reproduire et de diffuser un travail de recherche de cycles supérieurs* (SDU-522 – Rév.01-2006). Cette autorisation stipule que «conformément à l'article 11 du Règlement no 8 des études de cycles supérieurs, [l'auteur] concède à l'Université du Québec à Montréal une licence non exclusive d'utilisation et de publication de la totalité ou d'une partie importante de [son] travail de recherche pour des fins pédagogiques et non commerciales. Plus précisément, [l'auteur] autorise l'Université du Québec à Montréal à reproduire, diffuser, prêter, distribuer ou vendre des copies de [son] travail de recherche à des fins non commerciales sur quelque support que ce soit, y compris l'Internet. Cette licence et cette autorisation n'entraînent pas une renonciation de [la] part [de l'auteur] à [ses] droits moraux ni à [ses] droits de propriété intellectuelle. Sauf entente contraire, [l'auteur] conserve la liberté de diffuser et de commercialiser ou non ce travail dont [il] possède un exemplaire.»

# REMERCIEMENT

Tout d'abord, je tiens à exprimer ma profonde gratitude à ma directrice de mémoire, Madame Juli Atherton, pour la confiance qu'elle m'a témoignée en acceptant la direction de mon mémoire, pour ses multiples conseils et pour son soutien sans faille. Sa grande disponibilité, sa patience et ses qualités humaines m'ont énormément aidé à réaliser ce travail.

Je tiens à remercier vivement le Docteur Sherry Hsiang-Yi Chou, neurologue à Brigham and Women's Hospital et Professeur adjoint en neurologie à Harvard Medical School, pour m'avoir fourni les données et les informations dont j'avais besoin. Sans sa confiance, sa collaboration et ses conseils, ce travail n'aurait pu être mené à bien.

Ma gratitude s'adresse aussi à Adrian Roshan Vetta, Professeur associé à l'Université McGill, pour l'aide qu'il m'a apportée dans l'étude des modèles graphiques probabilistes. Mes discussions avec lui ont été particulièrement enrichissantes.

Je remercie également tous les professeurs du Département de mathématiques de l'UQAM, et tout particulièrement Serge Alalouf, Sorana Froda et Geneviève Lefebvre pour leur aide et leurs conseils durant mes années de maîtrise.

Un grand merci également à Gisèle Legault, analyste de l'informatique au laboratoire des cycles supérieurs en mathématiques, pour son assistance informatique.

Pour finir, je tiens à remercier tous mes collègues de classe, en particulier Seydou Sall avec qui j'ai partagé des moments agréables dans notre espace de travail partagé.

[Cette page a été laissée intentionnellement blanche]

TABLE DES MATIÈRES

# LISTE DES TABLEAUX

[Cette page a été laissée intentionnellement blanche]

# LISTE DES FIGURES

# LISTE DES ACRONYMES

| | |
|---|---|
| wICH | Warfarin-associated Intracerebral Hemorrhage : hémorragie intracérébrale sous warfarine |
| INR | International Normalized Ratio : rapport international normalisé |
| AVC | Accident Vasculaire Cérébral |
| AVK | Anti-vitamines K |
| FFP | Fresh Frozen Plasma : plasma frais congelé |
| rFVIIa | recombinant Factor VIIa : facteur VIIa recombinant |
| PCC | prothrombin complex concentrates : concentré de complexe prothrombique |
| SUTVA | Stable Unit Treatment Value Assumption : hypothèse d'absence d'effets de diffusion du traitement |
| DAG | Directed Acyclic Graph : graphe orienté acyclique |
| CART | Classification and Regression Trees : arbres de classification et de régression |

# RÉSUMÉ

L'estimation de l'effet causal dans une étude observationnelle reste l'une des questions les plus controversées dans le domaine de l'épidémiologie. Diverses méthodes ont été développées à cette fin. Parmi ces méthodes, nous citons : la stratification, la régression et l'appariement sur les caractéristiques observables des individus, ainsi que la technique du score de propension. Cependant, aucune méthode ne peut garantir un résultat sans équivoque. En effet, le risque de biais demeure présent. L'objectif de notre travail est d'estimer l'effet du traitement par le concentré de complexe prothrombique (PCC) sur la mortalité à 6 mois et sur l'expansion de l'hématome chez les patients atteints d'une hémorragie intracérébrale sous warfarine. Afin d'atteindre cet objectif, nous avons combiné deux principales méthodes : l'analyse graphique de la causalité d'une part et l'estimation de l'effet causal par la technique du score de propension d'autre part. La première sert à sélectionner l'ensemble de covariables à inclure dans l'estimation du score de propension. Elle consiste à représenter les éventuelles relations causales liant les variables, par un graphe orienté acyclique, et appliquer ensuite le critère porte-arrière sur le graphe obtenu afin de sélectionner l'ensemble recherché. La deuxième méthode consiste en premier lieu à estimer le score de propension à partir des variables déterminées par la méthode graphique, puis à estimer l'effet causal du traitement par une méthode utilisant le score de propension estimé. L'application des deux méthodes montre un effet statistiquement non significatif du traitement PCC sur les deux réponses considérées.

Mots-clés : étude observationnelle, effet causal, score de propension, graphe orienté acyclique, critère porte-arrière.

# CHAPITRE I

# INTRODUCTION

L'objectif principal de ce travail est de présenter une approche d'estimation de l'effet causal sur des données provenant d'une étude épidémiologique non expérimentale. Dans cette introduction, nous commençons tout d'abord par présenter le contexte de l'étude ainsi que la problématique considérée. Nous présentons par la suite l'approche que nous avons retenue pour répondre à cette problématique.

## 1.1 Contexte épidémiologique de l'étude

L'hémorragie intracérébrale sous warfarine (en anglais *Warfarin-associated intracerebral hemorrhage*, wICH) est considérée comme l'une des formes d'accidents vasculaires cérébraux (AVC) les plus mortelles (Cai *et al.*, 2014). Cette maladie iatrogène est causée par l'administration d'un anticoagulant, à savoir la warfarine, lors du traitement des caillots de sang.

La prise de warfarine doit être contrôlée par la mesure d'un indice connu sous le nom de rapport international normalisé (*international normalized ratio*, INR) qui désigne la vitesse à laquelle le sang forme les caillots. Une faible valeur de cet indice correspond à une vitesse de formation des caillots élevée. Chez les personnes normales, l'INR se situe entre 0.8 et 1.2. Concernant les personnes sous traitement anticoagulant, on vise généralement un INR compris entre 2 et 3. Néanmoins, pour

les personnes ayant subi une chirurgie, l'INR visé est de 3 à 4.5 (Schulman *et al.*, 2008). Lorsque la valeur de l'INR est supérieure à 5, le risque hémorragique est particulièrement accru.

Les anticoagulants sont également connus sous le nom d'anti-vitamines K (AVK) du fait qu'ils produisent l'effet inverse de la vitamine K, qui, quant à elle, favorise la coagulation sanguine. Il paraît donc naturel d'utiliser la vitamine K comme traitement de l'hémorragie intracérébrale sous warfarine (wICH). Ce traitement est souvent combiné à d'autres traitements comme le plasma frais congelé (*fresh frozen plasma*, FFP), le facteur VIIa recombinant (*recombinant factor VIIa*, rFVIIa) ou le concentré de complexe prothrombique (*prothrombin complex concentrates*, PCC).

Le Docteur Sherry Hsiang-Yi Chou est l'un des chercheurs qui s'intéressent à l'évolution et au traitement de la wICH. Neurologue à Brigham and Women's Hospital et Professeur adjoint en neurologie à Harvard Medical School aux États-Unis, elle est auteure de plusieurs publications dans son domaine, notamment sur l'hémorragie intracérébrale et l'hémorragie méningée.

## 1.2 Problématique

Habituellement, pour traiter la wICH, le Docteur Chou combine les deux coagulants suivants : la vitamine K et le FFP. L'utilisation de ces coagulants nécessite environ 30 heures pour pouvoir renverser l'effet de la warfarine. L'utilisation du facteur VIIa recombinant (rFVIIa), en plus du traitement habituel permet de réduire ce temps (Cai *et al.*, 2014). De plus, il présente l'avantage de ne pas augmenter le risque thrombo-embolique (risque de formation des caillots). Toutefois, ce coagulant ne permet pas de contrôler tous les facteurs affectés par l'utilisation de la warfarine.

Une alternative au facteur VIIa recombinant est d'utiliser le concentré de complexe prothrombique (PCC) en addition au traitement standard (la vitamine K et le FFP). Cai *et al.* (2014) ont montré que, comme pour le rFVIIa, le PCC permet de diminuer le temps de renversement de l'effet de la warfarine ainsi que la quantité du FFP nécessaire au traitement. Mais aussi, il ne présente aucun risque thrombo-embolique. Cependant, l'utilisation du PCC permet-elle d'aboutir à de meilleurs résultats en matière de réduction du taux de mortalité des patients et d'amélioration de leur état de santé ? C'est à partir de cette interrogation que nous tentons, dans ce travail, d'explorer l'effet du PCC sur la mortalité à 6 mois ainsi que sur l'expansion de l'hématome chez les patients atteints d'une hémorragie intracérébrale sous warfarine. Pour cela, nous disposons d'une base de données résumant les différentes caractéristiques des patients.

## 1.3 Les données

Les données que nous allons utiliser dans cette étude proviennent de Brigham and Women's Hospital. Elles consistent en 28 variables mesurées sur 89 patients souffrant d'une hémorragie intracérébrale sous warfarine (wICH). Parmi eux, 49 patients ont reçu le PCC en plus du traitement standard et 40 patients ont reçu uniquement le traitement standard. Initialement, la base de données contenait des informations relatives à 130 patients, 33 d'entre eux sont exclus de l'analyse en raison de leur état de santé très critique et 8 patients sont retirés de la base de données pour cause d'antécédents de traumatisme crânien. D'autres patients seront également exclus de certaines de nos analyses en raison de données manquantes. La liste détaillée des variables retenues pour réaliser l'analyse est présentée dans le tableau 1.1.

Tableau 1.1: Liste et description des variables

| Variable | Libellé de la variable | Format |
|---|---|---|
| Subject_ID | Identifiant du patient | Nominale |
| age | Âge du patient | Continue |
| gender | Sexe biologique du patient : elle est égale à 1 pour les femmes et 0 pour les hommes | Dichotomique |
| Group | Indicatrice de traitement par le PCC : elle est égale à 1 si traitement par PCC et 0 sinon | Dichotomique |
| afib | Indicatrice d'antécédent de fibrillation auriculaire : elle est égale à 1 si antécédent et 0 sinon | Dichotomique |
| dvt_pe | Indicatrice d'antécédent de thrombose veineuse profonde ou d'embolie pulmonaire : elle est égale à 1 si antécédent et 0 sinon | Dichotomique |
| valve | Indicatrice d'antécédent de port de valve cardiaque mécanique : elle est égale à 1 si antécédent et 0 sinon | Dichotomique |
| cmy | Indicatrice d'antécédent de cardiomyopathie : elle est égale à 1 si antécédent et 0 sinon | Dichotomique |
| antiplatelets | Indicatrice indiquant si le patient est sous traitement antiplaquettaire ou non : elle est égale à 1 si oui et 0 sinon | Dichotomique |
| cancer | Indicatrice d'antécédent de cancer : elle est égale à 1 si antécédent et 0 sinon | Dichotomique |
| cad | Indicatrice d'antécédent de coronaropathie : elle est égale à 1 si antécédent et 0 sinon | Dichotomique |
| renal_failure | Indicatrice d'antécédent d'insuffisance rénale : elle est égale à 1 si antécédent et 0 sinon | Dichotomique |

| Variable | Libellé de la variable | Format |
|---|---|---|
| alcohol | Indicatrice d'antécédent de consommation d'alcool : elle est égale à 1 si antécédent et 0 sinon | Dichotomique |
| osh | Indicatrice indiquant si le patient est transféré ou non d'un autre hôpital : elle est égale à 1 si oui et 0 sinon | Dichotomique |
| sbp | La pression artérielle systolique du patient à son arrivée à l'hôpital en millimètres de mercure (mmHg) | Continue |
| inr_bwh | L'INR du patient à son arrivée à l'hôpital | Continue |
| inr_3_to_6hr | L'INR du patient mesuré entre 3 et 6 heures après son arrivée à l'hôpital | Continue |
| ffp | Nombre d'unités de FFP transfusées | Continue |
| location | L'emplacement de l'hémorragie intracérébrale : elle a 5 modalités (1, 2, 3, 4 et 5) | Nominale |
| postfossa | Indicatrice indiquant si l'hémorragie intracérébrale se situe dans la fosse postérieure : elle est égale à 1 si oui et 0 sinon | Dichotomique |
| volume | Volume de l'hémorragie intracérébrale en centimètres cubes (cc) | Continue |
| ivh | Indicatrice indiquant si le patient est touché ou non par une hémorragie intraventriculaire : elle est égale à 1 si oui et 0 sinon | Dichotomique |
| expansion | Indicatrice indiquant s'il y a eu ou non expansion de l'hématome lié à l'hémorragie intracérébrale : elle est égale à 1 si oui et 0 sinon | Dichotomique |
| evd | Indicatrice indiquant si le patient a bénéficié ou non d'un drainage ventriculaire externe (DVE) : elle est égale à 1 si oui et 0 sinon | Dichotomique |

| Variable | Libellé de la variable | Format |
|---|---|---|
| surgery | Indicatrice indiquant si le patient a subi ou non une neurochirurgie pour l'hémorragie intracérébrale : elle est égale à 1 si oui et 0 sinon | Dichotomique |
| troponin | Indicatrice indiquant si le patient a subi ou non une élévation de troponine après l'hémorragie intracérébrale : elle est égale à 1 si oui et 0 sinon | Dichotomique |
| Cdvt_pe | Indicatrice indiquant si le patient est touché ou non par une thrombose veineuse profonde ou par une embolie pulmonaire après l'hémorragie intracérébrale : elle est égale à 1 si oui et 0 sinon | Dichotomique |
| 6mo_mortality | Indicatrice indiquant si le patient est décédé ou non 6 mois après l'apparition de l'hémorragie intracérébrale : elle est égale à 1 si oui et 0 sinon | Dichotomique |

## 1.4 Contexte statistique et approche proposée

Afin de répondre à la problématique posée, nous cherchons à déterminer, à partir des données disponibles l'effet causal du traitement par PCC (variable *Group*) sur la mortalité à 6 mois (*6mo_mortality*), et sur l'expansion de l'hématome (*expansion*) chez les patients. Dans notre étude, le médecin affecte le traitement par PCC selon les caractéristiques des patients, nous sommes donc dans le cas de ce que l'on appelle une « étude observationnelle », ou une « étude non expérimentale », différente des études expérimentales comme les essais randomisés où chaque patient est affecté de manière aléatoire à un groupe traité ou à un groupe contrôle.

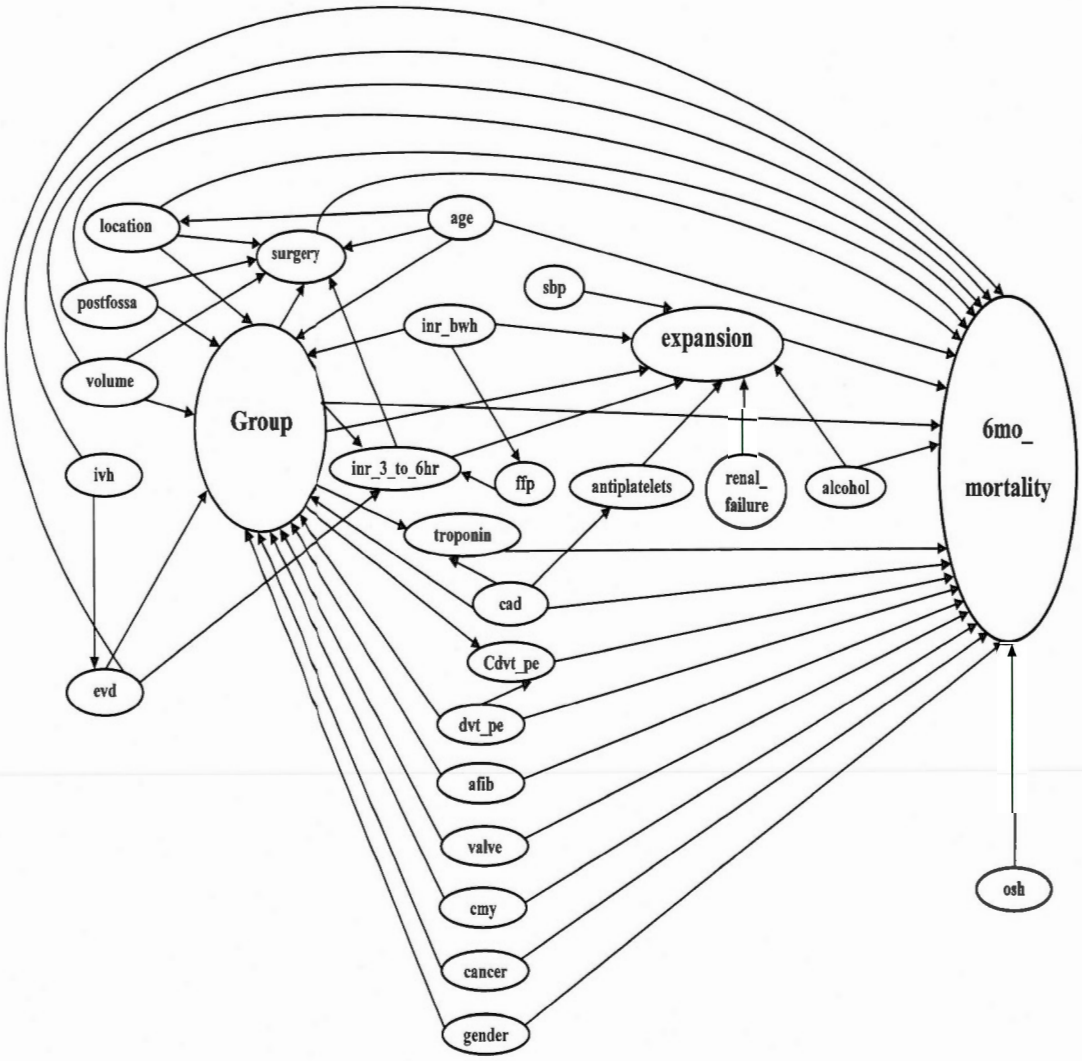
La randomisation permet d'éliminer les différences entre les groupes de traitement

et d'interpréter des associations statistiques comme étant des relations de cause à effet. Toutefois, tel n'est pas le cas dans les études non expérimentales, où l'effet de la variable de traitement sur la variable de réponse est confondu, en raison notamment de la présence de variables affectant à la fois la variable de traitement et la variable de réponse. Dans ce cas, l'association statistique ne veut pas dire causation, et un biais apparaît. Ces variables sont également présentes dans les études expérimentales, mais l'assignation aléatoire du traitement permet de créer des groupes de traitement similaires en matière de caractéristiques des individus.

Grâce à certaines hypothèses, il devient possible d'identifier l'effet causal à partir des quantités observables et de l'estimer ensuite par une méthode adéquate, telle que la stratification, la régression et l'appariement sur les caractéristiques observables, ou encore par la technique du score de propension. Dans ce travail, nous appliquons les méthodes qui répondent le mieux à la problématique de départ. Ainsi, il paraît raisonnable d'opter pour l'approche contrefactuelle de la causalité qui permet de mieux exprimer les hypothèses nécessaires pour l'identification de l'effet causal.

Dans cette étude, nous avons opté pour la technique du score de propension afin d'estimer l'effet causal du traitement. Cette méthode donne généralement de bons résultats et elle demeure la plus appropriée lorsque le nombre de variables est grand (Rubin, 1997). Le score de propension d'un individu est la probabilité que l'individu reçoive le traitement conditionnellement à un vecteur de covariables. Son intérêt réside dans le fait qu'il permet de remplacer le vecteur des covariables lors d'une stratification ou d'un appariement sur ces covariables. À cet égard, il est donc un puissant outil de réduction de dimension. Néanmoins, dans une étude non expérimentale, le score de propension n'est pas connu et doit être estimé. La principale difficulté pour l'estimation du score de propension réside dans le choix des covariables à inclure dans le modèle. Pour ce faire, les méthodes basées sur les

Figure 1.1: Graphe causal représentant les relations entre les variables mesurées




graphes causaux permettent de sélectionner les covariables pertinentes pour l'estimation. Un graphe causal est un graphe orienté représentant les relations de cause à effet entre des variables. Le graphe de la figure 1.1 représente les relations de cause à effet entre les variables décrites dans la section 1.3. Il est obtenu en se basant sur les connaissances du Docteur Chou en matière de liens causaux existants entre les différentes variables. Ainsi, par exemple, la flèche allant de la variable

*antiplatelets* à la variable *expansion* signifie que le traitement antiplaquettaire a un éventuel effet causal sur l'expansion de l'hématome. Ce graphe nous sera d'un grand intérêt pour la suite de notre travail.

## 1.5 Organisation du document

Après cette introduction générale, nous présentons en détails dans le prochain chapitre l'approche contrefactuelle de la causalité. En premier lieu, nous introduisons le modèle causal de Rubin (1974), ensuite nous exposons les conditions nécessaires pour l'identification de l'effet causal dans une étude non expérimentale, notamment l'hypothèse d'ignorabilité forte du traitement. Enfin, nous exposons quelques méthodes d'estimation de l'effet causal sur la base des caractéristiques observables des individus.

Le troisième chapitre est consacré aux modèles graphiques probabilistes, notamment, aux réseaux bayésiens, et leurs propriétés. Nous mettons principalement l'accent sur une notion graphique connue sous le nom de « d-séparation », laquelle permet de déterminer si deux variables aléatoires sont indépendantes conditionnellement à une autre variable. Nous présentons également les réseaux bayésiens causaux qui permettent de prendre en considération les relations de cause à effet.

Dans le quatrième chapitre, nous introduisons d'autres modèles utilisant les graphes orientés, à savoir les modèles causaux à équations structurelles. Ces modèles nous permettent de sélectionner les variables pertinentes pour l'estimation du score de propension et de l'effet causal, ainsi que de faire le lien entre l'approche graphique et l'approche contrefactuelle de la causalité.

Le cinquième chapitre est dédié à la technique du score de propension. Nous commençons par donner les conditions nécessaires à l'identification de l'effet causal dans une étude non expérimentale, basées sur le score de propension. Ensuite,

nous présentons quelques méthodes permettant d'estimer le score de propension. En dernier lieu, nous exposons quelques méthodes d'estimation de l'effet causal, basées sur le score de propension estimé, à savoir : l'appariement sur le score de propension, la stratification sur le score de propension, la pondération inverse et l'ajustement sur le score de propension.

Le sixième chapitre est consacré à l'application des différentes méthodes mentionnées sur nos données afin de répondre à la problématique posée. Ainsi, nous procédons à l'estimation de l'effet causal du traitement PCC sur la mortalité à 6 mois et sur l'expansion de l'hématome lié à la wICH.

Ce mémoire se termine par une conclusion dressant le bilan de ce travail et des perspectives envisageables pour les travaux futurs.

# CHAPITRE II

# ANALYSE CONTREFACTUELLE DE LA CAUSALITÉ

L'un des problèmes majeurs que l'on rencontre souvent en inférence causale est de savoir estimer l'effet d'un traitement dans une étude non expérimentale. Une étude est dite non expérimentale si les sujets recevant le traitement sont déterminés d'une façon non aléatoire, à l'opposé par exemple d'une étude randomisée au cours de laquelle les sujets sont répartis de façon aléatoire entre le groupe de traitement et le groupe de contrôle.

## 2.1 Définition de l'effet causal par la notion de résultat contrefactuel

L'analyse contrefactuelle de la causalité a été essentiellement développée par Rubin (1974) dans son modèle d'évaluation de l'effet causal d'un traitement particulier sur les individus d'une population donnée. L'effet causal du traitement associé à chaque individu de la population est défini comme étant une comparaison (souvent une différence) entre deux résultats potentiels du traitement. Pour illustrer ce propos, nous commençons par présenter le modèle d'évaluation de Rubin et son cadre statistique. Supposons que l'on dispose d'une population de $N$ individus indexés par $i = 1, \ldots, N$. Soit $W_i$ une variable aléatoire qui prend la valeur 1 si l'individu $i$ est traité et 0 s'il ne l'est pas (témoin) et $Y_i$ la variable de réponse (le résultat) de l'individu $i$ au traitement. On définit $Y_{(1)i}$ comme étant la réponse potentielle de l'individu $i$ au traitement et $Y_{(0)i}$ sa réponse potentielle

au non-traitement. L'effet du traitement sur l'individu $i$ est donné par :

$$\tau_i = Y_{(1)i} - Y_{(0)i}. \tag{2.1}$$

L'effet $\tau_i$ est appelé effet causal individuel. Selon Hernán et Robins (2013), pour désigner les variables aléatoires $Y_{(1)i}$ et $Y_{(0)i}$, certains auteurs préfèrent l'expression « résultats contrefactuels » afin de souligner que ces deux résultats représentent des situations qui ne peuvent pas réellement se produire simultanément, et que pour chaque individu, seulement le contrefactuel qui correspond au traitement que l'individu a réellement reçu est observé. D'autres auteurs utilisent l'expression « résultats potentiels » pour dire que l'un de ces résultats peut être potentiellement observé selon le traitement reçu. Il est à noter que dans ce travail, nous ne faisons pas de distinction entre ces deux expressions.

En pratique, l'effet causal individuel est inobservable en raison de l'impossibilité d'observer simultanément les deux résultats potentiels (Holland, 1986). Néanmoins, sous certaines conditions, on peut toujours identifier à partir des variables observables une autre mesure de l'effet causal à savoir « l'effet causal moyen dans la population » que nous notons $\tau$ et qui se définit comme suit :

$$\tau = \mathbb{E}(Y_{(1)} - Y_{(0)}), \tag{2.2}$$

où $Y_{(0)}$ et $Y_{(1)}$ sont deux variables aléatoires mesurant les réponses potentielles au traitement dans la population. Ces variables sont liées à la variable mesurant le résultat observé $Y$ par la relation :

$$Y = WY_{(1)} + (1 - W)Y_{(0)}. \tag{2.3}$$

où $W$ est une variable aléatoire mesurant le traitement. Dans le cas d'une variable de réponse dichotomique, Rosenbaum (2010) donne trois façons de mesurer l'effet causal :

– Différence de risque causale

$$\mathbb{P}(Y_{(1)} = 1) - \mathbb{P}(Y_{(0)} = 1),$$

- Risque relatif causal

$$\frac{\mathbb{P}(Y_{(1)} = 1)}{\mathbb{P}(Y_{(0)} = 1)},$$

- Rapport de cotes causal

$$\frac{\mathbb{P}(Y_{(1)} = 1)/\mathbb{P}(Y_{(1)} = 0)}{\mathbb{P}(Y_{(0)} = 1)/\mathbb{P}(Y_{(0)} = 0)}.$$

Il est parfois plus commode de mesurer l'effet moyen du traitement uniquement dans la sous-population des traités (Rubin, 1977; Heckman *et al.*, 1997). Pour cela, il existe une autre mesure de l'effet causal nommée « effet causal moyen sur les traités » que nous notons $\tau_{(1)}$ et qui est donné par :

$$\tau_{(1)} = \mathbb{E}(Y_{(1)} - Y_{(0)} \mid W = 1). \tag{2.4}$$

Dans le cas d'une étude randomisée, les deux paramètres $\tau$ et $\tau_{(1)}$ sont équivalents (voir proposition 2.2.1), mais dans une étude non expérimentale, elles peuvent donner des résultats différents.

## 2.2 Le rôle de la randomisation dans l'identification de l'effet causal moyen

L'identification de l'effet causal moyen est un point crucial en inférence causale. Cette tâche n'est pas aisée en raison de l'impossibilité d'observer simultanément les deux résultats contrefactuels $Y_{(1)}$ et $Y_{(0)}$. Dans une expérience randomisée, on a que $(Y_{(1)}, Y_{(0)})$ indépendant de $W$, et cela permet d'identifier à la fois $\tau$ et $\tau_{(1)}$.

Proposition 2.2.1. *Si $(Y_{(1)}, Y_{(0)})$ est indépendant de $W$, alors l'effet causal moyen et l'effet causal moyen chez les traités sont identifiables et ils sont équivalents.*

Preuve. Nous avons :

$$\begin{aligned}
\mathbb{E}(Y \mid W = 1) &= \mathbb{E}\left(\left(WY_{(1)} + (1 - W)Y_{(0)}\right) \mid W = 1\right) \\
&= \mathbb{E}(Y_{(1)} \mid W = 1) \\
&= \mathbb{E}(Y_{(1)}),
\end{aligned}$$

et

$$\begin{aligned}\mathbb{E}(Y \mid W=0) &= \mathbb{E}\left(\left(WY_{(1)} + (1-W)Y_{(0)}\right) \mid W=0\right) \\ &= \mathbb{E}(Y_{(0)} \mid W=0) \\ &= \mathbb{E}(Y_{(0)});\end{aligned}$$

alors, en vertu de (2.2) nous avons :

$$\tau = \mathbb{E}(Y \mid W=1) - \mathbb{E}(Y \mid W=0). \tag{2.5}$$

Notons $\mu_1 = \mathbb{E}(Y \mid W=1)$ et $\mu_0 = \mathbb{E}(Y \mid W=0)$. Nous pouvons également exprimer $\tau_{(1)}$ en termes de quantités observables. Puisque nous avons $(Y_{(1)}, Y_{(0)})$ est indépendant de $W$, alors :

$$\begin{aligned}\tau_{(1)} &= \mathbb{E}(Y_{(1)} \mid W=1) - \mathbb{E}(Y_{(0)} \mid W=1) \\ &= \mathbb{E}(Y_{(1)} \mid W=1) - \mathbb{E}(Y_{(0)} \mid W=0) \\ &= \mathbb{E}\left(\left(WY_{(1)} + (1-W)Y_{(0)}\right) \mid W=1\right) - \mathbb{E}\left(\left(WY_{(1)} + (1-W)Y_{(0)}\right) \mid W=0\right) \\ &= \mathbb{E}(Y \mid W=1) - \mathbb{E}(Y \mid W=0) \\ &= \mu_1 - \mu_0. \end{aligned} \tag{2.6}$$

Par (2.5) et (2.6) nous avons :

$$\tau = \tau_{(1)} = \mathbb{E}(Y \mid W=1) - \mathbb{E}(Y \mid W=0) = \mu_1 - \mu_0. \tag{2.7}$$

L'équation (2.7) ne contient aucune quantité contrefactuelle, ce qui montre que $\tau$ et $\tau_{(1)}$ sont identifiables. □

D'après (2.7), $\tau$ et $\tau_{(1)}$ correspondent à la différence des moyennes de la variable du résultat observable dans le groupe des traités et le groupe de contrôle, qui est une mesure d'association entre $Y$ et $W$. Cela signifie que dans une étude randomisée, causalité égale association. Il est à noter qu'en effet, l'indépendance entre $Y_{(0)}$ et $W$ est une condition suffisante pour l'identification de $\tau_{(1)}$.

Dans une étude non expérimentale, $(Y_{(1)}, Y_{(0)})$ et $W$ ne sont généralement pas

indépendants, on a donc :

$$\mathbb{E}(Y_{(1)} \mid W = 1) = \mathbb{E}(Y \mid W = 1) \neq \mathbb{E}(Y_{(1)}), \tag{2.8}$$

et

$$\mathbb{E}(Y_{(0)} \mid W = 0) = \mathbb{E}(Y \mid W = 0) \neq \mathbb{E}(Y_{(0)}). \tag{2.9}$$

Dans ce cas l'égalité (2.7) n'est donc pas vérifiée, ce qui ne permet pas l'identification de l'effet causal moyen et un biais apparaît.

## 2.3 La notion de biais

Un biais est une erreur systématique qui survient lors d'une étude et qui peut fausser les résultats d'analyse s'il n'est pas pris en compte. On distingue trois types de biais : les biais de sélection, les biais de confusion et les biais de mesure (ou d'information).

### 2.3.1 Le biais de sélection

Il s'agit d'une erreur systématique due à la méthode adoptée pour choisir les participants à l'étude. Ce biais surgit lorsque les groupes de traités et de contrôle ne sont pas représentatifs de la population cible. Comme nous l'avons montré dans la section précédente, dans une étude randomisée $\tau$ et $\tau_{(1)}$ coïncident avec la différence des moyennes de la variable du résultat observable dans le groupe des traités et le groupe de contrôle (équation (2.7)). De manière générale, cette différence des moyennes peut s'écrire sous la forme :

$$\begin{aligned}
\mu_1 - \mu_0 &= \mathbb{E}(Y \mid W = 1) - \mathbb{E}(Y \mid W = 0) \\
&= \mathbb{E}(Y_{(1)} \mid W = 1) - \mathbb{E}(Y_{(0)} \mid W = 0) \\
&= \mathbb{E}(Y_{(1)} \mid W = 1) - \mathbb{E}(Y_{(0)} \mid W = 0) + \mathbb{E}(Y_{(0)} \mid W = 1) - \mathbb{E}(Y_{(0)} \mid W = 1) \\
&= \mathbb{E}(Y_{(1)} \mid W = 1) - \mathbb{E}(Y_{(0)} \mid W = 1) + \mathbb{E}(Y_{(0)} \mid W = 1) - \mathbb{E}(Y_{(0)} \mid W = 0) \\
&= \tau_{(1)} + B.
\end{aligned}$$

La quantité $B$ représente le biais de sélection lié à l'étude. Il est différent de zéro lorsque :

$$\mathbb{E}(Y_{(0)} \mid W = 1) \neq \mathbb{E}(Y_{(0)} \mid W = 0).$$

Une telle situation se produit lorsque les groupes de traités et de non traités sont statistiquement dissemblables à l'origine sur certaines caractéristiques observables qui peuvent influencer les résultats, et que la sélection des participants à l'étude n'est pas indépendante de ces caractéristiques. Pour remédier à ce problème, les méthodes de score de propension et d'appariement permettent de réduire le biais et donnent de bons estimateurs de l'effet causal moyen.

### 2.3.2 Le biais de confusion

Ce biais apparaît généralement lorsque la variable du traitement et la variable de réponse partagent une même cause appelée variable confondante. À vrai dire, il n'existe pas de définition claire d'une variable confondante (Hernán et Robins, 2013). Dans le chapitre 4, nous donnerons la définition de la variable confondante que nous allons adopter dans ce travail. En l'absence de cette dernière, l'entière association entre le traitement et le résultat est expliquée par l'effet causal du traitement sur le résultat. Dans ce cas, l'effet causal moyen est identifiable (l'association égale la causalité). La présence d'une variable confondante crée une association supplémentaire entre le traitement et le résultat, ce qui complique l'identification de l'effet causal moyen.

Afin d'éliminer le biais de confusion, il est donc nécessaire de neutraliser toutes les variables confondantes potentielles intervenant dans l'étude. Pour cela, les méthodes graphiques (Pearl, 2009) demeurent un outil efficace pour construire des ensembles de covariables permettant d'éliminer le biais de confusion, en ajustant par exemple sur ces covariables à l'aide d'un modèle de régression, ou encore en utilisant d'autres méthodes telles que l'appariement ou la stratification sur

les variables confondantes ainsi que les techniques de score de propension (voir section 2.5).

### 2.3.3 Le biais d'information

Le biais d'information regroupe toutes les erreurs systématiques qui affectent la mesure de l'exposition au traitement, le résultat, ainsi que les autres facteurs intervenant dans l'étude. Ce biais est également connu sous le nom de « biais de classement », lorsque les variables de traitement et de résultat sont catégorielles. En effet, le biais est le résultat d'un mauvais classement des sujets, selon les modalités des variables. Par exemple, considérer les individus traités comme non traités et vice-versa.

On distingue généralement deux types de biais d'information : le biais différentiel et le biais non différentiel. Le biais différentiel est une erreur de mesure qui affecte d'une manière différente les groupes comparés. Dans ce cas, la proportion des individus classés à tort dans une catégorie est différente selon les groupes. Le biais non différentiel est une erreur de nature aléatoire qui affecte de façon identique les deux groupes comparés. Par exemple, un instrument de mesure défaillant aboutira à des mesures incorrectes dans les deux groupes. Il est à noter qu'une fois les données recueillies, il est pratiquement impossible de réduire le biais d'information, ce qui conduit à des estimateurs peu fiables, voire, à la remise en cause de la validité de l'étude, notamment dans le cas d'un biais différentiel.

