UNIVERSITÉ DU QUÉBEC À MONTRÉAL

SÉLECTION DE VARIABLES EN MÉDIATION CAUSALE : MÉTHODES BASÉES SUR LE CHANGEMENT D'ESTIMATION ET SUR LA DIFFÉRENCE EN ERREUR QUADRATIQUE MOYENNE

MÉMOIRE
PRÉSENTÉ
COMME EXIGENCE PARTIELLE
DE LA MAÎTRISE EN MATHÉMATIQUES

PAR
JESSE GERVAIS

DÉCEMBRE 2021

REMERCIEMENTS

Je tiens à remercier tout particulièrement ma directrice de recherche, Geneviève Lefebvre, pour sa participation active dans la réalisation de ce mémoire. Au cours des deux dernières années, elle a été en mesure d'apporter ses conseils et son soutien dans les différentes étapes de mon parcours académique. Merci pour l'accompagnement dont j'ai eu la chance de bénéficier.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

RÉSUMÉ

Les méthodes de sélection de variables basées sur les données pour l'estimation de l'effet total de l'exposition sur la réponse, comme le changement d'estimation (CIE) et la différence de l'erreur quadratique moyenne (ΔMSE), sont fréquemment employées dans plusieurs domaines de recherche lorsque la connaissance du domaine d'application est insuffisante pour identifier un ensemble d'ajustement adéquat pour l'analyse des données. Alors que des défis de modélisation sont similairement présents en analyse de médiation causale, il n'y a actuellement que très peu de connaissances et d'outils pour la sélection de variables basées sur les données dans ce cadre d'analyse de plus en plus populaire. L'objectif principal du mémoire est de modifier les procédures CIE et ΔMSE pour qu'elles soient adaptées à la sélection de variables en médiation causale. Nous avons considéré six procédures de sélection de variables, soit quatre algorithmes qui ont été développés pour cibler les variables appropriées pour l'estimation des effets direct et indirect de l'exposition sur la réponse (CIE_{max}, CIE_{effets}, ΔMSE_{max} et ΔMSE_{effets}) et deux procédures utilisées à des fins de sélection de variables pour l'effet total (CIE_{total} et ΔMSE_{total}). Les six algorithmes ont été évalués par simulation Monte-Carlo et ceux-ci ont par la suite été illustrés sur les données de la vague 1999 de l'étude sur l'alcool du *Harvard School of Public Health College*. L'étude réalisée dans le cadre de ce mémoire a permis de montrer qu'il est possible d'utiliser des méthodes basées sur les données pour faire de la sélection de variables en médiation causale. Les résultats obtenus suggèrent qu'une des adaptations proposées du CIE (CIE_{max}) est appropriée si l'objectif est de rejeter un grand nombre de covariables, alors que les méthodes basées sur le ΔMSE sont préférables pour réduire la variance des estimateurs de l'effet direct et indirect de l'exposition sur la réponse.

Mots-clés : Changement d'estimation, différence de l'erreur quadratique moyenne, médiation par inférence causale, régression, sélection de variables, simulation

INTRODUCTION

L'analyse de médiation est une méthode statistique utilisée pour aider à répondre à la question de savoir comment une variable d'exposition A transmet son effet sur une variable réponse Y à travers une variable de médiation M (Hayes, 2013). Par exemple, la violence subie durant l'enfance peut être associée à la perpétration de violence conjugale chez les jeunes (Berzenski & Yates, 2010; Dardis, Dixon, Edwards, & Turchik, 2015). De plus, la violence subie durant l'enfance est reliée à une augmentation de la consommation d'alcool (Tonmyr, Thornton, Draca, & Wekerle, 2010), ce qui pourrait accroître les comportements violents dans le couple chez les jeunes (Rothman, McNaughton Reyes, Johnson, & LaValley, 2012; Shorey, Stuart, & Cornelius, 2011). Ainsi, des chercheurs pourraient vouloir déterminer si la violence subie durant l'enfance (A) est associée à la perpétration de violence conjugale chez les jeunes (Y) indirectement à travers l'impact de la violence subie durant l'enfance sur la consommation d'alcool (M). La figure 0.1 schématise le concept de l'analyse de médiation.

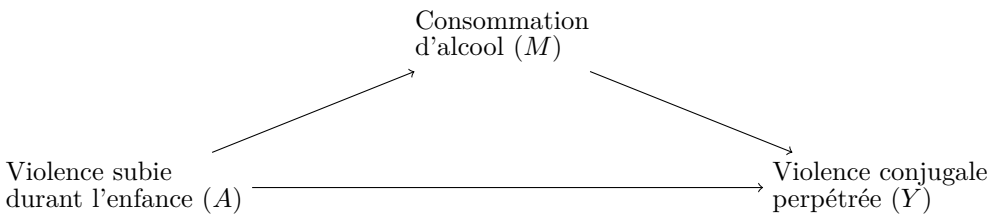


FIGURE 0.1 : Illustration de l'analyse de médiation

Un des objectifs de l'analyse de médiation est de décomposer l'effet total de l'ex-

position (A) sur la réponse (Y) comme la somme de l'effet direct (l'effet de A sur Y qui ne passe pas par M ; $A \rightarrow Y$) et de l'effet indirect (l'effet de A sur Y qui passe par M ; $A \rightarrow M \rightarrow Y$), où l'effet indirect permet de quantifier l'effet de médiation. Une des difficultés de l'analyse de médiation est que l'on cherche à établir des effets direct et indirect causaux. Conséquemment, cette méthode statistique impose des hypothèses fortes concernant l'inclusion de plusieurs types de variables de confusion, c'est-à-dire une variable qui brouille l'interprétation causale de l'association entre deux variables. En effet, l'omission de variables de confusion occasionne généralement une estimation erronée de la relation causale entre deux variables (Nguyen, Schmid, Ogburn, & Stuart, 2020). En particulier, l'analyse de médiation sur la base de données observationnelles nécessite d'ajuster adéquatement sur trois types de variables de confusion, soit les variables de confusion entre A et Y, les variables de confusion entre A et M et les variables de confusion entre M et Y (VanderWeele, 2015). Pour des données expérimentales, la randomisation de l'exposition assure que la relation entre A et Y et la relation entre A et M ne sont pas confondues ; l'association entre M et Y reste toutefois généralement confondue sauf si le médiateur est lui-même aussi randomisé (VanderWeele, 2015). De ce fait, peu importe la méthodologie utilisée, l'identification et l'inclusion des variables de confusion est un enjeu important de l'analyse de médiation.

Les hypothèses de l'analyse de médiation impliquent qu'il faut choisir un ensemble adéquat de variables pour obtenir une interprétation causale des effets direct et indirect. Toutefois, il peut arriver que l'on inclut plus de variables que nécessaires pour obtenir une estimation sans biais. Ainsi, il est parfois intéressant de réduire la dimension de l'ensemble des variables d'ajustement. En effet, dans le contexte de la régression, la réduction du nombre de variables peut être un objectif pertinent (Witte & Didelez, 2019), et certains chercheurs vont préférer des modèles

plus simples et plus utiles en pratique, si la réduction de ces modèles n'a pas d'effet négatif marqué sur le biais des estimateurs (Dunkler & Heinze, 2014). De plus, la sélection de variables peut servir à obtenir des estimateurs plus précis des coefficients de régression (Heinze, Wallisch, & Dunkler, 2018; Witte & Didelez, 2019). En particulier, l'exclusion de variables de bruit, de prédicteurs faibles (Heinze et al., 2018) et de variables qui prédisent uniquement la variable d'exposition (Lefebvre, Delaney, & Platt, 2008) peut diminuer la variance des estimateurs. Similairement, dans le contexte de la médiation, une étude récente a montré que l'exclusion des prédicteurs qui causent uniquement l'exposition permet de diminuer la variance des estimateurs pour les effets direct et indirect (Diop, Lefebvre, Duchaine, Laurin, & Talbot, 2021).

Contrairement aux travaux réalisés dans le domaine de la médiation, plusieurs méthodes ont été proposées pour faire de la sélection de variables pour l'effet total de l'exposition sur la variable réponse (Heinze et al., 2018; Witte & Didelez, 2019). Tel que présenté dans Witte et Didelez (2019), il existe plusieurs stratégies basées sur les données pour faire de la sélection de variables, dont deux que nous considérerons dans le présent travail. La première procédure est le changement d'estimation (*Change-in-estimate*; CIE), qui est une technique largement utilisée en épidémiologie (Talbot & Massamba, 2019). La démarche du changement d'estimation est discutée par plusieurs auteurs (Dunkler & Heinze, 2014; Maldonado & Greenland, 1993; Wang, 2007; Weng, Hsueh, Messam, & Hertz-Picciotto, 2009), en plus d'être une composante importante de l'algorithme de sélection ciblé proposé par Hosmer, Lemeshow et Sturdivant (2013) dans le cadre de la régression logistique. Une seconde approche basée sur les données est une méthode qui repose sur la différence en erreur quadratique moyenne (*Mean squared error*; ΔMSE) (Greenland, Daniel, & Pearce, 2016).

Bien que ces méthodes sont fréquemment appliquées pour faire de la sélection

de variables dans certains domaines de recherche, les conditions nécessaires pour identifier l'effet total causal de l'exposition sur la réponse diffèrent des hypothèses demandées pour obtenir des estimations causales des effets direct et indirect. De ce fait, le CIE et le ΔMSE ne sont pas nécessairement des approches optimales pour faire la sélection de variables dans le contexte de médiation. Ainsi, l'objectif principal de ce mémoire est d'adapter et d'évaluer par simulation Monte-Carlo le CIE et le ΔMSE pour faire la sélection de variables dans le cadre de la médiation par inférence causale lorsque la variable de médiation et la réponse sont continues. Un objectif secondaire de ce travail est d'offrir un exemple concret sur un jeu de données réelles de l'application des méthodes proposées, et de mettre à la disposition des chercheurs un code `R`[1] pour faire la sélection de variables dans le contexte de la médiation.