## 2.4 Conditions d'identification de l'effet causal dans une étude non expérimentale

Dans une étude non randomisée, l'identification de l'effet causal moyen nécessite d'autres conditions et hypothèses. Supposons que pour chaque individu, on observe un vecteur de covariables $\mathbf{Z}$ qui précède le traitement.

### 2.4.1 L'hypothèse d'absence d'effets de diffusion du traitement

Une hypothèse nécessaire pour l'identification de l'effet causal moyen est l'hypothèse d'absence d'effets de diffusion du traitement (*Stable unit treatment value assumption SUTVA*) (Rubin, 1980). Elle stipule que la réponse d'un individu au traitement ne doit pas être affectée par l'affectation d'un traitement à un autre individu. Contrairement à ce qu'il paraît, cette hypothèse n'est pas toujours vérifiée. Par exemple, si la variable de traitement représente la participation à un programme d'apprentissage dans une école, et que la variable de réponse mesure le score dans un test après la fin du programme, si on laisse les élèves communiquer entre eux, il y aura échange de connaissances. Dans ce cas, la participation d'un élève au programme peut affecter le score d'un autre élève.

### 2.4.2 Hypothèse d'ignorabilité forte du traitement

Une hypothèse centrale dans les études non expérimentales est celle d'ignorabilité forte de l'affectation du traitement (Rosenbaum et Rubin, 1983). Cette hypothèse est basée sur la notion probabiliste d'indépendance conditionnelle qui se définit comme suit :

Définition 2.4.1. *Soit $X, Y, Z$ trois variables aléatoires (possiblement multidimensionnelles) de loi de probabilité jointe $\mathbb{P}$. On dit que $X$ est indépendant de $Y$ sachant $Z$ et l'on écrit $X \perp\!\!\!\perp Y \mid Z$ si :*

$$\mathbb{P}(X \mid Y, Z) = \mathbb{P}(X \mid Z). \tag{2.10}$$

Cette équation signifie que connaissant $Z$, la connaissance de $Y$ n'apporte rien sur la connaissance de $X$.

On dit que l'affectation du traitement est fortement ignorable si :

(i) l'exposition au traitement est indépendante des contrefactuels étant données

les covariables, et que

(ii) la probabilité de recevoir chacun des traitements est strictement positive pour tous les individus étant données les covariables.

Plus formellement, cette hypothèse stipule que :

$$(Y_{(1)}, Y_{(0)}) \perp\!\!\!\perp W \mid \mathbf{Z}, \tag{2.11}$$

et que

$$0 < \mathbb{P}(W = 1 \mid \mathbf{Z}) < 1. \tag{2.12}$$

La première partie de l'hypothèse (équation (2.11)) est souvent appelée « hypothèse d'ignorabilité » ou « hypothèse d'indépendance conditionnelle », et elle stipule aussi qu'il n'y a pas de variables non observables affectant à la fois le traitement et la réponse. Concernant la deuxième partie de l'hypothèse (équation (2.12)), elle est connue sous le nom de « hypothèse de positivité » ou « hypothèse de support commun ». Pour l'identification de l'effet causal moyen sur les traités $\tau(1)$, il existe une version plus faible de l'hypothèse d'ignorabilité forte (Heckman *et al.*, 1997) qui est donnée par :

$$Y_{(0)} \perp\!\!\!\perp W \mid \mathbf{Z}, \tag{2.13}$$

et

$$\mathbb{P}(W = 1 \mid \mathbf{Z}) < 1. \tag{2.14}$$

Si les deux hypothèses présentées ci-dessus sont vérifiées, alors, il serait possible d'identifier les deux quantités $\mathbb{E}(Y_{(1)})$ et $\mathbb{E}(Y_{(0)})$ à partir des données observées de $Y$, $W$ et $\mathbf{Z}$, comme l'indique le théorème suivant :

Théorème 2.4.1. *Soit les hypothèses :*

1. *Absence d'effets de diffusion du traitement (SUTVA),*
2. *Ignorabilité forte :* $(Y_{(1)}, Y_{(0)}) \perp\!\!\!\perp W \mid \mathbf{Z}$ *et* $0 < \mathbb{P}(W = 1 \mid \mathbf{Z}) < 1$.

*Alors, l'effet causal moyen est identifiable, et on a :*

$$\tau = \mathbb{E}(Y_{(1)} - Y_{(0)}) = \mathbb{E}_{\mathbf{Z}}\left(\mathbb{E}(Y \mid W = 1, \mathbf{Z}) - \mathbb{E}(Y \mid W = 0, \mathbf{Z})\right). \tag{2.15}$$

Preuve. Nous avons :

$$\begin{aligned}\mathbb{E}(Y_{(1)}) &= \mathbb{E}_{\mathbf{Z}}\left(\mathbb{E}(Y_{(1)} \mid \mathbf{Z})\right) = \mathbb{E}_{\mathbf{Z}}\left(\mathbb{E}(Y_{(1)} \mid W = 1, \mathbf{Z})\right) \\ &= \mathbb{E}_{\mathbf{Z}}\left(\mathbb{E}\left(WY_{(1)} + (1-W)Y_{(0)} \mid W = 1, \mathbf{Z}\right)\right) \\ &= \mathbb{E}_{\mathbf{Z}}\left(\mathbb{E}(Y \mid W = 1, \mathbf{Z})\right),\end{aligned}$$

et

$$\begin{aligned}\mathbb{E}(Y_{(0)}) &= \mathbb{E}_{\mathbf{Z}}\left(\mathbb{E}(Y_{(0)} \mid \mathbf{Z})\right) = \mathbb{E}_{\mathbf{Z}}\left(\mathbb{E}(Y_{(0)} \mid W = 0, \mathbf{Z})\right) \\ &= \mathbb{E}_{\mathbf{Z}}\left(\mathbb{E}\left(WY_{(1)} + (1-W)Y_{(0)} \mid W = 0, \mathbf{Z}\right)\right) \\ &= \mathbb{E}_{\mathbf{Z}}\left(\mathbb{E}(Y \mid W = 0, \mathbf{Z})\right),\end{aligned}$$

alors

$$\tau = \mathbb{E}(Y_{(1)} - Y_{(0)}) = \mathbb{E}_{\mathbf{Z}}\left(\mathbb{E}(Y \mid W = 1, \mathbf{Z}) - \mathbb{E}(Y \mid W = 0, \mathbf{Z})\right). \qquad \square$$

L'hypothèse de positivité (équation (2.12)) est aussi nécessaire dans l'identification de l'effet causal moyen, car dans le cas où elle ne serait pas vérifiée pour $\mathbf{Z} = \boldsymbol{z}$, cela veut dire que pour cette valeur $\boldsymbol{z}$, on a seulement des traités ou des non-traités, ce qui ne permet pas d'estimer à la fois $\mathbb{E}(Y \mid W = 1, \mathbf{Z} = \boldsymbol{z})$ et $\mathbb{E}(Y \mid W = 0, \mathbf{Z} = \boldsymbol{z})$.

Comme dans Imbens (2004), dans ce qui suit, nous adoptons les notations :

$$\mu_1(\boldsymbol{z}) = \mathbb{E}(Y \mid W = 1, \mathbf{Z} = \boldsymbol{z}) \quad \text{et} \quad \mu_0(\boldsymbol{z}) = \mathbb{E}(Y \mid W = 0, \mathbf{Z} = \boldsymbol{z}).$$

## 2.5 Estimation de l'effet du traitement

Une fois l'effet causal moyen $\tau$ identifié à partir des quantités observables, il devient possible de l'estimer à partir des observations du triplet $(Y, W, \mathbf{Z})$. Pour pouvoir estimer $\tau$, il faut disposer de $n$ réalisations $(y_1, w_1, \boldsymbol{z}_1), \ldots, (y_n, w_n, \boldsymbol{z}_n)$ du triplet $(Y, W, \mathbf{Z})$. Ensuite, une méthode appropriée est utilisée pour effectuer

cette estimation. Dans cette section, nous présentons trois méthodes d'estimation de l'effet causal moyen : une méthode d'estimation par stratification, une autre par régression et, finalement, une méthode d'appariement sur les caractéristiques observables.

### 2.5.1 Stratification

L'une des méthodes les plus utilisées pour l'estimation de l'effet causal moyen est la stratification basée sur les caractéristiques observables des individus. Cette méthode consiste à regrouper dans une même classe les individus ayant les mêmes caractéristiques. On obtient donc $K$ classes $S_1, S_2, \ldots, S_K$. À l'intérieur de chaque classe $S_j$, $j = 1, \ldots, K$ on estime l'effet causal moyen $\hat{\tau}_j$. Ensuite, on combine les estimateurs obtenus pour obtenir une estimation globale de l'effet causal moyen. L'estimateur lié à chaque strate est donné par :

$$\begin{aligned}\hat{\tau}_j &= \hat{\mathbb{E}}(Y \mid W = 1, \mathbf{Z} = \boldsymbol{z}) - \hat{\mathbb{E}}(Y \mid W = 0, \mathbf{Z} = \boldsymbol{z}) \\ &= \hat{\mu}_1(\boldsymbol{z}) - \hat{\mu}_0(\boldsymbol{z}).\end{aligned}$$

Un estimateur convergent de l'effet causal moyen est donné par la moyenne pondérée des estimateurs par strate :

$$\hat{\tau}_{\text{strat}} = \sum_{\boldsymbol{z}} \hat{\tau}_j \hat{\mathbb{P}}(\mathbf{Z} = \boldsymbol{z}).$$

L'un des premiers travaux sur la technique de stratification a été effectué par Cochran (1968). Il consistait à établir un lien entre le tabagisme et le cancer du poumon. Dans cette étude, l'auteur a effectué une stratification sur une seule variable confondante, à savoir l'âge de l'individu. Il a montré que le choix de 5 strates permet de réduire de 90% le biais attribué à la variable confondante considérée. La réduction du biais est de 79% et de 86% pour le choix de 3 strates et de 4 strates respectivement.

Lorsque le nombre de variables confondantes augmente, le nombre de strates né-

cessaires pour prendre en compte la variation entre individus devient très grand. Par exemple, pour 15 variables confondantes dichotomiques, le nombre de strates nécessaires est de $2^{15} = 32768$ strates. De plus, chaque strate doit contenir au moins un individu traité et un individu non traité afin de pouvoir les comparer, ce qui nécessite un échantillon d'au moins 65536 individus. Des échantillons de cette taille sont généralement difficiles à obtenir, il faut donc utiliser d'autres méthodes telles que la régression et la stratification sur le score de propension que nous présenterons en détail dans la section 5.3.2.

### 2.5.2 Régression

L'équation (2.15) permet d'identifier la distribution de l'effet causal moyen dans la population sous les hypothèses de l'ignorabilité forte du traitement et de la SUTVA. Il est donc naturel de l'estimer à partir d'une quantité échantillonnale analogue donnée par :

$$\hat{\tau}_{\text{reg}} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_1(\mathbf{Z}_i) - \hat{\mu}_0(\mathbf{Z}_i)\right), \tag{2.16}$$

où $\hat{\mu}_1(\mathbf{Z}_i)$ et $\hat{\mu}_0(\mathbf{Z}_i)$ sont les estimateurs de $\mu_1(\mathbf{Z}) = \mathbb{E}(Y \mid W = 1, \mathbf{Z})$ et de $\mu_0(\mathbf{Z}) = \mathbb{E}(Y \mid W = 0, \mathbf{Z})$ respectivement.

Supposons maintenant que pour $w \in \{0, 1\}$, $\mu_w(\boldsymbol{z})$ est linéaire en $w$ et $\boldsymbol{z}$ :

$$\mu_w(\boldsymbol{z}) = \alpha + \beta w + \gamma^t \boldsymbol{z}. \tag{2.17}$$

Le modèle de régression linéaire correspondant est donné par :

$$Y_i = \alpha + \beta w_i + \gamma^t \boldsymbol{z}_i + \varepsilon_i, \tag{2.18}$$

où $i = 1, \ldots, n$ et $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. Admettons maintenant que l'effet causal est le même pour tous les individus :

$$\tau_i = \tau, \;\; i = 1, \ldots, n. \tag{2.19}$$

Nous avons donc :

$$\begin{aligned}\tau_i &= \mathbb{E}_{\mathbf{Z}_i}\left(\mathbb{E}(Y_i \mid W_i = 1, \mathbf{Z}_i = \boldsymbol{z}_i) - \mathbb{E}(Y_i \mid W_i = 0, \mathbf{Z}_i = \boldsymbol{z}_i)\right) \\ &= \mathbb{E}_{\mathbf{Z}_i}(\mu_1(\boldsymbol{z}_i) - \mu_0(\boldsymbol{z}_i)) \\ &= \mathbb{E}_{\mathbf{Z}_i}(\alpha + \beta + \gamma^t \boldsymbol{z}_i - \alpha - \gamma^t \boldsymbol{z}_i) \\ &= \mathbb{E}_{\mathbf{Z}_i}(\beta) \\ &= \beta.\end{aligned}$$

Cela signifie que le coefficient de régression $\beta$ coïncide avec l'effet causal moyen. Le modèle peut donc se réécrire :

$$Y_i = \alpha + \tau w_i + \gamma^t \boldsymbol{z}_i + \varepsilon_i. \tag{2.20}$$

Il est également possible d'estimer les deux quantités $\mu_1(\boldsymbol{z})$ et $\mu_0(\boldsymbol{z})$ séparément (Imbens, 2004) :

$$\begin{cases} \mu_1(\boldsymbol{z}) = \alpha_1 + \gamma_1^t \boldsymbol{z} \quad si \quad W = 1 \\ \mu_0(\boldsymbol{z}) = \alpha_0 + \gamma_0^t \boldsymbol{z} \quad si \quad W = 0. \end{cases} \tag{2.21}$$

Les deux modèles de régression correspondants sont donnés par :

$$\begin{cases} Y_i = \alpha_1 + \gamma_1^t \boldsymbol{z}_i + \varepsilon_i \quad si \quad W = 1 \\ Y_i = \alpha_0 + \gamma_0^t \boldsymbol{z}_i + \varepsilon_i \quad si \quad W = 0. \end{cases} \tag{2.22}$$

À partir de ces deux modèles, on détermine les deux quantités $\hat{\mu}_0(\mathbf{Z}_i)$ et $\hat{\mu}_1(\mathbf{Z}_i)$, que l'on utilise ensuite dans la formule de l'estimateur de l'effet causal moyen donné par (2.16).

Certains auteurs se sont intéressés à l'estimation non paramétrique de l'effet causal moyen. Heckman *et al.* (1997) ont proposé un estimateur à noyau de $\mu_w(\boldsymbol{z})$ qui est donné par :

$$\hat{\mu}_w(\boldsymbol{z}) = \frac{\sum\limits_{i|W_i=w} Y_i K\left(\dfrac{\mathbf{Z}_i - \boldsymbol{z}}{h}\right)}{\sum\limits_{i|W_i=w} K\left(\dfrac{\mathbf{Z}_i - \boldsymbol{z}}{h}\right)}, \tag{2.23}$$

où $K$ est un noyau et $h$ un paramètre appelé fenêtre, lequel détermine le degré de lissage. L'effet causal moyen est ensuite mesuré par l'équation (2.16).

### 2.5.3 Appariement sur les caractéristiques observables

Cette méthode consiste à apparier chaque individu traité avec un ou plusieurs individus non traités qui se rapprochent de lui en matière de caractéristiques observables. Il existe plusieurs façons de faire l'appariement sur les covariables. Premièrement, il faut faire la distinction entre l'appariement sans remise et l'appariement avec remise. Le premier stipule qu'un individu non traité déjà apparié avec un traité ne sera plus disponible pour l'appariement avec les autres traités, ce qui n'est pas le cas pour le deuxième (Rosenbaum, 2002). En deuxième lieu, il faut également choisir entre l'appariement par la méthode du plus proche voisin et l'appariement optimal (Rosenbaum, 2002). Le premier consiste à sélectionner aléatoirement un individu traité, et lui apparier ensuite l'individu non traité le plus proche de lui en matière des caractéristiques observables, et ce malgré le fait que le non traité sélectionné pourrait être plus proche d'un autre traité. Dans l'appariement optimal, on choisit l'appariement qui minimise une mesure de distance entre les traités et les non traités dans toutes les combinaisons de paires possibles.

Pour mesurer la distance entre un individu traité $i$ et un individu non traité $j$, on opte généralement pour la distance de Mahalanobis $D_{ij}$ donnée par :

$$D_{ij} = (\mathbf{Z}_i - \mathbf{Z}_j)^t \Sigma^{-1} (\mathbf{Z}_i - \mathbf{Z}_j), \tag{2.24}$$

où $\Sigma^{-1}$ désigne la matrice variance covariance de $\mathbf{Z}$. Lorsque l'individu à apparier se situe à une distance éloignée de son plus proche voisin, l'utilisation d'une mesure de distance peut conduire à un appariement de mauvaise qualité. Pour remédier à ce problème, on fixe un certain seuil connu sous le nom de « caliper » qui correspond à la distance maximale tolérée entre deux membres d'une paire. Si un traité se trouve à une distance supérieure au caliper par rapport à tous les non-

traités, alors il sera exclu de l'appariement. Cochran et Rubin (1973) ont effectué un appariement sur une seule variable confondante continue $Z$ en utilisant un caliper de la forme $a\sqrt{(\sigma_1^2 + \sigma_0^2)/2}$, où $\sigma_1^2$ et $\sigma_0^2$ sont respectivement la variance de $Z$ dans le groupe des traités et la variance de $Z$ dans le groupe contrôle. Ils ont montré que lorsque $\sigma_1^2 = \sigma_0^2$, donner à $a$ les valeurs 0.2, 0.4 et 0.6 permet d'éliminer respectivement 99%, 95% et 89% du biais induit par $Z$. Cette amélioration est encore plus forte lorsque la quantité $\sigma_1^2/\sigma_0^2$ diminue.

Une fois que l'appariement est effectué, on passe à l'estimation de l'effet causal moyen à partir des groupes constitués. Abadie et Imbens (2002) proposent un estimateur basé sur l'estimation des réponses potentielles manquantes. Les deux auteurs ont opté pour un appariement par la méthode du plus proche voisin, de sorte qu'à chaque individu traité soit associé un nombre $m$ prédéfini d'individus non traités. Soit $\mathcal{I}_m(i)$ l'ensemble des indices des $m$ individus les plus proches de l'individu $i$. Les réponses potentielles estimées sont données par :

$$\hat{Y}_{(1)i} = \begin{cases} Y_i & si \quad W_i = 1 \\ \dfrac{1}{m} \displaystyle\sum_{j \in \mathcal{I}_m(i)} Y_j & si \quad W_i = 0, \end{cases}$$

et

$$\hat{Y}_{(0)i} = \begin{cases} \dfrac{1}{m} \displaystyle\sum_{j \in \mathcal{I}_m(i)} Y_j & si \quad W_i = 1 \\ Y_i & si \quad W_i = 0. \end{cases}$$

À partir de ces deux quantités, les deux auteurs ont proposé un estimateur par appariement qui est donné par :

$$\hat{\tau}_{appar} = \frac{1}{n} \sum_{i=1}^{n} (\hat{Y}_{(1)i} - \hat{Y}_{(0)i}).$$

Cependant, de manière générale, une fois l'appariement réalisé, on utilise certaines méthodes adaptées aux données appariées pour l'estimation de l'effet causal moyen, comme le test de Student pour données appariées si la variable de

réponse est continue, ou le test de McNemar dans le cas d'une variable réponse dichotomique.

Les hypothèses d'ignorabilité forte et d'absence d'effets de diffusion du traitement (SUTVA) indiquent les conditions d'identification de l'effet causal moyen, mais elles ne précisent pas le mécanisme de choix de l'ensemble des covariables **Z**. Comme nous l'avons déjà mentionné, il est impossible d'observer simultanément les deux résultats potentiels $Y_{(0)}$ et $Y_{(1)}$, ce qui ne permet pas de vérifier les relations d'indépendance conditionnelle impliquant ces variables. Par conséquent, l'hypothèse d'ignorabilité forte reste donc non vérifiable. Pour cela, certains auteurs ont tendance à supposer que cette hypothèse est automatiquement vérifiée lorsqu'on inclut un grand nombre de covariables. Cependant, cette supposition n'est pas réellement valide, car en incluant certaines covariables, notamment des variables post-traitement, on risque d'introduire un nouveau biais lors de l'estimation de l'effet causal, en créant une association supplémentaire entre le traitement et le résultat (voir section 4.3.1). Pour remédier à cette situation, les méthodes basées sur les graphes orientés acycliques se révèlent plus pertinentes pour sélectionner l'ensemble des covariables à inclure dans le modèle et déterminer les sources du biais.

# CHAPITRE III

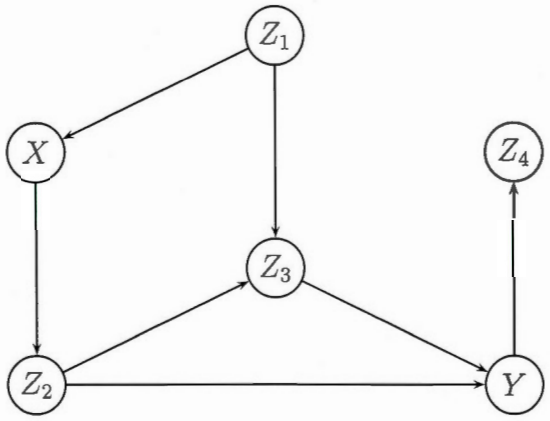
# MODÈLES GRAPHIQUES PROBABILISTES ET CAUSALITÉ

Comme nous l'avons expliqué à la fin du chapitre précédent, l'hypothèse d'ignorabilité est non vérifiable. Il est donc primordial de s'assurer du bon choix des covariables à inclure dans le modèle pour permettre de satisfaire cette hypothèse. Une solution à ce problème est de tenir compte des connaissances a priori d'experts sur les relations causales existantes entre la variable du traitement, la variable du résultat et les autres variables intervenant dans l'étude. Ces connaissances peuvent être exprimées sous forme d'hypothèses causales et représentées par un graphe orienté acyclique.

## 3.1 Quelques notions de la théorie des graphes

Un graphe est un couple $(V, E)$, où $V$ est un ensemble de nœuds (sommets) et $E \subseteq V \times V$ un ensemble d'arêtes permettant de relier certaines paires de $V$. Une arête est donc définie par une paire $(X, Y)$, où $X$ et $Y$ sont deux nœuds. On appelle sous-graphe de $G$ engendré par $\acute{V}$ tout graphe $\acute{G} = (\acute{V}, \acute{E})$ tel que $\acute{V} \subset V$ et $\acute{E} = (\acute{V} \times \acute{V}) \cap E$. Cela veut dire que $\acute{G}$ est composé d'un sous-ensemble de nœuds de $G$ et de leurs arêtes correspondantes. Dans l'analyse graphique que nous allons présenter, nous nous baserons principalement sur les travaux de Pearl (2009), Lauritzen (1996) et Kowell *et al.* (1999). L'ensemble $V$ regroupe toutes les variables qui interviennent dans l'étude, et l'ensemble $E$ représente les liens qui

Figure 3.1: Exemple d'un graphe orienté acyclique




existent entre les paires de variables. Signalons que dans ce chapitre et le chapitre suivant, nous manipulons principalement des ensembles de variables et de nœuds. Pour des raisons pratiques, nous avons décidé de ne pas faire usage de caractères gras pour désigner ces ensembles.

Une arête peut être orientée ou non. Une arête orientée est appelée arc et celui-ci est représenté par une flèche à une pointe liant un nœud de départ (origine) à un nœud d'arrivée (extrémité). On utilise généralement les deux notations $(X, Y)$ et $X \rightarrow Y$ pour désigner un arc allant de $X$ à $Y$. On distingue deux types de graphes : les graphes orientés et les graphes non orientés.

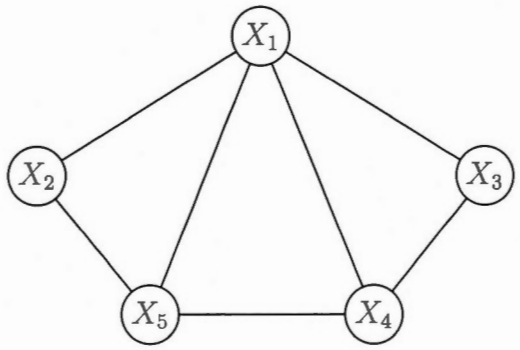
Un graphe orienté est un graphe dont toutes les arêtes sont des arcs (voir Figure 3.1). Un chemin entre deux nœuds est une suite d'arêtes consécutives telle que chaque arête a un nœud en commun avec l'arête suivante. Deux nœuds liés par un chemin sont dits connectés. Dans la figure 3.1, $\{(Z_1, X), (Z_1, Z_3), (Z_3, Y)\}$ est un exemple de chemin entre $X$ et $Y$, que l'on peut exprimer également sous la forme $X \leftarrow Z_1 \rightarrow Z_3 \rightarrow Y$. Pour la suite de ce travail, nous adopterons la deuxième forme afin d'exprimer un chemin reliant deux nœuds, car cette forme

est plus facile à lire. Dans un graphe orienté $G = (V, E)$, un chemin orienté est un chemin dont l'extrémité de chaque arc coïncide avec l'origine de l'arc suivant. Considérons à nouveau la figure 3.1, nous constatons que $X \rightarrow Z_2 \rightarrow Z_3 \rightarrow Y$ est un chemin orienté allant de $X$ à $Y$. Les nœuds $X$ et $Y$ sont respectivement l'origine et l'extrémité du chemin. On appelle cycle tout chemin orienté dont l'origine et l'extrémité coïncident. Un graphe orienté qui ne contient aucun cycle orienté est appelé graphe orienté acyclique (en anglais *directed acyclic graph, DAG*). Le graphe présenté dans la figure 3.1 ne contient aucun cycle et il est orienté, il est donc un exemple simple d'un graphe orienté acyclique. Pour ce travail, nous allons utiliser l'acronyme anglais DAG pour désigner un graphe orienté acyclique, car il s'agit du type le plus couramment utilisé.

Les arcs dans un graphe orienté acyclique permettent également de définir des relations de type parent/enfant, ancêtre/descendant entre les différents nœuds du graphe. On dit que $X$ est parent de $Y$ s'il existe un arc de $X$ vers $Y$. On dit alors que $Y$ est enfant de $X$. Le nœud $X$ est ancêtre de $Y$ ou, en d'autres termes, $Y$ est descendant de $X$ s'il existe un chemin orienté allant de $X$ à $Y$. Dans la figure 3.1, $X$ a un seul parent qui est $Z_1$, qui est aussi son seul ancêtre, un seul enfant ($Z_2$) et quatre descendants ($Z_2, Z_3, Y$ et $Z_4$). Un puits dans un graphe orienté est un nœud qui n'a pas d'enfants. Un nœud qui n'a pas de parents est appelé source. Dans un graphe orienté acyclique, il existe au moins une source et un puits. Les nœuds $Z_1$ et $Z_4$ sont respectivement l'unique source et l'unique puits du DAG présenté dans la figure 3.1.

Un graphe non orienté est un graphe dont toutes les arêtes sont non orientées (voir figure 3.2). Ce type de graphes est adapté à la modélisation des relations symétriques. Un sous-graphe non orienté $\acute{G} = (\acute{V}, \acute{E})$ d'un graphe non orienté $G = (V, E)$ est dit complet si chaque nœud de $\acute{G}$ est relié à tous les autres nœuds

Figure 3.2: Exemple d'un graphe non orienté




de $\acute{G}$. Dans ce cas, le sous-ensemble de nœuds $\acute{V}$ est aussi dit complet. L'ensemble

$$\mathcal{A} = \Big\{\{X_1\},\{X_2\},\{X_3\},\{X_4\},\{X_5\},\{X_1,X_2\},\{X_1,X_3\},\{X_1,X_4\},\{X_1,X_5\},$$
$$\{X_2,X_5\},\{X_3,X_4\},\{X_4,X_5\},\{X_1,X_2,X_5\},\{X_1,X_3,X_4\},\{X_1,X_4,X_5\}\Big\}$$

représente l'ensemble des sous-ensembles complets liés au graphe de la figure 3.2. On appelle clique un sous-graphe complet qui n'est pas contenu dans un autre sous-graphe complet. Cette appellation est donnée aussi au sous-ensemble de nœuds qui n'est pas contenu dans un autre sous-ensemble de nœuds complet. L'ensemble des cliques relatif au graphe présenté dans la figure 3.2 est donné par :

$$\mathcal{C} = \Big\{\{X_1,X_2,X_5\},\{X_1,X_3,X_4\},\{X_1,X_4,X_5\}\Big\}.$$

Il est à noter que la notion de clique est propre aux graphes non orientés. Néanmoins, certains auteurs l'utilisent également pour les graphes orientés. Cette notion nous sera utile pour la suite pour démontrer certaines propriétés liées aux graphes.

Comme dans les graphes orientés, les notions de chemin et de cycle sont définies de la même façon dans les graphes non orientés. Soulignons que les termes chemin et cycle que nous avons utilisés sont des adaptations des termes « *path* » et « *cycle* » du vocabulaire anglais. En effet, certains auteurs francophones (Berge,

1983) préfèrent utiliser le terme chaîne plutôt que chemin, et le terme chemin au lieu de chemin orienté. Le terme cycle est généralement réservé aux graphes non orientés, tandis que pour les graphes orientés, ils utilisent le terme circuit.

Revenons maintenant à notre étude de cas. Le graphe représentant nos variables (voir figure 1.1) est orienté car toutes ses arêtes sont orientées. Il reste maintenant à vérifier s'il est acyclique. En raison de sa taille, il n'est pas aisé d'affirmer à première vue la présence ou non de cycles. Par contre, il existe une technique qui permet de détecter des cycles dont le principe est le suivant : nous commençons tout d'abord par supprimer tous les nœuds sources du graphe ainsi que tous les arcs qui leur sont liés. Cela est justifié par le fait qu'un cycle ne peut pas contenir un arc lié à une source. Ensuite, nous recommençons le scénario sur le graphe obtenu ainsi que sur chaque nouveau graphe jusqu'à l'obtention d'un graphe simplifié nous permettant de mieux détecter les cycles. Cette technique est illustrée à la figure A.1 de l'annexe A.1. Le graphe terminal ne contient aucun cycle, ce qui signifie l'absence de cycles dans le graphe de départ. Cela veut dire que notre graphe de départ est un graphe orienté acyclique.

Les relations entre les variables d'un DAG, notamment les relations causales, sont souvent probabilisées. La représentation graphique de ces variables permet essentiellement de prendre en compte leurs dépendances et indépendances conditionnelles. Une telle spécification à la fois probabiliste et graphique nécessite l'introduction d'autres types de modèles à savoir « les modèles graphiques probabilistes ».

Les modèles graphiques probabilistes souvent appelés « modèles graphiques » sont des modèles qui représentent des variables aléatoires sous forme de graphes. On distingue deux classes de modèles graphiques : les modèles graphiques orientés qui sont basés sur les graphes orientés acycliques et les modèles graphiques non

orientés appelés aussi champs de Markov, et qui sont associés aux graphes non orientés. Dans ce chapitre, nous nous intéresserons davantage aux modèles graphiques orientés, car ils sont les plus appropriés pour modéliser les relations de cause à effet. Cependant, nous aborderons également les modèles graphiques non orientés, lesquels seront utiles pour bien assimiler certaines propriétés liées aux modèles graphiques orientés.

Dans les deux modèles, nous supposons que l'on dispose d'un ensemble de variables aléatoires $X_V = \{X_1, X_2, \ldots, X_k\}$ à valeurs dans un ensemble noté $\mathcal{X}_V$ et qui sont représentées dans un graphe (orienté ou non) $G = (V, E)$. Les deux notations $X_V$ et $V$ représentent les mêmes éléments, à savoir $\{X_1, X_2, \ldots, X_k\}$. Nous avons opté pour deux notations différentes comme dans Lauritzen (1996) et Kowell *et al.* (1999) pour mieux différencier les nœuds du graphe des variables aléatoires correspondantes. Pour un sous-ensemble $A$ de $V$, nous notons $X_A$ le sous-ensemble de variables aléatoires correspondant qui prend des valeurs dans $\mathcal{X}_A$. Pour simplifier, nous utilisons les notations $X$ et $\mathcal{X}$ au lieu de $X_V$ et $\mathcal{X}_V$ et nous gardons les notations $X_A$ et $\mathcal{X}_A$ pour le sous-ensemble de variables aléatoires associé au sous-ensemble de nœuds $A$. Mais avant toute chose, commençons tout d'abord par revenir sur quelques notions de probabilité qui nous seront utiles pour la suite de ce travail.

## 3.2 Quelques notions de probabilité

### 3.2.1 Propriétés d'indépendance conditionnelle

Dans le chapitre précédent, nous avons introduit la notion probabiliste d'indépendance conditionnelle entre trois variables aléatoires (définition 2.4.1). Dawid (1979) a rassemblé les propriétés de cette notion que nous pouvons résumer comme suit :

1. Symétrie

$$X \perp\!\!\!\perp Y \mid Z \iff Y \perp\!\!\!\perp X \mid Z;$$

2. Décomposition

$$X \perp\!\!\!\perp (Y,W) \mid Z \implies X \perp\!\!\!\perp Y \mid Z \quad \text{et} \quad X \perp\!\!\!\perp W \mid Z; \tag{3.1}$$

3. Union faible

$$X \perp\!\!\!\perp (Y,W) \mid Z \implies X \perp\!\!\!\perp Y \mid (Z,W) \quad \text{et} \quad X \perp\!\!\!\perp W \mid (Z,Y);$$

4. Contraction

$$X \perp\!\!\!\perp Y \mid (Z,W) \quad \text{et} \quad X \perp\!\!\!\perp W \mid Z \implies X \perp\!\!\!\perp (Y,W) \mid Z.$$

La démonstration de ces propriétés apparaît de façon détaillée dans la section A.2 de l'annexe A.

### 3.2.2 Théorème de multiplication

Le théorème de multiplication, aussi connu sous le nom de règle de la chaîne (*chain rule*) joue un grand rôle dans la représentation des lois de probabilité jointes. Ce théorème s'énonce comme suit :

Théorème 3.2.1. *Soient $A_1, A_2, \ldots, A_k$ des événements de probabilité non nulle d'un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$. On a :*

$$\mathbb{P}(A_1 \cap A_2 \cap \ldots \cap A_k) = \mathbb{P}(A_1)\mathbb{P}(A_2 \mid A_1)\mathbb{P}(A_3 \mid A_1 \cap A_2)\ldots\mathbb{P}(A_k \mid A_1 \cap \ldots A_{k-1}). \tag{3.2}$$

L'équation (3.2) n'est qu'une simple généralisation à $k > 2$ événements de la formule des probabilités conditionnelles donnée par :

$$\mathbb{P}(A_1 \mid A_2) = \frac{\mathbb{P}(A_1 \cap A_2)}{\mathbb{P}(A_2)},$$

où $A_1$ et $A_2$ sont deux événements sur un espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$ tels que $\mathbb{P}'(A_2) > 0$.

## 3.3 Modèles graphiques non orientés

L'intérêt d'une représentation graphique des variables aléatoires est multiple : elle permet de tenir compte des connaissances a priori d'experts et de mieux exprimer les hypothèses, elle fournit un moyen efficace d'exprimer les dépendances et les indépendances conditionnelles entre les variables, et elle permet de mieux représenter la loi de probabilité jointe des variables aléatoires.

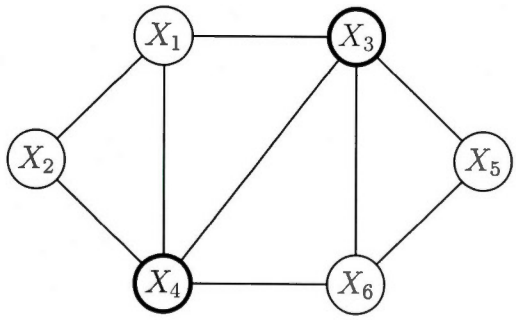
### 3.3.1 Critère de séparation dans les graphes non orientés

Comme nous l'avons mentionné ci-haut, les modèles graphiques probabilistes permettent de spécifier les relations d'indépendance conditionnelle des variables représentées dans le graphe. En effet, lorsque le nombre de variables est grand, il devient difficile de déterminer toutes les relations d'indépendance conditionnelle existantes entre les différents sous-ensembles de variables. Pour remédier à cette situation, il existe un critère graphique nommé « séparation », aussi connu sous le nom de « u-séparation », qui permet d'obtenir l'ensemble des liens existants entre les nœuds d'un graphe, que l'on peut interpréter sous certaines conditions comme étant des relations d'indépendance conditionnelle entre les variables liées au graphe, tel que nous le verrons dans cette section. La notion de séparation est purement graphique et elle est principalement utilisée dans les graphes non orientés (Lauritzen, 1996; Lauritzen *et al.*, 1990).