Le chapitre I présente l'état des connaissances concernant la médiation et la sélection de variables. Après avoir établi certaines notations importantes (section 1.1), ce chapitre introduit les approches classiques de médiation (section 1.2), la médiation par inférence causale (section 1.3), les fondements de la sélection de variables pour l'effet total de l'exposition sur la réponse et les méthodes CIE et ΔMSE (section 1.4). Le chapitre II discute des limites (section 2.1) des méthodes CIE et ΔMSE standards, tout en proposant des modifications à ces procédures (section 2.2), pour faire de la sélection de variables dans un contexte de médiation. Le chapitre III est consacré à la planification (section 3.1) et à la réalisation (section 3.2) d'une étude de simulation Monte-Carlo conçue pour évaluer les propriétés des algorithmes de sélection de variables proposés. Le chapitre IV illustre les méthodes de sélection de variables discutées sur des données provenant de la vague 1999 de l'étude sur l'alcool du *Harvard School of Public Health College*. Enfin, le dernier chapitre de ce mémoire conclut par une discussion générale des résultats

1. Au moment d'écrire ces lignes, un paquetage `R` est toujours en construction.

obtenus. En outre, les limites et les forces de la présente étude sont énoncées et des recommandations pour les chercheurs sont avancées.

CHAPITRE I

ÉTAT DES CONNAISSANCES

Ce chapitre présente l'état des connaissances concernant la médiation et la sélection de variables. Plus spécifiquement, après une brève présentation des notations utilisées, les méthodes classiques de la médiation, la médiation par inférence causale et les principes de la sélection de variables pour l'effet total de l'exposition sur la réponse sont détaillés. En particulier, cette dernière sous-section présente de manière détaillée deux méthodes de sélection de variables pour l'estimation de l'effet total, soit le CIE et le ΔMSE.

1.1 Notations

Pour l'ensemble de la présentation, à moins d'indication contraire, on pose A la variable d'exposition (aussi appelée variable de traitement), M la variable de médiation et Y la variable de réponse. Pour simplifier la rédaction, on prend A une variable binaire avec deux conditions, soit $A = a$ ou $A = a^*$. On considère également que $a > a^*$. Typiquement, les valeurs de A sont $a = 1$ et $a^* = 0$. De plus, on note $\boldsymbol{C}$ le vecteur de *covariables*. Dans le contexte actuel, l'appellation *covariable* fait référence à la totalité des variables qui peuvent être incluses dans le modèle de médiation, à l'exception de l'exposition, du médiateur et de la ré-

ponse. Lorsque cela est pertinent, il est possible de spécifier davantage le rôle des covariables dans le modèle de médiation. Plus précisément, on peut noter $\boldsymbol{C}_{am}$ le vecteur de variables de confusion [1] entre A et M, $\boldsymbol{C}_{ay}$ le vecteur de variables de confusion entre A et Y et $\boldsymbol{C}_{my}$ le vecteur de variables de confusion entre M et Y. Une covariable est considérée comme un facteur de confusion entre deux variables si elle est une cause commune de ces deux variables (Nguyen, Schmid, Ogburn, & Stuart, 2020). De plus, on note $\boldsymbol{B}$ le vecteur de variables qui n'est pas associé aux autres variables dans le modèle (variables de bruit), $\boldsymbol{P}_a$ le vecteur de variables qui prédit uniquement la variable A (prédicteurs purs de l'exposition), $\boldsymbol{P}_m$ le vecteur de variables qui prédit uniquement la variable M (prédicteurs purs du médiateur) et $\boldsymbol{P}_y$ le vecteur de variables qui prédit uniquement la variable Y (prédicteurs purs de la variable réponse). Enfin, on remplace les notations $\boldsymbol{P}$ et $\boldsymbol{C}$ par $\boldsymbol{U}$ pour indiquer que ces variables ne sont pas mesurées et ne peuvent donc pas être incluses dans les analyses. Les variables de l'ensemble de ces vecteurs sont notées de manière similaire, mais elles sont également indicées par leur position dans le vecteur. Par exemple, on note $\boldsymbol{P}_y^T = (P_{y_1}, P_{y_2}, \ldots, P_{y_k})$, avec k la dimension du vecteur $\boldsymbol{P}_y$. La figure 1.1 permet d'éclaircir ces notations.

1.2 Approche classique de la médiation

L'analyse de médiation remonte aux travaux de Wright (1934) mais a été popularisée par Baron et Kenny (1986), et développée ultérieurement par plusieurs chercheurs (voir Hayes (2013) et MacKinnon (2008) pour une introduction). Il est fréquent d'opérer une distinction entre les approches classiques de la médiation (méthode de Baron et Kenny, méthode du produit, méthode de la différence) et

1. On utilise également les termes variable confondante, facteur confondant et facteur de confusion pour décrire les variables de confusion.

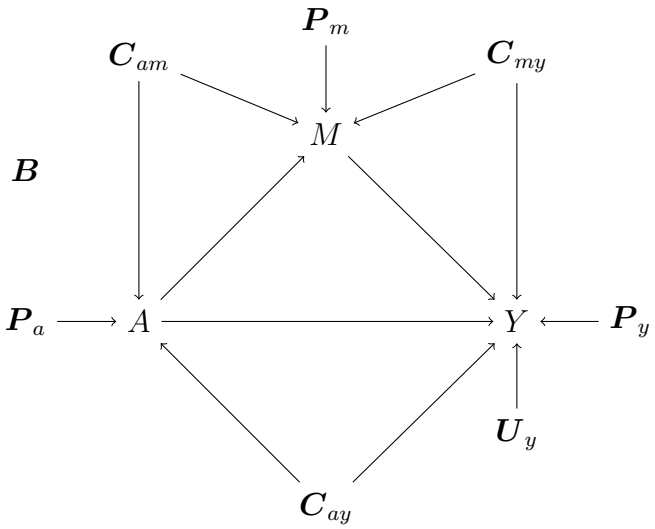


FIGURE 1.1 : Modèle de médiation et autres variables : clarification de la notation

la médiation par inférence causale. Dans les sous-sections suivantes, l'approche classique de la médiation est présentée et ses limites sont explicitées. Par la suite, une description détaillée de la médiation par inférence causale est effectuée.

1.2.1 Baron et Kenny

La première approche classique de la médiation est la méthode de Baron et Kenny (Baron & Kenny, 1986). La section qui suit est basée sur l'interprétation de Hayes (2013) de cette méthode. Supposons que le médiateur et que la variable réponse sont continus. De plus, admettons que les modèles de régression linéaire suivants correspondent aux données observées

$$\mathbb{E}[Y|A=a,\boldsymbol{C}=\boldsymbol{c}] = \gamma_0 + \gamma_1 a + \boldsymbol{\gamma}_2^T \boldsymbol{c}, \tag{1.2.1}$$

$$\mathbb{E}[M|A=a,\boldsymbol{C}=\boldsymbol{c}] = \beta_0 + \beta_1 a + \boldsymbol{\beta}_2^T \boldsymbol{c}, \tag{1.2.2}$$

$$\mathbb{E}[Y|A=a,M=m,\boldsymbol{C}=\boldsymbol{c}] = \theta_0 + \theta_1 a + \theta_2 m + \boldsymbol{\theta}_3^T \boldsymbol{c}, \tag{1.2.3}$$

avec $\boldsymbol{\theta}_3^T$ (similairement pour $\boldsymbol{\gamma}_2^T$) et $\boldsymbol{\beta}_2^T$ qui représentent, respectivement, les vecteurs lignes des coefficients de régression pour les covariables $\boldsymbol{C}$ dans le modèle pour Y et M. Selon Baron et Kenny (1986), on admet que M est un médiateur de l'effet de l'exposition sur la réponse si les trois conditions suivantes sont respectées : (1) dans le modèle 1.2.1, l'exposition est associée à la variable réponse ($\gamma_1 \neq 0$) ; (2) dans le modèle 1.2.2, l'exposition est associée au médiateur ($\beta_1 \neq 0$) ; et (3) dans le modèle 1.2.3, le médiateur est associé à la réponse, conditionnellement à l'exposition ($\theta_2 \neq 0$). La condition (1) réfère au cas où l'effet total de l'exposition est non nul, alors que les conditions (2) et (3) encodent un effet de l'exposition sur la réponse qui est explicable totalement ou partiellement par le médiateur (c'est-à-dire, l'effet indirect).

Plusieurs critiques peuvent être apportées à cette procédure. Premièrement, les trois étapes sont basées sur une série de tests de signification statistique. Or, il est possible que des associations existent entre certaines variables, même s'il n'est pas permis de rejeter l'hypothèse nulle. Cette problématique est grandement amplifiée par le fait que trois tests statistiques doivent être complétés (Hayes, 2013; VanderWeele, 2015). Deuxièmement, plusieurs méthodologistes (Hayes, 2013; MacKinnon, 2008) argumentent qu'il n'est pas nécessaire que l'effet total de l'exposition soit différent de 0. En effet, il est possible d'avoir une médiation incohérente, c'est-à-dire que l'effet direct et l'effet indirect ont une magnitude similaire, mais dans des directions opposées, ce qui réduit l'effet total vers la nulle (Hayes, 2013; VanderWeele, 2015). Enfin, la méthode de Baron et Kenny (1986) ne quantifie pas l'effet indirect. Ainsi, la présence d'un médiateur est inférée qualitativement à la suite de plusieurs tests statistiques. Conséquemment, il est difficile de comparer l'importance de médiateurs distincts, ou encore d'effectuer des analyses de médiation plus sophistiquées (par exemple, déterminer si un effet de médiation est le même pour différents groupes) (Hayes, 2013).