Définition 3.3.1 (Séparation). *Soit $G = (V, E)$ un graphe non orienté et soit $(A, B, D)$ un triplet de sous-ensembles de $V$ disjoints deux à deux. On dit que $D$ sépare $A$ et $B$ si tout chemin reliant un nœud de $A$ à un nœud de $B$ passe par un nœud de $D$.*

En guise d'illustration, considérons le graphe non orienté de la figure 3.3. Nous constatons que les deux sous-ensembles $A = \{X_1, X_2\}$ et $B = \{X_5, X_6\}$ sont

Figure 3.3: Séparation dans un graphe non orienté




séparés par le sous-ensemble $D = \{X_3, X_4\}$. Dans la suite de ce travail, nous adopterons la notation $(A \perp\!\!\!\perp_u B \mid D)_G$ pour dire que les nœuds $A$ et $B$ sont séparés par $D$ dans le graphe $G$. Grâce à la notion de séparation, il est possible de représenter des indépendances conditionnelles entre les variables aléatoires d'un modèle graphique non orienté, ce qui se traduit par la propriété de Markov globale pour les modèles graphiques non orientés.

Définition 3.3.2 (Propriété de Markov globale pour les graphes non orientés). *Soit $X$ un ensemble de variables aléatoires représentées par un graphe non orienté $G = (V, E)$ et soit $\mathbb{P}$ une mesure de probabilité sur $\mathcal{X}$. On dit que $\mathbb{P}$ satisfait la propriété de Markov globale par rapport à $G$, si pour tout triplet $(A, B, D)$ de sous-ensembles de $V$ disjoints deux à deux tel que $D$ sépare $A$ de $B$, on a : $X_A \perp\!\!\!\perp X_B \mid X_D$. Le graphe $G$ est appelé réseau Markovien de $\mathbb{P}$.*

La propriété de Markov globale est liée à une autre caractéristique des modèles graphiques non orientés, à savoir la factorisation d'une loi de probabilité jointe selon un graphe.

### 3.3.2 Factorisation d'une loi de probabilité jointe selon un graphe non orienté

Définition 3.3.3. *Soit $G = (V, E)$ un graphe non orienté représentant un ensemble de variables aléatoires $X = \{X_1, X_2, \ldots, X_k\}$ et soit $\mathbb{P}$ une mesure de probabilité sur $\mathcal{X}$ de densité (ou fonction de masse) $f$. On dit que $\mathbb{P}$ se factorise selon $G$ si pour tout sous-ensemble complet $A$ de l'ensemble de tous les sous-ensembles complets $\mathcal{A}$ de $V$ on a :*

$$f(x_1, x_2, \ldots, x_k) = \prod_{A \in \mathcal{A}} \psi_A(x_A), \tag{3.3}$$

*où $\psi_A$ sont des fonctions non négatives appelées potentiels.*

Une autre définition existe en se basant sur la notion de clique (voir section 3.1 pour les définitions de clique et de sous-ensemble complet).

Définition 3.3.4. *Soit $G = (V, E)$ un graphe non orienté représentant un ensemble de variables aléatoires $X = \{X_1, X_2, \ldots, X_k\}$ et soit $\mathcal{C}$ l'ensemble des cliques de $V$. Une mesure de probabilité $\mathbb{P}$ sur $\mathcal{X}$ de densité (ou fonction de masse) $f$ est dite se factoriser selon $G$ si pour toute clique $C$ de $\mathcal{C}$ on a :*

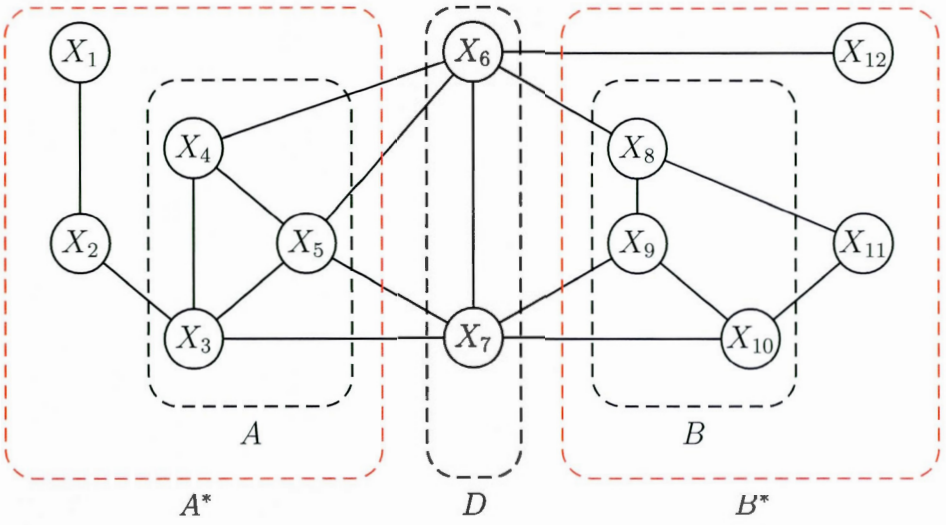
$$f(x_1, x_2, \ldots, x_k) = \prod_{C \in \mathcal{C}} \psi_C(x_C), \tag{3.4}$$

*où $\psi_C$ sont des potentiels.*

Proposition 3.3.1. *Soit $G = (V, E)$ un graphe non orienté dont les nœuds représentent un ensemble de variables aléatoires $X = \{X_1, X_2, \ldots, X_k\}$ et soit $\mathbb{P}$ une mesure de probabilité sur $\mathcal{X}$ de densité (ou fonction de masse) $f$. Si $\mathbb{P}$ se factorise selon $G$ alors elle satisfait la propriété de Markov globale par rapport à $G$.*

Preuve. Soit $(A, B, D)$ un triplet de sous-ensembles de $V$ disjoints deux à deux tel que $D$ sépare $A$ de $B$. Posons $A^* = A \cup (V \setminus D)_{G\text{-}A}$, où $(V \setminus D)_{G\text{-}A}$ représente tous les nœuds de $V \setminus D$ qui sont connectés à $A$ dans le sous-graphe engendré par $V \setminus D$. Le graphe de la figure 3.4 illustre la composition des différents sous-ensembles pour 12 variables que nous pouvons généraliser à un nombre quelconque de variables

Figure 3.4: Illustration pour la preuve de la proposition 3.3.1




aléatoires. Par séparation, nous avons $A^* \cap B = \emptyset$. Posons $B^* = V \setminus (A^* \cup D)$, nous avons donc $A^*$, $B^*$ et $D$ sont disjoints deux à deux et $D$ sépare $A^*$ de $B^*$ avec $A^* \cup B^* \cup D = V$. Il est donc clair qu'une clique de $V$ appartient soit à $A^* \cup D$ soit à $B^* \cup D$. Notons $\mathcal{C}$ l'ensemble des cliques de $A^* \cup D$ et $\widetilde{\mathcal{C}}$ l'ensemble des cliques de $B^* \cup D$, par (3.4) nous avons

$$f(x) = f(x_{A^*}, x_{B^*}, x_D) = \prod_{C \in \mathcal{C}} \psi_C(x_C) \prod_{\widetilde{C} \in \widetilde{\mathcal{C}}} \psi_{\widetilde{C}}(x_{\widetilde{C}}). \tag{3.5}$$

Nous avons aussi

$$\begin{aligned} f(x_{A^*}, x_{B^*} \mid x_D) &= \frac{f(x_{A^*}, x_{B^*}, x_D)}{\int_{\mathcal{X}_{B^*}} \int_{\mathcal{X}_{A^*}} f(x_{A^*}, x_{B^*}, x_D) dx_{A^*} dx_{B^*}} \\ &= \frac{\prod_{C \in \mathcal{C}} \psi_C(x_C) \prod_{\widetilde{C} \in \widetilde{\mathcal{C}}} \psi_{\widetilde{C}}(x_{\widetilde{C}})}{\int_{\mathcal{X}_{B^*}} \int_{\mathcal{X}_{A^*}} \left( \prod_{C \in \mathcal{C}} \psi_C(x_C) \prod_{\widetilde{C} \in \widetilde{\mathcal{C}}} \psi_{\widetilde{C}}(x_{\widetilde{C}}) \right) dx_{A^*} dx_{B^*}} \qquad (3.6) \\ &= \frac{\prod_{C \in \mathcal{C}} \psi_C(x_C)}{\int_{\mathcal{X}_{A^*}} \prod_{C \in \mathcal{C}} \psi_C(x_C) dx_{A^*}} \times \frac{\prod_{\widetilde{C} \in \widetilde{\mathcal{C}}} \psi_{\widetilde{C}}(x_{\widetilde{C}})}{\int_{\mathcal{X}_{B^*}} \prod_{\widetilde{C} \in \widetilde{\mathcal{C}}} \psi_{\widetilde{C}}(x_{\widetilde{C}}) dx_{B^*}} \qquad (3.7) \end{aligned}$$

L'équation (3.6) est obtenue en appliquant (3.5), concernant le passage à (3.7), il est justifié par le fait $C \cap B^* = \emptyset$ et $\widetilde{C} \cap A^* = \emptyset$. Enfin l'application de (3.4) nous donne :

$$\begin{aligned} f(x_{A^*}, x_{B^*} \mid x_D) &= \frac{f(x_{A^*}, x_D)}{\int_{\mathcal{X}_{A^*}} f(x_{A^*}, x_D) dx_{A^*}} \times \frac{f(x_{B^*}, x_D)}{\int_{\mathcal{X}_{B^*}} f(x_{B^*}, x_D) dx_{B^*}} \\ &= f(x_{A^*} \mid x_D) f(x_{B^*} \mid x_D). \end{aligned}$$

alors

$$X_{A^*} \perp\!\!\!\perp X_{B^*} \mid X_D.$$

Puisque $A \subset A^*$ et $B \subset B^*$, alors, en appliquant deux fois la propriété de décomposition de l'indépendance conditionnelle (équation (3.1)) nous pouvons écrire :

$$X_{A^*} \perp\!\!\!\perp X_{B^*} \mid X_D \implies X_A \perp\!\!\!\perp X_{B^*} \mid X_D \implies X_A \perp\!\!\!\perp X_B \mid X_D. \qquad \square$$

## 3.4 Modèles graphiques orientés

Les modèles graphiques orientés, aussi appelés réseaux bayésiens, sont basés sur les graphes orientés acycliques. Ces modèles sont adaptés à la modélisation des relations asymétriques, notamment les relations de cause à effet. Par asymétrie, on entend que si un phénomène $X$ cause un autre phénomène $Y$, il est rare que $Y$ puisse causer $X$. Le but de cette section est de présenter trois propriétés importantes des réseaux bayésiens, à savoir la compatibilité de Markov, la condition de Markov et la propriété de Markov globale pour les graphes orientés. Ces trois propriétés sont équivalentes (voir théorème 3.4.1), mais pour pouvoir en faire la démonstration, nous devons introduire d'autres notions liées aux graphes non orientés.

Comme pour les modèles graphiques non orientés, les modèles graphiques orientés permettent de mieux représenter la loi de probabilité jointe des variables aléatoires. En effet, grâce à la représentation graphique, il serait possible de factoriser la loi de

probabilité jointe en produit de plusieurs lois conditionnelles, chacune dépendant d'un sous-ensemble de variables.

Définition 3.4.1 (Compatibilité de Markov). *Soit $G = (V, E)$ un graphe orienté acyclique dont les nœuds représentent un ensemble de variables aléatoires $X = \{X_1, X_2, \ldots, X_k\}$. On note $PA_i$ l'ensemble des parents de $X_i$, $i = 1, \ldots, k$. Soit $\mathbb{P}$ une mesure de probabilité sur $\mathcal{X}$ de densité (fonction de masse) $f$. Si $f$ admet une factorisation sous la forme :*

$$f(x_1, x_2, \ldots, x_k) = \prod_{i=1}^{k} f(x_i \mid x_{pa_i}), \tag{3.8}$$

*alors on dit que $\mathbb{P}$ et $G$ sont compatibles, $G$ est compatible avec $\mathbb{P}$, ou $\mathbb{P}$ est compatible avec $G$.*

La formule (3.8) permet de réduire le nombre de variables aléatoires intervenant dans la loi conditionnelle associée à chaque variable. Cela permet d'économiser de la mémoire et du temps de traitement dans le tableau de la loi de probabilité jointe. L'ensemble $X_{PA_i}$ est appelé « parents Markoviens » (Pearl, 2009). Nous sommes maintenant en mesure de définir les réseaux bayésiens à partir des éléments exposés ci-dessus.

Définition 3.4.2 (Réseau bayésien). *Soit $G = (V, E)$ un graphe orienté acyclique représentant un ensemble de variables aléatoires $X = \{X_1, X_2, \ldots, X_k\}$ et soit $\mathbb{P}$ une mesure de probabilité sur $\mathcal{X}$ de densité (fonction de masse) $f$. On dit que $G$ est un réseau bayésien s'il est compatible avec $\mathbb{P}$.*

On peut donc dire que la condition nécessaire pour que le DAG $G$ présenté dans la définition 3.4.2 soit un réseau bayésien est que $f$ se factorise sous la forme de la formule (3.8). On peut donc voir un réseau bayésien comme un DAG représentant un ensemble de variables aléatoires $X = \{X_1, X_2, \ldots, X_k\}$, où un arc de $X_j$ vers $X_i$ signifie que $X_j$ est un parent Markovien de $X_i$. Reprenons le DAG de la figure 3.1 en supposant que les nœuds du graphe représentent des variables aléatoires de

densité (fonction de masse) $f$. Le DAG est un réseau bayésien si :

$$f(x, z_1, z_2, z_3, y, z_4) = f(z_1)f(x \mid z_1)f(z_2 \mid x)f(z_3 \mid z_1, z_2)f(y \mid z_2, z_3)f(z_4 \mid y).$$

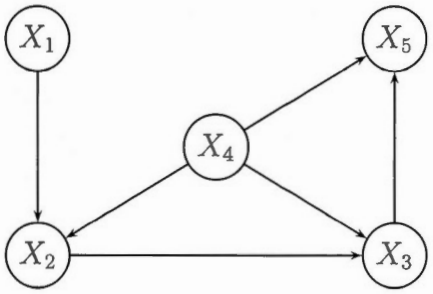
Une autre caractéristique des réseaux bayésiens est que les modèles peuvent encoder des relations d'indépendance conditionnelle. Nous avons vu que pour les modèles graphiques non orientés, grâce à leur propriété de Markov globale (définition 3.3.2), il devient possible d'interpréter les séparations entre des sous-ensembles de nœuds d'un graphe comme des relations d'indépendance conditionnelle entre les sous-ensembles de variables aléatoires correspondants. La notion de u-séparation est aussi adaptable aux réseaux bayésiens. Néanmoins, elle n'implique pas toujours l'indépendance conditionnelle, notamment lorsque le graphe contient une V-structure. Une V-structure est une chaîne convergente en $D$ $A \rightarrow D \leftarrow B$, où $A$, $B$ et $D$ sont trois nœuds tels que $A$ et $B$ ne sont pas liés directement par un arc. Dans une telle structure, $A$ et $B$ sont séparés par $D$, mais les variables aléatoires $X_A$ et $X_B$ sont dépendantes étant donné $X_D$. Plus formellement, on a $X_A \not\perp\!\!\!\perp X_B \mid X_D$ alors que $(A \perp\!\!\!\perp_u B \mid D)_G$. Pour tenir compte des V-structures, Pearl (1988) a introduit un nouveau critère graphique nommé « d-séparation » (« d » provient du mot en anglais *directional*).

### 3.4.1 D-séparation dans un réseau bayésien

La notion de d-séparation est plus complexe que la u-séparation. Elle permet de déterminer le mécanisme de la circulation de l'information dans un graphe orienté acyclique. Nous commençons par rappeler la définition de la d-séparation telle qu'elle est donnée par Pearl (2009), avant examiner comment elle est liée à la notion d'indépendance conditionnelle.

Définition 3.4.3. *Un chemin $C$ est dit d-séparé (ou bloqué) par un ensemble de nœuds $D$ si et seulement si :*

Figure 3.5: d-séparation dans un DAG



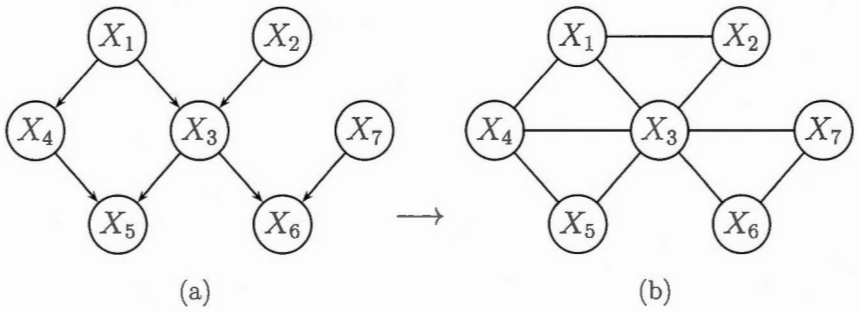

1. *$C$ contient une chaîne en série $i \to m \to j$ ou une chaîne divergente en $m$ $i \leftarrow m \to j$ tel que $m$ appartient à $D$, ou*
2. *$C$ contient une chaîne convergente en $m$ $i \to m \leftarrow j$ tel que $m$ n'appartient pas à $D$ et qu'aucun descendant de $m$ ne soit dans $D$.*

*On dit qu'un ensemble $D$ d-sépare les deux ensembles $A$ et $B$ si, et seulement si, $D$ bloque tout chemin allant d'un nœud de $A$ à un nœud de $B$.*

Examinons le graphe orienté acyclique de la figure 3.5. Selon la définition 3.4.3, l'ensemble $\{X_3, X_4\}$ d-sépare l'ensemble $\{X_5\}$ de l'ensemble $\{X_1, X_2\}$, car il bloque tous les chemins entre $X_2$ et $X_5$ et ceux entre $X_1$ à $X_5$. Par contre, il n'existe aucun ensemble (y compris l'ensemble vide) qui d-sépare l'ensemble $\{X_1\}$ de l'ensemble $\{X_4, X_3\}$. En effet, le seul ensemble qui bloque le chemin $X_1 \to X_2 \to X_3$ est $\{X_2\}$, puisque ce chemin est une chaîne en série avec comme nœud milieu $X_2$. Mais, en même temps, le chemin $X_1 \to X_2 \leftarrow X_4$ est une chaîne convergente en $X_2$, un ensemble d-séparant $X_1$ de $X_4$ ne doit donc pas contenir $X_2$. Par conséquent, le seul ensemble qui d-sépare $X_1$ de $X_3$ débloque un chemin bloqué allant de $X_1$ à $X_4$.

Comme pour la séparation, nous adopterons la notation $(A \perp\!\!\!\perp_d B \mid D)_G$ pour dire que les nœuds $A$ et $B$ sont d-séparés par $D$ dans le DAG $G$.

Figure 3.6: Passage d'un DAG à un graphe moral




Lauritzen *et al.* (1990) ont introduit une autre version de la d-séparation équivalente à celle que nous venons de présenter. Elle est basée sur les notions de graphe moral et d'ensemble ancestral.

Définition 3.4.4 (Graphe moral). *Soit $G = (V, E)$ un graphe orienté acyclique, on appelle graphe moral associé à $G$, le graphe non orienté $G^m = (V, E^m)$ où $E^m$ est composé de :*

- *toutes les arêtes $(i, j)$ telles que $(i, j) \in E$ ou $(j, i) \in E$,*
- *toutes les arêtes $(i, j)$ telles que les nœuds $i$ et $j$ ont un enfant en commun dans $G$.*

Selon la définition 3.4.4, pour obtenir un graphe moral à partir d'un DAG, on ajoute une arête entre chaque paire de nœuds ayant un enfant en commun, ensuite on transforme tous les arcs en arêtes non orientées (voir figure 3.6).

Définition 3.4.5 (Ensemble ancestral). *Soit $G = (V, E)$ un graphe orienté acyclique et soit $A$ un ensemble de nœuds tel que $A \subset V$, on dit que $A$ est ancestral si pour tout $X_i \in A$ on a $PA_i \subset A$. Un graphe engendré par un ensemble ancestral est appelé graphe ancestral.*

La notion de sous-ensemble engendré est présentée dans la section 3.1. L'ensemble $\{X_1, X_2, X_4\}$ lié au graphe de la figure 3.5 est ancestral.

Définition 3.4.6 (Autre définition de la d-séparation). *Soit $G = (V, E)$ un graphe orienté acyclique et soit $(A, B, D)$ un triplet de sous-ensembles de $V$ disjoints deux à deux. On dit que $D$ d-sépare $A$ et $B$ si $D$ sépare $A$ et $B$ dans le graphe moral associé au graphe ancestral engendré par le plus petit sous-ensemble ancestral contenant $A \cup B \cup D$.*

La définition 3.4.6 permet de lier la d-séparation à la notion de séparation dans les graphes non orientés.

### 3.4.2 Relations d'indépendance conditionnelle dans un réseau bayésien

Comme nous l'avons déjà mentionné, les réseaux bayésiens encodent des relations d'indépendance conditionnelle, ce qui se traduit par deux propriétés de Markov pour les modèles graphiques orientés, à savoir la condition de Markov et la propriété de Markov globale pour les graphes orientés.

Définition 3.4.7 (Condition de Markov). *On dit qu'une mesure de probabilité $\mathbb{P}$ satisfait la condition de Markov par rapport à un graphe orienté acyclique $G$, si chaque nœud $X_i$ de $G$ est indépendant de ses non-descendants étant donné ses parents Markoviens $X_{PA_i}$, et on écrit $X_i \perp\!\!\!\perp X_{ND_i} \mid X_{PA_i}$, où $ND_i$ représentent les non-descendants de $X_i$ dans $G$ ($X_i$ et $X_{PA_i}$ ne sont pas inclus dans $X_{ND_i}$).*

Cette propriété est connue aussi sous le nom de propriété de Markov locale et elle est souvent assimilée à la définition de la notion du réseau bayésien comme dans Williamson (2005). Il est à noter que dans notre définition des réseaux bayésiens, nous avons utilisé la notion de compatibilité de Markov (définition 3.4.1).

Définition 3.4.8 (Propriété de Markov globale pour les graphes orientés). *Soit $\mathbb{P}$ une mesure de probabilité et soit $X$ un ensemble de variables aléatoires représentées par un graphe orienté acyclique $G = (V, E)$. On dit que $\mathbb{P}$ satisfait la propriété de Markov globale par rapport à $G$, si pour tout triplet $(A, B, D)$ de sous-ensembles*

*de $V$ disjoints deux à deux tel que $D$ d-sépare $A$ de $B$, on a $X_A$ est indépendant de $X_B$ sachant $X_D$. De manière formelle on a :*

$$(A \perp\!\!\!\perp_d B \mid D)_G \implies X_A \perp\!\!\!\perp X_B \mid X_D.$$

Cette propriété permet d'établir le premier lien entre la d-séparation dans un DAG et l'indépendance conditionnelle des variables liées à ce DAG. Le critère de d-séparation permet donc de lire directement à partir d'un DAG des relations d'indépendance conditionnelle satisfaites par une loi de probabilité qui satisfait la propriété de Markov globale par rapport au DAG (définition 3.4.8). Cela est également valable pour les lois de probabilité satisfaisant la condition de Markov (définition 3.4.7) et la compatibilité de Markov (définition 3.4.1). En effet, les trois propriétés sont équivalentes comme nous le verrons avec le théorème 3.4.1. Cependant, nous présenterons tout d'abord quelques résultats utiles à la preuve de ce théorème.

Supposons que l'on dispose d'un graphe orienté acyclique $G = (V, E)$ dont les nœuds représentent un ensemble de variables aléatoires $X = \{X_1, X_2, \ldots, X_k\}$ et soit $\mathbb{P}$ une mesure de probabilité sur $\mathcal{X}$ de densité (fonction de masse) $f$.

Lemme 3.4.1. *Si $\mathbb{P}$ est compatible avec le DAG $G$, alors elle vérifie la propriété de Markov globale par rapport au graphe moral $G^m$ lié à $G$.*

Preuve. Dans le graphe moral $G^m$ (définition 3.4.4), chaque sous-ensemble $C_i = \{X_i\} \cup \{PA_i\}$ est complet. Alors $\forall\ i = 1, \ldots, k$, $\exists\ \psi_{C_i}$ tel que :

$$f(x_1, x_2, \ldots, x_k) = \prod_{C_i} \psi_{C_i}(x_{C_i}).$$

Nous avons donc que $\mathbb{P}$ se factorise par rapport à $G^m$ et, d'après la proposition 3.3.1, $\mathbb{P}$ satisfait la propriété de Markov globale par rapport à $G^m$. □

Proposition 3.4.1. *Si $\mathbb{P}$ est compatible avec le DAG $G$, alors elle vérifie la propriété de Markov globale par rapport à $G$.*

Preuve. Soit $(A, B, D)$ un triplet de sous-ensembles de $V$ disjoints deux à deux et soit $S$ le plus petit sous-ensemble ancestral de $V$ contenant $A \cup B \cup D$. Posons $\widetilde{S} = V \setminus S$ et $\acute{G} = (S, \acute{E})$ le DAG engendré par $S$. Nous avons $\mathbb{P}$ est compatible avec $G$, alors selon la définition 3.4.1 :

$$\begin{aligned} f(x_1, x_2, \ldots, x_k) &= \prod_{i=1}^{k} f(x_i \mid x_{PA_i}) \\ &= \prod_{x_i \in x_S} f(x_i \mid x_{PA_i}) \prod_{x_j \in x_{\widetilde{S}}} f(x_j \mid x_{PA_j}), \end{aligned}$$

et puisque $\acute{G}$ est un graphe ancestral (définition 3.4.5) alors $\prod_{x_i \in x_S} f(x_i \mid x_{PA_i})$ ne dépend pas de $x_{\widetilde{S}}$, nous avons donc :

$$\begin{aligned} f(x_S) &= \int_{\mathcal{X}_{\widetilde{S}}} f(x_1, x_2, \ldots, x_k) dx_{\widetilde{S}} \\ &= \int_{\mathcal{X}_{\widetilde{S}}} \prod_{x_i \in x_S} f(x_i \mid x_{PA_i}) \prod_{x_j \in x_{\widetilde{S}}} f(x_j \mid x_{PA_j}) dx_{\widetilde{S}} \\ &= \prod_{x_i \in x_S} f(x_i \mid x_{PA_i}) \int_{\mathcal{X}_{\widetilde{S}}} \prod_{x_j \in x_{\widetilde{S}}} f(x_j \mid x_{PA_j}) dx_{\widetilde{S}} \\ &= \prod_{x_i \in x_S} f(x_i \mid x_{PA_i}) \times 1 \\ &= \prod_{x_i \in x_S} f(x_i \mid x_{PA_i}), \end{aligned}$$

ce qui signifie que $f(x_S)$ est aussi compatible avec $\acute{G}$, alors d'après le lemme 3.4.1, $f(x_S)$ satisfait la propriété de Markov globale selon le graphe moral $\acute{G}^m$. Nous avons donc la relation d'implication suivante :

$$(A \perp\!\!\!\perp_u B \mid D)_{\acute{G}^m} \implies X_A \perp\!\!\!\perp X_B \mid X_D,$$

et selon la définition 3.4.6, nous pouvons écrire aussi :

$$(A \perp\!\!\!\perp_d B \mid D)_G \implies X_A \perp\!\!\!\perp X_B \mid X_D,$$

ce qui signifie que $\mathbb{P}$ satisfait la propriété de Markov globale selon $G$. □

Proposition 3.4.2. *Si $\mathbb{P}$ satisfait la propriété de Markov globale par rapport au DAG $G$, alors elle vérifie la condition de Markov par rapport à $G$.*

Preuve. Nous avons pour tout $i = 1, \ldots, k$, $\{X_i\} \cup PA_i \cup ND_i$ est un sous-ensemble ancestral de $V$. Nous avons aussi $(X_i \perp\!\!\!\perp_d ND_i \mid PA_i)_G$, alors d'après la proposition 3.4.1 $X_i \perp\!\!\!\perp X_{ND_i} \mid X_{PA_i}$. □

Proposition 3.4.3. *Si $\mathbb{P}$ satisfait la condition de Markov par rapport au DAG $G$, alors elle est compatible avec $G$.*

Preuve. Nous nommons les variables aléatoires représentées par le graphe de sorte que si $X_i$ est un descendant de $X_j$, alors $i > j$. D'après le théorème de multiplication nous avons

$$f(x_1, x_2, \ldots, x_k) = f(x_1) f(x_2 \mid x_1) f(x_3 \mid x_1, x_2) \ldots f(x_k \mid x_1, x_2, \ldots, x_{k-1}).$$

Chaque sous-ensemble $\{X_1, \ldots, X_{i-1}\}$ contient tous les parents de $X_i$ et aucun de ses descendants. D'après la condition de Markov nous avons :

$$\begin{aligned} f(x_1, x_2, \ldots, x_k) &= f(x_1) f(x_2 \mid x_1) f(x_3 \mid x_1, x_2) \ldots f(x_k \mid x_1, x_2, \ldots, x_{k-1}) \\ &= f(x_1) \prod_{i=2}^{k} f(x_i \mid x_{PA_i}). \end{aligned}$$

La variable $X_1$ n'a pas de prédécesseur dans l'ordre spécifié, cela signifie que l'ensemble $PA_1$ est vide, alors $f(x_1) = f(x_1 \mid x_{PA_1})$. Par conséquent, la loi de probabilité jointe $f(x_1, x_2, \ldots, x_k)$ peut s'écrire sous la forme de :

$$f(x_1, x_2, \ldots, x_k) = \prod_{i=1}^{k} f(x_i \mid x_{PA_i}),$$

ce qui signifie que $\mathbb{P}$ est compatible avec $G$. □

Théorème 3.4.1. *Soit $G = (V, E)$ un graphe orienté acyclique dont les nœuds représentent un ensemble de variables aléatoires $X = \{X_1, X_2, \ldots, X_k\}$ et soit $\mathbb{P}$ une mesure de probabilité sur $\mathcal{X}$ de densité (fonction de masse) $f$. Les trois conditions suivantes :*

1. *$\mathbb{P}$ est compatible avec $G$ (définition 3.4.1) ;*
2. *$\mathbb{P}$ satisfait la condition de Markov par rapport à $G$ (définition 3.4.7) ;*
3. *$\mathbb{P}$ satisfait la propriété de Markov globale par rapport à $G$ (définition 3.4.8),*

*sont équivalentes.*

Preuve. La démonstration du théorème découle directement des trois propositions précédentes. En effet, d'après la proposition 3.4.1, la compatibilité de Markov implique la propriété de Markov globale, qui implique à son tour la condition de Markov d'après la proposition 3.4.2. Enfin, selon la proposition 3.4.3 la condition de Markov implique la compatibilité de Markov. □

Étant donnée l'équivalence des trois conditions du théorème 3.4.1, on a le choix d'utiliser l'une ou l'autre selon le besoin. En résumé, on peut dire qu'un DAG dont les nœuds sont des variables aléatoires est un réseau bayésien s'il satisfait l'une des propriétés précédentes. Celles-ci permettent d'interpréter les d-séparations dans le DAG comme étant des relations d'indépendance conditionnelle qui sont vérifiées pour chaque loi compatible avec le DAG. On parle dans ce cas de relations d'indépendance conditionnelle engendrée par un DAG.

Définition 3.4.9 (Indépendance conditionnelle engendrée par un DAG). *Soit* $G = (V, E)$ *un graphe orienté acyclique représentant un ensemble de variables aléatoires* $X$ *et soit* $(A, B, D)$ *un triplet de sous-ensembles de* $V$ *disjoints deux à deux. On dit que le DAG* $G$ *engendre la relation d'indépendance conditionnelle* $X_A \perp\!\!\!\perp X_B \mid X_D$ *si cette relation est vérifiée pour tout* $\mathbb{P} \in \mathcal{P}$*, où* $\mathcal{P}$ *représente l'ensemble de toutes les lois de probabilité compatibles avec* $G$*.*

Le lien entre d-séparation et indépendance conditionnelle est en réalité plus fort comme l'indique le théorème suivant :

Théorème 3.4.2. *Soit* $V$ *un ensemble de variables aléatoires représenté par un graphe orienté acyclique* $G = (V, E)$ *et soit* $\mathcal{P}$ *l'ensemble de toutes les lois de probabilité compatibles avec* $G$*. Pour tout triplet* $(A, B, D)$ *de sous-ensembles de* $V$ *disjoints deux à deux, on a :*

- $(A \perp\!\!\!\perp_d B \mid D)_G \implies X_A \perp\!\!\!\perp X_B \mid X_D$ *pour tout* $\mathbb{P} \in \mathcal{P}$*, et*

- *si pour tout* $\mathbb{P} \in \mathcal{P}$, $X_A \perp\!\!\!\perp X_B \mid X_D$, *alors* $(A \perp\!\!\!\perp_d B \mid D)_G$.

Pour la suite de ce travail, nous aurions besoin uniquement des résultats démontrés précédemment pour justifier notre démarche. La démonstration du théorème 3.4.2 est très longue et nécessite l'introduction d'autres théorèmes et concepts qui ne font pas l'objet de ce travail. La preuve est donnée dans Geiger *et al.* (1990), Verma et Pearl (1988) ainsi que dans Spirtes *et al.* (2000). Le théorème 3.4.2 indique qu'un réseau bayésien engendre toutes les indépendances conditionnelles identifiées par le critère de d-séparation dans le DAG et que chaque indépendance conditionnelle vérifiée par toutes les lois de probabilité compatibles avec le DAG est une d-séparation dans le DAG. Cependant, il se pourrait qu'une loi de probabilité particulière compatible avec un DAG inclut une relation d'indépendance qui ne soit pas identifiée par la d-séparation dans le DAG. On parle alors d'une loi de probabilité non fidèle au DAG (Spirtes *et al.*, 2000).

Définition 3.4.10 (Fidélité). *Soit* $G = (V, E)$ *un graphe orienté acyclique représentant un ensemble de variables aléatoires* $X$ *et soit* $\mathbb{P}$ *une loi de probabilité compatible avec* $G$. *On dit que* $\mathbb{P}$ *satisfait la condition de fidélité si toute indépendance conditionnelle vérifiée par* $\mathbb{P}$ *est engendrée par le DAG* $G$.

La condition de fidélité, aussi connue sous le nom de condition de stabilité (Pearl, 1988) est souvent considérée vérifiée car en pratique, les lois de probabilité non fidèles sont rares.

## 3.5 Réseau bayésien causal et calculs d'interventions

Les relations d'indépendance et de dépendance conditionnelle encodées par un réseau bayésien peuvent être de natures différentes. Néanmoins, elles sont souvent causales ou chronologiques. L'intérêt des modèles graphiques orientés pour modéliser les relations de cause à effet est dû au fait que les relations de cause

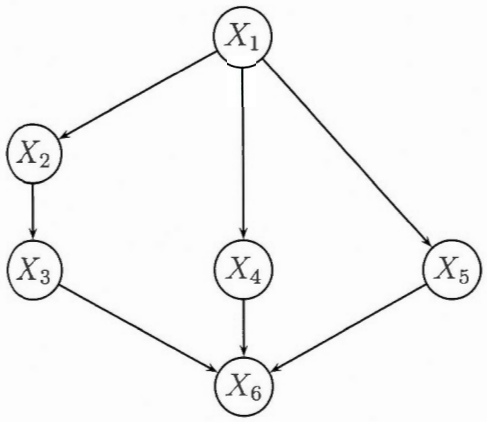
à effet sont souvent intuitives et asymétriques, ce qui correspond à la propriété d'acyclicité des réseaux bayésiens contrairement aux relations d'association qui sont généralement quantitatives et symétriques.

Supposons que l'on dispose d'un graphe orienté acyclique représentant un ensemble de variables aléatoires où chaque arc se retrouvant entre deux nœuds représente une relation de cause à effet directe. Un tel DAG est appelé graphe causal. Nous allons illustrer ce cas à l'aide d'un exemple inspiré de De Oliveira *et al.* (2010), où nous représentons une théorie selon laquelle le brossage des dents peut réduire le risque de maladies cardiaques. En effet, ces dernières années, des chercheurs cherchent à établir un lien entre la maladie des gencives et les maladies cardiaques. L'une des hypothèses évoquées est que l'infection de la gencive favorise l'introduction de bactéries dans le sang ce qui peut provoquer des problèmes cardiaques. Les hypothèses de cette théorie sont représentées dans le graphe de la figure 3.7 où les nœuds du DAG sont des variables aléatoires définies comme suit :

- $X_1$ : une variable qui mesure le niveau de prise de conscience d'un individu de sa santé.
- $X_2$ : variable mesurant la fréquence de brossage des dents par un individu.
- $X_3$ : une variable qui indique l'état de la gencive d'un individu, elle prend la valeur 1 si la gencive est atteinte et 0 sinon.
- $X_4$ : variable mesurant la fréquence de l'activité physique d'un individu.
- $X_5$ : une variable qui mesure la quantité de viande rouge et grasse dans l'alimentation de l'individu.
- $X_6$ : variable mesurant l'état cardiaque de l'individu, cette variable prend la valeur 1 si l'individu est atteint d'une maladie cardiaque et 0 sinon.