1.2.2 Méthode du produit

La méthode du produit, largement discutée par Hayes (2013) et MacKinnon (2008), est utilisée dans plusieurs domaines de recherche, comme la psychologie, la sociologie, l'économie, etc. Cette approche est intégrée dans les modèles d'équations structurelles (Hair, Hult, Ringle, & Sarstedt, 2016; Kline, 2015), et elle est implémentée dans plusieurs logiciels comme Mplus (Muthén & Muthén, 2017), R (paquetage *lavaan* ; Rosseel (2012)), Stata (commande *sem* ; StataCorp (2019)), etc. Reprenons les modèles 1.2.3, 1.2.2 et 1.2.1. Dans la méthode du produit, on définit l'effet indirect comme $\beta_1\theta_2$. L'idée intuitive de cette définition est que l'effet de médiation existe dans la mesure où l'exposition affecte le médiateur (effet mesuré par β_1), et que le médiateur influence la réponse, conditionnellement à l'exposition (effet mesuré par θ_2). L'effet direct est donné par θ_1, alors que l'effet total est obtenu par la somme de l'effet direct et de l'effet indirect, c'est-à-dire que l'effet total correspond à $\theta_1 + \beta_1\theta_2$.

1.2.3 Méthode de la différence

Une dernière technique de médiation qui fait partie de l'approche classique est la méthode de la différence (VanderWeele, 2015). Encore une fois, reprenons les modèles 1.2.3, 1.2.2 et 1.2.1. Comme c'est le cas pour la méthode du produit, on peut considérer que θ_1 correspond à l'effet direct de l'exposition sur la réponse. De plus, dans cette approche, on considère que γ_1 représente l'effet total de l'exposition sur la réponse. Conséquemment, on peut estimer l'effet indirect comme la différence entre ces deux quantités. Ainsi, on a que l'effet indirect est donné par $\gamma_1 - \theta_1$. Lorsque M et Y sont évalués par des modèles de régression linéaire, les méthodes du produit et de la différence obtiennent les mêmes valeurs pour les

effets direct et indirect (MacKinnon, Warsi, & Dwyer, 1995).

1.2.4 Critiques des approches classiques de la médiation

Bien que simples, ces approches classiques présentent plusieurs limites. Premièrement, il n'y a pas de distinction entre la définition des effets et la méthode d'estimation (Nguyen, Schmid, & Stuart, 2020). Conséquemment, ce que représente l'effet direct et l'effet indirect demeure incertain. Une seconde limite des approches classiques est la difficulté d'introduire un terme d'interaction entre la variable d'exposition et le médiateur dans la prédiction de la variable réponse (VanderWeele, 2015). Troisièmement, les approches classiques se généralisent généralement difficilement aux cas où la variable de médiation ou la variable réponse n'est pas continue (VanderWeele, 2015). Enfin, une des limites importantes des travaux classiques sur la médiation est l'absence d'un cadre formel de la causalité (VanderWeele, 2015). Bien que cela ne soit pas inhérent aux approches classiques, cela se traduit généralement par une omission des conditions nécessaires pour l'identification de l'effet direct et de l'effet indirect. Plus particulièrement, peu d'attention est accordée à l'importance d'ajuster adéquatement pour les variables de confusion. Pour répondre à ces difficultés, des chercheurs ont développé un modèle contrefactuel pour servir de base à la médiation par inférence causale.

1.3 Médiation et inférence causale

Plusieurs méthodes pour l'analyse de médiation ont été développées afin d'obtenir des estimations causales des effets direct et indirect (Imai, Keele, & Tingley, 2010; Lange, Vansteelandt, & Bekaert, 2012; VanderWeele, 2015). L'approche par l'inférence causale se distingue des méthodes classiques de médiation, notamment

par une définition générale et causale des effets direct et indirect, ainsi que par une clarification des postulats nécessaires à l'identification de ces effets (Nguyen, Schmid, & Stuart, 2020). Cette approche procède généralement en trois étapes, soit (1) la définition rigoureuse des divers effets, (2) l'identification des conditions nécessaires à l'évaluation des effets, et (3) par le développement d'une méthode d'estimation pour quantifier les effets (Nguyen, Schmid, Ogburn, & Stuart, 2020). Ces différentes étapes sont présentées dans les sections suivantes.

1.3.1 Effet total : définition selon le modèle contrefactuel

Le modèle contrefactuel est un cadre de travail fréquemment utilisé pour l'inférence causale. Dans cette approche, on compare des variables réponses potentielles [2] , c'est-à-dire les réponses qu'on aurait observées, potentiellement contrairement au fait, si la variable d'exposition avait pris différentes valeurs. Par exemple, supposons que la variable réponse, Y, correspond au fait d'avoir ($Y = 1$) ou de ne pas avoir ($Y = 0$) une maladie, alors que la variable d'exposition, A, représente le fait d'avoir reçu un traitement ($A = 1$) ou de ne pas l'avoir reçu ($A = 0$). Dans ce contexte, on a deux réponses potentielles pour un individu i, soit $Y_i(1)$ (la réponse si l'individu i avait reçu le traitement, pour $i = 1, 2, \ldots, N$) et $Y_i(0)$ (la réponse si l'individu i n'avait pas reçu le traitement). Plus généralement, on note $Y_i(a)$ la réponse contrefactuelle pour l'individu i s'il possédait une valeur de $A = a$ sur la variable d'exposition. Évidemment, en pratique, il n'est pas possible d'avoir accès aux deux réponses contrefactuelles pour un même individu, puisqu'on ne peut pas observer la réponse d'un individu sous plus d'un niveau de traitement sur une même période de temps donnée. Cette difficulté est connue comme le *problème*

2. Dans l'ensemble de la présentation, on utilise également l'expression *variables contrefactuelles* pour faire référence aux *variables potentielles*. Il est à noter que certains auteurs opèrent une distinction entre ces deux concepts (voir VanderWeele (2015), p.461, pour une discussion.).

fondamental de l'inférence causale (Holland, 1986, cité dans Nguyen, Schmid, Ogburn et Stuart (2020)). Avec cette conceptualisation, on peut maintenant définir l'effet total individuel et l'effet total populationnel de l'exposition sur la réponse.

Définition 1.3.1. Effet total individuel si l'exposition varie de a à a^*

$$ET_i = Y_i(a) - Y_i(a^*).$$

Définition 1.3.2. Effet total populationnel si l'exposition varie de a à a^*

$$ET = \mathbb{E}[Y(a)] - \mathbb{E}[Y(a^*)].$$

L'effet total populationnel correspond donc au changement moyen qui serait obtenu sur la réponse si le niveau de l'exposition était fixé à $A = a$ comparativement à $A = a^*$ pour l'ensemble des individus de la population. Le tableau 1.1, inspiré de Nguyen, Schmid et Stuart (2020), schématise le concept du modèle contrefactuel, avec $Y(1)$ et $Y(0)$ les variables contrefactuelles, Y et A les variables observées et ET l'effet total de l'exposition sur la réponse. Pour bien comprendre ce tableau, il est important de mentionner que l'on émet généralement une hypothèse de cohérence (VanderWeele, 2015), c'est-à-dire que $Y_i(a) = Y_i$ si $A_i = a$, et $Y_i(a^*) = Y_i$ si $A_i = a^*$. Ainsi, on suppose que la réponse observée pour un individu i avec un traitement de niveau $A_i = a$ (respectivement $A_i = a^*$) correspond à la réponse contrefactuelle $Y_i(a)$ (respectivement $Y_i(a^*)$).

1.3.2 Effets direct et indirect : définition selon le modèle contrefactuel

Dans le contexte de la médiation, on ne s'intéresse pas simplement qu'à établir l'effet total de la variable d'exposition sur la réponse. En effet, on cherche égale-

TABLEAU 1.1 : Exemple du modèle contrefactuel et de l'effet total de l'exposition sur la réponse

Individu	$Y(1)$	$Y(0)$	Y	A	ET
1	3.79	3.52	3.52	0	0.27
2	5.28	3.00	3.00	0	2.28
4	2.65	4.06	4.06	0	-1.41
7	4.43	3.49	3.49	0	0.94
10	4.11	6.42	6.42	0	-2.31
3	6.08	3.22	6.08	1	2.86
5	5.43	4.96	5.43	1	0.47
6	5.51	3.89	5.51	1	1.62
8	4.45	3.09	4.45	1	1.36
9	4.44	3.16	4.44	1	1.28
Moyenne	4.62	3.88	4.64	0.5	0.74

ment à comprendre comment l'exposition peut influencer la variable Y à travers la variable de médiation (M). Conséquemment, on doit maintenant définir deux nouveaux types de variables contrefactuelles, soit les variables contrefactuelles pour le médiateur et les variables contrefactuelles emboîtées pour la réponse. Plus spécifiquement, on note $M_i(a)$ la valeur du médiateur de l'individu i si l'exposition était fixée à $A = a$ et $M_i(a^*)$ la valeur du médiateur si l'exposition était fixée à $A = a^*$. Avec cette information, on peut maintenant définir la variable potentielle emboîtée pour la réponse, soit $Y_i(a, M_i(a^*))$, c'est-à-dire la réponse de l'individu i si l'exposition était fixée à $A = a$ et que le médiateur était fixé au niveau qu'il devrait être si la variable d'exposition était fixée à $A = a^*$. Il est maintenant pos-

sible de construire les effets direct et indirect dits *naturels*[3] à partir de contrastes des variables potentielles emboîtées de la réponse, soit $Y_i(a, M_i(a)), Y_i(a, M_i(a^*))$ et $Y_i(a^*, M_i(a^*))$. On débute par la présentation de l'effet naturel direct individuel et de l'effet naturel direct populationnel.

Définition 1.3.3. Effet naturel direct individuel si l'exposition varie de a à a^*

$$END_i = Y_i(a, M_i(a^*)) - Y_i(a^*, M_i(a^*)).$$

Définition 1.3.4. Effet naturel direct populationnel si l'exposition varie de a à a^*

$$END = \mathbb{E}[Y(a, M(a^*))] - \mathbb{E}[Y(a^*, M(a^*))].$$

L'effet naturel direct populationnel correspond donc au changement moyen qui serait obtenu sur la réponse si le niveau de l'exposition était fixé à $A = a$ comparativement à $A = a^*$ et que le médiateur était fixé, pour chaque individu, au niveau qu'il devrait être si la variable d'exposition était fixée à $A = a^*$. Ainsi, l'effet naturel direct traduit l'impact de l'exposition sur la réponse si on pouvait faire abstraction de l'effet de l'exposition sur le médiateur. Les autres quantités d'intérêt sont l'effet naturel indirect individuel et l'effet naturel indirect populationnel.

Définition 1.3.5. Effet naturel indirect individuel si l'exposition varie de a à a^*

$$ENI_i = Y_i(a, M_i(a)) - Y_i(a, M_i(a^*)).$$

---

3. La terminologie *effets naturels* réfère au fait que le médiateur n'est pas fixé à une valeur unique, mais varie à son niveau naturellement observé pour chaque individu, selon la valeur du traitement.

Définition 1.3.6. Effet naturel indirect populationnel si l'exposition varie de a à a^*

$$ENI = \mathbb{E}[Y(a, M(a))] - \mathbb{E}[Y(a, M(a^*))].$$

L'effet indirect populationnel traduit le changement moyen qui serait obtenu sur la réponse si l'exposition était fixée à $A = a$ et que le médiateur varierait du niveau qu'il devrait être si l'exposition était fixée à $A = a$ au niveau qu'il devrait être si l'exposition était fixée à $A = a^*$. Ainsi, rect exprime l'impact de l'exposition sur la réponse qui est transmis par le médiateur. Le tableau 1.2 présente le concept des variables potentielles dans le cadre de la médiation, de l'effet total, l'effet naturel direct et l'effet naturel indirect. Encore une fois, pour comprendre ce tableau, il est nécessaire de mentionner que l'on émet généralement certaines hypothèses de cohérence pour le médiateur (VanderWeele, 2015), c'est-à-dire que :

$$\begin{aligned} M_i(a) &= M_i \quad \text{si} \quad A_i = a, \\ M_i(a^*) &= M_i \quad \text{si} \quad A_i = a^*. \end{aligned}$$

De plus, on formule une hypothèse de composition, c'est-à-dire que $Y_i(a, M_i(a)) = Y_i(a)$ si $A_i = a$ et $Y_i(a^*, M_i(a^*)) = Y_i(a^*)$ si $A_i = a^*$ (VanderWeele, 2015). Une conséquence directe de cette hypothèse est que l'effet total populationnel peut se décomposer additivement comme la somme de l'effet naturel indirect populationnel et de l'effet naturel direct populationnel. En effet, on a que :

$$
\begin{aligned}
ET &= \mathbb{E}[Y(a)] - \mathbb{E}[Y(a^*)] \\
&= \mathbb{E}[Y(a, M(a))] - \mathbb{E}[Y(a^*, M(a^*))] \text{ (\textit{hypothèse de composition})} \\
&= \Big(\mathbb{E}[Y(a, M(a))] - \mathbb{E}[Y(a, M(a^*))]\Big) + \Big(\mathbb{E}[Y(a, M(a^*))] - \mathbb{E}[Y(a^*, M(a^*))]\Big) \\
&= ENI + END.
\end{aligned}
$$

TABLEAU 1.2 : Exemple du modèle contrefactuel pour l'analyse de médiation

Identification	$Y(1, M(1))$	$Y(1, M(0))$	$Y(0, M(0))$	$M(1)$	$M(0)$	Y	M	A	ET	ENI	END
1	3.79	3.94	3.52	2.22	-1.07	3.52	-1.07	0	0.27	-0.15	0.42
2	5.28	4.27	3.00	1.36	-0.22	3.00	-0.22	0	2.28	1.01	1.27
4	2.65	4.57	4.06	1.11	-0.73	4.06	-0.73	0	-1.41	-1.92	0.51
7	4.43	4.96	3.49	1.50	0.84	3.49	0.84	0	0.94	-0.53	1.47
10	4.11	4.05	6.42	0.53	1.25	6.42	1.25	0	-2.31	0.06	-2.37
3	6.08	6.06	3.22	1.40	-1.03	6.08	1.40	1	2.86	0.02	2.84
5	5.43	4.63	4.96	0.44	-0.63	5.43	0.44	1	0.47	0.80	-0.33
6	5.51	6.22	3.89	2.79	-1.69	5.51	2.79	1	1.62	-0.71	2.33
8	4.45	3.23	3.09	-0.97	0.15	4.45	-0.97	1	1.36	1.22	0.14
9	4.44	3.81	3.16	1.70	-1.14	4.44	1.70	1	1.28	0.63	0.65
Moyenne	4.62	4.57	3.88	1.21	-0.43	4.64	0.54	0.5	0.74	0.04	0.69

Il est à noter que les effets marginaux ont été présentés, mais que des versions analogues existent pour les effets conditionnels. Les effets conditionnels correspondent aux effets (naturel direct, naturel indirect et total) pour une sous-population définie par un niveau fixé de covariables. Ainsi, on a les effets conditionnels suivants :

$$
\begin{aligned}
ENI(\boldsymbol{c}) &= \mathbb{E}[Y(a, M(a))|\boldsymbol{C} = \boldsymbol{c}] - \mathbb{E}[Y(a, M(a^*))|\boldsymbol{C} = \boldsymbol{c}], \\
END(\boldsymbol{c}) &= \mathbb{E}[Y(a, M(a^*))|\boldsymbol{C} = \boldsymbol{c}] - \mathbb{E}[Y(a^*, M(a^*))|\boldsymbol{C} = \boldsymbol{c}], \\
ET(\boldsymbol{c}) &= \mathbb{E}[Y(a)|\boldsymbol{C} = \boldsymbol{c}] - \mathbb{E}[Y(a^*)|\boldsymbol{C} = \boldsymbol{c}] = ENI(\boldsymbol{c}) + END(\boldsymbol{c}).
\end{aligned}
$$

Maintenant que les définitions rigoureuses des différents effets ont été énoncées, il est possible d'établir les hypothèses nécessaires pour identifier l'effet naturel direct populationnel et l'effet naturel indirect populationnel. Pour alléger la présentation,

l'appellation populationnelle pour désigner les effets naturels direct et indirect est désormais omise.

1.3.3 Médiation causale : hypothèses et identification des effets

Dans un premier temps, posons $Y(a, m)$ la réponse si l'exposition était fixée à $A = a$ et que le médiateur était fixé à $M = m$. On émet généralement cinq hypothèses pour identifier l'effet naturel direct et l'effet naturel indirect sur la base de données observationnelles. Les quatre premières hypothèses (Nguyen, Schmid, Ogburn, & Stuart, 2020; Steen, Loeys, Moerkerke, & Vansteelandt, 2017; VanderWeele, 2015) sont

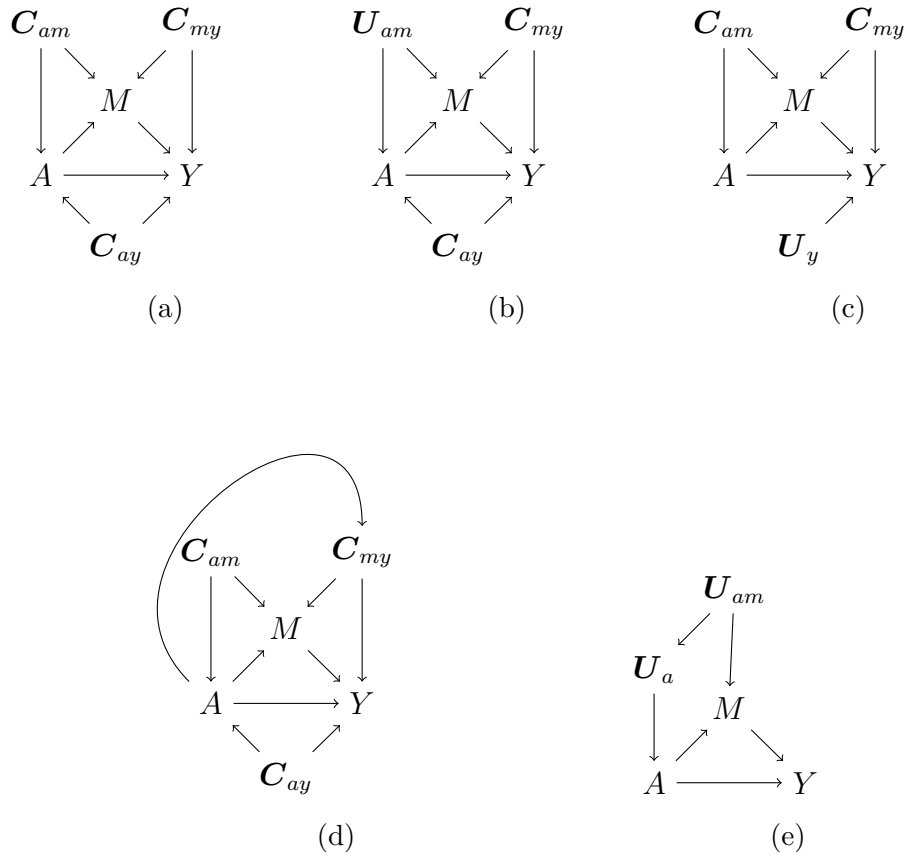
$$
\begin{aligned}
&H_1 : Y(a, m) \perp\!\!\!\perp A | \boldsymbol{C}, \text{ pour tous les niveaux de } a \text{ et } m,\\
&H_2 : M(a) \perp\!\!\!\perp A | \boldsymbol{C}, \text{ pour tous les niveaux de } a,\\
&H_3 : Y(a, m) \perp\!\!\!\perp M | \{A, \boldsymbol{C}\}, \text{ pour tous les niveaux de } a \text{ et } m,\\
&H_4 : Y(a, m) \perp\!\!\!\perp M(a^*) | \boldsymbol{C}, \text{ pour tous les niveaux de } a, a^* \text{ et } m,
\end{aligned}
$$

où le symbole $\perp\!\!\!\perp$ représente l'indépendance entre variables aléatoires. Par exemple, $M(a) \perp\!\!\!\perp A | \boldsymbol{C}$ signifie que $M(a)$ est indépendante de A, conditionnellement aux covariables $\boldsymbol{C}$. Les trois premières hypothèses se traduisent par le fait que l'on ajuste pour l'ensemble des variables de confusion entre A et Y (H_1), entre A et M (H_2) et entre M et Y (H_3). La quatrième hypothèse (H_4) est l'hypothèse la plus forte et la plus difficile à respecter. Un cas classique où H_4 n'est pas satisfaite est lorsqu'il y a une variable de confusion entre M et Y qui est causée par A[4]. Enfin, on émet généralement des hypothèses de cohérence (H_5), comme cela est