Le DAG de la figure 3.7 est un exemple de graphe causal et il est construit à l'aide des intuitions causales. Par exemple, l'arc allant de $X_1$ à $X_4$ signifie que le niveau de prise de conscience d'un individu a un effet direct sur son activité

Figure 3.7: Un Réseau bayésien pour relations causales



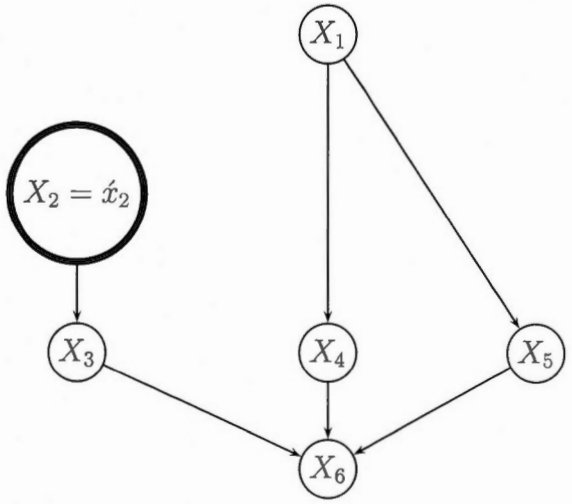

physique. L'absence d'un arc allant directement de $X_2$ à $X_6$ signifie que l'effet du brossage des dents de l'individu sur son état cardiaque dépend s'il est atteint d'une maladie des gencives. Cela veut dire que connaissant $X_3$ rend $X_6$ indépendant de $X_2$, on peut voir également que connaissant $\{X_3, X_4, X_5\}$ rend $X_6$ indépendant de $\{X_1, X_2\}$. Nous retrouvons donc les relations d'indépendance conditionnelle et la condition de Markov qui caractérisent les réseaux bayésiens. Le DAG de la figure 3.7 représente donc un réseau bayésien. Soit $f$ la loi de probabilité jointe de l'ensemble des variables aléatoires $X = \{X_1, \ldots, X_6\}$, $f$ peut donc s'écrire sous la forme de l'équation (3.8)

$$f(x) = f(x_1)f(x_2 \mid x_1)f(x_3 \mid x_2)f(x_4 \mid x_1)f(x_5 \mid x_1)f(x_6 \mid x_3, x_4, x_5). \qquad (3.9)$$

### 3.5.1 Calcul d'interventions

Pour Pearl (2009), une intervention notée $do(X_i = \acute{x}_i)$ consiste à intervenir de l'extérieur sur la variable $X_i$ et lui imposer une valeur $\acute{x}_i$, ensuite regarder l'effet de ce changement sur d'autres variables. Graphiquement, cela consiste à retirer tous les arcs pointant sur $X_i$. Reprenons l'exemple de la figure 3.7 et imaginons une intervention externe $do(\acute{x}_2)$ qui fait en sorte que tous les individus de la

Figure 3.8: Exemple d'intervention




population brossent leurs dents à une fréquence $\acute{x}_2$. On peut penser, par exemple, à une campagne de sensibilisation hors norme dont on est certain qu'elle incitera tout le monde à brosser ses dents à une fréquence $\acute{x}_2$. Le DAG de la figure 3.8 est le résultat de cette intervention. Il a été obtenu en retirant l'arc pointant sur $X_2$ du graphe de la figure 3.7. Ce nouveau graphe est Markov compatible avec une nouvelle mesure de probabilité notée $\mathbb{P}_{do(\acute{x}_2)}$ dont la densité $f_{do(\acute{x}_2)}$ s'écrit sous la forme :

$$f_{do(\acute{x}_2)}(x) = f(x_1)f(x_3 \mid X_2 = \acute{x}_2)f(x_4 \mid x_1)f(x_5 \mid x_1)f(x_6 \mid x_3, x_4, x_5). \quad (3.10)$$

Si nous comparons l'équation (3.10) à l'équation (3.9), la seule différence entre les deux est dans le facteur $f(x_2 \mid x_1)$, les autres facteurs sont restés les mêmes. En effet, le fait qu'on soit intervenu sur $X_2$, son passé n'a plus d'influence. Le fait que les autres facteurs n'aient pas changé montre également l'autonomie et la stabilité des relations de cause à effet. L'exemple précédent illustre bien le rôle des réseaux bayésiens dans la modélisation des relations de cause à effet. Néanmoins, les propriétés de ces modèles sont insuffisantes pour faire de l'inférence et déterminer l'effet causal à partir des équations de la forme de la formule (3.10).

### 3.5.2 Réseaux bayésiens causaux

Les réseaux bayésiens causaux sont des réseaux bayésiens avec quelques conditions supplémentaires à respecter pour pouvoir faire de l'inférence causale et calculer l'effet d'une intervention sur une variable sur les autres variables. Soit $G = (V, E)$ un graphe orienté acyclique représentant un ensemble de variables aléatoires $X = \{X_1, \ldots, X_k\}$ et soit $\mathbb{P}$ une mesure de probabilité sur $\mathcal{X}$ de densité (fonction de masse) $f$. On définit une intervention $do(x_A)$ qui impose la valeur $\acute{x}_A$ à $X_A$ où $A \subset V$ et notons $\mathbb{P}_{do(\acute{x}_A)}$ la mesure de probabilité qui résulte de cette intervention et $f_{do(\acute{x}_A)}$ sa fonction de densité correspondante. Soit $\mathcal{P}$ l'ensemble de toutes les probabilités de type $\mathbb{P}_{do(\acute{x}_A)}$ incluant $\mathbb{P}$ qui correspond au cas où on n'effectue pas d'intervention. Selon Pearl (2009), $G$ est un réseau bayésien causal dit compatible avec $\mathcal{P}$ si, et seulement si, pour tout $\mathbb{P}_{do(\acute{x}_A)} \in \mathcal{P}$ on a :

1. $\mathbb{P}_{do(\acute{x}_A)}$ est compatible avec $G$ ;
2. Pour tout $x_i \in X_A$, $f_{do(\acute{x}_A)}(x_i) = 1$ ;
3. Pour tout $x_i \notin X_A$, $f_{do(\acute{x}_A)}(x_i \mid x_{PA_i}) = f(x_i \mid x_{PA_i})$.

La condition (1) signifie que $f_{do(\acute{x}_A)}$ admet une factorisation de la forme suivante :

$$f_{do(\acute{x}_A)}(x_1, \ldots, x_k) = \prod_{i|X_i \notin X_A}^{k} f(x_i \mid x_{PA_i}). \tag{3.11}$$

Sous les trois conditions énumérées ci-dessus, il est possible de déterminer l'effet causal d'une variable $X_i$ sur une autre variable $X_j$. En premier lieu, on détermine la factorisation de la loi de $f_{do(x_i)}$ qui résulte d'une intervention $do(\acute{x}_i)$, ensuite on calcule la densité (ou fonction de masse) marginale $f_{do(\acute{x}_i)}(x_j)$ de $X_j$ et on regarde le comportement de cette densité marginale vis-à-vis des différentes valeurs possibles de $\acute{x}_i$.

Les réseaux bayésiens causaux sont un puissant outil de modélisation des relations de cause à effet, mais ils ne permettent pas de faire le lien entre l'analyse graphique

et l'analyse contrefactuelle présentée au chapitre 2. Pour cela, nous allons présenter un autre modèle, en l'occurrence le modèle causal à équations structurelles, qui permet de faire ce lien.

[Cette page a été laissée intentionnellement blanche]

# CHAPITRE IV

# MODÉLISATION CAUSALE PAR ÉQUATIONS STRUCTURELLES

Dans un réseau bayésien causal, les relations parents/enfants sont probabilistes et non déterministes. Par relation causale déterministe, on entend que la valeur d'une variable est déterminée d'une façon fonctionnelle par celles de ses parents. Dans le cas opposé, elle est dite indéterministe ou probabiliste. Ainsi, les valeurs prises par les parents ne déterminent pas celles de leurs enfants, mais elles modifient uniquement la probabilité de les observer. Les modèles causaux à équations structurelles souvent appelés modèles causaux, considèrent les relations parents/enfants comme déterministes, modélisées par des équations dites structurelles avec une composante aléatoire liée aux variables non observées.

## 4.1 Équations structurelles

Les modèles causaux sont basés essentiellement sur un ensemble d'équations structurelles exprimant des relations fonctionnelles et causales entre des variables aléatoires. Pearl (2009) exprime ces équations sous la forme :

$$x_i = h_i(pa_i, u_i), \quad i = 1, \ldots, k, \tag{4.1}$$

où $X_1, X_2, \ldots, X_k$ sont des variables aléatoires endogènes et observables, $PA_i$ représente l'ensemble des variables aléatoires qui déterminent de manière déterministe la valeur de $X_i$, et $U_i$ est une variable aléatoire exogène (non expliquée par le

modèle) représentant l'ensemble des variables non observables (ou non observées) qui sont susceptibles d'influencer $X_i$. Nous sommes maintenant prêts à définir le modèle causal à partir des éléments exposés ci-dessus.

## 4.2 Modèle causal

Définition 4.2.1 (modèle causal). *Un modèle causal est un triplet* $M = (U, X, H)$ *où :*

- $U$ *est un ensemble de variables exogènes (erreurs) qui sont déterminées par des facteurs externes au modèle ;*
- $X = \{X_1, X_2, \ldots, X_k\}$ *est un ensemble de variables endogènes qui sont déterminées par des variables du modèle, à savoir des variables dans* $X \cup U$ *;*
- $H = \{h_1, h_2, \ldots, h_k\}$ *est un ensemble de fonctions tel que chaque fonction* $h_i$ *est une application de* $PA_i \cup U_i$ *dans* $X$ *:*

$$x_i = h_i(pa_i, u_i), i = 1, \ldots, k, \tag{4.2}$$

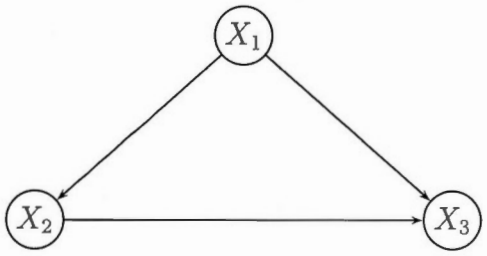
*où* $U_i \in U$ *et* $PA_i \subseteq X \setminus X_i$.

À titre d'illustration, soit $X = \{X_1, X_2, X_3\}$ un ensemble de variables endogènes et $U = \{U_1, U_2, U_3\}$ un ensemble de variables exogènes, et supposons que ces variables sont liées par les équations structurelles suivantes :

$$\begin{aligned} x_1 &= h_1(u_1), \\ x_2 &= h_2(x_1, u_2), \\ x_3 &= h_3(x_1, x_2, u_3). \end{aligned} \tag{4.3}$$

Le système d'équations (4.3) définit donc un modèle causal. Comme nous pouvons le voir dans cet exemple, le sous-ensemble $PA_i$ décrit dans la définition 4.2.1 peut être vide. Ainsi, selon le système d'équations (4.3), la valeur de $x_1$ est déterminée uniquement par $u_1$, ce qui signifie que $PA_1 = \emptyset$.

Figure 4.1: Exemple de diagramme causal



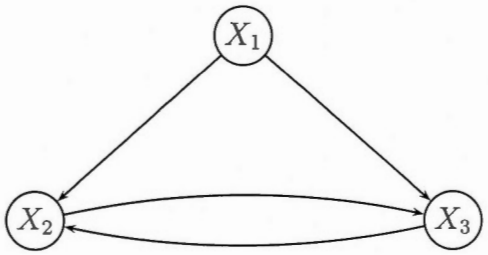

### 4.2.1 Diagramme causal

Dans la définition 4.2.1, pour désigner le sous-ensemble de $X$ déterminant la valeur de $X_i$, nous avons opté pour la notation $PA_i$ que nous avons utilisée dans les modèles graphiques orientés. Ce choix est motivé par le fait qu'à chaque modèle causal $M$ correspond un graphe orienté $G(M)$ appelé diagramme causal (Pearl, 2009), dont le sous-ensemble de variables $PA_i$ est représenté par un sous-ensemble de nœuds $PA_i$ qui correspond à l'ensemble des parents de $X_i$ dans le graphe. Le diagramme causal $G(M) = (X, E)$ associé au modèle causal $M$ présenté dans la définition 4.2.1 est obtenu en traçant un arc de chaque élément de $PA_i$ vers $X_i$. Le diagramme causal associé au système d'équations (4.3) est présenté dans la figure 4.1. Pour des raisons pratiques, chaque ensemble de variables et sous-ensemble de nœuds correspondants seront notés, dans ce chapitre, par la même lettre. Par exemple, l'ensemble $X$ désigne à la fois l'ensemble des variables endogènes du modèle et l'ensemble des nœuds du diagramme causal, ce qui n'était pas le cas dans le chapitre précédent.

À chaque modèle causal $M$, est également associée une loi de probabilité jointe des variables représentées dans le diagramme causal $G(M)$. Nous notons $\mathbb{P}$ cette loi de probabilité et $f(x)$ sa densité ou sa fonction de masse.

Contrairement à un réseau bayésien, un diagramme causal n'est pas nécessai-

Figure 4.2: Diagramme causal associé à un modèle causal non-Markovien




rement acyclique. En effet, comme nous l'avons déjà mentionné, les relations de cause à effet sont souvent asymétriques. Néanmoins, certaines d'entre elles peuvent être symétriques. Par symétrie, on entend que si un phénomène $X$ cause un autre phénomène $Y$, il est possible que $Y$ puisse également causer $X$ directement ou à travers un autre phénomène $Z$. Reprenons l'exemple du modèle causal associé au système d'équations (4.3) et supposons, en outre, que la variable $X_2$ pourrait avoir un effet causal sur la variable $X_1$. Il en résulte donc un nouveau modèle causal dont les équations structurelles sont les suivantes :

$$\begin{aligned} x_1 &= h_1(u_1), \\ x_2 &= h_2(x_1, x_3, u_2), \\ x_3 &= h_3(x_1, x_2, u_3). \end{aligned} \tag{4.4}$$

Le diagramme causal associé à ce nouveau modèle est présenté dans la figure 4.2.

Les relations causales symétriques conduisent à des cycles dans le diagramme causal et le résultat est un diagramme causal cyclique (voir figure 4.2) auquel correspond un modèle causal appelé modèle causal non-Markovien. Si le diagramme causal est acyclique, le modèle causal qui lui correspond est appelé modèle causal semi-Markovien, et si en plus les variables exogènes sont mutuellement indépendantes, on parle d'un modèle causal Markovien. Une caractéristique importante des modèles causaux Markoviens et semi-Markoviens est que chaque variable en-

dogène peut s'exprimer uniquement en fonction des variables exogènes. Reconsidérons le système d'équations (4.3) ; une autre façon d'exprimer les variables endogènes $x_1$, $x_2$ et $x_3$ est :

$$\begin{aligned} x_1 &= h_1(u_1), \\ x_2 &= h_2(h_1(u_1), u_2), \\ x_3 &= h_3(h_1(u_1), h_2(h_1(u_1), u_2), u_3). \end{aligned} \tag{4.5}$$

Cette formulation nous indique que pour les modèles causaux Markoviens et semi-Markoviens, la loi de probabilité jointe des variables endogènes $\mathbb{P}$ est entièrement déterminée par celle des variables exogènes que nous notons $\mathbb{P}_U$.

### 4.2.2 Condition de Markov causale

Dans la section 3.4.2, nous avons présenté une importante propriété des réseaux bayésiens, en l'occurrence la condition de Markov (définition 3.4.7). Comme nous l'avons déjà mentionné, cette propriété est souvent assimilée à la définition du réseau bayésien. Une autre propriété analogue à celle que nous venons d'évoquer caractérise les diagrammes causaux et elle se définit comme suit :

Définition 4.2.2 (condition de Markov causale). *Soit $M$ un modèle causal auquel est associé un diagramme causal $G(M) = (X, E)$ et une mesure de probabilité $\mathbb{P}$ de densité (fonction de masse) $f(x)$. On dit que $\mathbb{P}$ satisfait la condition de Markov causale par rapport à $G(M)$ si, et seulement si, elle satisfait la condition de Markov par rapport à $G(M)$.*

Le point sur lequel la condition de Markov causale diffère de la condition de Markov est que la première impose que les relations parents/enfants soient des relations de cause à effet directes. Il est à noter qu'une mesure de probabilité associée à un diagramme causal ne satisfait pas nécessairement la condition de Markov causale. Mais cette propriété est principalement liée aux modèles causaux

Markoviens, comme l'indique le théorème suivant :

Théorème 4.2.1. *Soit $M$ un modèle causal Markovien auquel est associé un diagramme causal $G(M) = (X, E)$ et une loi de probabilité $\mathbb{P}$ de densité de probabilité (fonction de masse) $f(x)$. Posons $V = X \cup U$, où $X = \{X_1, X_2, \ldots, X_k\}$ représente l'ensemble des variables endogènes de $M$ et $U = \{U_1, U_2, \ldots, U_k\}$ désigne l'ensemble des variables exogènes du modèle. Sous ces conditions, la loi de probabilité $\mathbb{P}$ satisfait la condition de Markov causale par rapport au diagramme causal $G(M)$.*

Preuve. Considérons le DAG $\acute{G} = (V, \acute{E})$ où les variables exogènes apparaissent explicitement. Nous nommons les variables endogènes de sorte que si $X_i$ est un descendant de $X_j$, alors $i > j$. D'après le théorème de multiplication, nous avons :

$$f(u_1, \ldots, u_k, x_1, \ldots, x_k) = f(u_1)f(u_2 \mid u_1)f(u_3 \mid u_1, u_2) \ldots f(u_k \mid u_1, u_2, \ldots, u_{k-1})$$
$$\times f(x_1 \mid u_1, \ldots, u_k)f(x_2 \mid u_1, \ldots, u_k, x_1) \ldots f(x_k \mid u_1, \ldots, u_k, x_1, \ldots, x_{k-1}).$$

Chaque $X_i$ est déterminé fonctionnellement par l'ensemble $\{PA_i, U_i\}$. Nous savons également que $M$ est un modèle causal Markovien, cela signifie que les variables exogènes sont mutuellement indépendantes. Par conséquent, nous avons :

$$f(u_1, \ldots, u_k, x_1, \ldots, x_k) = \prod_{i=1}^{k} f(u_i) \prod_{i=1}^{k} f(x_i \mid pa_i, u_i).$$

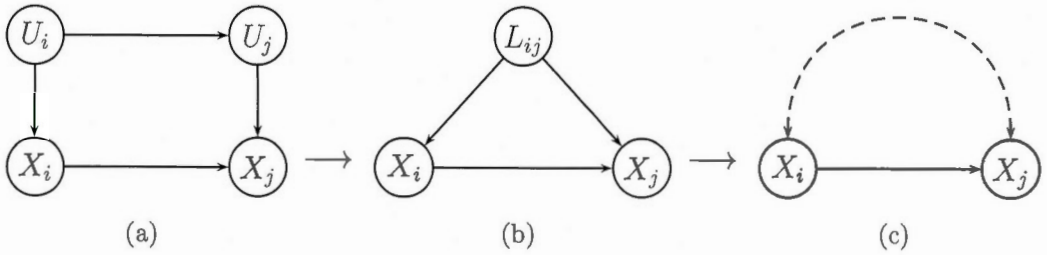
Notons $PAU_i$ l'ensemble des parents de $U_i$ dans $\acute{G}$. Nous avons $PAU_i = \emptyset$, $\forall$ $i = 1, \ldots, k$, alors $f(u_i) = f(u_i \mid pau_i)$. Nous avons donc :

$$f(u_1, \ldots, u_k, x_1, \ldots, x_k) = \prod_{i=1}^{k} f(u_i \mid pau_i) \prod_{i=1}^{k} f(x_i \mid pa_i, u_i). \tag{4.6}$$

L'équation (4.6) signifie que la loi de probabilité jointe des variables de $V$ notée $\mathbb{P}_V$ est compatible avec le DAG $\acute{G}$ et, selon le théorème 3.4.1 $\mathbb{P}_V$ satisfait également la propriété de Markov globale selon $\acute{G}$. Toute d-séparation dans le DAG peut donc être considérée comme une relation d'indépendance conditionnelle.

Notons $ND_i$ l'ensemble des non-descendants de $X_i$ dans $\acute{G}$ appartenant à $X$.

Figure 4.3: Représentation d'une variable latente dans un diagramme causal




Nous avons pour tout $i = 1, \ldots, k$, $(X_i \perp\!\!\!\perp_d ND_i \mid PA_i)_{\mathcal{G}}$. Cela implique que $X_i \perp\!\!\!\perp ND_i \mid PA_i$. L'ensemble $ND_i$ correspond également à l'ensemble des non-descendants de $X_i$ dans le diagramme causal $G(M)$, ce qui signifie que $\mathbb{P}$ satisfait la condition de Markov par rapport à $G(M)$. Et puisque les relations parents/enfants dans $G(M)$ sont des relations de cause à effet directes, alors $\mathbb{P}$ satisfait également la condition de Markov causale par rapport à $G(M)$. □

Dans le cas d'un modèle causal semi-Markovien, certaines variables exogènes ne sont pas indépendantes et la condition de Markov causale n'est donc pas satisfaite. Pour pouvoir la restaurer, il devient nécessaire de représenter ces dépendances dans le diagramme causal. Pearl (2009) suggère de modéliser chaque dépendance entre deux variables exogènes $U_i$ et $U_j$ par une variable latente $L_{ij}$ non observable qui affecte à la fois $X_i$ et $X_j$ et d'inclure cette variable dans les deux ensembles $PA_i$ et $PA_j$. Cependant, comme dans les modèles causaux Markoviens les variables non observables n'apparaissent pas souvent dans le diagramme causal, la variable latente $L_{ij}$ est souvent représentée dans le diagramme causal par un arc en pointillé liant $X_i$ et $X_j$ comme indiqué sur la figure 4.3.

Dans ce qui suit, nous nous intéresserons uniquement aux modèles causaux Markoviens et aux modèles causaux semi-Markoviens dont les variables latentes sont représentées dans le diagramme causal. Par conséquent, lorsqu'on parle de modèle

causal, on exclut les modèles causaux non-Markoviens.

L'intérêt du théorème 4.2.1 est qu'il permet d'utiliser les résultats établis dans les réseaux bayésiens, et de les appliquer aux modèles causaux Markoviens et semi-Markoviens. Ainsi, il devient possible de factoriser la loi de probabilité jointe des variables représentées dans le diagramme causal sous la forme :

$$f(x) = \prod_{i=1}^{k} f(x_i \mid pa_i). \tag{4.7}$$

Cette factorisation est possible sans même spécifier les formes des fonctions $h_i$ et celle de la loi de probabilité jointe des variables exogènes $\mathbb{P}'_U$, qui, comme nous l'avons mentionné auparavant, détermine entièrement la loi de probabilité jointe des variables endogènes $\mathbb{P}$.

Druzdzel et Simon (1993) ont montré pour leur part qu'à chaque réseau bayésien ayant une loi de probabilité jointe $\mathbb{P}$ qui se factorise sous la forme de la formule (3.8) est associé un modèle causal Markovien $M$ ayant une loi de probabilité jointe identique à $\mathbb{P}$. Considérons de nouveau la figure 3.7 et assumons que les relations de cause à effet entre les variables aléatoires représentées dans le DAG sont des relations fonctionnelles, le graphe est donc un diagramme causal. Les équations structurelles qui lui correspondent sont les suivantes :

$$\begin{aligned} x_1 &= h_1(u_1), \\ x_2 &= h_2(x_1, u_2), \\ x_3 &= h_3(x_2, u_3), \\ x_4 &= h_4(x_1, u_4), \\ x_5 &= h_5(x_1, u_5), \\ x_6 &= h_6(x_3, x_4, x_5, u_6). \end{aligned} \tag{4.8}$$

Les variables aléatoires $U_1, U_2, \ldots, U_6$ sont considérées mutuellement indépendantes et le modèle causal est donc Markovien.

## 4.3 Interventions et identification de l'effet causal

Dans la section 3.5, nous avons introduit le mécanisme de calcul de l'effet d'une intervention sur une variable $X_i$ notée $do(\acute{x}_i)$ sur une autre variable $X_j$ dans le cas des réseaux bayésiens causaux. Dans cette section, nous allons élargir la notion d'intervention aux modèles causaux à équations structurelles, sur laquelle se base le calcul de l'effet causal.

Soit $M$ un modèle causal tel qu'il est donné dans la définition 4.2.1. Une intervention consiste à modifier certaines fonctions de $H$ tout en gardant les autres fonctions inchangées. Le résultat est un autre modèle causal ayant une nouvelle loi de probabilité jointe. Nous nous intéressons plus particulièrement à l'intervention $do(\acute{x}_i)$ qui consiste à imposer la valeur $\acute{x}_i$ à une certaine variable $X_i$ comme nous l'avons déjà vu dans la section 3.5. Il résulte de cette intervention un nouveau modèle causal que nous notons $M_{do(\acute{x}_i)}$. Ce modèle est obtenu en supprimant l'équation $x_i = h_i(pa_i, u_i)$ de l'ensemble des équations structurelles de $M$ et en la remplaçant par la fonction constante $X_i = \acute{x}_i$, et également, fixant la valeur de $X_i$ à $\acute{x}_i$ dans les autres équations. À ce nouveau modèle causal est associée une loi de probabilité jointe que nous notons $\mathbb{P}_{do(\acute{x}_i)}$ de densité (fonction de masse) $f_{do(\acute{x}_i)}(x)$ que nous notons également $f(x \mid do(\acute{x}_i))$ pour des raisons pratiques et qui vérifie :

$$f(x \mid do(\acute{x}_i)) = \prod_{j \neq i} f(x_j \mid pa_j). \tag{4.9}$$

Reprenons le système d'équations (4.8) associé à l'exemple de la figure 3.7, et supposons une intervention $do(\acute{x}_2)$ qui fait en sorte que tous les individus de la population brossent leurs dents à une fréquence $\acute{x}_2$. Les équations structurelles

liées au modèle issu de cette intervention $M_{do(\acute{x}_2)}$ sont données par :

$$\begin{aligned} x_1 &= h_1(u_1), \\ X_2 &= \acute{x}_2, \\ x_3 &= h_3(\acute{x}_2, u_3), \\ x_4 &= h_4(x_1, u_4), \\ x_5 &= h_5(x_1, u_5), \\ x_6 &= h_6(x_3, x_4, x_5, u_6). \end{aligned} \tag{4.10}$$

À ce modèle causal est associé une loi de probabilité jointe $\mathbb{P}_{do(\acute{x}_2)}$ dont la densité $f_{do(\acute{x}_2)}$ s'écrit sous la forme :

$$f(x \mid do(\acute{x}_2)) = f(x_1)f(x_3 \mid X_2 = \acute{x}_2)f(x_4 \mid x_1)f(x_5 \mid x_1)f(x_6 \mid x_3, x_4, x_5). \tag{4.11}$$

À partir de la nouvelle loi de probabilité jointe $\mathbb{P}_{do(\acute{x}_i)}$ liée à l'intervention $do(\acute{x}_i)$, il devient possible de calculer l'effet causal de la variable $X_i$ sur une autre variable $X_j$. Nous allons nous concentrer sur le cas discret, mais on peut étendre les résultats au cas continu. Selon Pearl (2009), l'effet causal de la variable $X_i$ sur la variable $X_j$ correspond à la loi de probabilité marginale qui s'écrit :

$$f(x_j \mid do(\acute{x}_i)) = \mathbb{P}_{do(\acute{x}_i)}(X_j = x_j). \tag{4.12}$$

Le théorème suivant indique le principe de calcul de l'effet causal dans un modèle causal à équations structurelles Markovien.

Théorème 4.3.1. *Soit $M$ un modèle causal auquel est associée une loi de probabilité $\mathbb{P}$ de fonction de masse $f$ et un diagramme causal $G(M) = (X, E)$. Soit $do(\acute{x}_i)$ une intervention externe qui engendre un nouveau modèle causal $M_{do(\acute{x}_i)}$ auquel est associée une loi de probabilité $\mathbb{P}_{do(\acute{x}_i)}$ de fonction de masse $f_{do(\acute{x}_i)}(x)$. Pour tout $X_j \notin \{X_i \cup PA_i\}$ l'effet causal de $X_i$ sur $X_j$ est donné par :*

$$f(x_j \mid do(\acute{x}_i)) = \sum_{pa_i} f(x_j \mid \acute{x}_i, pa_i)f(pa_i). \tag{4.13}$$

Preuve. Par (4.7) et (4.9) nous avons :

$$f(x \mid do(\acute{x}_i)) = \frac{f(x)}{f(\acute{x}_i \mid pa_i)}$$
$$= f(x \mid \acute{x}_i, pa_i) f(pa_i). \tag{4.14}$$

Posons $\tilde{X} = X \setminus \{X_i, X_j\}$, par marginalisation des variables appartenant à $\tilde{X}$ nous obtenons :

$$\sum_{\tilde{x}} f(x \mid do(\acute{x}_i)) = \sum_{\tilde{x}} f(x \mid \acute{x}_i, pa_i) f(pa_i),$$

ce qui implique que :

$$f(x_j \mid do(\acute{x}_i)) = \sum_{pa_i} f(x_j \mid \acute{x}_i, pa_i) f(pa_i).$$

□

### 4.3.1 Identification de l'effet causal dans un modèle causal Markovien

La question clé dans les modèles causaux est l'identification de l'effet causal, c'est-à-dire la possibilité d'estimer $f(x_j \mid do(\acute{x}_i))$ qui représente la densité (ou fonction de masse) après intervention à partir des données des variables observées pour lesquelles on connait la loi de probabilité avant intervention $\mathbb{P}$. Les termes de droite de l'équation (4.13) représentent des probabilités avant intervention. Il apparaît donc clairement que pour identifier l'effet causal, il faut pouvoir mesurer les variables $X_i$ et $X_j$ ainsi que tous les parents de $X_i$. Tout cela se résume par le théorème suivant :

Théorème 4.3.2. *Soit $M$ un modèle causal Markovien auquel est associée une loi de probabilité $\mathbb{P}$ de fonction de masse $f$ et un diagramme causal $G(M) = (X, E)$. Soit $do(\acute{x}_i)$ une intervention externe qui engendre un nouveau modèle causal $M_{do(\acute{x}_i)}$ auquel est associée une loi de probabilité $\mathbb{P}_{do(\acute{x}_i)}$ de fonction de masse $f_{do(\acute{x}_i)}(x)$. Si $O \subset X$ représente l'ensemble des variables observées appartenant à $X$, alors pour tout ensemble $\{X_i, X_j, PA_i\} \subseteq O$, l'effet causal de $X_i$ sur $X_j$ $f(x_j \mid do(\acute{x}_i))$ est identifiable et il est donné par l'équation* (4.13).

L'équation (4.13) est connue sous le nom de « Ajustement sur les parents observables ».

Dans notre étude de cas, nous cherchons à estimer l'effet du traitement PCC (*Group*) sur la mortalité à 6 mois (*6mo_mortality*) et sur l'expansion de l'hématome (*expansion*). Nous considérons que le graphe de la figure 1.1 est un diagramme causal issu d'un modèle causal Markovien. Les erreurs du modèle sont donc considérées indépendantes et toutes les variables du graphe sont observées. Nous pouvons donc effectuer un ajustement sur l'ensemble des parents de la variable *Group* qui sont donnés par :

$$PA_{Group} = \{cad, dvt\_pe, afib, valve, cmy, cancer, gender, evd, volume, postfossa, location, inr\_bwh, age\}.$$

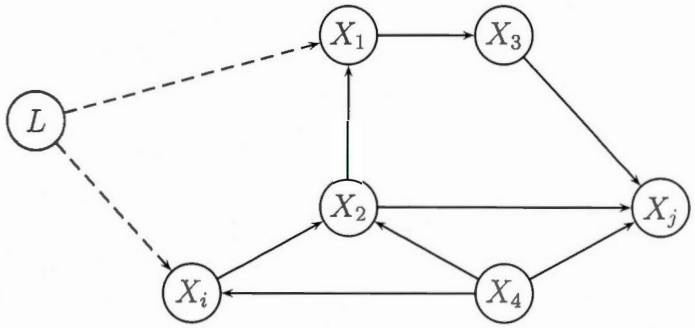
L'ensemble $PA_{Group}$ est représenté par des nœuds colorés sur la figure B.1 de l'annexe B.

Lorsqu'une ou plusieurs variables de $PA_i$ ne sont pas observées, il devient difficile, et parfois impossible, d'identifier l'effet causal. C'est le cas par exemple lorsque $X_i$ et $X_j$ ont une cause commune non observée $L$ ($X_i \leftarrow L \rightarrow X_j$). L'effet causal de $X_i$ sur $X_j$ se confond avec celui de $L$ qui crée une association supplémentaire non causale entre $X_i$ et $X_j$ (biais de confusion). Pour remédier à cette situation, il existe un critère graphique appelé critère porte-arrière qui permet d'éliminer parfois cette association non causale et d'identifier l'effet causal.

### 4.3.2 Critère porte-arrière

Dans cette section, nous supposons que l'on dispose d'un diagramme causal $G = (V, E)$ auquel est associée une loi de probabilité discrète $\mathbb{P}$ de fonction de masse $f$. Le but ici est de déterminer les conséquences d'une intervention $do(x_i)$ qui engendre un nouveau diagramme causal et une nouvelle loi de probabilités $\mathbb{P}_{do(x_i)}$

Figure 4.4: Illustration du critère porte-arrière




de fonction de masse $f(x \mid do(\acute{x}_i))$. Le critère porte-arrière permet de sélectionner un ensemble de variables observées $Z$, tel qu'un conditionnement par ses variables permet d'éliminer le biais de confusion dans l'estimation de l'effet causal. Mais avant d'aborder ce critère, nous commençons par définir la notion de « chemin porte-arrière ».

Définition 4.3.1 (chemin porte-arrière). *Soit $G = (X, E)$ un graphe orienté acyclique et soient $X_i$ et $X_j$ deux variables de $G$. On appelle chemin porte-arrière allant de $X_i$ à $X_j$ tout chemin entre $X_i$ et $X_j$ ayant un arc qui pointe sur la variable $X_i$.*

Considérons le DAG de la figure 4.4 qui inclut une variable non observée $L$. Le chemin $X_i \dashleftarrow L \dashrightarrow X_1 \longrightarrow X_3 \longrightarrow X_j$ est un exemple de chemin porte-arrière allant de $X_i$ à $X_j$. En se basant sur la définition 4.3.1, Pearl (1993) introduit le critère porte-arrière comme suit :

Définition 4.3.2 (critère porte-arrière). *Soit $G = (X, E)$ un graphe orienté acyclique et soient $X_i$ et $X_j$ deux variables de $G$. On dit qu'un sous-ensemble $Z$ de $X$ satisfait le critère porte-arrière par rapport à la paire ordonnée $(X_i, X_j)$ si :*

*(i) aucun descendant de $X_i$ n'est dans $Z$ ;*

*(ii) $Z$ d-sépare tous les chemins porte-arrière allant de $X_i$ à $X_j$.*

Reprenons à nouveau le DAG de la figure 4.4. Le sous-ensemble $Z_1 = \{X_1, X_2, X_4\}$ satisfait le critère porte-arrière par rapport à la paire ordonnée $(X_i, X_j)$, alors que le sous-ensemble $Z_2 = \{X_1, X_4\}$ ne satisfait pas ce critère par rapport à la même paire car il ne d-sépare le chemin porte-arrière $X_i \dashleftarrow L \dashrightarrow X_1 \longleftarrow X_2 \longrightarrow X_j$.

Lorsqu'il existe un sous-ensemble $Z$ satisfaisant le critère porte-arrière par rapport à une paire ordonnée $(X_i, X_j)$, il devient possible d'identifier l'effet causal de $X_i$ sur $X_j$ comme l'indique le théorème suivant :

Théorème 4.3.3. *S'il existe un sous-ensemble de variables observées $Z$ qui satisfait le critère porte-arrière par rapport à une paire ordonnée de variables $(X_i, X_j)$, alors l'effet causal de $X_i$ sur $X_j$ est identifiable et il est donné par :*

$$f(x_j \mid do(\acute{x}_i)) = \sum_z f(x_j \mid \acute{x}_i, z) f(z). \tag{4.15}$$

Preuve. Nous avons vu que dans le cas d'un modèle causal Markovien, l'effet causal de la variable $X_i$ sur $X_j$ est donné par l'équation (4.13). Supposons maintenant que certaines variables de $PA_i$ ne sont pas observées. Nous commençons tout d'abord par exprimer de manière plus formelle les deux conditions (i) et (ii) de la définition 4.3.2. Selon la condition de Markov la condition (ii) implique que :

$$X_i \perp\!\!\!\perp Z \mid PA_i. \tag{4.16}$$

La condition (i) implique $(X_j \perp\!\!\!\perp_d PA_i \mid X_i, Z)_G$. Pour le voir, soit $C$ un chemin entre $X_j$ et $PA_i$. Si $C$ traverse $X_i$, cela signifie qu'il contient un chemin porte-arrière allant de $X_i$ à $X_j$. Alors, un ensemble $Z$ qui d-sépare ce chemin porte-arrière d-sépare également le chemin $C$, ce qui signifie que $C$ est aussi d-séparé par $\{X_i, Z\}$. Si $C$ ne traverse pas $X_i$, alors il est contenu dans un chemin porte-arrière allant de $X_i$ à $X_j$. Aussi, il est facile de voir qu'un ensemble $Z$ qui d-sépare ce chemin porte-arrière d-sépare $C$, et que $\{X_i, Z\}$ le d-sépare également. Par conséquent, nous avons $(X_j \perp\!\!\!\perp_d PA_i \mid X_i, Z)_G$, ce qui implique que :

$$X_j \perp\!\!\!\perp PA_i \mid X_i, Z. \tag{4.17}$$

D'après (4.13) nous avons :

$$\begin{aligned} f(x_j \mid do(\acute{x}_i)) &= \sum_{pa_i} f(x_j \mid \acute{x}_i, pa_i) f(pa_i) \\ &= \sum_{pa_i} f(pa_i) \sum_{z} f(x_j, z \mid \acute{x}_i, pa_i) \\ &= \sum_{pa_i} f(pa_i) \sum_{z} f(x_j \mid \acute{x}_i, pa_i, z) f(z \mid \acute{x}_i, pa_i). \end{aligned}$$

En utilisant (4.16) et (4.17) nous obtenons :

$$\begin{aligned} f(x_j \mid do(\acute{x}_i)) &= \sum_{pa_i} f(pa_i) \sum_{z} f(x_j \mid \acute{x}_i, z) f(z \mid pa_i) \\ &= \sum_{z} f(x_j \mid \acute{x}_i, z) \sum_{pa_i} f(pa_i) f(z \mid pa_i) \\ &= \sum_{z} f(x_j \mid \acute{x}_i, z) \sum_{pa_i} f(z, pa_i) \\ &= \sum_{z} f(x_j \mid \acute{x}_i, z) f(z). \end{aligned}$$

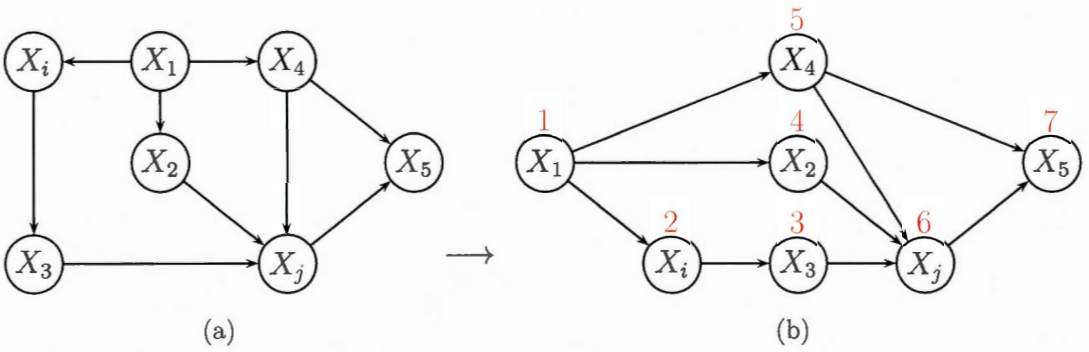
Ce qui correspond bien à l'équation (4.15). □

D'autres démonstrations du théorème sont données dans Pearl (2009) et Pearl (1993).

Le sous-ensemble $Z$ est dit suffisant pour l'ajustement de la confusion. Il est dit suffisant minimal pour l'ajustement de la confusion s'il n'est pas inclus dans un autre ensemble suffisant (Greenland *et al.*, 1999). Il est à noter que si $Z_1$ est suffisant pour l'ajustement de la confusion, un ensemble $Z_2$ contenant $Z_1$ ($Z_1 \subset Z_2$) n'est pas forcement suffisant. Pour illustrer ceci, considérons le graphe de la figure 4.4. L'ensemble $Z_1 = \{X_3, X_4\}$ est suffisant pour l'ajustement de la confusion, car il d-sépare tous les chemins porte-arrière allant de $X_i$ et $X_j$. Par contre, l'ensemble $Z_1 = \{X_1, X_3, X_4\}$ n'est pas suffisant, car il ne permet pas de d-séparer le chemin porte-arrière $X_i \dashleftarrow L \dashrightarrow X_1 \longleftarrow X_2 \longrightarrow X_j$ qui était initialement bloqué par $X_1$.

Comme nous l'avons déjà mentionné, il n'existe pas de définition claire d'une

Figure 4.5: Passage à un tri topologique



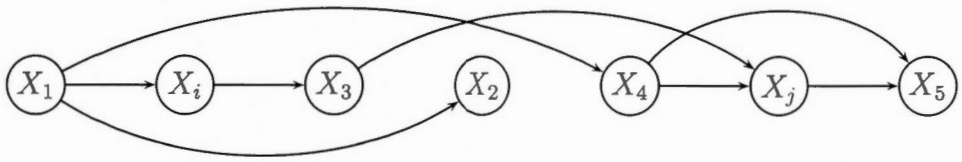

variable confondante. Pour certains auteurs (Hernán et Robins, 2013), est une variable confondante, toute variable appartenant à un ensemble suffisant minimal pour l'ajustement de la confusion. Nous adoptons cette définition pour la suite de ce travail.

### 4.3.3 Recherche d'ensembles suffisants pour l'ajustement de la confusion

Lorsque la taille du diagramme causal augmente, il devient difficile de déterminer les ensembles suffisants pour l'ajustement de la confusion dans l'examen de l'effet causal d'une variable $X_i$ sur une variable $X_j$. Pour les modèles causaux Markoviens, nous avons vu qu'il est toujours possible d'éliminer la confusion en ajustant sur les parents observables de la variable du traitement $X_i$. Néanmoins cette technique peut s'avérer parfois très coûteuse en raison de la nécessité d'observer toutes ces variables. Pour remédier à ces situations, il existe une technique permettant de réduire le nombre de variables à inclure dans l'ensemble suffisant pour l'ajustement de la confusion. Cette méthode est basée sur ce que l'on appelle « tri topologique d'un DAG » qui se définit comme suit :

Définition 4.3.3 (tri topologique). *Soit $G = (V, E)$ un graphe orienté acyclique. Un tri topologique de $G$ est un ordre des nœuds de $G$ tel que si $(v_1, v_2) \in E$ alors $v_1$ apparaît avant $v_2$ dans cet ordre.*

Figure 4.6: Autre représentation d'un tri topologique




Pour réaliser un tri topologique d'un DAG, on commence par regrouper les nœuds du DAG en plusieurs niveaux verticaux, de sorte qu'on obtient un nouveau DAG dont tous les arcs sont orientés de gauche à droite, comme indiqué sur la figure 4.5b. Ensuite, on procède à une numérotation ascendante de gauche à droite des nœuds du graphe obtenu (voir figure 4.5b). Pour les nœuds qui se trouvent sur un même niveau vertical, on choisit l'ordre qui convient le mieux à l'objectif recherché, ce qui signifie qu'un tri topologique n'est pas unique. Une autre façon de représenter un tri topologique consiste en un DAG dont les nœuds forment une ligne horizontale et qui sont liés par des arcs allant de gauche à droite. Un graphe de ce type lié à l'exemple de la figure 4.5 est présenté dans la figure 4.6.

Supposons maintenant que l'on dispose d'un diagramme causal $G = (V, E)$ associé à un modèle causal $M$ et que l'on souhaite déterminer un ensemble $Z$ suffisant minimal permettant d'éliminer la confusion dans l'estimation de l'effet causal d'une variable $X_i$ sur une autre variable $X_j$. Pour cela, il existe une technique simple qui consiste dans les manipulations suivantes :

(i) effectuer un tri topologique du graphe $G$ de façon à ce que les variables $X_i$ et $X_j$ aient les plus petits rangs possibles ;

(ii) supprimer du graphe trié toutes les variables qui sont classées après $X_j$ ainsi que les arcs qui leur sont liés ;

(iii) supprimer du graphe trié toute variable non traversée par un chemin entre $X_i$ et $X_j$ et les arcs qui lui correspondent ;

(iv) inclure dans l'ensemble $Z$ toute variable qui a moins de deux parents dans le graphe trié si elle correspond au nœud milieu d'une chaîne divergente de la forme $X_i \leftarrow X_k \rightarrow X_j$. Supprimer ensuite ces variables du graphe ainsi que les arcs qui leur sont liés.

La manipulation (ii) est motivée par le fait que tous les chemins traversant les variables concernées sont bloqués par une variable ayant deux parents correspondant à l'une de ces variables. Concernant la manipulation (iv), les nœuds évoqués doivent être absolument inclus dans l'ensemble $Z$, car c'est l'unique façon de bloquer les chemins porte-arrières sous forme de chaîne divergente.

À l'issue des manipulations susmentionnées, on obtient un nouveau graphe sur lequel on effectue à nouveau les manipulations (iii) et (iv) ainsi que sur chaque nouveau graphe jusqu'à ce qu'il n'y ait plus de simplifications possibles. Après avoir obtenu un graphe simplifié, on recherche un ensemble de variables suffisant minimal permettant de bloquer tous les chemins porte-arrières restés ouverts dans ce graphe. Cet ensemble, ainsi que toutes les variables sélectionnées suite aux manipulations de type (iv), forment l'ensemble suffisant minimal pour l'ajustement de la confusion que l'on recherche.

L'application de la technique que nous venons de voir à notre étude de cas sera présentée en détails dans la section 6.2.1. Nous allons déterminer des ensembles suffisants minimaux pour l'ajustement de la confusion dans les deux cas, à savoir dans l'étude de l'effet de la variable *Group* sur *6mo_mortality* et dans l'effet de *Group* sur *expansion*.

## 4.4 Lien avec la théorie contrefactuelle

Nous allons maintenant spécifier le lien entre l'analyse graphique de la causalité et l'analyse contrefactuelle que nous avons présentée au chapitre 2. Considérons

une quantité contrefactuelle notée $Y_{(\acute{d})i}$ qui désigne la valeur qu'aurait pu prendre la variable $Y$ pour l'unité $i$ si la variable $D$ avait pris la valeur $\acute{d}$. Pour Pearl (2009), il est possible de représenter cette quantité à l'aide d'un modèle causal $M$, décrit par un ensemble d'équations structurelles de la forme de l'équation (4.1) et dans lequel $Y$ et $D$ représentent deux variables endogènes dont les équations sont données par :

$$y = h_y(pa_y, u_y) \quad \text{et} \quad d = h_d(pa_d, u_d). \tag{4.18}$$

Imaginons maintenant une intervention $do(\acute{d})$ qui impose la valeur $\acute{d}$ à la variable $D$. Et soit $M_{do(\acute{d})}$ le modèle causal issu de cette intervention. Pour $U = u$, où $U$ représente le vecteur des variables exogènes dans le modèle, notons $Y_{do(\acute{d})}(u)$ la solution de $Y$ dans le système d'équations lié au modèle causal modifié $M_{do(\acute{d})}$. Pearl (2009) propose d'interpréter l'expression contrefactuelle « si la variable $D$ avait pris la valeur $\acute{d}$ » comme une action qui modifie le modèle de base $M$ en remplaçant l'équation $d = h_d(pa_d, u_d)$ par une constante $\acute{d}$. Il propose également de dresser un parallèle entre l'unité $i$ sur laquelle la quantité contrefactuelle est mesurée et la réalisation $u$ du vecteur des variables exogènes dans le modèle causal. Habituellement, en analyse contrefactuelle, l'unité $i$ correspond à un individu dans la population étudiée, mais on peut également l'assimiler à un ensemble de caractéristiques individuelles d'où le choix du vecteur des variables exogènes $U$ ayant une loi de probabilité jointe $\mathbb{P}_U$ qui détermine entièrement celle des variables endogènes. On peut donc donner un sens contrefactuel à la solution de $Y$ dans le modèle causal modifié $M_{do(\acute{d})}$, ainsi on a :

$$Y_{(\acute{d})u} \stackrel{def}{=} Y_{do(\acute{d})}(u). \tag{4.19}$$

Puisque $U$ est un vecteur aléatoire alors $Y_{(\acute{d})u}$ est aussi une variable aléatoire notée $Y_{(\acute{d})}$. On peut alors imaginer un ensemble de variables $\tilde{X}$ contenant l'ensemble des variables endogènes $X$ et les variables contrefactuelles. Notons $\tilde{\mathbb{P}}$ la loi de probabilité jointe des variables de $\tilde{X}$; cela signifie que la loi de probabilité jointe

des variables endogènes $\mathbb{P}$ est une loi marginale de $\tilde{\mathbb{P}}$. Par conséquent, en utilisant (4.19) l'effet causal de la variable $D$ sur la variable $Y$ donné par $f(y \mid do(\acute{d})) = \mathbb{P}_{do(\acute{d})}(Y = y)$, est également donnée par $\tilde{\mathbb{P}}(Y_{(\acute{d})} = y)$.

Supposons maintenant que l'on dispose d'un ensemble de variables $Z \subset X$ vérifiant :

$$Y_{(\acute{d})} \perp\!\!\!\perp D \mid Z. \tag{4.20}$$

L'équation (4.20) désigne l'hypothèse d'ignorabilité, mais avec une seule variable contrefactuelle. Sous (4.20) nous avons :

$$\begin{aligned}
\tilde{\mathbb{P}}(Y_{(\acute{d})} = y) &= \sum_z \tilde{\mathbb{P}}(Y_{(\acute{d})} = y \mid z) f(z) \\
&= \sum_z \tilde{\mathbb{P}}(Y_{(\acute{d})} = y \mid \acute{d}, z) f(z) \\
&= \sum_z \tilde{\mathbb{P}}(Y = y \mid \acute{d}, z) f(z) \\
&= \sum_z \mathbb{P}(Y = y \mid \acute{d}, z) f(z) \\
&= \sum_z f(y \mid \acute{d}, z) f(z).
\end{aligned} \tag{4.21}$$

Nous obtenons donc un résultat identique à celui de l'équation (4.15) obtenu en appliquant le critère porte-arrière. Cela signifie que l'ensemble suffisant pour l'ajustement de la confusion obtenu par le critère porte-arrière permet de satisfaire l'hypothèse d'ignorabilité et d'identifier l'effet causal dans l'analyse contrefactuelle.

# CHAPITRE V

# SCORE DE PROPENSION

L'une des méthodes les plus utilisées pour l'estimation de l'effet causal moyen est la méthode d'appariement sur un vecteur de covariables $\mathbf{Z}$ que nous avons vu dans le chapitre 2. Toutefois, lorsque le nombre de covariables augmente, il devient très difficile de trouver un apparié exact à chaque sujet traité. Pour remédier à ce problème, la théorie du score de propension permet de réduire la dimension de l'appariement et d'obtenir un bon estimateur de l'effet causal.

## 5.1 Score de propension : définition et propriétés

Introduite pour la première fois par Rosenbaum et Rubin (1983), la théorie du score de propension n'a pas cessé d'évoluer. Cette technique est basée sur un ensemble de méthodes d'estimation utilisant une probabilité particulière appelée score de propension.

Définition 5.1.1. *Le score de propension d'un individu noté* $e(\mathbf{Z})$ *est la probabilité que l'individu reçoive un traitement conditionnellement à un vecteur de covariables* $\mathbf{Z}$. *De manière formelle on a :*

$$e(\mathbf{Z}) = P(W = 1 \mid \mathbf{Z}). \tag{5.1}$$

### 5.1.1 Propriété d'équilibrage du score de propension

Pour Rosenbaum et Rubin (1983), le score de propension $e(\mathbf{Z})$ est avant tout un score d'équilibrage (balance) qui sert à équilibrer la distribution des covariables $\mathbf{Z}$ dans les deux groupes de traitement. Cette propriété est une conséquence directe du théorème suivant, attribué à Rosenbaum et Rubin (1983).

Théorème 5.1.1. *Les covariables sont indépendantes de l'affectation du traitement étant donné le score de propension.*

$$\mathbf{Z} \perp\!\!\!\perp W \mid e(\mathbf{Z}).$$

Preuve. il suffit de montrer que :

$$f(w \mid e(\boldsymbol{z})) = f(w \mid \boldsymbol{z}, e(\boldsymbol{z})),$$

ce qui revient à montrer que :

$$\mathbb{P}(W = 1 \mid \mathbf{Z}, e(\mathbf{Z})) = \mathbb{P}(W = 1 \mid e(\mathbf{Z})) \text{ et } \mathbb{P}(W = 0 \mid \mathbf{Z}, e(\mathbf{Z})) = \mathbb{P}(W = 0 \mid e(\mathbf{Z})).$$

Mais puisque $W$ est une variable dichotomique, nous avons :

$$\mathbb{P}(W = 0 \mid \mathbf{Z}, e(\mathbf{Z})) = 1 - \mathbb{P}(W = 1 \mid \mathbf{Z}, e(\mathbf{Z})),$$

et que :

$$\mathbb{P}(W = 0 \mid e(\mathbf{Z})) = 1 - \mathbb{P}(W = 1 \mid e(\mathbf{Z})).$$

Cela signifie que nous avons besoin de montrer uniquement que :

$$\mathbb{P}(W = 1 \mid \mathbf{Z}, e(\mathbf{Z})) = \mathbb{P}(W = 1 \mid e(\mathbf{Z})).$$

Nous avons $e(\mathbf{Z})$ est une fonction de $\mathbf{Z}$, alors :

$$\mathbb{P}(W = 1 \mid \mathbf{Z}, e(\mathbf{Z})) = \mathbb{P}(W = 1 \mid \mathbf{Z}) \tag{5.2}$$

$$= e(\mathbf{Z}). \tag{5.3}$$

Rappelons la loi des espérances étirées. Pour deux variables aléatoires $U$ et $V$ nous avons :

$$\mathbb{E}(U) = \mathbb{E}(\mathbb{E}(U \mid V)). \tag{5.4}$$

Maintenant, posons $U = W \mid e(\mathbf{Z})$ et $V = \mathbf{Z} \mid e(\mathbf{Z})$. Nous avons donc :

$$\begin{aligned}
\mathbb{P}(W = 1 \mid e(\mathbf{Z})) &= \mathbb{E}(U) \\
&= \mathbb{E}\left(\mathbb{E}(U \mid V)\right) \\
&= \mathbb{E}\left(\mathbb{E}(W \mid \mathbf{Z}, e(\mathbf{Z})) \mid e(\mathbf{Z})\right) \\
&= \mathbb{E}(e(\mathbf{Z}) \mid e(\mathbf{Z})) \\
&= e(\mathbf{Z}).
\end{aligned}$$

Cela est obtenu en appliquant (5.4) et (5.3).

$$\mathbb{P}(W = 1 \mid \mathbf{Z}, e(\mathbf{Z})) = \mathbb{P}(W = 1 \mid e(\mathbf{Z})) = e(\mathbf{Z}), \tag{5.5}$$

et cela implique que :

$$f(w \mid \boldsymbol{z}, e(\boldsymbol{z})) = f(w \mid e(\boldsymbol{z})),$$

alors :

$$f(\boldsymbol{z}, w \mid e(\boldsymbol{z})) = f(\boldsymbol{z} \mid e(\boldsymbol{z}))f(w \mid e(\boldsymbol{z})).$$

Ce qui signifie que :

$$\mathbf{Z} \perp\!\!\!\perp W \mid e(\mathbf{Z}).$$

Nous pouvons donc dire que les covariables sont indépendantes de l'affectation du traitement étant donné le score de propension. □

### 5.1.2 Hypothèse d'ignorabilité forte basée sur le score de propension

Dans la section 2.4, nous avons montré l'importance de l'hypothèse d'ignorabilité forte du traitement pour la réduction du biais. Rosenbaum et Rubin (1983) ont donné une autre version de cette hypothèse basée sur le score de propension, ce qui sera l'objet du théorème suivant :

Théorème 5.1.2. *Si l'affectation du traitement est fortement ignorable étant donné un ensemble de covariables* $\mathbf{Z}$*, alors il est fortement ignorable étant donné le score*

*de propension* $e(\mathbf{Z})$. *De manière formelle, si :*

$$(Y_{(1)}, Y_{(0)}) \perp\!\!\!\perp W \mid \mathbf{Z} \quad et \quad 0 < \mathbb{P}(W = 1 \mid \mathbf{Z}) < 1, \tag{5.6}$$

*alors*

$$(Y_{(1)}, Y_{(0)}) \perp\!\!\!\perp W \mid e(\mathbf{Z}) \quad et \quad 0 < \mathbb{P}(W = 1 \mid e(\mathbf{Z})) < 1. \tag{5.7}$$

Preuve. D'après (5.5) et la définition du score de propension, nous avons :

$$\mathbb{P}(W = 1 \mid \mathbf{Z}) = e(\mathbf{Z}) = \mathbb{P}(W = 1 \mid e(\mathbf{Z})),$$

cela signifie que :

$$\text{si} \quad 0 < \mathbb{P}(W = 1 \mid \mathbf{Z}) < 1 \quad \text{alors} \quad 0 < \mathbb{P}(W = 1 \mid e(\mathbf{Z})) < 1.$$

Nous allons maintenant montrer que sous l'hypothèse du théorème nous avons :

$$\mathbb{P}(W = w \mid Y_{(1)}, Y_{(0)}, e(\mathbf{Z})) = \mathbb{P}(W = w \mid e(\mathbf{Z})), w \in \{0, 1\}.$$

Nous savons déjà que :

$$\mathbb{P}(W = 1 \mid e(\mathbf{Z})) = e(\mathbf{Z}) \quad \text{et} \quad \mathbb{P}(W = 0 \mid e(\mathbf{Z})) = 1 - e(\mathbf{Z}). \tag{5.8}$$

La variable aléatoire $W$ est une variable dichotomique, alors :

$$\begin{aligned}
\mathbb{P}(W = 1 \mid Y_{(1)}, Y_{(0)}, e(\mathbf{Z})) &= \mathbb{E}(W = 1 \mid Y_{(1)}, Y_{(0)}, e(\mathbf{Z})) \\
&= \mathbb{E}(\mathbb{E}(W = 1 \mid Y_{(1)}, Y_{(0)}, e(\mathbf{Z}), \mathbf{Z}) \mid Y_{(1)}, Y_{(0)}, e(\mathbf{Z})) \\
&= \mathbb{E}(\mathbb{E}(W = 1 \mid Y_{(1)}, Y_{(0)}, \mathbf{Z}) \mid Y_{(1)}, Y_{(0)}, e(\mathbf{Z})) \\
&= \mathbb{E}(\mathbb{E}(W = 1 \mid \mathbf{Z}) \mid Y_{(1)}, Y_{(0)}, e(\mathbf{Z})) \\
&= \mathbb{E}(e(\mathbf{Z}) \mid Y_{(1)}, Y_{(0)}, e(\mathbf{Z})) \\
&= e(\mathbf{Z}).
\end{aligned}$$

Par conséquent :

$$\mathbb{P}(W = 1 \mid Y_{(1)}, Y_{(0)}, e(\mathbf{Z})) = e(\mathbf{Z}) \text{ et } \mathbb{P}(W = 0 \mid Y_{(1)}, Y_{(0)}, e(\mathbf{Z})) = 1 - e(\mathbf{Z}). \tag{5.9}$$

Par (5.8) et (5.9) nous avons :

$$\mathbb{P}(W = w \mid Y_{(1)}, Y_{(0)}, e(\mathbf{Z})) = \mathbb{P}(W = w \mid e(\mathbf{Z})), w \in \{0, 1\},$$

alors l'affectation du traitement est ignorable étant donné le score de propension $e(\mathbf{Z})$. □

La démonstration du théorème est également donnée dans Rosenbaum et Rubin (1983) et Imbens (2004). La première et la deuxième partie de l'équation (5.7) désignent respectivement les hypothèses d'ignorabilité et de positivité. Le théorème 2.4.1 du chapitre 2 stipule que sous les hypothèses d'gnorabilité forte et d'absence d'effets de diffusion du traitement (SUTVA), il est possible d'identifier l'effet causal moyen à partir des variables observables. Rosenbaum et Rubin (1983) ont montré également que sous les deux hypothèses, il est aussi possible d'identifier l'effet causal moyen en considérant le score de propension $e(\mathbf{Z})$ au lieu de l'ensemble des covariables $\mathbf{Z}$, et que cela est suffisant pour éliminer le biais induit par les variables confondantes.

Théorème 5.1.3. *Sous les hypothèses :*

1. *Ignorabilité forte du traitement étant donné le score de propension :*

$$(Y_{(1)}, Y_{(0)}) \perp\!\!\!\perp W \mid e(\mathbf{Z}) \quad et \quad 0 < \mathbb{P}(W = 1 \mid e(\mathbf{Z})) < 1;$$

2. *Absence d'effets de diffusion du traitement (SUTVA) ;*

*l'effet causal moyen est identifiable, et on a :*

$$\tau = \mathbb{E}(Y_{(1)} - Y_{(0)}) = \mathbb{E}_{e(\mathbf{Z})} \left[ \mathbb{E}(Y \mid W = 1, e(\mathbf{Z})) - \mathbb{E}(Y \mid W = 0, e(\mathbf{Z})) \right]. \quad (5.10)$$

Preuve. L'effet causal moyen peut s'écrire :

$$\begin{aligned}
\tau &= \mathbb{E}(Y_{(1)} - Y_{(0)}) \\
&= \mathbb{E}_{e(\mathbf{Z})} \left( \mathbb{E}(Y_{(1)} \mid e(\mathbf{Z})) - \mathbb{E}(Y_{(0)} \mid e(\mathbf{Z})) \right) \\
&= \mathbb{E}_{e(\mathbf{Z})} \left( \mathbb{E}(Y_{(1)} \mid W = 1, e(\mathbf{Z})) - \mathbb{E}(Y_{(0)} \mid W = 0, e(\mathbf{Z})) \right) \\
&= \mathbb{E}_{e(\mathbf{Z})} \left( \mathbb{E}(Y \mid W = 1, e(\mathbf{Z})) - \mathbb{E}(Y \mid W = 0, e(\mathbf{Z})) \right).
\end{aligned}$$

La dernière expression ne contient aucune quantité contrefactuelle, l'effet causal moyen est donc identifiable. □

L'hypothèse de positivité ($0 < \mathbb{P}(W = 1 \mid e(\mathbf{Z})) < 1$) intervient également dans l'identification de l'effet causal moyen. En effet, dans le cas où elle n'est pas satisfaite pour $e(\mathbf{Z}) = e(\boldsymbol{z})$ ($\mathbb{P}(W = 1 \mid e(\mathbf{Z})) = 0$ ou $\mathbb{P}(W = 1 \mid e(\mathbf{Z})) = 1$), il devient impossible d'estimer à la fois les deux quantités $\mathbb{E}(Y \mid W = 1, e(\mathbf{Z}) = e(\boldsymbol{z}))$ et $\mathbb{E}(Y \mid W = 0, e(\mathbf{Z}) = e(\boldsymbol{z}))$.

En pratique, le score de propension est rarement connu. Il est donc nécessaire de l'estimer au préalable à partir des données avant de procéder à l'estimation de l'effet causal moyen.

## 5.2 Estimation du score de propension

Pour estimer le score de propension, on utilise généralement la régression logistique, mais d'autres méthodes peuvent être également utilisées telles que la méthode d'arbres de classification et de régression (Luellen *et al.*, 2005; Lee *et al.*, 2009; Setoguchi *et al.*, 2008; Westreich *et al.*, 2010), le bagging (Luellen *et al.*, 2005; Lee *et al.*, 2009), le boosting (McCaffrey *et al.*, 2004; Westreich *et al.*, 2010; Lee *et al.*, 2009) et les forêts aléatoires (Lee *et al.*, 2009). Dans notre étude de cas, nous allons utiliser la régression logistique pour estimer le score de propension. Néanmoins, nous avons décidé de présenter aussi les autres méthodes, en raison de l'intérêt que portent de plus en plus les chercheurs à ces différentes techniques.

### 5.2.1 La régression logistique

La régression logistique est la méthode la plus couramment utilisée pour estimer les scores de propension. Il s'agit d'un modèle paramétrique utilisé pour prédire le score de propension $e(\mathbf{Z})$ à partir des observations de $W$ et du vecteur des covariables $\mathbf{Z}$, $\mathbf{Z}^t = (Z_1, Z_2, \ldots, Z_p)$ où $p$ désigne le nombre de covariables (le « $t$ » est pour la transposée). Un modèle linéaire ne permet pas d'expliquer le lien entre $e(\mathbf{Z})$ et $\mathbf{Z}$ lorsque ce dernier contient des covariables continues prenant des valeurs

dans $\mathbb{R}$, car $e(\mathbf{Z})$ est une probabilité et elle prend ses valeurs dans $[0,1]$. Pour pouvoir modéliser la relation entre $e(\mathbf{Z})$ et $\mathbf{Z}$, on utilise une fonction particulière appelée fonction logit et qui est définie comme suit :

$$\text{logit}(e(\mathbf{Z})) = \log\left(\frac{e(\mathbf{Z})}{1-e(\mathbf{Z})}\right).$$

Contrairement à $e(\mathbf{Z})$, $\text{logit}(e(\mathbf{Z}))$ prend ses valeurs dans $\mathbb{R}$. Il est maintenant possible d'exprimer la relation entre $e(\mathbf{Z})$ et $\mathbf{Z}$ par un modèle nommé modèle logit qui s'écrit sous la forme :

$$\text{logit}(e(\mathbf{Z})) = \log\left(\frac{e(\mathbf{Z})}{1-e(\mathbf{Z})}\right) = \boldsymbol{\beta}^t\tilde{\mathbf{Z}}, \tag{5.11}$$

où $\tilde{\mathbf{Z}}^t = (1, \mathbf{Z}^t) = (1, Z_1, Z_2, \ldots, Z_p)$, $\boldsymbol{\beta}^t = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)$ est un vecteur $(p+1) \times 1$ de paramètres inconnus à estimer. Le rapport $e(\mathbf{Z})/(1-e(\mathbf{Z}))$ est la cote de la régression logistique : il est le rapport de la probabilité d'être traité et la probabilité de ne pas recevoir le traitement. Le modèle peut également s'écrire sous une autre forme, en effet : d'après (5.11) nous avons :

$$\frac{e(\mathbf{Z})}{1-e(\mathbf{Z})} = \exp\left\{\boldsymbol{\beta}^t\tilde{\mathbf{Z}}\right\}.$$

Cela signifie que :

$$e(\mathbf{Z})\left(1+\exp\left\{\boldsymbol{\beta}^t\tilde{\mathbf{Z}}\right\}\right) = \exp\left\{\boldsymbol{\beta}^t\tilde{\mathbf{Z}}\right\}.$$

Par conséquent :

$$e(\mathbf{Z}) = \mathbb{P}(W=1 \mid \mathbf{Z}) = \frac{\exp\left\{\boldsymbol{\beta}^t\tilde{\mathbf{Z}}\right\}}{1+\exp\left\{\boldsymbol{\beta}^t\tilde{\mathbf{Z}}\right\}}. \tag{5.12}$$

#### 5.2.1.1 Estimation des paramètres du modèle

L'idée est d'estimer les paramètres du modèle $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ à partir des observations $(W_i, \mathbf{Z}_i) = (W_i, Z_{i1}, Z_{i2}, \ldots, Z_{ip})$, $i = 1, \ldots, n$, où $n$ est le nombre d'individus. En premier lieu, il faut déterminer la loi de $\mathbb{P}(W = w \mid \mathbf{Z} = \boldsymbol{z})$. Il est clair que cette loi est une loi Bernoulli, de paramètre $e(\boldsymbol{z})$, nous avons donc :

$$\mathbb{P}(W = w \mid \mathbf{Z} = \boldsymbol{z}) = e(\boldsymbol{z})^w(1-e(\boldsymbol{z}))^{1-w}. \tag{5.13}$$

L'une des méthodes les plus utilisées pour estimer les paramètres d'une loi est la méthode du maximum de vraisemblance. La fonction de vraisemblance liée au modèle que nous venons de présenter est donnée par :

$$\begin{aligned} L(w_i, \boldsymbol{z}_i, \boldsymbol{\beta}) &= \prod_{i=1}^{n} e(\boldsymbol{z}_i)^{w_i}(1 - e(\boldsymbol{z}_i))^{1-w_i} \\ &= \prod_{i=1}^{n} \left( \frac{\exp\{\boldsymbol{\beta}^t \tilde{\boldsymbol{z}}_i\}}{1 + \exp\{\boldsymbol{\beta}^t \tilde{\boldsymbol{z}}_i\}} \right)^{w_i} \left( 1 - \frac{\exp\{\boldsymbol{\beta}^t \tilde{\boldsymbol{z}}_i\}}{1 + \exp\{\boldsymbol{\beta}^t \tilde{\boldsymbol{z}}_i\}} \right)^{1-w_i} \\ &= \prod_{i=1}^{n} \left( \frac{\exp\{\boldsymbol{\beta}^t \tilde{\boldsymbol{z}}_i\}}{1 + \exp\{\boldsymbol{\beta}^t \tilde{\boldsymbol{z}}_i\}} \right)^{w_i} \left( \frac{1}{1 + \exp\{\boldsymbol{\beta}^t \tilde{\boldsymbol{z}}_i\}} \right)^{1-w_i}. \end{aligned}$$

Pour simplifier les calculs, il est préférable de travailler avec le logarithme de la fonction de vraisemblance qui s'écrit comme suit :

$$\begin{aligned} \log L(w_i, \boldsymbol{z}_i, \boldsymbol{\beta}) &= \log \left( \prod_{i=1}^{n} \left( \frac{\exp\{\boldsymbol{\beta}^t \tilde{\boldsymbol{z}}_i\}}{1 + \exp\{\boldsymbol{\beta}^t \tilde{\boldsymbol{z}}_i\}} \right)^{w_i} \left( \frac{1}{1 + \exp\{\boldsymbol{\beta}^t \tilde{\boldsymbol{z}}_i\}} \right)^{1-w_i} \right) \\ &= \sum_{i=1}^{n} \left( w_i \log \left( \frac{\exp\{\boldsymbol{\beta}^t \tilde{\boldsymbol{z}}_i\}}{1 + \exp\{\boldsymbol{\beta}^t \tilde{\boldsymbol{z}}_i\}} \right) + (1 - w_i) \log \left( \frac{1}{1 + \exp\{\boldsymbol{\beta}^t \tilde{\boldsymbol{z}}_i\}} \right) \right) \\ &= \sum_{i=1}^{n} \left( w_i \boldsymbol{\beta}^t \tilde{\boldsymbol{z}}_i - \log\left(1 + \exp\left\{\boldsymbol{\beta}^t \tilde{\boldsymbol{z}}_i\right\}\right) \right). \end{aligned}$$

Pour trouver les estimateurs du maximum de vraisemblance $\hat{\boldsymbol{\beta}}$, on maximise le logarithme de la fonction de vraisemblance par rapport à $\boldsymbol{\beta}$. Le vecteur $\hat{\boldsymbol{\beta}}$ est la valeur de $\boldsymbol{\beta}$ qui annule les dérivées partielles de $\log L(w_i, \boldsymbol{z}_i, \boldsymbol{\beta})$ par rapport à $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$. Il en résulte un système à $p+1$ d'équations :

$$\begin{aligned} \frac{\partial \log L(w_i, \boldsymbol{z}_i, \boldsymbol{\beta})}{\partial \beta_0} &= \sum_{i=1}^{n} w_i - \frac{\exp\{\boldsymbol{\beta}^t \tilde{\boldsymbol{z}}_i\}}{1 + \exp\{\boldsymbol{\beta}^t \tilde{\boldsymbol{z}}_i\}}, \\ \frac{\partial \log L(w_i, \boldsymbol{z}_i, \boldsymbol{\beta})}{\partial \beta_j} &= \sum_{i=1}^{n} z_{ij} \left( w_i - \frac{\exp\{\boldsymbol{\beta}^t \tilde{\boldsymbol{z}}_i\}}{1 + \exp\{\boldsymbol{\beta}^t \tilde{\boldsymbol{z}}_i\}} \right), \quad \text{pour } j = 1, \ldots, p. \end{aligned} \tag{5.14}$$

Le système d'équations (5.14) n'admet pas de solution analytique. Il faut donc faire appel à des méthodes numériques itératives.