4. Il existe d'autres conditions où la quatrième hypothèse n'est pas rencontrée. Le lecteur peut se référer au travail de Andrews et Didelez (2021) pour plus d'information.

présenté dans les sections 1.3.1 et 1.3.2 (Nguyen, Schmid, Ogburn, & Stuart, 2020; VanderWeele, 2015), c'est-à-dire que $Y_i(a, M(a)) = Y_i(a) = Y_i$ si $A = a$. On doit également ajouter deux hypothèses de cohérence supplémentaires, soit que $Y_i(a, m) = Y_i$ si $A_i = a$ et $M_i = m$, ainsi que $Y_i(a, M_i(a^*)) = Y_i(a, m)$ si $M_i(a^*) = m$. Il est à noter que la randomisation de l'exposition assure le respect des hypothèses H_1 et H_2, mais pas nécessairement le respect des hypothèses H_3 et H_4. La figure 1.2 illustre cinq représentations de données, avec $\boldsymbol{C}_i$ un vecteur de variables de confusion qui est mesuré et $\boldsymbol{U}_i$ un vecteur de variables de confusion ou de prédicteurs purs qui n'est pas mesuré.

FIGURE 1.2 : Analyse de médiation et identification



Les sous-figures (a) et (c) permettent d'estimer correctement les effets naturels direct et indirect en ajustant pour les covariables $\boldsymbol{C}_{am}$ et $\boldsymbol{C}_{my}$, mais pas les sous-figures (b) (d) et (e). En effet, l'hypothèse H_4 n'est pas respectée pour la sous-figure (d), alors que c'est l'hypothèse H_2 qui n'est pas rencontrée dans les sous-figures (b) et (e). Plus précisément, la sous-figure (e) ne rencontre pas l'hypothèse H_2, car $\boldsymbol{U}_{am}$ est une cause commune du médiateur et de l'exposition à travers $\boldsymbol{U}_a$. Toutefois, si $\boldsymbol{U}_a$ était observé, alors il serait possible d'identifier les effets naturels direct et indirect en ajustant pour les variables de ce vecteur.

1.3.4 Médiation causale : formule de médiation

Supposons que $\boldsymbol{C}$ ne contient pas de covariables causées par l'exposition. Sous les hypothèses H_1-H_5, on peut dériver la formule de médiation (Pearl, Glymour, & Jewell, 2016; Steen et al., 2017; VanderWeele, 2015) qui est utilisée pour estimer les effets naturels direct et indirect. Plus spécifiquement, cette formule permet d'exprimer l'espérance contrefactuelle $\mathbb{E}[Y(a, M(a^*))|\boldsymbol{C} = \boldsymbol{c}]$ en fonction des variables observées. Les démonstrations sont effectuées avec un médiateur (M) discret.

Proposition 1.3.1. La formule de médiation est donnée par

$$\mathbb{E}[Y(a, M(a^*))|\boldsymbol{C} = \boldsymbol{c}] = \sum_m \mathbb{E}[Y|A = a, M = m, \boldsymbol{C} = \boldsymbol{c}]P(M = m|A = a^*, \boldsymbol{C} = \boldsymbol{c})$$

Démonstration. (VanderWeele, 2015)

$$\begin{aligned}
&\mathbb{E}[Y(a, M(a^*))|\boldsymbol{C} = \boldsymbol{c}] \\
&= \sum_m P(M(a^*) = m|\boldsymbol{C} = \boldsymbol{c})\, \mathbb{E}[Y(a, m)|M(a^*) = m, \boldsymbol{C} = \boldsymbol{c}] \quad (H_5) \\
&= \sum_m P(M(a^*) = m|\boldsymbol{C} = \boldsymbol{c}, A = a^*)\, \mathbb{E}[Y(a, m)|\boldsymbol{C} = \boldsymbol{c}] \quad (H_2 \textit{ et } H_4)
\end{aligned}$$

$$= \sum_m P(M = m|\boldsymbol{C} = \boldsymbol{c}, A = a^*)\, \mathbb{E}[Y(a, m)|\boldsymbol{C} = \boldsymbol{c}] \quad (H_5)$$
$$= \sum_m P(M = m|\boldsymbol{C} = \boldsymbol{c}, A = a^*)\, \mathbb{E}[Y(a, m)|A = a, \boldsymbol{C} = \boldsymbol{c}] \quad (H_1)$$
$$= \sum_m P(M = m|\boldsymbol{C} = \boldsymbol{c}, A = a^*)\, \mathbb{E}[Y(a, m)|A = a, M = m, \boldsymbol{C} = \boldsymbol{c}] \quad (H_3)$$
$$= \sum_m P(M = m|\boldsymbol{C} = \boldsymbol{c}, A = a^*)\, \mathbb{E}[Y|A = a, M = m, \boldsymbol{C} = \boldsymbol{c}] \quad (H_5).$$

□

Ainsi, bien que la quantité $\mathbb{E}[Y(a, M(a^*))|\boldsymbol{C} = \boldsymbol{c}]$ ne soit pas observable, on peut néanmoins, sous H_1-H_5, représenter cette espérance en fonction des variables observées. Avec la formule de médiation, il est maintenant possible de dériver les équations pour l'effet naturel direct et l'effet naturel indirect.

Corollaire 1.3.1. L'effet naturel direct conditionnel lorsque l'exposition varie de a à a^* est donné par

$$END(\boldsymbol{c}) = \sum_m \Big(\mathbb{E}[Y|A = a, M = m, \boldsymbol{C} = \boldsymbol{c}] - \mathbb{E}[Y|A = a^*, M = m, \boldsymbol{C} = \boldsymbol{c}] \Big) \times P(M = m|A = a^*, \boldsymbol{C} = \boldsymbol{c}).$$

Corollaire 1.3.2. L'effet naturel indirect conditionnel lorsque l'exposition varie de a à a^* est donné par

$$ENI(\boldsymbol{c}) = \sum_m \Big(P(M = m|A = a, \boldsymbol{C} = \boldsymbol{c}) - P(M = m|A = a^*, \boldsymbol{C} = \boldsymbol{c}) \Big) \times \mathbb{E}[Y|A = a, M = m, \boldsymbol{C} = \boldsymbol{c}].$$

1.3.5 Médiation causale : estimation

Il existe plusieurs méthodes d'estimation pour évaluer l'effet naturel direct et l'effet naturel indirect, comme les méthodes par simulation (Tingley, Yamamoto, Hirose, Keele, & Imai, 2014), par pseudopopulation (Steen et al., 2017), par régression (VanderWeele, 2015), etc. Dans le cadre du présent projet, on considère uniquement l'approche par régression comme méthode d'estimation. En effet, cette approche est très utilisée en pratique, et elle possède une forme fonctionnelle fermée dans certains contextes, incluant lorsque M et Y sont des variables continues modélisées par régression linéaire. Enfin, bien que l'ensemble de ces méthodes émettent les hypothèses H_1-H_5, les répercussions de l'ajustement (ou l'absence d'ajustement) de certaines variables sur la variance des estimateurs des effets naturels direct et indirect est généralement mieux compris pour les techniques de régression classique.

1.3.5.1 Médiation causale : estimation par régression des effets naturels direct et indirect conditionnels

Supposons que les variables Y et M sont continues, alors que A est une variable binaire. Également, prenons $\boldsymbol{C}$ des variables continues ou binaires. De plus, supposons que les modèles de régression linéaire suivants sont correctement spécifiés, c'est-à-dire que :

$$\mathbb{E}[M|A=a,\boldsymbol{C}=\boldsymbol{c}] = \beta_0 + \beta_1 a + \boldsymbol{\beta}_2^T \boldsymbol{c}, \tag{1.3.1}$$

$$\mathbb{E}[Y|A=a,M=m,\boldsymbol{C}=\boldsymbol{c}] = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \boldsymbol{\theta}_4^T \boldsymbol{c}. \tag{1.3.2}$$

Dans ce contexte, il est possible de trouver des formes analytiques pour les dif-

férents effets conditionnels. Des expressions fermées des effets naturels direct et indirect conditionnels basées sur le modèle de régression du médiateur 1.3.1 et de la réponse 1.3.2 lorsque l'exposition varie de $A = a$ à $A = a^*$ sont présentées, respectivement, par la formule paramétrique 1.3.1 et 1.3.2. Pour simplifier la notation des démonstrations, on note $g(m|a^*, \boldsymbol{c})$ la densité de M conditionnelle à $A = a^*$ et $\boldsymbol{C} = \boldsymbol{c}$.