### 5.2.2 Arbres de classification et de régression

Cette technique appelée également « CART » est attribuée à Breiman *et al.* (1984). Il s'agit d'une méthode non paramétrique ; on n'a donc pas besoin d'hypothèses

sur la loi des variables aléatoires intervenant dans l'étude. Cette méthode permet de construire ce que l'on appelle un « arbre de décision binaire » qui a pour but de prédire une variable expliquée $Y$ qui prend des valeurs dans $\mathbb{Y}$ à partir d'un ensemble de variables explicatives $\mathbf{X} = (X_1, \ldots, X_p)$ prenant des valeurs dans $\mathbb{X}$. Lorsque $Y$ est continue, on est dans le cas d'une régression et l'arbre obtenu est connu sous le nom d'arbre de régression. Alors que si $Y$ est une variable catégorielle à $m$ classes (modalités) $\mathbb{Y} = \{1, 2, \ldots, m\}$, on est dans un problème de classification et on construit ce que l'on appelle « un arbre de classification ». Un arbre de classification est un type particulier de classifieur, ce dernier désigne une fonction $h : \mathbb{X} \to \mathbb{Y}$ qui associe à chaque observation $\boldsymbol{x}$ une classe $y$ dans $\mathbb{Y}$. En d'autres termes, si l'on observe $\boldsymbol{x}$ la prédiction de $Y$ est donnée par $y = h(\boldsymbol{x})$.

Dans notre cas, le but est d'estimer le score de propension $e(\mathbf{Z})$, et cela passe par la construction d'un arbre de classification avec $W$ comme variable expliquée et l'ensemble des covariables $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_p)$ comme variables explicatives. Ce choix est motivé par le fait que la variable $W$ est une variable dichotomique, ce qui signifie qu'on est dans un cas de classification. Soit $S = \{(w_1, \boldsymbol{z}_1), (w_2, \boldsymbol{z}_2), \ldots, (w_n, \boldsymbol{z}_n)\}$ $n$ réalisations du couple $(W, \mathbf{Z}) \in \{0, 1\} \times \mathbb{Z}$, où $\mathbb{Z}$ désigne le support de $\mathbf{Z}$. La construction d'un arbre de classification consiste en un partitionnement de l'ensemble $\mathbb{Z}$ en $k$ partitions. La partition s'effectue à l'aide de l'ensemble des observations $S$ représenté par un nœud $a_0$ appelé « racine » et qui constitue le point de départ de l'arbre.

La construction d'un arbre de classification s'effectue à l'aide d'une procédure itérative. On commence par diviser le nœud $a_0$ qui correspond à l'ensemble de l'échantillon en deux nœuds (enfants) $a_g$ et $a_d$ (« $g$ » et « $d$ » sont pour gauche et droit respectivement). Cette division est déterminée par une règle de décision basée sur les deux types de questions suivants :

- est-ce que $z_{ij} \leq c_j$ ? où $c_j \in \mathbb{R}$, si $Z_j$ est quantitative ;

– est-ce que $z_{ij} \in M_j$, où $M_j$ est un sous-ensemble de l'ensemble de modalités de $Z_j$, si $Z_j$ est qualitative.

Une question ne fait intervenir qu'une seule variable, le couple $(w_i, \boldsymbol{z}_i)$ est classé dans l'une des deux classes selon que l'on répond par oui ou par non à la question. On répète le processus pour les deux nœuds $a_1$ et $a_2$ ainsi que pour les nœuds qui en résultent et on arrête lorsque les nœuds atteignent une taille minimale, ou quand une règle d'arrêt prédéfinie le préconise. À chaque division, le choix de la variable de partage est déterminé par un critère nommé « mesure d'impureté ». Considérons la partition du nœud $a_0$ en deux nœuds $a_g$ et $a_d$. On note :

– $N(a)$ le nombre d'observations dans le nœud $a$ ;

– $N_1(a)$ le nombre d'individus traités dans le nœud $a$ ;

– $N_0(a)$ le nombre d'individus non traités dans le nœud $a$ ;

Soit $p_1(a) = N_1(a)/N(a)$ et $p_0(a) = N_0(a)/N(a)$ la proportion des traités et la proportion des non-traités dans le nœud $a$ respectivement. Un nœud contenant que des traités ou que des non traités est dit « pur », dans le cas contraire il est dit « impur ». Pour mesurer l'impureté d'un nœud $a$ notée $\mathrm{Imp}(a)$, on utilise généralement l'un des deux indices suivants :

– l'indice de Gini : $\mathrm{Imp}(a) = 1 - (p_1^2(a) + p_0^2(a))$ ;

– l'entropie de Shannon : $\mathrm{Imp}(a) = -(p_1(a)\log\{p_1(a)\} + p_0(a)\log\{p_0(a)\})$.

Parmi toutes les divisions possibles, on retient celle qui maximise la quantité suivante :

$$\Delta\mathrm{Imp}(a_0) = \mathrm{Imp}(a_0) - p_g\mathrm{Imp}(a_g) - p_d\mathrm{Imp}(a_d),$$

où $p_g$ et $p_d$ désignent les proportions d'observations qui vont respectivement vers $a_g$ et $a_d$. La même règle est utilisée pour la division des nœuds $a_g$ et $a_d$ ainsi que pour toutes les autres divisions qui suivent. À l'issue du processus de partition, on obtient un nombre $N$ de nœuds terminaux (sans enfant) qu'on note $D_1, D_2, \ldots, D_N$, et chaque nœud est étiqueté par sa classe majoritaire, ce qui per-

met d'effectuer la prévision pour la variable $W$. Cependant, notre objectif avec cette méthode est d'estimer le score de propension $e(\mathbf{Z})$ à partir du vecteur des covariables $\mathbf{Z}$. Les nœuds terminaux $D_1, D_2, \ldots, D_N$ sont par définition des strates de scores de propension. Les individus d'une même strate ont le même score de propension estimé qui correspond à la proportion des traités dans la strate. Et au sein de chaque strate, la distribution de chaque covariable est similaire entre les deux groupes de traitement. Cette méthode se généralise facilement au cas où la variable dépendante ($W$ dans notre cas) a plus de deux modalités, le cas général est donné en détail dans Breiman *et al.* (1984) et Hastie *et al.* (2009).

### 5.2.3 Techniques de bagging et de boosting

Ces méthodes sont applicables à toute méthode de modélisation mais elles trouvent leur utilité essentiellement dans les modèles qui présentent des résultats instables, notamment les arbres de classification et de régression.

La technique du bagging est introduite par Breiman (1996). Elle s'appuie sur la méthode de boostrap pour perfectionner les arbres de classification et de régression. Cette méthode peut être utilisée pour améliorer un arbre de classification dans l'estimation des scores de propension (Lee *et al.*, 2009; Luellen *et al.*, 2005). Soit $S = \{(w_1, \boldsymbol{z}_1), (w_2, \boldsymbol{z}_2), \ldots, (w_n, \boldsymbol{z}_n)\}$ $n$ réalisations du couple $(W, \mathbf{Z}) \in \{0, 1\} \times \mathbb{Z}$. Le point de départ du bagging consiste à ré-échantilloner $T$ fois l'échantillon initial $S$ par la méthode de bootstrap qui consiste à faire des tirages sans remise à partir de $S$ jusqu'à l'obtention de $T$ nouveaux échantillons $S_1, \ldots, S_T$. À partir de chaque nouvel échantillon $S_t$, on construit un classifieur $h_t$ de type arbre de classification, qui associe à chaque valeur $\boldsymbol{z}$ de $\mathbf{Z}$ une valeur prédite $w = h_t(\boldsymbol{z})$ de $W$ et qui permet d'estimer les scores de propension comme cela a été expliqué précédemment. On obtient donc $T$ arbres de classification. On commence d'abord par estimer le score de propension lié à chaque strate dans

chaque arbre, ensuite on agrège les résultats en calculant la moyenne des scores de propension sur tous les arbres.

La technique du boosting s'articule autour du concept de « classifieur faible », qui désigne un classifieur dont la probabilité de prédire correctement une variable qualitative est un peu meilleure que celle d'un choix aléatoire. L'algorithme du boosting le plus utilisé pour prédire une variable qualitative est connu sous le nom de « AdaBoost ». Cet algorithme proposé par Freund et Schapire (1997) sert à combiner plusieurs classifieurs faibles pour obtenir un classifieur fort dont la probabilité de prédire correctement la variable est proche de 1. On suppose toujours disposer d'un échantillon observé $S = \{(w_1, \boldsymbol{z}_1), (w_2, \boldsymbol{z}_2), \ldots, (w_n, \boldsymbol{z}_n)\}$. Contrairement au bagging, le boosting utilise l'échantillon $S$ au complet pour construire les classifieurs. On commence par construire un premier classifieur faible $h_1 : \mathbb{Z} \to \mathbb{W} = \{0, 1\}$ que l'on utilise ensuite pour construire un deuxième classifieur $h_2$ de manière adaptative. Ce dernier est à son tour utilisé pour construire un autre classifieur et on continue le processus jusqu'à l'obtention d'un nombre $T$ de classifieurs. Comme pour le bagging, le boosting permet d'obtenir des strates de scores de propension en calculant la moyenne des scores de propension sur tous les classifieurs.

### 5.2.4 Forêts aléatoires

Cette méthode a été introduite par Breiman (2001), comme une amélioration de la technique du bagging. L'idée est d'introduire la randomisation dans le choix des variables dans les modèles CART, afin rendre les arbres plus indépendants. Considérons un échantillon observé $S = \{(w_1, \boldsymbol{z}_1), (w_2, \boldsymbol{z}_2), \ldots, (w_n, \boldsymbol{z}_n)\}$. Comme pour le bagging, cette méthode consiste à construire un ensemble de classifieurs de type arbres de classifications $\{h_1, h_2, \ldots, h_T\}$ où chaque arbre $h_t$ est obtenu à partir d'un échantillon bootstrap $S_t$ issu de $S$. Le point sur lequel les deux

méthodes diffèrent est la façon de choisir les variables de partitionnement lors de la construction de chaque arbre de décision. En effet, pour le bagging, chaque variable de partitionnement est choisie parmi toutes les autres variables, selon une fonction d'homogénéité. Alors que pour les forêts aléatoires, lors de chaque division d'un nœud, on génère aléatoirement un ensemble de $q$ variables parmi les $p$ variables existantes. On utilise ensuite une fonction d'homogénéité pour choisir la variable de partitionnement parmi les $q$ variables. La construction d'une forêt aléatoire s'effectue comme suit : Pour $t = 1, \ldots, T$ :

- Tirer un échantillon bootstrap $S_t$ de l'échantillon initial $S$.
- Construire un classifieur $h_t$ de type arbre de décision, tel qu'à chaque division d'un nœud, on sélectionne aléatoirement un ensemble de $q$ variables parmi les $p$ variables exogènes $Z_1, \ldots, Z_p$, ensuite on choisit la variable de partitionnement parmi les $q$ variables sélectionnées.

À l'issue de ce processus, on obtient $T$ classifieurs. Comme pour le bagging et le boosting, il est possible d'obtenir les scores de propension à l'aide des forêts aléatoires en calculant la moyenne des scores de propension sur tous les arbres.

Excepté la régression logistique, les autres techniques que nous venons de présenter sont des méthodes non paramétriques basées essentiellement sur les arbres. L'avantage des méthodes basées sur les arbres est qu'elles permettent de sélectionner automatiquement les variables ainsi que les termes d'interaction à inclure dans le modèle, ce qui n'est pas le cas de la régression logistique (Luellen *et al.*, 2005). De nombreux auteurs ont tenté de comparer les performances de ces différentes méthodes dans l'estimation du score de propension. Lee *et al.* (2009) ont examiné l'ensemble des méthodes que nous venons de présenter et ils ont conclu qu'en présence de non-linéarité ou d'interactions complexes, les méthodes basées sur les arbres donnent de meilleurs résultats en matière de balance des covariables dans les groupes de traitement comparativement à la régression logistique qui ne

permet pas de détecter des interactions dans le modèle. Toutefois, un bon choix de covariables dans un modèle de régression logistique permet d'obtenir une bonne estimation du score de propension.

Dans ce travail, nous avons opté pour la méthode graphique pour déterminer le vecteur des covariables à inclure dans notre modèle ; nous allons inclure toutes les variables confondantes dans l'estimation du score de propension. Selon Rubin et Thomas (1996), il faut également inclure les variables confondantes potentielles qui sont des variables affectant le résultat mais pas le traitement.

## 5.3 Estimation de l'effet du traitement

Après l'estimation des scores de propension, quatre méthodes peuvent être utilisées pour l'estimation de l'effet causal moyen : appariement sur le score de propension, stratification sur le score de propension, pondération inverse et ajustement sur le score de propension.

### 5.3.1 Appariement sur le score de propension

L'appariement sur le score de propension est l'une des méthodes les plus utilisées pour estimer l'effet causal. Elle consiste généralement à créer des paires d'individus, un traité et un non-traité qui ont un même score de propension (Rosenbaum et Rubin, 1985). Ainsi donc, au lieu d'effectuer un appariement sur les caractéristiques observables, il suffit d'apparier les individus sur la base de leurs scores de propensions, qui résument l'ensemble de leurs caractéristiques. Néanmoins, les principes des deux méthodes restent fondamentalement les mêmes. En effet, il existe aussi plusieurs façons de faire l'appariement sur le score de propension. On distingue entre l'appariement avec remise et l'appariement sans remise, et on a également le choix entre l'appariement par la méthode du plus proche voisin et l'appariement optimal (voir section 2.5.3). Le premier consiste à sélectionner

aléatoirement un individu traité, et lui imputer ensuite l'individu non traité le plus proche de lui en termes de score de propension, et ce malgré le fait que le non-traité sélectionné pourrait être plus proche d'un autre traité. Alors que dans l'appariement optimal, on prend l'appariement qui minimise les différences des scores de propension entre les traités et les non-traités dans les paires formées. Gu et Rosenbaum (1993) ont montré que les deux dernières méthodes donnent des résultats similaires en matière de balance entre les groupes appariés créés.

Le point sur lequel l'appariement sur le score de propension diffère de l'appariement sur les covariables observables est le choix de la mesure de distance. Pour le dernier, on utilise généralement la distance de Mahalanobis (voir section 2.5). Concernant l'appariement sur le score de propension, usuellement on utilise deux mesures pour déterminer la distance $D_{ij}$ entre deux individus $i$ et $j$ :

1. valeur absolue de la différence entre les scores de propension

$$D_{ij} = \mid e(\mathbf{Z}_i) - e(\mathbf{Z}_j) \mid ;$$

2. valeur absolue de la différence entre les logits des scores de propension

$$D_{ij} = \mid \text{logit}(e(\mathbf{Z}_i)) - \text{logit}(e(\mathbf{Z}_j)) \mid .$$

Mais, lorsque l'individu à apparier se situe à une distance éloignée de son plus proche voisin, l'utilisation de l'une des mesures citées ci-dessus peut conduire à un appariement de mauvaise qualité. Pour remédier à ce problème, on fixe un caliper, lequel correspond à la distance maximale tolérée entre deux membres d'une paire (voir section 2.5.3). Rosenbaum et Rubin (1985) ont examiné le choix du caliper lorsque la distance basée sur le logit du score de propension est utilisée. Ils ont conclu que l'utilisation d'un caliper égal à 0.2 fois l'écart-type du logit du score de propension permet de réduire de 98% le biais lors de l'estimation de l'effet causal moyen. De façon générale, ils suggèrent d'utiliser un caliper égal à 0.25 fois l'écart-type du logit du score de propension. Ces résultats sont une généralisation de ceux

obtenus par Cochran et Rubin (1973) dans leur étude sur le choix d'un caliper dans le cas d'un appariement sur une seule variable confondante. Austin (2011) a montré qu'un caliper de 0.2 à 0.55 fois l'écart-type du logit du score de propension minimise l'erreur quadratique moyenne (MSE) si au moins une des covariables est continue. Un caliper de 0.8 fois l'écart-type du logit du score de propension donne un résultat similaire lorsque toutes les covariables sont dichotomiques.

Une fois que l'appariement est réalisé, on procède à l'estimation de l'effet causal moyen à partir des deux groupes appariés, en utilisant l'une des méthodes présentées dans la section 2.5.3.

### 5.3.2 Stratification sur le score de propension

La stratification sur le score de propension consiste à créer des strates homogènes renfermant des individus de même score de propension. Il existe plusieurs approches utilisant la stratification sur le score de propension pour estimer l'effet causal moyen (Rosenbaum et Rubin, 1983, 1984; Imbens, 2004). L'approche la plus utilisée a été présentée en détail par Lunceford et Davidian (2004) et elle se résume comme suit :

1. on commence par calculer le score de propension estimé $\hat{e}(\boldsymbol{z}_i)$ de chaque individu $i$, $(i = 1, \ldots, n)$, on obtient donc $n$ valeurs du score de propension $\hat{e}(\boldsymbol{z}) = \{\hat{e}(\boldsymbol{z}_1), \hat{e}(\boldsymbol{z}_2), \ldots, \hat{e}(\boldsymbol{z}_n)\}$ ;
2. on découpe l'ensemble des valeurs du score de propension en $K$ strates (blocs) $S_1, S_2, \ldots, S_K$ à l'aide des quantiles empiriques du score de propension $\hat{q}_j, j = 1, \ldots, K-1$, ainsi, on a pour $j = 1, \ldots, K$, $\hat{e}(\boldsymbol{z}_i) \in S_j$ si et seulement si $\hat{e}(\boldsymbol{z}_i) \in ]\hat{q}_{j-1}, \hat{q}_j]$, où $\hat{q}_0 = 0$ et $\hat{q}_K = 1$ ;
3. à l'intérieur de chaque strate $S_j$, on estime l'effet causal moyen $\tau_j$ qui correspond à la différence des moyennes observées de la variable de réponse $Y$

entre les deux groupes de traitement :

$$\hat{\tau}_j = \frac{1}{n_{(1)j}} \sum_{i=1}^{n} w_i y_i \mathbf{1}_{\hat{e}(z_i) \in S_j} - \frac{1}{n_{(0)j}} \sum_{i=1}^{n} (1 - w_i) y_i \mathbf{1}_{\hat{e}(z_i) \in S_j},$$

où $n_{(1)j}$ et $n_{(0)j}$ sont respectivement le nombre de traités et le nombre de non-traités dans la strate $S_j$ ;

4. en dernier lieu, on procède à l'estimation de l'effet causal total par la moyenne pondérée des estimateurs par strate :

$$\hat{\tau}_{\text{strat-sp}} = \sum_{j=1}^{K} \frac{n_j}{n} \hat{\tau}_j;$$

où $n_j$ représente le nombre d'individus dans la strate $S_j$.

Rosenbaum et Rubin (1984) suggèrent d'opter pour 5 strates de même taille qu'il faudrait déterminer à partir des quintiles du score de propension estimé. Ils ont montré que cela permet de réduire de 90% le biais lié aux covariables de confusion dans le cas de l'estimation de l'effet d'un traitement. Les deux auteurs ont ainsi étendu les résultats obtenus par Cochran (1968) (voir section 2.5.1) à la théorie du score de propension. Néanmoins, pour les échantillons de grande taille, le choix d'un grand nombre de strates (jusqu'à 20 strates) permet de réduire davantage le biais (Lunceford et Davidian, 2004). D'autres études sont nécessaires pour pouvoir déterminer le choix optimal du nombre de strates.

L'estimateur que nous venons de présenter nécessite une bonne spécification du modèle de score de propension, notamment veiller à ce que la propriété de balance dans les strates soit vérifiée. En effet, si cette propriété n'est pas vérifiée, cela entraîne un biais dans l'estimation de l'effet causal moyen car des facteurs de confusion pourraient subsister même après la construction des strates. Pour atténuer l'effet de ces facteurs de confusion résiduels, Lunceford et Davidian (2004) proposent d'utiliser une méthode combinant stratification et régression pour estimer l'effet causal moyen. Le principe de cette méthode consiste à construire des strates de score de propension en appliquant les étapes 1 et 2 de la démarche

décrite ci-haut, ensuite, on estime à l'intérieur de chaque strate $S_j$ l'effet causal moyen $\hat{\tau}_{\text{reg},j}$ à l'aide d'un modèle de régression reliant la variable de réponse à $Y$ à la variable du traitement en ajoutant certaines covariables (voir section 2.5.2). L'estimateur pour l'effet causal moyen est obtenu en calculant la moyenne pondérée des estimateurs par strates :

$$\hat{\tau}_{\text{strat-reg}} = \frac{1}{K}\sum_{j=1}^{K}\hat{\tau}_{\text{reg},j}.$$

Lunceford et Davidian (2004) ont montré $\hat{\tau}_{\text{strat-reg}}$ est un estimateur convergent pour $\tau$ sous condition que le modèle de régression de chaque strate soit bien spécifié.

### 5.3.3 Pondération inverse avec le score de propension

La pondération inverse basée sur les scores de propension est la méthode la moins utilisée dans l'estimation de l'effet causal moyen. Elle consiste à attribuer aux individus des poids ($p_i$) basés sur leurs scores de propension estimés pour mieux représenter la population. La forme de pondération la plus utilisée est donnée par la formule suivante :

$$p_i = \frac{w_i}{\hat{e}(\boldsymbol{z}_i)} + \frac{1-w_i}{1-\hat{e}(\boldsymbol{z}_i)}. \tag{5.15}$$

L'équation (5.15) montre que le poids alloué à un individu traité est égal à $1/\hat{e}(\boldsymbol{z})$ et celui d'un individu non traité est égal à $1/(1-\hat{e}(\boldsymbol{z}))$.

Lunceford et Davidian (2004) ont présenté plusieurs estimateurs pour l'effet causal moyen par la méthode de pondération inverse avec le score de propension. Le plus simple de ces estimateurs a été initialement proposé par Rosenbaum (1987) et s'écrit comme suit :

$$\hat{\tau}_{\text{PI1}} = \frac{1}{n}\sum_{i=1}^{n}\frac{w_i y_i}{\hat{e}(\boldsymbol{z}_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-w_i)y_i}{1-\hat{e}(\boldsymbol{z}_i)}.$$

Le problème avec cet estimateur est que la somme des poids dans chaque groupe de traitement ne donne pas 1. Mais nous savons que :

$$\mathbb{E}\left(\frac{W}{e(\mathbf{Z})}\right) = \mathbb{E}\left(\frac{\mathbb{E}(W \mid \mathbf{Z})}{e(\mathbf{Z})}\right) = 1,$$

et

$$\mathbb{E}\left(\frac{1-W}{e(\mathbf{Z})}\right) = \mathbb{E}\left(\frac{\mathbb{E}(1-W \mid \mathbf{Z})}{1-e(\mathbf{Z})}\right) = 1.$$

Par conséquent, un nouvel estimateur est donné par :

$$\hat{\tau}_{\text{PI2}} = \left(\sum_{i=1}^{n} \frac{w_i}{\hat{e}(\boldsymbol{z}_i)}\right)^{-1} \sum_{i=1}^{n} \frac{w_i y_i}{\hat{e}(\boldsymbol{z}_i)} - \left(\sum_{i=1}^{n} \frac{1-w_i}{1-\hat{e}(\boldsymbol{z}_i)}\right)^{-1} \sum_{i=1}^{n} \frac{(1-w_i)y_i}{1-\hat{e}(\boldsymbol{z}_i)}.$$

Il est également possible d'estimer l'effet causal moyen à l'aide d'un modèle de régression simple qui utilise des pondérations.

### 5.3.4 Ajustement sur les scores de propension

Ici, il s'agit de faire une régression avec comme variable dépendante la variable de réponse $Y$ et comme variables indépendantes, la variable liée au traitement $W$ et le score de propension $e(\mathbf{Z})$. Le choix du modèle dépend de la nature de la variable $Y$, si elle est continue, on opte pour le modèle linéaire suivant :

$$Y_i = \alpha + \tau w_i + \beta \hat{e}(\boldsymbol{z}_i) + \varepsilon_i, \;\; \varepsilon_i \sim \mathcal{N}(0, \sigma^2),$$

où $\tau$ représente l'effet causal moyen. Dans le cas où $Y$ est une variable dichotomique, on opte pour la régression logistique.

## 5.4 Vérification des hypothèses et des propriétés liées à l'estimation par score de propension

Nous allons maintenant présenter quelques méthodes permettant de vérifier les hypothèses et les propriétés évoquées.

### 5.4.1 Propriété de balance du score de propension

Comme nous l'avons évoqué dans la section 5.1.1, le score de propension est avant tout un score de balance. En effet, étant donné le score de propension, la distribution des covariables est indépendante de l'affectation du traitement. De ce fait, une façon de vérifier si le score de propension est adéquatement estimé est de vérifier la propriété de balance des différents sous-échantillons créés par le score de propension (échantillons appariés dans le cas d'appariement, strates dans le cas de stratification). Pour cela, il existe un critère nommé « différence standardisée, DS » qui permet de comparer les moyennes des variables continues et dichotomiques entre les deux groupes de traitement. Pour une covariable $X$ continue, cette différence est donnée par :

$$d_x = \frac{\left(\bar{x}_{(1)} - \bar{x}_{(0)}\right)}{\sqrt{\frac{S^2_{(1)} + S^2_{(0)}}{2}}},$$

où $\bar{x}_{(1)}$ et $\bar{x}_{(0)}$ désignent respectivement les moyennes de la variable $x$ dans le groupe des traités et le groupe des non-traités dans l'échantillon considéré, et $S^2_{(1)}$ et $S^2_{(0)}$ sont respectivement les variances de $x$ dans le groupe des traités et le groupe des non-traités. Si $X$ est une covariable dichotomique, la différence standardisée est donnée par :

$$d_x = \frac{\left(\hat{p}_{(1)x} - \hat{p}_{(0)x}\right)}{\sqrt{\frac{\hat{p}_{(1)x}(1 - \hat{p}_{(1)x}) + \hat{p}_{(0)x}(1 - \hat{p}_{(0)x})}{2}}},$$

où $\hat{p}_{(1)x}$ et $\hat{p}_{(0)x}$ désignent les fréquences observées de la variable $x$ dans le groupe des traités et le groupe des non-traités dans l'échantillon considéré.

### 5.4.2 Vérification de positivité

Une autre hypothèse qu'on doit vérifier est l'hypothèse de positivité, ou de support commun. Il suffit juste de représenter la distribution du score de propension

dans les deux groupes de traitement, pour vérifier les zones de support commun. Lorsque la positivité n'est pas vérifiée sur tout l'échantillon, Dehejia et Wahba (1999) proposent d'utiliser un critère nommé « Min-Max » qui consiste à éliminer les traités ayant un score de propension supérieur au maximum du score de propension des non-traités, et d'éliminer également les non-traités ayant un score de propension inférieur au minimum du score de propension des traités.

[Cette page a été laissée intentionnellement blanche]

# CHAPITRE VI

# ESTIMATION DE L'EFFET DU TRAITEMENT PCC

Alors que dans le chapitre précédent nous avons présenté une analyse théorique de la technique du score de propension, ce présent chapitre est consacré à l'application de cette technique sur notre étude de cas. Nous avons deux groupes de traitement : un groupe qui a reçu le traitement standard (vitamine K et le FFP) et un autre groupe qui a reçu en plus du traitement standard le PCC (voir section 1.3). Pour vérifier l'effet de ce traitement, nous allons présenter deux cas de figure :

- premier cas : examen de l'effet causal du traitement PCC (*Group*) sur la mortalité à 6 mois (*6mo_mortality*) ;
- deuxième cas : examen de l'effet causal du traitement PCC sur l'expansion de l'hématome lié à la wICH (*expansion*).

Les analyses que nous allons présenter dans ce chapitre sont réalisées à l'aide du logiciel SAS version 9.3 voir SAS Institute Inc (2013).

## 6.1 Comparaison entre les deux groupes de traitement

Avant toute chose, nous dressons une comparaison entre les deux groupes de traitement sous la base des caractéristiques observées chez les différents patients. La deuxième et la troisième colonnes du tableau 6.1 indiquent la répartition de certaines caractéristiques des sujets, entre le groupe traité et le groupe témoin. Pour une variable continue, les chiffres reportés dans les deux colonnes indiquent les

Tableau 6.1: Comparaison des deux groupes dans l'échantillon initial

| Variable | Groupe traité (N=49) | Groupe témoin (N=40) | p-value | DS |
|---|---|---|---|---|
| *age* | 70.71(14.93) | 72.95(11.95) | 0.4986 | 0.1653 |
| *inr_bwh* | 2.28(0.63) | 2.20(0.83) | 0.1634 | 0.1050 |
| *volume* | 25.92(34.68) | 16.81(23.38) | 0.0224 | 0.3077 |
| *sbp* | 153.30(32.81) | 156.00(25.38) | 0.6728 | 0.0918 |
| *inr_3_to_6hr* | 1.37(0.15) | 1.65(0.59) | <.0001 | 0.6350 |
| *ffp* | 3.27(1.64) | 6.85(4.84) | <.0001 | 0.9913 |
| *valve* | 8(16.33%) | 5(12.50%) | 0.6111 | 0.1091 |
| *cmy* | 2(4.08%) | 2(5.00%) | 1.0000 | 0.0441 |
| *cancer* | 10(20.41%) | 11(27.50%) | 0.4331 | 0.1667 |
| *gender* | 18(36.73%) | 17(42.50%) | 0.5797 | 0.1181 |
| *cad* | 17(34.69%) | 14(35.00%) | 0.9759 | 0.0064 |
| *postfossa* | 8(16.33%) | 6(15.00%) | 0.8642 | 0.0365 |
| *afib* | 31(63.27%) | 21(52.50%) | 0.3053 | 0.2193 |
| *dvt_pe* | 10(20.41%) | 12(30.00%) | 0.2967 | 0.2223 |
| *evd* | 20(40.82%) | 3(7.50%) | 0.0004 | 0.8450 |
| *location* | 17(34.69%) | 16(40.00%) | 0.6950 | 0.1099 |
| *renal_failure* | 5(10.20%) | 1(2.50%) | 0.2174 | 0.3199 |
| *alcohol* | 4(8.16%) | 1(2.50%) | 0.3739 | 0.2541 |
| *Cdvt_pe* | 5(10.20%) | 8(20.00%) | 0.1930 | 0.2762 |
| *6mo_mortality* | 15(32.61%) | 14(35.00%) | 0.8150 | 0.0506 |
| *expansion* | 5(10.20%) | 5(12.50%) | 0.7487 | 0.0724 |

moyennes de la variable dans les deux groupes, suivies des écarts-type (chiffres entre parenthèses). Pour une variable binaire, les chiffres représentent les effectifs

de la modalité 1 suivis de ses fréquences. Concernant la variable *location*, initialement, elle avait 5 modalités, mais nous avons décidé de la transformer en variables binaires en regroupant les modalités 2, 3, 4 et 5 sous la modalité 0. Le but de cette manipulation est de remédier au problème de séparation quasi complète des points de données lors de l'estimation du score de propension.

Pour pouvoir comparer les deux groupes de traitement, nous avons effectué le test de comparaison de deux échantillons indépendants pour les variables continues et le test d'indépendance pour les variables catégorielles. Concernant le test de comparaison de deux échantillons indépendants, dans le cas de normalité et d'égalité des variances de la variable considérée (tests de normalité et d'égalité des variances effectués), le test réalisé est celui de Student pour la comparaison de deux moyennes. Dans le cas contraire, nous avons effectué le test de Wilcoxon pour les échantillons indépendants. Pour le test d'indépendance, nous avons opté pour le test d'indépendance du Khi-deux, sauf dans le cas où l'effectif d'une modalité dans un groupe est inférieur à 5, où nous avons effectué le test exact de Fisher. Les *p-values* des tests que nous venons d'évoquer sont reportées dans le tableau 6.1. Dans tout ce chapitre, nous optons pour un seuil de signification de 5% pour tous les tests effectués.

La dernière colonne du tableau représente les différences standardisées (DS) liées aux variables considérées. C'est une autre façon de comparer les moyennes des variables continues et binaires entre les deux groupes de traitement.

L'examen du tableau 6.1 montre que les deux groupes sont différents en matière de quantité de plasma frais congelé (*ffp*) nécessaire pour le traitement (3.27 unités pour les traités contre 6.85 unités pour les témoins), avec une p-value presque nulle. Nous constatons également que le groupe traité et le groupe témoin sont différents en ce qui concerne la valeur de l'INR mesurée entre 3 et 6 heures après l'arrivée

du patient à l'hôpital (*inr_3_to_6hr*), avec une p-value également presque nulle. D'autre part, nous constatons que la distribution de la variable *Cdvt_pe* (indicateur d'une maladie thrombo-embolique) est la même dans les deux groupes de traitement (p-value = 0.1930). Ces résultats sont cohérents avec ceux obtenus par Cai *et al.* (2014) (voir section 1.2).

L'analyse du tableau 6.1 montre également que le groupe traité par PCC se caractérise par un volume plus élevé d'hémorragie intracérébrale (*volume*) que dans le groupe qui a reçu le traitement standard (25.92 pour les traités contre 16.81). Cette différence observée est statistiquement significative (p-value = 0.0224). La fréquence de patients ayant bénéficié d'un drainage ventriculaire externe (*evd*) dans le groupe traité est supérieure à celle du groupe témoin (40.82% contre 7.50%, avec une p-value égale à 0.0004). En revanche, les deux groupes présentent des fréquences proches et élevées en termes d'antécédents de fibrillation auriculaire (*afib*).

## 6.2 Estimation des scores de propension

Nous allons maintenant procéder à l'estimation du score de propension dans nos deux études de cas. Pour cela, nous avons opté pour des modèles de régression logistique avec *Group* comme variable expliquée.

### 6.2.1 Sélection de variables

Pour notre premier cas d'analyse, nous avons choisi les variables explicatives suivantes :

$age, inr\_bwh, volume, sbp, valve, cmy, cancer, gender, cad, postfossa, afib, dvt\_pe,$
$evd, location, renal\_failure, alcohol.$

En plus de ces variables, nous avons ajouté les deux termes d'interaction $age * valve$ et $sbp * postfossa$, ainsi que le carré de la variable *volume* noté *volume_square*.

Concernant notre deuxième cas d'analyse, nous avons retenu les variables explicatives :

$$inr\_bwh, sbp, evd, renal\_failure, alcohol, cad,$$

ainsi que le terme d'interactions $inr\_bwh * cad$.

Le choix de nos variables est essentiellement déterminé par l'analyse graphique que nous avons présentée en détails dans la section 4.3.3. Pour chaque cas d'étude, nous avons sélectionné les variables confondantes (l'ensemble suffisant minimal pour l'ajustement de la confusion) comme suit :

En premier lieu, nous avons effectué un tri topologique de notre graphe de départ selon la manipulation (i). Le résultat de ce tri est présenté dans la figure B.2a qui montre huit niveaux verticaux de variables. Ce graphe est ensuite utilisé pour déterminer un ensemble suffisant minimal pour l'ajustement de la confusion dans l'estimation de l'effet causal de la variable *Group* sur *6mo_mortality*.

Considérons la figure B.2a, l'ensemble de variables {*valve*, *cmy*, *cancer*, *gender*, *cad*, *age*, *postfossa*, *afib*, *volume*, *dvt_pe*, *evd*, *location*} (nœuds colorés en bleu) doivent absolument être incluses dans notre ensemble (manipulation (iv)). Les nœuds colorés en jaune représentent les variables qui ne se trouvent pas sur un chemin entre *Group* et *6mo_mortality* (manipulation (iii)). Concernant les nœuds colorés en rouge, ils indiquent les variables à ne pas inclure dans l'ensemble recherché car elles sont toutes des descendantes de la variable du traitement *Group* (voir définition 4.3.2). L'application des manipulations (ii), (iii) et (iv) sur notre graphe trié nous donne le graphe de la figure B.2b sur lequel nous avons effectué à nouveau la manipulation (iii) qui a abouti au graphe final de la figure B.2c. Dans

ce graphe final, la variable *inr_bwh* (colorée en bleu) bloque tous les chemins porte-arrières et elle doit absolument apparaître dans l'ensemble recherché. En résumé, notre ensemble suffisant minimal pour l'ajustement de la confusion dans l'estimation de l'effet causal de *Group* sur *6mo_mortality* est donné par :

$$Z_1 = \{valve, cmy, cancer, gender, cad, age, postfossa, afib, volume, dvt\_pe, evd, location, inr\_bwh\}.$$

Cet ensemble correspond aux nœuds colorés en bleu du graphe de la figure B.2d.

Nous avons suivi la même démarche pour déterminer un ensemble suffisant minimal pour l'ajustement de la confusion dans l'estimation de l'effet causal de *Group* sur *expansion*. La manipulation (iv) a permis de sélectionner une variable à inclure dans notre ensemble, à savoir *inr_bwh* (voir figure B.3a), et pour bloquer les chemins porte-arrières dans le graphe simplifié (figure B.3c), nous avons abouti à deux ensembles minimaux, à savoir $\{antiplatelets, evd\}$ et $\{cad, evd\}$. En effet, pour bloquer les chemins $Group \longleftarrow cad \longrightarrow antiplatelets \longrightarrow expansion$ nous avons le choix entre les deux variables *antiplatelets* et *cad*. C'est la raison pour laquelle nous avons utilisé la couleur bleu foncé pour les différencier des autres variables à inclure dans l'ensemble recherché. Nous avons donc deux ensembles suffisants minimaux pour l'ajustement de la confusion dans l'estimation de l'effet causal de *Group* sur *expansion*, et ils sont donnés par :

$$Z_2 = \{inr\_bwh, antiplatelets, evd\},$$

et

$$Z_3 = \{inr\_bwh, cad, evd\}.$$

Nous avons également sélectionné les variables confondantes potentielles, excepté la variable *osh* qui diminue la qualité de l'appariement qui sera l'objet de la section 6.3.1. L'ajout des termes d'interaction et le carré de la variable *volume* servent à créer des groupes appariés qui satisfont mieux la propriété de balance

du score de propension (voir section 6.3.1). Cette propriété de balance du score de propension est également utilisée pour choisir l'ensemble suffisant minimal à retenir dans notre deuxième cas d'analyse parmi les deux ensembles trouvés.

Les résultats de nos estimations sont donnés dans les tableaux 6.2 et 6.3. Il est à noter que dans l'estimation du score de propension, on n'interprète pas la signification des coefficients du modèle. Le plus important réside dans le fait de savoir quelle covariable inclure dans le modèle pour satisfaire la propriété de balance du score de propension.

### 6.2.2 Vérification de l'hypothèse de positivité

Après l'estimation du score de propension, nous allons procéder à la vérification de l'hypothèse de positivité et déterminer la région de support commun des distributions du score de propension des deux groupes de traitement, à laquelle nous allons restreindre notre analyse. La figure 6.1 laisse apparaître des zones de non-chevauchement des deux distributions dans les deux cas d'analyse. L'application du critère Min-Max permet de déterminer la région de support commun où les deux distributions se chevauchent.

La figure 6.1a montre que pour le premier cas d'analyse, le minimum du logit du score de propension est égal à $-1.48$ dans le groupe des traités, ce qui correspond à un score de propension de 0.19. D'autre part, le maximum du logit du score de propension est égal à 2.18 dans le groupe témoin, ce qui donne un score de propension de 0.90. Cela signifie que la positivité est satisfaite dans la région de $[0.19; 0.90]$. De la même façon, nous obtenons la région de support commun du logit du score de propension pour notre deuxième cas d'analyse, qui est $[-1.06; 1.83]$ (voir figure 6.1b), et qui correspond à $[0.26; 0.86]$ en termes de score de propension.

Tableau 6.2: Estimation du score de propension (premier cas)

| Variable | Valeur estimée | Erreur standard | Khi-2 de Wald | p-value |
|---|---|---|---|---|
| *Intercept* | 3.7954 | 3.5781 | 1.1252 | 0.2888 |
| *age* | -0.0369 | 0.0402 | 0.8421 | 0.3588 |
| *inr_bwh* | 1.0530 | 0.5133 | 4.2077 | 0.0402 |
| *volume* | 0.0003 | 0.0263 | 0.0002 | 0.9896 |
| *sbp* | -0.0092 | 0.0119 | 0.6047 | 0.4368 |
| *valve* | -1.1177 | 2.7018 | 0.1711 | 0.6791 |
| *cmy* | 0.3691 | 0.9631 | 0.1469 | 0.7016 |
| *cancer* | 0.4516 | 0.3899 | 1.3418 | 0.2467 |
| *gender* | 0.2270 | 0.3346 | 0.4602 | 0.4975 |
| *cad* | -0.2133 | 0.3634 | 0.3445 | 0.5573 |
| *postfossa* | 0.2416 | 2.0476 | 0.0139 | 0.9061 |
| *afib* | -0.8246 | 0.6031 | 1.8693 | 0.1716 |
| *dvt_pe* | -0.3688 | 0.6363 | 0.3360 | 0.5621 |
| *evd* | -1.1469 | 0.5156 | 4.9474 | 0.0261 |
| *location* | 0.1215 | 0.3717 | 0.1068 | 0.7438 |
| *renal_failure* | -0.7617 | 0.7595 | 1.0058 | 0.3159 |
| *alcohol* | -0.8480 | 0.7282 | 1.3562 | 0.2442 |
| *age * valve* | 0.0116 | 0.0357 | 0.1049 | 0.7461 |
| *sbp * postfossa* | -0.0009 | 0.0118 | 0.0057 | 0.9400 |
| *volume_square* | 0.0000 | 0.0001 | 0.0384 | 0.8447 |

## 6.3 Estimation de l'effet du traitement PCC

Pour évaluer l'effet du traitement PCC sur la mortalité à 6 mois et sur l'expansion de l'hématome, nous avons utilisé la technique d'appariement sur le score

Tableau 6.3: Estimation du score de propension (deuxième cas)

| Variable | Valeur estimée | Erreur standard | Khi-2 de Wald | p-value |
|---|---|---|---|---|
| *Intercept* | 2.3880 | 1.7835 | 1.7927 | 0.1806 |
| *inr_bwh* | 0.3231 | 0.3643 | 0.7868 | 0.3751 |
| *sbp* | -0.0050 | 0.0084 | 0.3612 | 0.5478 |
| *evd* | -1.1234 | 0.3464 | 10.5179 | 0.0012 |
| *renal_failure* | -0.7259 | 0.6020 | 1.4542 | 0.2279 |
| *alcohol* | -0.9113 | 0.6281 | 2.1051 | 0.1468 |
| *cad* | -1.2622 | 0.8648 | 2.1302 | 0.1444 |
| *inr_bwh* $*$ *cad* | 0.6044 | 0.3654 | 2.7360 | 0.0981 |

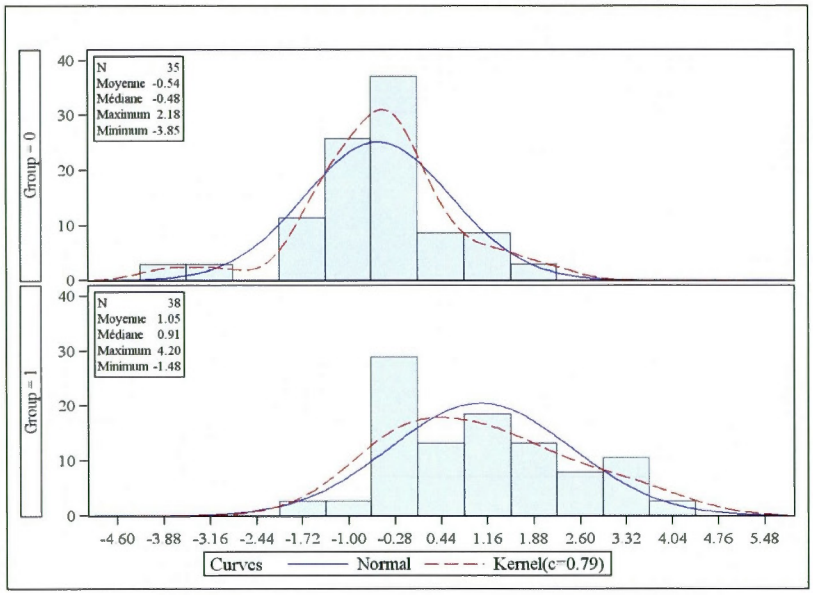
de propension. Nous avons également effectué une brève analyse basée sur la méthode d'ajustement sur le score de propension et la pondération inverse dans le but de confirmer nos résultats. La méthode de stratification sur le score de propension est exclue de cette analyse, en raison de la taille relativement faible de notre échantillon.

### 6.3.1 Appariement sur le score de propension

Dans nos deux cas d'étude, nous avons opté pour la méthode du plus proche voisin sans remise pour réaliser l'appariement sur le score de propension. Dans le but d'améliorer la qualité de notre appariement, nous avons utilisé un caliper égal à 0.2 fois l'écart-type du logit du score de propension lors du calcul des distances entre les traités et les non-traités (voir section 5.3.1). À l'issue de cette étape, nous avons obtenu 28 paires dans notre premier cas d'analyse, et 35 paires dans le deuxième. Nous constatons que nous avons obtenu moins de paires dans le premier cas que dans le deuxième. Cela est dû au fait que dans notre analyse de l'effet du traitement PCC sur la mortalité à 6 mois, nous avons inclus un nombre

Figure 6.1: Distribution du logit du score de propension dans l'échantillon initial

(a) Premier cas




(b) Deuxième cas

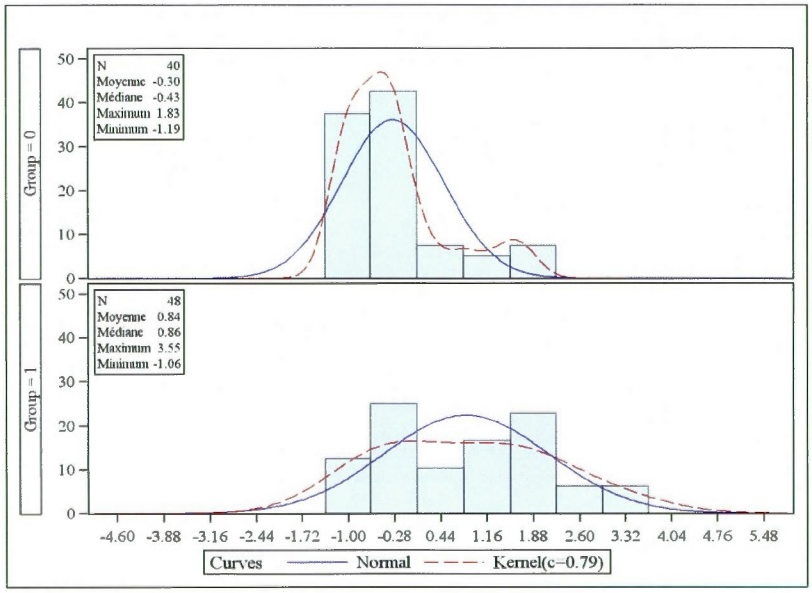
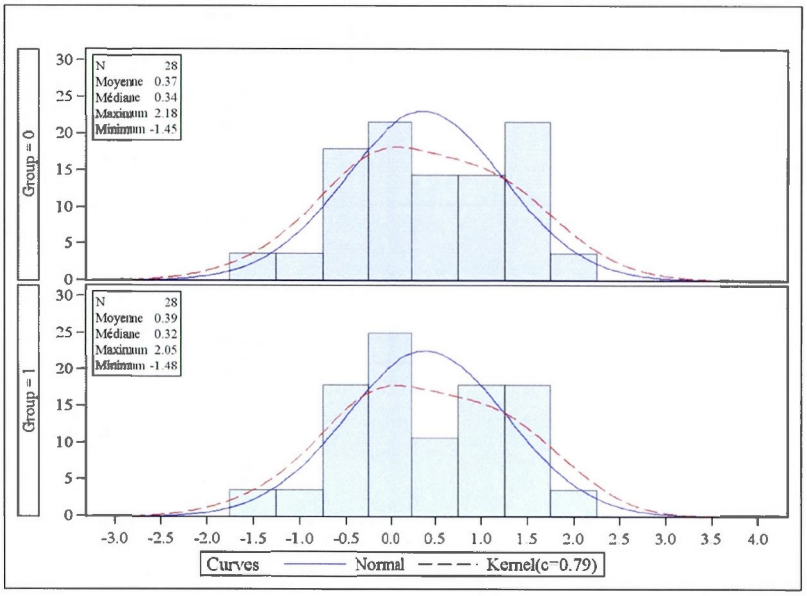


Figure 6.2: Distribution du logit du score de propension dans les échantillons appariés

(a) Premier cas




(b) Deuxième cas

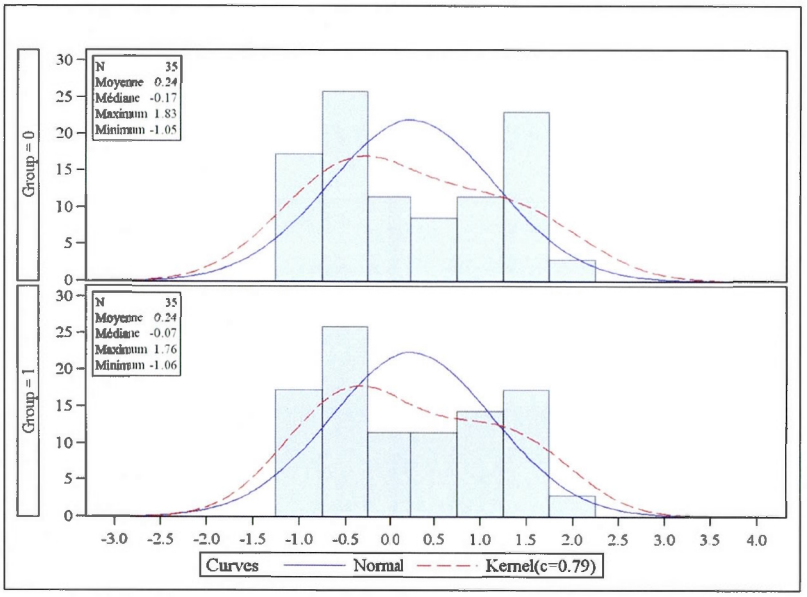


Tableau 6.4: Comparaison des deux groupes de traitement dans les échantillons appariés (premier cas)

| Variable | Groupe traité (N=22) | Groupe témoin (N=22) | DS après appariement | DS avant appariement |
|---|---|---|---|---|
| *age* | 71.89(12.73) | 69.79(12.84) | 0.1648 | 0.1653 |
| *inr_bwh* | 2.20(0.55) | 2.20(0.72) | 0.0000 | 0.1050 |
| *volume* | 27.79(38.99) | 32.36(33.67) | 0.1255 | 0.3077 |
| *sbp* | 161.50(33.11) | 160.30(23.63) | 0.0410 | 0.0918 |
| *valve* | 5(17.86%) | 4(14.29%) | 0.0974 | 0.1091 |
| *cmy* | 0.00(0.00%) | 0.00(0.00%) | 0.0000 | 0.0441 |
| *cancer* | 4(14.29%) | 5(17.86%) | 0.0974 | 0.1667 |
| *gender* | 11(39.29%) | 11(39.29%) | 0.0000 | 0.1181 |
| *cad* | 9(32.14%) | 9(32.14%) | 0.0000 | 0.0064 |
| *postfossa* | 5(17.86%) | 3(10.71%) | 0.2052 | 0.0365 |
| *afib* | 15(53.57%) | 13(46.43%) | 0.1432 | 0.2193 |
| *dvt_pe* | 8(28.57%) | 7(25.00%) | 0.0807 | 0.2223 |
| *evd* | 5(17.86%) | 7(25.00%) | 0.1747 | 0.8450 |
| *location* | 12(42.86%) | 12(42.86%) | 0.0000 | 0.1099 |
| *renal_failure* | 2(7.14%) | 2(7.14%) | 0.0000 | 0.3199 |
| *alcohol* | 2(7.14%) | 3(10.71%) | 0.1255 | 0.2541 |

important de covariables dans l'estimation du score de propension, et certaines de ces covariables présentent des valeurs manquantes. La figure 6.2 montre la distribution du logit du score de propension dans les échantillons appariés formés.

L'étape qui suit consiste à vérifier la propriété de balance des scores de propension estimés. Pour cela, nous avons calculé à partir des échantillons appariés les différences standardisées des variables intervenant dans l'estimation du score de

Tableau 6.5: Comparaison des deux groupes de traitement dans les échantillons appariés (deuxième cas)

| Variable | Groupe traité (N=28) | Groupe témoin (N=28) | DS après appariement | DS avant appariement |
|---|---|---|---|---|
| *inr_bwh* | 2.22(0.57) | 2.25(0.88) | 0.0463 | 0.1050 |
| *sbp* | 157.70(34.41) | 156.60(26.40) | 0.0335 | 0.0918 |
| *evd* | 8(22.86%) | 6(17.14%) | 0.1432 | 0.8450 |
| *renal_failure* | 2(5.71%) | 1(2.86%) | 0.1414 | 0.3199 |
| *alcohol* | 2(5.71%) | 3(8.57%) | 0.1111 | 0.2541 |
| *cad* | 11(31.43%) | 10(28.57%) | 0.0624 | 0.0064 |

propension, pour pouvoir comparer leurs moyennes dans les groupes appariés obtenus. Nous avons également comparé ces différences à celles obtenues sur la base de l'échantillon initial sur les mêmes variables. Les résultats obtenus lors de cette étape sont récapitulés dans les tableaux 6.4 et 6.5.

À partir de ces deux tableaux, nous constatons que toutes les variables ont des différences standardisées après appariement, inférieures ou égales à 0.2 (valeur maximale recommandée par Rosenbaum et Rubin (1985)) dans les deux cas d'étude, ce qui signifie que les moyennes de ces covariables sont très similaires entre les sujets traités et les sujets non traités. Pour beaucoup de covariables, nous remarquons une diminution significative des différences standardisées liées aux échantillons appariés, par rapport à celles obtenues à partir de l'échantillon initial. Il est à noter que l'ajout des termes d'interaction et le terme de degrés deux de la variable *volume* lors de l'estimation des scores de propension était dans le but de réduire ces distances.

### 6.3.2 Effet du traitement

Tout d'abord, nous commençons par examiner la répartition des paires formées présentées dans le tableau 6.6. Dans notre première étude de cas, nous avons obtenu 28 paires dont :

- 5 paires concordantes où les deux sujets sont décédés,
- 10 paires concordantes où les deux sujets ne sont pas décédés,
- 4 paires discordantes pour lesquelles le sujet traité est décédé et le non-traité ne l'est pas,
- 9 paires discordantes où seuls les sujets non traités sont décédés.

Concernant notre deuxième étude de cas, les 35 paires obtenues sont réparties comme suit :

- 25 paires concordantes où les deux sujets n'ont pas connu d'expansion de leurs hématomes,
- 3 paires discordantes pour lesquelles seuls les sujets traités ont connu une expansion de leurs hématomes,
- 7 paires discordantes où l'on a enregistré une expansion de l'hématome uniquement chez les sujets non traités.

Nous allons maintenant procéder à l'estimation proprement dite, de l'effet du traitement PCC sur la mortalité à 6 mois et sur l'expansion de l'hématome lié à l'hémorragie intracérébrale sous warfarine. D'après le tableau 6.6, le risque de mortalité à 6 mois est égal à $9/28 = 32.14\%$ chez le groupe traité, et à $14/28 = 50.00\%$ chez les non-traités, ce qui donne un risque relatif estimé de mortalité à 6 mois du groupe traité par rapport au groupe témoin d'une valeur de $32.14/50 = 0.64$, et une différence des risques de 18%. De la même façon, nous obtenons un risque relatif estimé de l'expansion de l'hématome pour le groupe traité comparativement au groupe témoin, égal à 0.43 et une différence des risques de 11.43%. Comme

Tableau 6.6: Répartition des paires issues de l'appariement

| Premier cas | | | |
|---|---|---|---|
| Traité \ Témoin | non-décès (0) | décès (1) | Total |
| non-décès (0) | 10 | 9 | 19 |
| décès (1) | 4 | 5 | 9 |
| Total | 14 | 14 | 28 |
| **Deuxième cas** | | | |
| Traité \ Témoin | non-expansion (0) | expansion (1) | Total |
| non-expansion (0) | 25 | 7 | 32 |
| expansion (1) | 3 | 0 | 3 |
| Total | 28 | 7 | 35 |

l'appariement sur le score de propension permet d'éliminer les différences existantes entre les traités et les non-traités, il devient possible d'estimer l'effet du traitement par l'un des risques calculés, à savoir le risque relatif ou la différence des risques. En effet, un risque relatif différent de 1 ou une différence des risques différente de 0 signifie l'existence d'un effet du traitement sur la variable réponse.

Pour savoir si cet effet est statistiquement significatif, nous allons appliquer le test de McNemar qui est le plus adapté aux échantillons appariés lorsque la variable de réponse est binaire. Les résultats du test sont présentés dans le tableau 6.7 pour les deux cas d'analyse. Selon le test de McNemar, l'effet du traitement estimé n'est pas significatif dans nos deux cas d'analyse au seuil de 5%. Les *p-values* du test sont respectivement 0.2668 et 0.3437 pour le premier et le deuxième cas d'analyse.

Tableau 6.7: Effet du traitement PCC : test de McNemar

| Cas d'étude | Statistique du test | p-value exacte |
|---|---|---|
| Premier cas | 1.9231 | 0.2668 |
| Deuxième cas | 1.6000 | 0.3437 |

Tableau 6.8: Estimation de l'effet du traitement par intervalle de confiance

| Premier cas | | | | |
|---|---|---|---|---|
| Valeur estimée | Erreur standard | p-value | Borne inf | Borne sup |
| 0.1786 | 0.1243 | 0.1507 | -0.06499 | 0.4221 |
| Deuxième cas | | | | |
| Valeur estimée | Erreur standard | p-value | Borne inf | Borne sup |
| 0.1143 | 0.08826 | 0.1954 | -0.05870 | 0.2873 |

Le test de McNemar ne donne pas de valeur estimée de l'effet du traitement. Pour trouver la valeur estimée de l'effet du traitement exprimé en termes de la différence des risques, ainsi que son intervalle de confiance, nous appliquons un modèle à mesures répétées. Les résultats de l'estimation sont donnés dans le tableau 6.8. Ainsi, nous retrouvons les valeurs de la différence des risques que nous avons déjà calculées, ainsi que leurs intervalles de confiance.

## 6.4 Ajustement sur le score de propension et pondération inverse

Pour confirmer les résultats d'estimation obtenus, nous avons appliqué brièvement deux autres méthodes, à savoir l'ajustement sur le score de propension et la pondération inverse.

Dans la première méthode, nous avons utilisé la régression logistique pour estimer l'effet de la variable du traitement sur les deux variables de réponse. Nous avons in-

Tableau 6.9: Estimation de l'effet du traitement par ajustement

| Premier cas | | | | |
|---|---|---|---|---|
| Variable | Valeur estimée | Erreur standard | Khi-2 de Wald | p-value |
| Intercept | 2.2549 | 0.9183 | 6.0300 | 0.0141 |
| Group | -0.2153 | 0.3190 | 0.4554 | 0.4998 |
| pr_s | -2.6160 | 1.5741 | 2.7620 | 0.0965 |
| Deuxième cas | | | | |
| Variable | Valeur estimée | Erreur standard | Khi-2 de Wald | p-value |
| Intercept | 3.7219 | 1.3469 | 7.6357 | 0.0057 |
| Group | -0.1300 | 0.4477 | 0.0844 | 0.7715 |
| pr_s | -2.5271 | 2.2266 | 1.2881 | 0.2564 |

clus le score de propension comme variable explicative continue dans les modèles : Comme pour l'analyse précédente, l'effet du traitement n'est pas significatif dans les deux cas (voir tableau 6.9).

Concernant la méthode de pondération inverse, nous avons effectué une régression logistique avec un poids $w$ donné par :

$$w = \frac{1}{pr\_s} \text{ si } \quad Group = 1$$
$$w = \frac{1}{1 - pr\_s} \text{ si } \quad Group = 0.$$

Les résultats de cette analyse sont représentés dans le tableau 6.10. Nous constatons également que l'effet du traitement n'est pas significatif. Nous aboutissons donc au même résultat avec les trois méthodes, à savoir l'absence de l'effet du traitement PCC sur les deux variables de réponse.

Tableau 6.10: Estimation de l'effet du traitement par la technique de pondération inverse

| Premier cas | | | | |
|---|---|---|---|---|
| Valeur estimée | Erreur standard | p-value | Borne inf | Borne sup |
| 0.1786 | 0.1243 | 0.1507 | -0.06499 | 0.4221 |
| Deuxième cas | | | | |
| Valeur estimée | Erreur standard | p-value | Borne inf | Borne sup |
| 0.1143 | 0.08826 | 0.1954 | -0.05870 | 0.2873 |

# CONCLUSION ET PERSPECTIVES

Dans cette partie, nous dressons en conclusion une synthèse de notre travail et des résultats obtenus dans les analyses que nous avons effectuées. Enfin, nous terminons par quelques perspectives pour les travaux futurs.

## Synthèse du mémoire

L'objectif de ce mémoire était de proposer une méthodologie pour estimer l'effet du traitement par le concentré de complexe prothrombique (PCC) chez les patients atteints d'une hémorragie intracérébrale sous warfarine. Les données dont nous disposons sont issues d'une étude clinique non expérimentale, où les sujets qui ont reçu le traitement sont déterminés de manière non aléatoire. De plus, c'est le médecin qui a choisi le groupe de traitement dans lequel le patient est assigné. Ainsi, nous nous sommes placés dans le contexte des études non expérimentales pour examiner l'effet du traitement PCC sur la mortalité à 6 mois et sur l'expansion de l'hématome lié à l'hémorragie.

Dans un premier temps, nous avons présenté les hypothèses nécessaires pour l'identification de l'effet causal, notamment les hypothèses d'ignorabilité du traitement et de positivité, dans le cadre de l'analyse contrefactuelle de la causalité. La première hypothèse stipule que conditionnellement à un ensemble de covariables observées, les contrefactuels sont indépendants de l'exposition au traitement. Cette hypothèse suppose l'absence de variables confondantes inobservées. L'hypothèse de positivité, aussi connue sous le nom d'hypothèse de support commun, permet d'assurer que pour chaque individu traité, il existe un non-traité ayant les mêmes caractéristiques, et il devient alors possible d'estimer l'effet causal moyen.

Nous avons ensuite présenté les modèles graphiques probabilistes et leurs propriétés, ainsi que leur rôle dans le calcul de lois de probabilités. Nous nous sommes intéressés plus particulièrement aux réseaux bayésiens, et leurs trois propriétés, à savoir, la compatibilité de Markov, la propriété de Markov globale et la condition de Markov. Ces propriétés permettent d'interpréter les d-séparations dans un graphe orienté acyclique comme étant des relations d'indépendance conditionnelle entre les variables représentées dans le graphe. Nous nous sommes intéressés également à des réseaux bayésiens particuliers, à savoir, les réseaux bayésiens causaux, qui permettent de modéliser les relations de cause à effet. Nous avons introduit l'opérateur *do* qui permet de déterminer l'effet d'une manipulation d'une variable $X$ sur une autre variable $Y$, en évaluant la quantité $\mathbb{P}(Y \mid do(x))$.

Afin de faire le lien entre l'analyse graphique et l'analyse contrefactuelle de la causalité, nous avons fait appel aux modèles causaux à équations structurelles. Le graphe causal représentant nos variables est considéré comme étant un diagramme causal auquel est lié un ensemble d'équations structurelles. À partir de ce diagramme causal, nous avons sélectionné, en appliquant le critère porte-arrière, l'ensemble de variables suffisant minimal pour l'ajustement de la confusion dans l'estimation de l'effet causal du traitement PCC sur la mortalité à 6 mois et sur l'expansion de l'hématome lié à l'hémorragie. Nous avons ensuite montré le lien entre cet ensemble minimal et l'hypothèse d'ignorabilité du traitement nécessaire pour l'identification de l'effet causal.

Pour l'estimation de l'effet causal, nous avons adopté la technique du score de propension. Ainsi, au lieu de considérer toutes les covariables dans l'estimation de l'effet causal, nous les remplaçons par le score de propension estimé. En premier lieu, nous avons montré que si le traitement est fortement ignorable étant donné un ensemble de covariables, alors il est fortement ignorable étant donné le score de propension lié à ces variables. Nous avons ensuite présenté les différentes méthodes

utilisées pour estimer le score de propension. Nous nous sommes intéressés plus particulièrement à la régression logistique qui est la méthode la plus utilisée à cette fin. En dernier lieu, nous avons rappelé les différentes techniques utilisées pour estimer l'effet causal en se basant sur le score de propension estimé.

Enfin, nous avons appliqué les différentes méthodes développées tout au long de ce mémoire à notre base de données. Nous avons procédé alors à l'estimation de l'effet du traitement PCC sur la mortalité à 6 mois et sur l'expansion de l'hématome chez les patients atteints d'une hémorragie intracérébrale sous warfarine. Nous sommes parvenus à la conclusion que le traitement PCC n'a pas d'effet sur les deux réponses considérées. Lors de cette estimation, nous avons procédé également à la vérification des hypothèses d'ignorabilité du traitement et de positivité, ainsi que la propriété de balance du score de propension.

Limites et perspectives

Notre étude est limitée par la taille relativement faible de notre échantillon. De plus, certaines variables présentent des valeurs manquantes, ce qui conduit à exclure certains individus de l'analyse. En outre, les relations causales que nous avons considérées dans notre analyse sont asymétriques. Néanmoins, selon la Docteure Chou, certaines d'entre elles pourraient être symétriques. Il en résulte donc un modèle causal non-Markovien, auquel il ne serait pas possible d'appliquer le critère porte-arrière pour déterminer un ensemble de variables suffisant minimal pour l'ajustement de la confusion.

Compte tenu des limites citées ci-dessus, ce travail nécessite d'être poursuivi. Il serait intéressant d'appliquer l'approche développée dans ce mémoire sur un ensemble de données plus large, pour pouvoir confirmer les résultats que nous avons obtenus. Une autre idée consisterait à remplacer les valeurs manquantes par des valeurs plausibles à l'aide d'une méthode adéquate telle que l'imputation multiple

(*multiple imputation*) ou la méthode basée sur les patrons des données manquantes (*missingness pattern*). Une analyse préliminaire basée sur l'imputation multiple sur notre ensemble de données montre l'absence d'effet du traitement PCC sur la mortalité à 6 mois et sur l'expansion de l'hématome, mais une analyse plus approfondie est nécessaire pour pouvoir confirmer nos résultats. Enfin, il serait aussi envisageable d'exploiter les méthodes permettant de détecter un ensemble suffisant pour l'ajustement de la confusion dans le cadre des modèles causaux non-Markoviens.

# ANNEXE A

# COMPLÉMENT SUR LES MODÈLES GRAPHIQUES

## A.1 Recherche de cycles dans un graphe orienté

Les figures ci-dessous montrent les étapes de détection de cycles dans notre graphe causal.

## A.2 Preuve des propriétés d'indépendance conditionnelle

Dans la section 3.2.1, nous avons présenté quatre propriétés d'indépendance conditionnelle à savoir la symétrie, la décomposition, l'union faible et la contraction. Dans cette section, nous allons les montrer dans le cas continu. Nous supposons que toutes nos variables sont continues de densité de probabilité jointe $f(.)$.

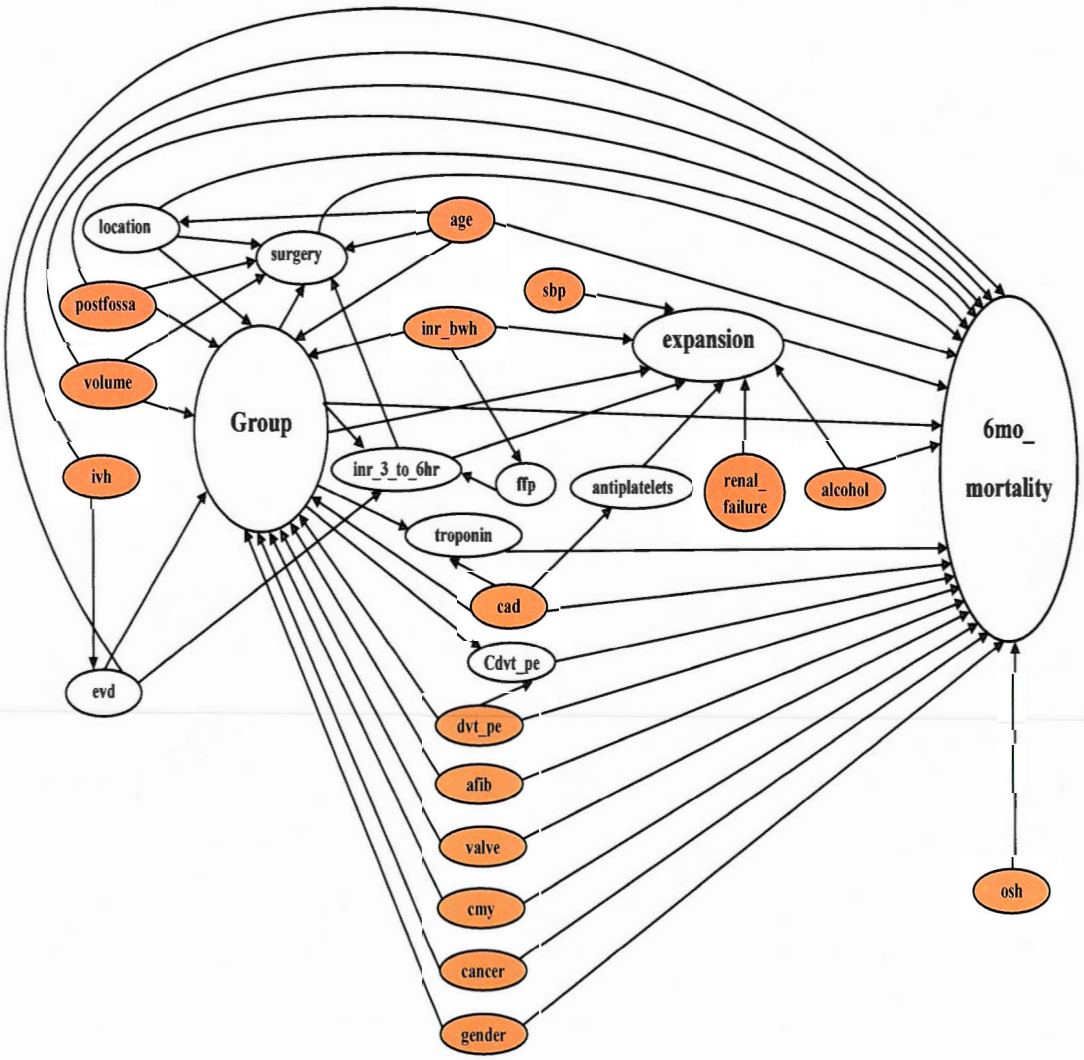
1. Symétrie : $X \perp\!\!\!\perp Y \mid Z \implies f(x \mid y, z) = f(x \mid z)$. Nous avons

$$\begin{aligned} f(y \mid x, z) &= \frac{f(x, y \mid z)}{f(x \mid z)} \\ &= \frac{f(x \mid y, z) f(y \mid z)}{f(x \mid z)} \\ &= \frac{f(x \mid z) f(y \mid z)}{f(x \mid z)} \\ &= f(x \mid z). \end{aligned}$$

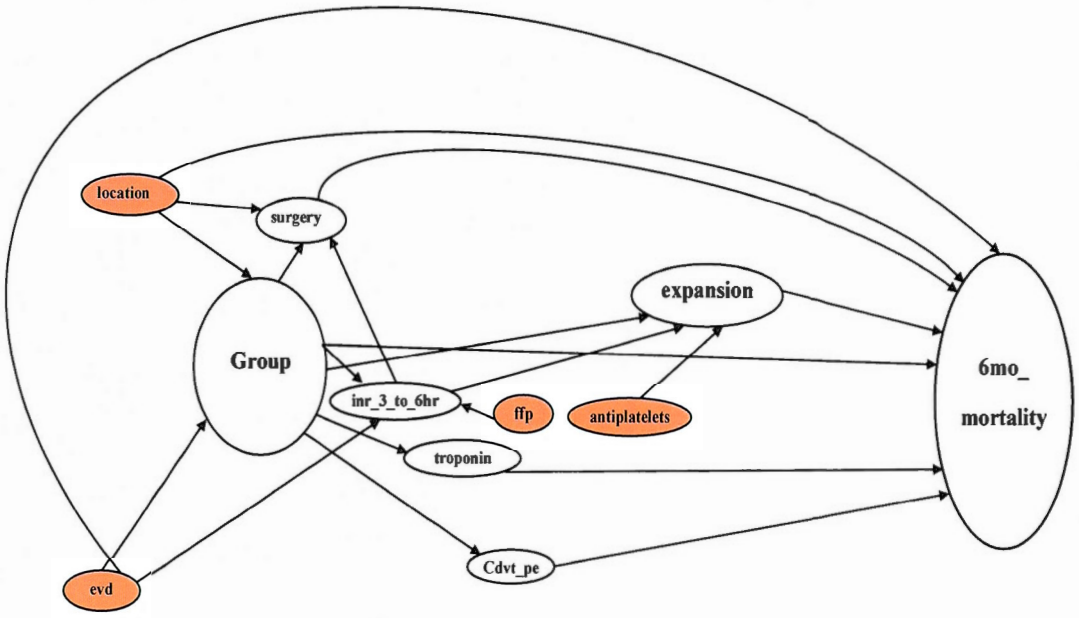
Alors $Y \perp\!\!\!\perp X \mid Z$.