Formule paramétrique 1.3.1. L'effet naturel direct conditionnel obtenu par régression lorsque l'exposition varie de a à a^* est donné par

$$END(\boldsymbol{c}) = \Big(\theta_1 + \theta_3(\beta_0 + \beta_1 a^* + \boldsymbol{\beta}_2^T \boldsymbol{c})\Big)\,(a - a^*)\,.$$

Démonstration. (VanderWeele, 2015)

$$\begin{aligned}
END(\boldsymbol{c}) &= \mathbb{E}[Y(a, M(a^*))|\boldsymbol{C} = \boldsymbol{c}] - \mathbb{E}[Y(a^*, M(a^*))|\boldsymbol{C} = \boldsymbol{c}] \\
&= \int \Big(\mathbb{E}[Y|A = a, M = m, \boldsymbol{C} = \boldsymbol{c}] - \mathbb{E}[Y|A = a^*, M = m, \boldsymbol{C} = \boldsymbol{c}]\Big) \\
&\qquad \times g(m|a^*, \boldsymbol{c})dm \\
&= \int \Big(\theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \boldsymbol{\theta}_4^T \boldsymbol{c} - \theta_0 - \theta_1 a^* - \theta_2 m - \theta_3 a^* m - \boldsymbol{\theta}_4^T \boldsymbol{c}\Big) \\
&\qquad \times g(m|a^*, \boldsymbol{c})dm \\
&= \int \Big(\theta_1 a + \theta_3 am - \theta_1 a^* - \theta_3 a^* m\Big) g(m|a^*, \boldsymbol{c})dm \\
&= \theta_1(a - a^*) \int g(m|a^*, \boldsymbol{c})dm + \theta_3(a - a^*) \int m g(m|a^*, \boldsymbol{c})dm \\
&= \theta_1(a - a^*) + \theta_3(a - a^*)\,\mathbb{E}[M|A = a^*, \boldsymbol{C} = \boldsymbol{c}] \\
&= \theta_1(a - a^*) + \theta_3(a - a^*)(\beta_0 + \beta_1 a^* + \boldsymbol{\beta}_2^T \boldsymbol{c}) \\
&= \Big(\theta_1 + \theta_3(\beta_0 + \beta_1 a^* + \boldsymbol{\beta}_2^T \boldsymbol{c})\Big)\,(a - a^*)\,.
\end{aligned}$$

□

Formule paramétrique 1.3.2. L'effet naturel indirect conditionnel obtenu par régression lorsque l'exposition varie de a à a^* est donné par

$$ENI(\boldsymbol{c}) = (\theta_2\beta_1 + \theta_3\beta_1 a)\,(a - a^*)\,.$$

Démonstration. (VanderWeele, 2015)

$$\begin{aligned}
ENI(\boldsymbol{c}) =& \mathbb{E}[Y(a, M(a))|\boldsymbol{C} = \boldsymbol{c}] - \mathbb{E}[Y(a, M(a^*))|\boldsymbol{C} = \boldsymbol{c}] \\
=& \int \mathbb{E}[Y|A = a, M = m, \boldsymbol{C} = \boldsymbol{c}]\Big(g(m|a, \boldsymbol{c}) - g(m|a^*, \boldsymbol{c})\Big)dm \\
=& \int \Big(\theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \boldsymbol{\theta}_4^T\boldsymbol{c}\Big)g(m|a, \boldsymbol{c})dm \\
& - \int \Big(\theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \boldsymbol{\theta}_4^T\boldsymbol{c}\Big)g(m|a^*, \boldsymbol{c})dm \\
=& \Big(\theta_0 + \theta_1 a + \boldsymbol{\theta}_4^T\boldsymbol{c}\Big) + (\theta_2 + \theta_3 a)\int mg(m|a, \boldsymbol{c})dm \\
& - \Big(\theta_0 + \theta_1 a + \boldsymbol{\theta}_4^T\boldsymbol{c}) - (\theta_2 + \theta_3 a)\int mg(m|a^*, \boldsymbol{c})dm \\
=& (\theta_2 + \theta_3 a)\,\mathbb{E}[M|A = a, \boldsymbol{C} = \boldsymbol{c}] - (\theta_2 + \theta_3 a)\,\mathbb{E}[M|A = a^*, \boldsymbol{C} = \boldsymbol{c}] \\
=& (\theta_2 + \theta_3 a)\Big[(\beta_0 + \beta_1 a + \boldsymbol{\beta}_2^T\boldsymbol{c}) - (\beta_0 + \beta_1 a^* + \boldsymbol{\beta}_2^T\boldsymbol{c})\Big] \\
=& (\theta_2 + \theta_3 a)\Big[\beta_1(a - a^*)\Big] \\
=& \ (\theta_2\beta_1 + \theta_3\beta_1 a)\,(a - a^*)\,.
\end{aligned}$$

□

Il y a deux éléments à remarquer. Dans un premier temps, lorsque l'effet d'interaction (θ_3) n'est pas nul, l'effet naturel direct dépend de la valeur de conditionnement des covariables. En règle générale, à moins de vouloir calculer un effet naturel direct pour des valeurs particulières de certaines covariables, on prend $\boldsymbol{c} = (\bar{c}_1, \bar{c}_2, \ldots, \bar{c}_p)$, avec $\bar{c}_i = \mathbb{E}[C_i]$. Dans le cas présent, cela revient à évaluer l'effet naturel direct marginal (voir Annexe A). Dans un second temps, on peut noter que si le terme d'interaction est nul, alors on retrouve les résultats de la méthode

du produit. De ce fait, pour le cas d'un modèle linéaire pour le médiateur et la réponse, on peut considérer la méthode du produit comme un cas particulier de la médiation par inférence causale.

Il est maintenant possible de proposer des estimateurs pour $END(\boldsymbol{c})$, $ENI(\boldsymbol{c})$ et $ET(\boldsymbol{c})$, soit

$$\begin{aligned}
\widehat{END}(\boldsymbol{c}) &= \left(\hat{\theta}_1 + \hat{\theta}_3(\hat{\beta}_0 + \hat{\beta}_1 a^* + \hat{\boldsymbol{\beta}}_2^T \boldsymbol{c})\right)(a - a^*), \\
\widehat{ENI}(\boldsymbol{c}) &= (\hat{\theta}_2\hat{\beta}_1 + \hat{\theta}_3\hat{\beta}_1 a)(a - a^*), \\
\widehat{ET}(\boldsymbol{c}) &= \widehat{END}(\boldsymbol{c}) + \widehat{ENI}(\boldsymbol{c}),
\end{aligned}$$

avec $\hat{\beta}_0$, $\hat{\beta}_1$ et $\hat{\boldsymbol{\beta}}_2^T$ les estimateurs des coefficients de régression pour le modèle 1.3.1, et $\hat{\theta}_1$, $\hat{\theta}_2$ et $\hat{\theta}_3$ les estimateurs des coefficients de régression pour le modèle 1.3.2.

1.3.5.2 Médiation causale : variance des estimateurs obtenus par régression des effets naturels direct et indirect conditionnels

On peut utiliser la méthode delta pour obtenir la variance des estimateurs, soit $V[\widehat{END}(\boldsymbol{c})]$, $V[\widehat{ENI}(\boldsymbol{c})]$ et $V[\widehat{ET}(\boldsymbol{c})]$. Selon cette procédure, la variance des effets naturels direct et indirect est obtenue par

$$\Gamma^T \Sigma \Gamma,$$

où Γ est le vecteur colonne des dérivées partielles de $ENI(\boldsymbol{c})$ ou $END(\boldsymbol{c})$ selon $(\beta_0, \beta_1, \boldsymbol{\beta}_2^T, \theta_0, \theta_1, \theta_2, \theta_3, \boldsymbol{\theta}_4^T)$, et Σ est la matrice de covariance des coefficients. Pour l'effet naturel direct, on a

$$\Gamma^T = \left[\frac{\partial END(\boldsymbol{c})}{\partial \boldsymbol{\beta}^T}, \frac{\partial END(\boldsymbol{c})}{\partial \boldsymbol{\theta}^T} \right]$$

$$= \left[\theta_3, \theta_3 a^*, \theta_3 \boldsymbol{c}^T, 0, 1, 0, \beta_0 + \beta_1 a^* + \boldsymbol{\beta}_2^T \boldsymbol{c}, \mathbf{0}^T \right] (a - a^*),$$

avec $\mathbf{0}$ un vecteur colonne nul de dimension égale au nombre de covariables, $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \boldsymbol{\beta}_2^T)$, et $\boldsymbol{\theta}^T = (\theta_0, \theta_1, \theta_2, \theta_3, \boldsymbol{\theta}_4^T)$. Similairement, pour l'effet naturel indirect, on a

$$\Gamma^T = \left[\frac{\partial ENI(\boldsymbol{c})}{\partial \boldsymbol{\beta}^T}, \frac{\partial ENI(\boldsymbol{c})}{\partial \boldsymbol{\theta}^T} \right]$$

$$= \left[0, \theta_2 + \theta_3 a, \mathbf{0}^T, 0, 0, \beta_1, \beta_1 a, \mathbf{0}^T \right] (a - a^*).$$

Dans les deux cas, on a

$$\Sigma = \begin{pmatrix} \Sigma_{\boldsymbol{\beta}} & \Sigma_{\boldsymbol{\beta\theta}} \\ \\ \Sigma_{\boldsymbol{\beta\theta}} & \Sigma_{\boldsymbol{\theta}} \end{pmatrix}.$$

Proposition 1.3.2. La covariance entre les estimateurs $\hat{\boldsymbol{\beta}}$ et $\hat{\boldsymbol{\theta}}$ est donnée par

$$Cov(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}) = \Sigma_{\boldsymbol{\beta\theta}} = \mathbf{0}.$$