Figure A.1: Vérification de cycles dans le graphe causal

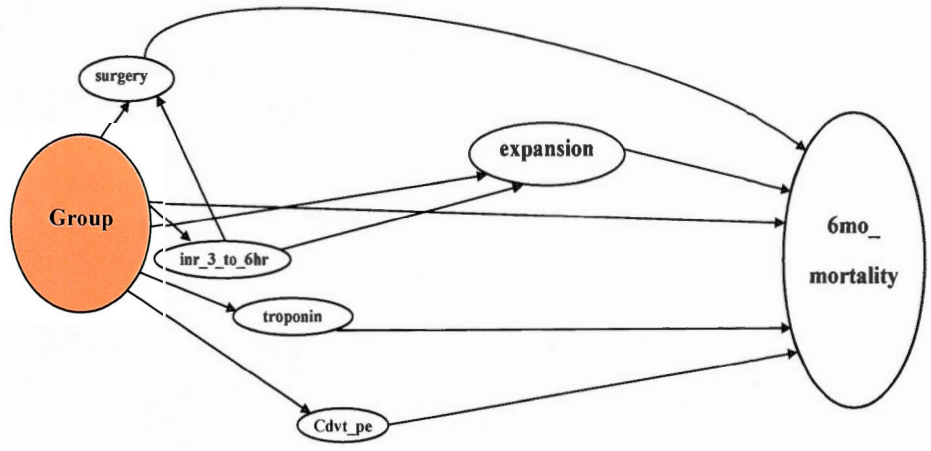
(a)



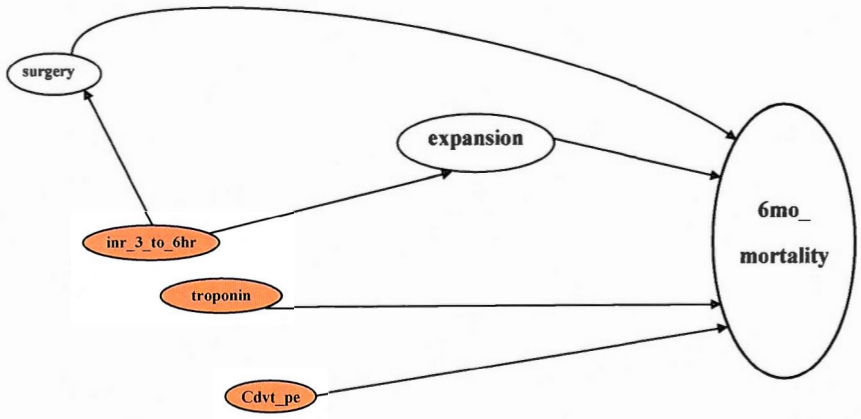
(b)


location
surgery
Group
expansion
6mo_
mortality
inr_3_to_6hr
ffp
antiplatelets
troponin
Cdvt_pe
evd

(c)


surgery
expansion
Group
6mo_
mortality
inr_3_to_6hr
troponin
Cdvt_pe

(d)




2. Décomposition : nous avons

$$X \perp\!\!\!\perp (Y, W) \mid Z \implies f(x \mid y, z, w) = f(x \mid z). \tag{A.1}$$

Nous avons également

$$\begin{aligned} f(x \mid y, z) &= \int f(x, w \mid y, z) dw \\ &= \int f(x \mid y, z, w) f(w \mid y, z) dw \\ &= f(x \mid z) \int f(w \mid y, z) dw \\ &= f(x \mid z). \end{aligned} \tag{A.2}$$

Alors $X \perp\!\!\!\perp Y \mid Z$. De la même façon que ci-dessus, nous avons

$$\begin{aligned} f(x \mid w, z) &= \int f(x, y \mid w, z) dy \\ &= \int f(x \mid y, z, w) f(y \mid w, z) dy \\ &= f(x \mid z) \int f(y \mid w, z) dy \\ &= f(x \mid z). \end{aligned} \tag{A.3}$$

Par conséquent, nous avons aussi $X \perp\!\!\!\perp W \mid Z$.

3. Union faible : la preuve de cette propriété découle directement de celle de la propriété précédente. En effet, lorsque $X \perp\!\!\!\perp (Y, W) \mid Z$ nous avons d'après

(A.1) et (A.3) :

$$f(x \mid y, z, w) = f(x \mid w, z).$$

Par conséquent, $X \perp\!\!\!\perp Y \mid (Z, W)$. De la même façon, nous savons que d'après (A.1) et (A.2) :

$$f(x \mid y, z, w) = f(x \mid y, z).$$

Alors $X \perp\!\!\!\perp W \mid (Z, Y)$.

4. Contraction : nous avons

$$X \perp\!\!\!\perp Y \mid (Z, W) \implies f(x \mid y, z, w) = f(x \mid z, w). \tag{A.4}$$

D'autre part nous avons

$$X \perp\!\!\!\perp W \mid Z \implies f(x \mid z, w) = f(x \mid z). \tag{A.5}$$

En combinant (A.4) et (A.6) nous concluons que

$$f(x \mid y, z, w) = f(x \mid z). \tag{A.6}$$

Ce qui montre que $X \perp\!\!\!\perp (Y, W) \mid Z$.

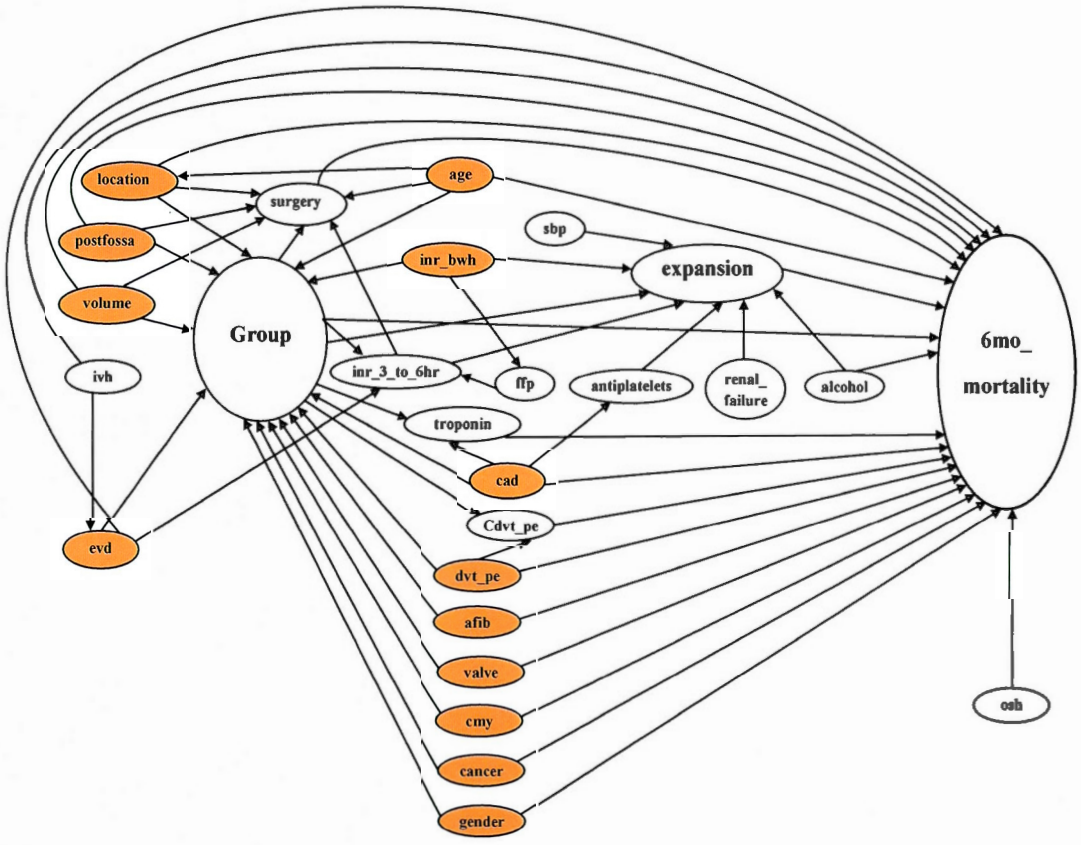
[Cette page a été laissée intentionnellement blanche]

# ANNEXE B

## ENSEMBLES ÉLIMINANT LA CONFUSION

### B.1 Ajustement sur les parents observables

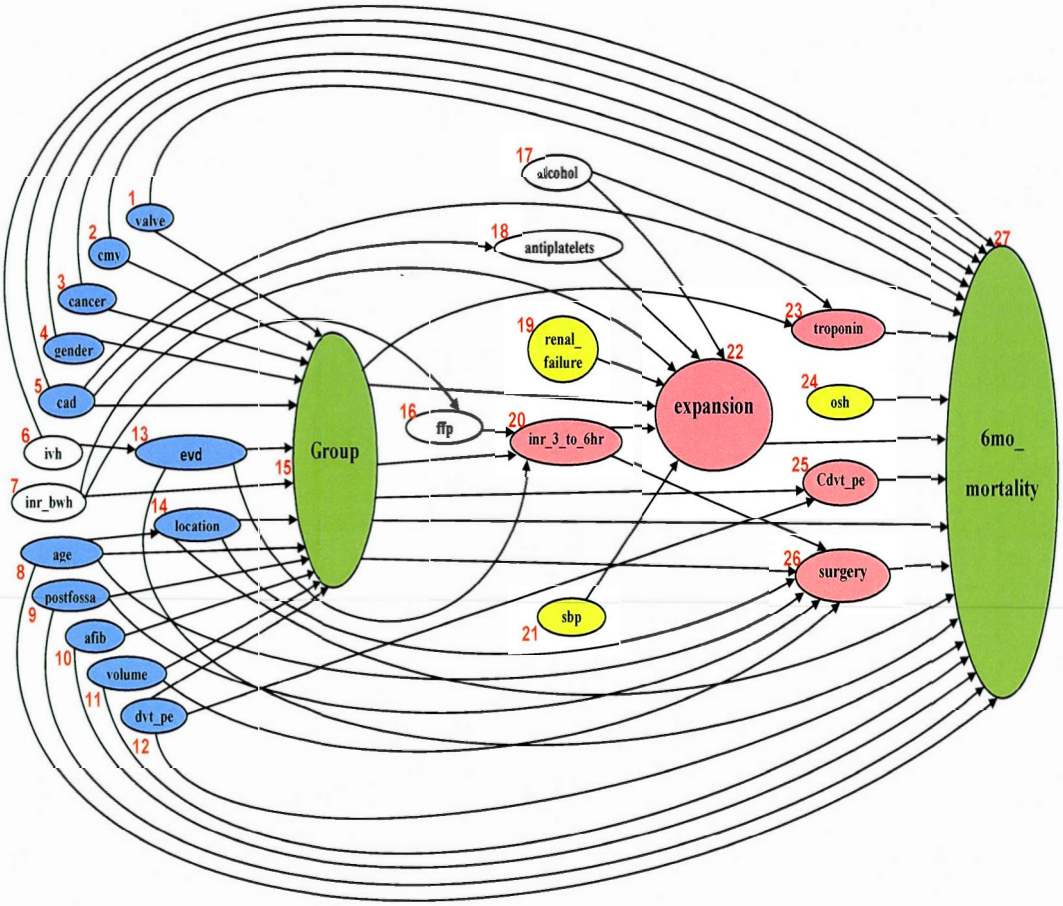
Figure B.1: Ajustement sur les parents observables



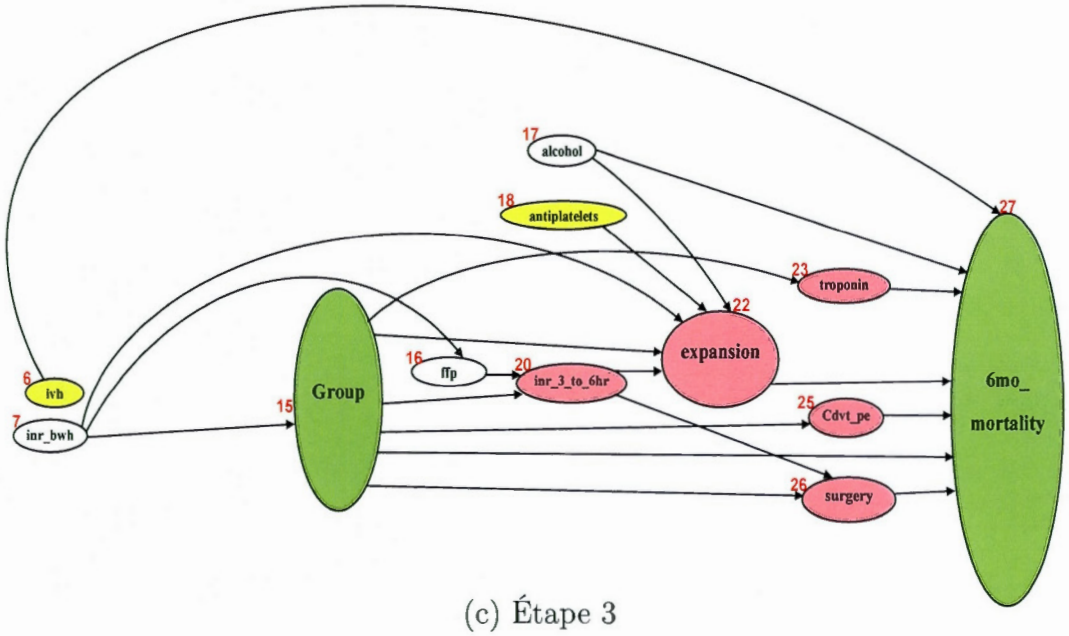
## B.2 Ensemble suffisant pour ajustement de la confusion

Figure B.2: Ensemble suffisant pour ajustement (*6mo_mortality*)

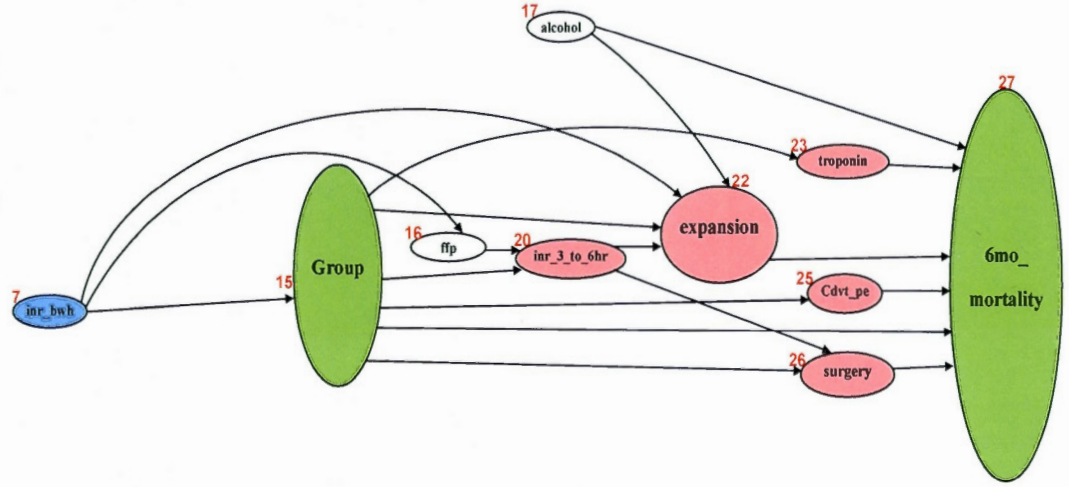
(a) Étape 1



(b) Étape 2




(c) Étape 3



(d) Résumé

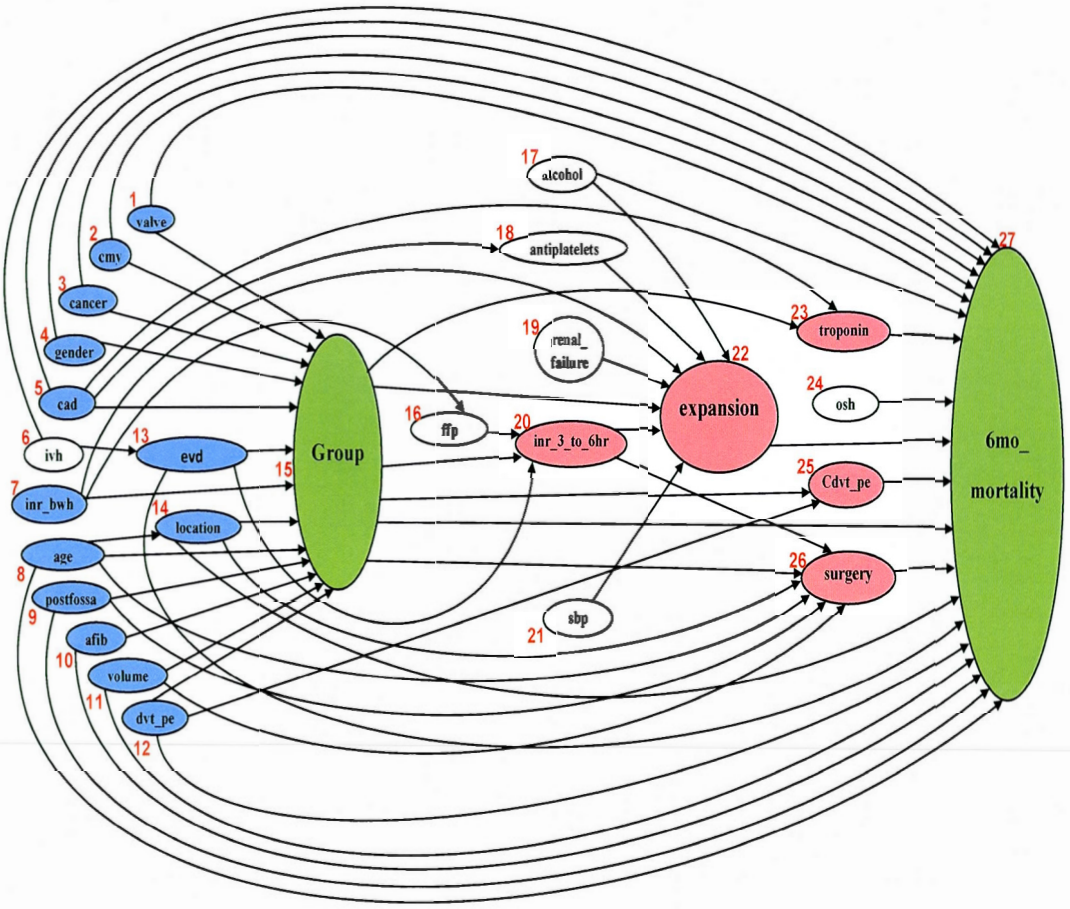
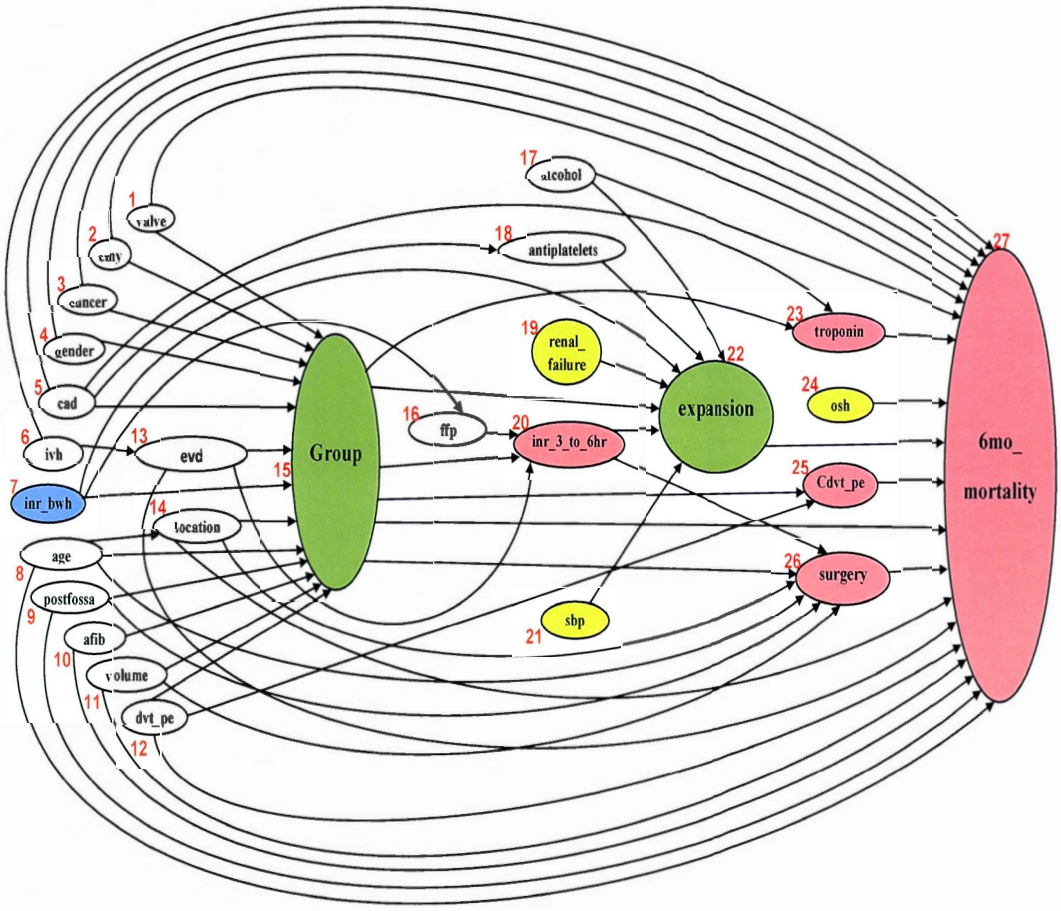
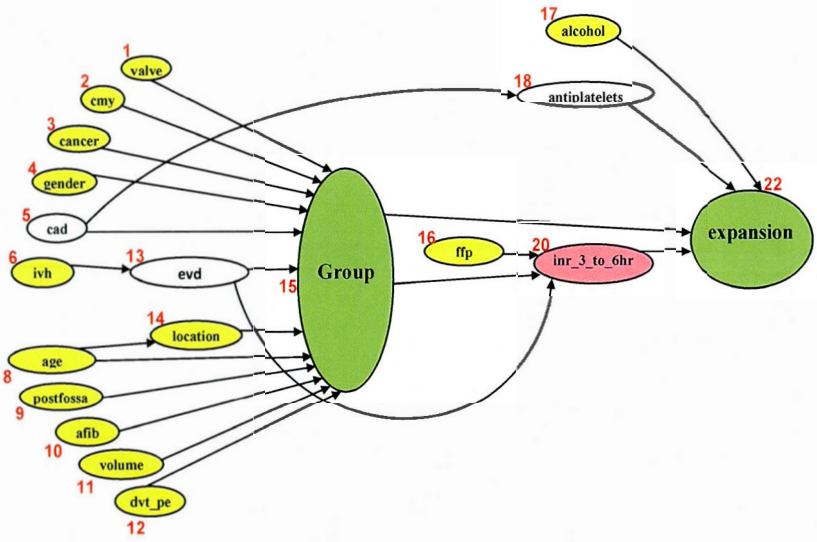

1
valve
2
cmv
3
cancer
4
gender
5
cad
6
ivh
7
inr_bwh
8
age
9
postfossa
10
afib
11
volume
12
dvt_pe
13
evd
14
location
15
Group
16
ffp
17
alcohol
18
antiplatelets
19
renal_
failure
20
inr_3_to_6hr
21
sbp
22
expansion
23
troponin
24
osh
25
Cdvt_pe
26
surgery
27
6mo_
mortality

Figure B.3: Ensemble suffisant pour ajustement (*expansion*)

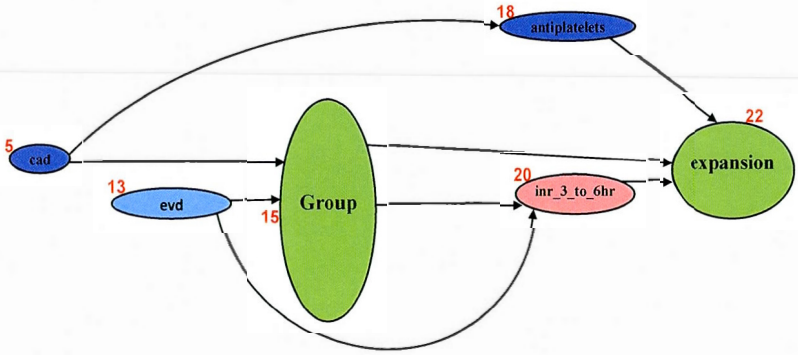
(a) Étape 1



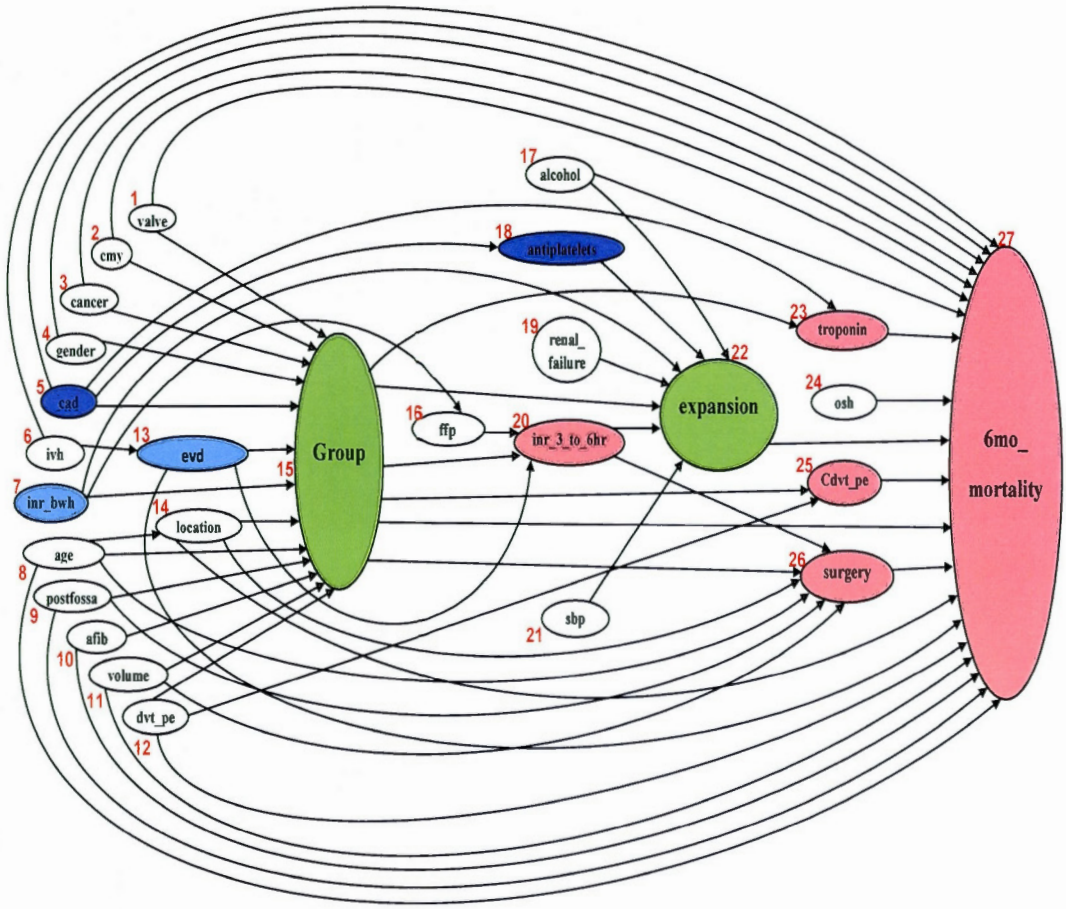
(b) Étape 2




(c) Étape 3



(d) Étape Résumé



[Cette page a été laissée intentionnellement blanche]

# BIBLIOGRAPHIE


Abadie, A. et Imbens, G. W. (2002). *Simple and bias-Corrected matching estimators for average treatment effects.* Technical Working Paper 283, National Bureau of Economic Research.

Abadie, A. et Imbens, G. W. (2009). *Matching on the estimated propensity score.* Working Paper 15301, National Bureau of Economic Research.

Austin, P. C. (2010). The performance of different propensity score methods for estimating difference in proportions (risk differences or absolute risk reductions) in observational studies. *Statistics in Medicine*, *29*, 2137–2148.

Austin, P. C. (2011). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics*, *10*, 150–161.

Austin, P. C. (2012). Using ensemble-based methods for directly estimating causal effects : An investigation of tree-based g-computation. *Multivariate Behavioral Research*, *47*(1), 115–135.

Berge, C. (1983). *Graphes.* Paris : Gauthier-Villars, Bordas.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*, 123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L., Friedman, J., Olshen, R. A. et Stone, C. J. (1984). *Classification and Regression Trees.* Monterey, Calif : Wadsworth and Brooks/Cole.

Cai, X., Orzell, S. C., Oussaid, E., Bresette, L. M., Sorond, F. A., Henderson, G. V., Atherton, J. K., Feske, S. K. et Chou, S. H. (2014). Reversing Coagulopathy : Prothrombin Complex Concentrate in Warfarin-associated Intracerebral Hemorrhage. [Document non publié]. Harvard Medical School, Boston, USA.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, *24*(2), 295–313.

Cochran, W. G. et Rubin, D. B. (1973). Controlling bias in observational studies : A review. *Sankhya : The Indian Journal of Statistics*, *35*(4), 417–446.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society, 41*(1), 1–31.

De Oliveira, C., Watt, R. et Hamer, M. (2010). Toothbrushing, inflammation, and risk of cardiovascular disease : results from scottish health surveyd. *BMJ, 340.*

Dehejia, R. H. et Wahba, S. (1999). Causal effects in nonexperimental studies : Reevaluating the evaluation of training programs. *Journal of the American Statistical Association, 94*(448), 1053–1062.

Dehejia, R. H. et Wahba, S. (2002). Propensity score matching methods for non-experimental causal studies. *Review of Economics and Statistics, 84*(1), 151–161.

Druzdzel, M. J. et Simon, H. A. (1993). Causality in bayesian belief networks. Dans *Proceedings of the Ninth Annual Conference on Uncertainty in Artificial Intelligence (UAI-93)*, 3–11., San Francisco, CA, USA. Morgan Kaufmann Publishers, Inc.

Faries, D. E., Leon, A. C., Maria Haro, J. et Obenchain, R. L. (2010). *Analysis of observational health care data using SAS.* Cary, NC : SAS Institute Inc.

Freund, Y. et Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences, 55*, 119–139.

Geiger, D., Verma, T. et Pearl, J. (1990). Identifying independence in bayesian networks. *Networks, 20*, 507–534.

Greenland, S., Pearl, J. et Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology, 10*(1), 37–48.

Gu, X. S. et Rosenbaum, P. R. (1993). Comparison of multivariate matching methods : Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics, 2*(4), 405–420.

Hastie, T., Tibshirani, R. et Friedman, J. (2009). *The Elements of Statistical Learning : Data Mining, Inference, and Prediction* (second éd.). Springer.

Heckman, J., Ichimura, H. et Todd, P. (1997). Matching as an econometric evaluation estimator : Evidence from evaluating a job training program. *Review of Economic Studies, 64*, 605–654.

Heckman, J., Ichimura, H. et Todd, P. (1998). Matching as an econometric evaluation estimator. *Review of Economic Studies, 65*, 261–294.

Hernán, M. A. et Robins, J. M. (2013). *Causal Inference.* Récupéré de `http://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/`

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*, 945–960.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity : A review. *The Review of Economics and Statistics*, *86*(1), 4–29.

Kowell, R. G., Dawid, A. P., Lauritzen, S. L. et Spiegelhalter, D. J. (1999). *Probabilistic Networks and Expert Systems.* New York : Springer-Verlag.

Lauritzen, S. L. (1996). *Graphical Models.* Oxford : Clarendon Press.

Lauritzen, S. L., Kowell, R. G., Dawid, A. P., Larsen, B. N. et Leimer, H. G. (1990). Independence properties of directed markov fields. *Networks*, *20*(5), 491–505.

Lee, B. K., Lessler, J. et Stuart, E. A. (2009). Improving propensity score weighting using machine learning. *Statistics in Medicine*, *29*, 337–346.

Luellen, J. K., Shadish, W. R. et Clark, M. H. (2005). Propensity scores : An introduction and experimental test. *Evaluation Review*, *29*(6), 530–558.

Lunceford, J. K. et Davidian, M. (2004). Stratication and weighting via the propensity score in estimation of causal treatment effects : A comparative study. *Statistics in Medicine*, *23*, 2937–2960.

McCaffrey, D. F., Ridgeway, G. et Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, *9*, 403–425.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference.* Morgan Kaufmann.

Pearl, J. (1993). Comment : Graphical models, causality, and intervention. *Statistical Science*, *8*(3), 266–269.

Pearl, J. (1998). Graphs, causality, and structural equation models. *Sociological Methods and Research*, *27*(2), 226–284.

Pearl, J. (2009). *Causality : Models, Reasoning and Inference* (second éd.). Cambridge University Press.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, *82*(398), 387–394.

Rosenbaum, P. R. (2002). *Observational studies* (second éd.). New York : Springer-Verlag.

Rosenbaum, P. R. (2010). *Design of observational studies* (second éd.). Springer.

Rosenbaum, P. R. et Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Rosenbaum, P. R. et Rubin, D. B. (1984). Rducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*, 516–524.

Rosenbaum, P. R. et Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, *39*(1), 33–38.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, *66*, 688–701.

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, *2*(1), 1–26.

Rubin, D. B. (1978). Bayesian inference for causal effects : The role of randomization. *The Annals of Statistics*, *6*(1), 34–58.

Rubin, D. B. (1980). Analysis of experimental data : The fisher randomization test comment. *Journal of the American Statistical Association*, *75*(371), 591–593.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*, *127*, 757–763.

Rubin, D. B. (2001). Using propensity scores to help design observational studies : Application to the tobacco litigation. *Health Services & Outcomes Research Methodology*, *2*, 169–188.

Rubin, D. B. et Thomas, N. (1996). Matching using estimated propensity score : relating theory to practice. *Biometrics*, *52*, 249–64.

SAS Institute Inc. (2013). *SAS (Version9.3) [Logiciel]*. Cary, North Carolina : SAS Institute.

Schulman, S., Beyth, R. J., Kearon, C. et Levine, M. N. (2008). Hemorrhagic complications of anticoagulant and thrombolytic treatment : American college of chest physicians evidence-based clinical practice guidelines (8th edition). *CHEST*, *133*, 257–298.

Setoguchi, S., Schneeweiss, S., AlanBrookhart, Glynn, R. J. et Cook, E. F. (2008). Evaluating uses of data mining techniques in propensity score estimation : A simulation study. *Pharmacoepidemiology and Drug Safety*, *17*, 546–555.

Spirtes, P., Glymour, C. et Scheines, R. (2000). *Causation, Prediction and Search* (second éd.). Cambridge : The MIT Press.

Verma, T. et Pearl, J. (1988). Causal networks : Semantics and expressiveness. Dans *Proceedings of the Fourth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-88)*, 352–359., Corvallis, Oregon. AUAI Press.

Westreich, D., Lessler, J. et Funk, M. J. (2010). Propensity score estimation : Neural networks, support vector machines, decision trees (cart), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology*, *63*, 826–833.

Williamson, J. (2005). *Bayesian Nets and Causality*. New-York : Oxford University Press.