Démonstration. (VanderWeele, 2015)
Posons $\boldsymbol{Z} = (M, A, \boldsymbol{C})$. Selon la loi de la covariance totale, on a que :

$$\begin{aligned}
\mathrm{Cov}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}) =& \mathbb{E}[\mathrm{Cov}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})|\boldsymbol{Z} = \boldsymbol{z}] + \mathrm{Cov}(\mathbb{E}[\hat{\boldsymbol{\beta}}|\boldsymbol{Z} = \boldsymbol{z}], \mathbb{E}[\hat{\boldsymbol{\theta}}|\ \boldsymbol{Z} = \boldsymbol{z}]) \\
=& \mathbf{0} + \mathrm{Cov}(\mathbb{E}[\hat{\boldsymbol{\beta}}|\boldsymbol{Z} = \boldsymbol{z}], \mathbb{E}[\hat{\boldsymbol{\theta}}|\ \boldsymbol{Z} = \boldsymbol{z}]) \quad (\hat{\boldsymbol{\beta}} \text{ et } \hat{\boldsymbol{\theta}} \text{ dépendent uniquement de } \boldsymbol{Z}) \\
=& \mathrm{Cov}(\boldsymbol{\beta}, \boldsymbol{\theta}) \\
=& \mathbf{0}.
\end{aligned}$$

□

Ainsi, on a que

$$\Sigma = \begin{pmatrix} \Sigma_{\beta} & \mathbf{0} \\ \mathbf{0} & \Sigma_{\boldsymbol{\theta}} \end{pmatrix}.$$

Le calcul de la variance pour l'effet total s'obtient directement en remarquant que l'effet total est la somme des effets naturels direct et indirect, et que la dérivée d'une somme est la somme des dérivées. Les estimateurs $\hat{V}[\widehat{END}(\boldsymbol{c})]$, $\hat{V}[\widehat{ENI}(\boldsymbol{c})]$ et $\hat{V}[\widehat{ET}(\boldsymbol{c})]$ sont obtenus en remplaçant les paramètres $\boldsymbol{\beta}^T$ et $\boldsymbol{\theta}^T$ par leur estimateur respectif calculé par des modèles de régression linéaire. Enfin, si on pose $Z_{1-\alpha/2}$ le quantile d'ordre $1-\alpha/2$ d'une loi normale centrée et réduite, alors l'intervalle de confiance pour $\widehat{END}(\boldsymbol{c})$ (similairement pour $\widehat{ENI}(\boldsymbol{c})$ et $\widehat{ET}(\boldsymbol{c})$) est donné par $\widehat{END}(\boldsymbol{c}) \pm \sqrt{\hat{V}[\widehat{END}(\boldsymbol{c})]} Z_{1-\alpha/2}$.

1.4 Sélection de variables

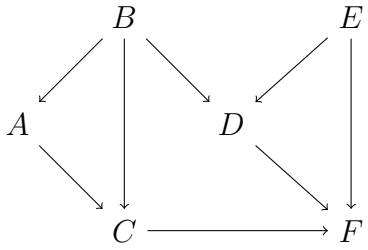
La sélection de variables est une étape importante dans l'analyse causale, autant pour évaluer l'effet total de l'exposition sur la réponse que pour identifier les

effets naturels direct et indirect. Plusieurs des méthodes de sélection de variables procèdent en deux étapes, soit : (1) l'identification d'un ensemble de covariables qui permet l'interprétation causale des effets et (2) la diminution de la dimension de cet ensemble, tout en s'assurant que l'ensemble réduit est toujours adéquat pour obtenir une interprétation causale. Ces deux étapes sont distinctes et requièrent des approches différentes. Plus spécifiquement, alors que la première phase est souvent réalisée à l'aide de connaissances avancées sur le sujet étudié et de graphes orientés acycliques (*directed acyclic graph* ; DAG), la seconde étape s'accompagne généralement de divers outils statistiques. Puisque la majorité des travaux ne portent pas spécifiquement sur la sélection de variables en contexte de médiation, les concepts importants en lien avec la sélection de variables concernant l'effet total de l'exposition sur la réponse sont discutés dans les sous-sections suivantes.

1.4.1 Graphe orienté acyclique

On présente maintenant quelques notions fondamentales des DAGs qui sont nécessaires à la compréhension de la sélection de variables en inférence causale. À moins d'indication contraire, les informations qui suivent proviennent de Lewis et Kuerbis (2016). La figure 1.3, tirée de Lewis et Kuerbis (2016), illustre un graphe orienté acyclique.

FIGURE 1.3 : Graphe orienté acyclique



Un graphe orienté est un ensemble de *sommets* reliés par des *flèches*. Dans le

cadre de l'inférence causale, les sommets correspondent à des variables, alors que les flèches traduisent la relation causale entre les variables. Par exemple, la relation $X \to Y$ signifie que la variable X cause la variable Y. On dit que X est le *parent* de Y, alors que Y est l'*enfant* de X (Pearl et al., 2016). Un *chemin* correspond à une succession de variables reliées par des flèches. Un chemin est dit *orienté* si la totalité des flèches dans le chemin sont dans la direction $\boxed{\to}$. Inversement, un chemin est dit *non orienté* s'il existe au moins une flèche dans le chemin qui est dans la direction $\boxed{\leftarrow}$. Par exemple, dans la figure 1.3, le chemin $B \to A \to C \to F$ est orienté, alors que le chemin $B \to D \leftarrow E \to F$ est non orienté. Une variable Y est une *descendante* de X s'il existe un chemin orienté qui débute par X et qui se termine à Y. De plus, on dit que les variables, à l'exception du premier et du dernier sommet, *interceptent* un chemin. Un *cycle* correspond à un chemin orienté où une variable est à la fois le sommet de départ et le sommet d'arrivé. Une des contraintes importantes des DAGs est qu'ils ne peuvent pas contenir de cycles ; ils sont donc acycliques.

Un *chemin porte arrière* de X à Y est un chemin entre X et Y où il y a une flèche qui pointe sur X (Greenland, Pearl, & Robins, 1999). Par exemple, les chemins $C \leftarrow B \to D \to F$ et $C \leftarrow B \to D \leftarrow E \to F$ sont des chemins porte arrière de C à F, alors que le chemin $B \to D \leftarrow E \to F$ n'est pas un chemin porte arrière. Toutefois, la succession de variables $D \leftarrow E \to F$ forme un chemin porte arrière. Une variable est considérée comme un *collisionneur* dans un chemin s'il y a deux flèches qui pointent sur celle-ci. Ainsi, par exemple, D est un collisionneur dans le chemin $C \leftarrow B \to D \leftarrow E \to F$. Un chemin est dit *fermé* ou *bloqué* s'il y a un collisionneur qui est présent dans le chemin, mais il sera considéré comme *ouvert* ou *non bloqué* dans le cas contraire. De ce fait, le chemin porte arrière $C \leftarrow B \to D \to F$ est ouvert, alors que le chemin porte arrière $C \leftarrow B \to D \leftarrow E \to F$ est bloqué. Statistiquement, cela se traduit par

la présence d'une association entre C et F dans le chemin porte arrière ouvert $C \leftarrow B \rightarrow D \rightarrow F$, et par une absence d'association entre ces mêmes variables dans le chemin porte arrière fermé $C \leftarrow B \rightarrow D \leftarrow E \rightarrow F$. Avec l'ensemble de ces informations, il est désormais possible de proposer une définition graphique d'un chemin de confusion. Un *chemin de confusion* correspond aux sommets qui interceptent un chemin porte arrière ouvert. Par exemple, les variables B et D forment un chemin de confusion dans le chemin $C \leftarrow B \rightarrow D \rightarrow F$. Dans ce contexte, B est une cause commune de C (directement) et de F, indirectement à travers D.

Enfin, mentionnons qu'il est possible de fermer ou d'ouvrir un chemin en ajustant (en conditionnant) sur certaines variables. Plus précisément, on peut fermer un chemin en ajustant sur une variable qui intercepte un chemin si cette dernière n'est pas un collisionneur ou une descendante d'un collisionneur, alors qu'on peut ouvrir un chemin en ajustant sur un collisionneur ou un descendant d'un collisionneur (Etminan, Collins, & Mansournia, 2020). Par exemple, le chemin porte arrière $C \leftarrow B \rightarrow D \rightarrow F$ sera fermé si on ajuste sur au moins une des variables du chemin de confusion, c'est-à-dire B ou D. Inversement, le chemin porte arrière $C \leftarrow B \rightarrow D \leftarrow E \rightarrow F$ sera débloqué si on conditionne sur le collisionneur D. La sous-section qui suit discute des principes de la sélection de variables, ainsi que des caractéristiques qu'un ensemble d'ajustement doit posséder pour permettre une interprétation causale de la relation entre l'exposition et la réponse.

1.4.2 Sélection d'un ensemble suffisant pour l'effet total de l'exposition sur la réponse

Pour simplifier la notation, prenons Y une variable réponse binaire, A une variable d'exposition binaire et $\mathbf{C}^*$ un ensemble de covariables qui est mesuré. De plus,

notons $\mathbf{C}$ un sous-ensemble de $\mathbf{C}^*$. Enfin, inspiré des travaux de Pearl, Glymour et Jewell (2016), notons $do(Z = z)$ une variable Z qui est fixée au niveau z par une intervention (par exemple, étude expérimentale). Ainsi, $Y|do(Z = z)$ correspond à la variable réponse si on intervient pour imposer une valeur de $Z = z$ à l'ensemble de la population. En conséquence, l'effet total de l'intervention sur la réponse lorsque Z varie de z à z^* est donné par $P(Y|do(Z = z)) - P(Y|do(Z = z^*))$.

Les méthodes de sélection de variables dans le contexte de l'inférence causale reposent généralement sur trois hypothèses (Witte & Didelez, 2019). Dans un premier temps, on suppose que l'intervention sur la variable d'exposition n'affecte pas la distribution des covariables :

Hypothèse 1.4.1. Covariables de prétraitement

$$P(\mathbf{C}|do(A = 1)) = P(\mathbf{C}|do(A = 0)) = P(\mathbf{C}).$$

Si on utilise les DAGs, cette hypothèse revient à supposer que chaque élément de $\mathbf{C}$ est un non-descendant de A. Une deuxième hypothèse est que, conditionnellement à $\mathbf{C}$, l'association entre A et Y obtenue d'étude observationnelle peut être interprétée causalement. Plus précisément, on a que :

Hypothèse 1.4.2. Échangeabilité conditionnelle

$$P(Y|do(A = a), \mathbf{C}{=}\mathbf{c}) = P(Y|A = a, \mathbf{C}{=}\mathbf{c}), \text{ pour } a = 0, 1.$$

Si on utilise les DAGs, cela se traduit par le fait que tous les chemins portes arrières de A à Y sont bloqués par $\mathbf{C}$. Enfin, on suppose également que l'ensemble des valeurs des covariables est présent pour $A = 1$ et $A = 0$. Plus formellement, on émet l'hypothèse de positivité, c'est-à-dire que :

Hypothèse 1.4.3. Positivité

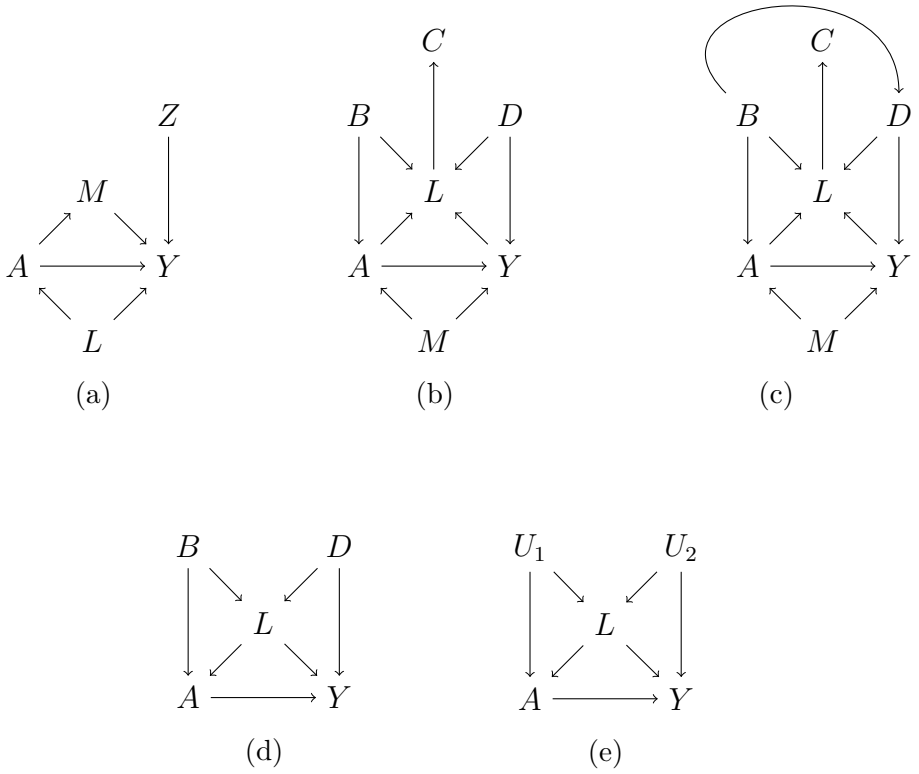
$P(A = a|\mathbf{C}{=}\mathbf{c}) > 0$, pour $a = 0, 1$ et pour toutes les valeurs possibles de $\mathbf{c}$.

Un sous-ensemble $\boldsymbol{C} \subseteq \mathbf{C}^*$ qui rencontre les trois hypothèses est un ensemble d'ajustement qui est dit *suffisant*. Dans les modèles de régression, le coefficient associé à l'exposition s'interprète comme l'effet total de l'exposition sur la réponse si on ajuste pour un ensemble de covariables suffisant. Un ensemble suffisant est considéré comme *globalement minimal* si sa cardinalité est la plus petite parmi tous les autres ensembles suffisants. Similairement, on dit que $\boldsymbol{C}$ est *localement minimal* si aucun sous-ensemble strict de $\mathbf{C}$ n'est suffisant. Un ensemble qui est globalement minimal est également localement minimal, mais l'inverse n'est pas nécessairement vrai. La figure 1.4 offre quelques exemples de DAGs. Pour les cinq représentations, on suppose que l'hypothèse de positivité est respectée.

Pour la sous-figure (a), les ensembles suffisants sont $\{\{L\}, \{L, Z\}\}$. En effet, L doit nécessairement être inclus, car son omission ouvre le chemin porte arrière $A \leftarrow L \rightarrow Y$. La variable M ne peut pas appartenir à l'ensemble $\boldsymbol{C}$, car elle est une descendante de A. Il n'y a pas de restriction sur Z. En effet, Z n'est pas un descendant de A. De plus, une fois que L est inclus, l'association conditionnelle entre A et Y s'interprète causalement indépendamment de l'inclusion ou non de la variable Z dans $\boldsymbol{C}$. L'ensemble globalement minimal est $\{L\}$.

Pour la sous-figure (b), la variable M doit obligatoirement être incluse pour fermer le chemin porte arrière $A \leftarrow M \rightarrow Y$. Les variables L et C sont des descendantes de A et ne peuvent pas être incluses dans $\mathbf{C}$. Plus spécifiquement, L est un collisionneur ($A \rightarrow L \leftarrow Y$) et C est un descendant du collisionneur L. Ainsi, les ensembles suffisants sont $\{\{M\}, \{M, B\}, \{M, D\}, \{M, B, D\}\}$, alors que l'ensemble globalement minimal est $\{M\}$.

FIGURE 1.4 : Exemples de graphes orientés acycliques et ensembles suffisants



La sous-figure (c) est similaire à la sous-figure (b), mais on a rajouté la relation $B \to D$. Conséquemment, on doit certainement inclure B et/ou D pour fermer le chemin porte arrière $A \leftarrow B \to D \to Y$. Par suite, les ensembles suffisants sont $\{\{M, B\}, \{M, D\}, \{M, B, D\}\}$. Pour ce cas, les ensembles globalement minimaux sont $\{\{M, B\}, \{M, D\}\}$.

La sous-figure (d) est légèrement plus compliquée. En effet, il est nécessaire d'inclure L dans $\boldsymbol{C}$, car L est la seule covariable qui intercepte, et donc peut fermer, le chemin porte arrière $A \leftarrow L \to Y$. Or, puisque L est un collisionneur ($B \to L \leftarrow D$), conditionner sur L ouvre le chemin porte arrière $A \leftarrow B \to L \leftarrow D \to Y$. Ainsi, il est impératif de fermer ce chemin en incluant B et/ou D. Par conséquent, les ensembles suffisants sont $\{\{L, M, B\}, \{L, M, D\}, \{L, M, B, D\}\}$, alors que les

ensembles globalement minimaux sont $\{\{L, M, B\}, \{L, M, D\}\}$.

Enfin, la sous-figure (e) est identique à la sous-figure (d), mais U_1 et U_2 sont des variables qui ne sont pas mesurées. De ce fait, on ne peut pas trouver un ensemble suffisant.

Il n'y a pas de méthodes statistiques pour déterminer un ensemble $\mathbf{C}^*$ de covariables suffisant. En pratique, les chercheurs doivent utiliser les théories existantes pour construire le DAG le plus approprié et déterminer l'ensemble de covariables suffisant. Cela dépasse l'objectif du présent projet, mais le lecteur peut se référer aux travaux suivants pour des exemples de construction de DAGs (Etminan et al., 2020; Shrier & Platt, 2008).

1.4.3 Sélection de variables : méthode de réduction de la dimension

Le CIE et le ΔMSE ont pour objectif de réduire la dimension d'un ensemble $\boldsymbol{C}^*$, avec la contrainte que l'ensemble réduit, $\boldsymbol{C} \subseteq \boldsymbol{C}^*$, est suffisant. Une des hypothèses fondamentales de ces deux méthodes est que $\boldsymbol{C}^*$ est lui-même un ensemble qui est suffisant (Witte & Didelez, 2019). Conséquemment, dans un modèle qui ajuste pour les variables de $\boldsymbol{C}^*$, on suppose que l'estimateur de l'effet total (causal) de l'exposition sur la réponse est sans biais.

1.4.3.1 Changement d'estimation

Le CIE est une méthode basée sur les données qui est fréquemment employée en épidémiologie pour faire de la sélection de variables (Talbot & Massamba, 2019). Il existe plusieurs variantes de cette approche. Posons $\hat{\theta}_{\boldsymbol{c}}$ l'estimateur du coefficient de régression de la variable d'exposition lorsque les covariables d'un

ensemble d'ajustement $\boldsymbol{C}$ sont présentes, et $\hat{\theta}_{\boldsymbol{c}_{(-i)}}$ l'estimateur du coefficient de régression de la variable d'exposition lorsque la $i^{ème}$ covariable (C_i) de l'ensemble d'ajustement est omise. De plus, notons p la cardinalité de $\boldsymbol{C}$. Dans une des versions classiques du CIE, on débute avec l'ensemble suffisant $\boldsymbol{C}^*$ et on pose $\boldsymbol{C} = \boldsymbol{C}^*$. Par la suite, on obtient des estimations $\hat{\theta}_{\boldsymbol{c}}$ et $\hat{\theta}_{\boldsymbol{c}_{(-i)}}$ pour $i = 1, 2, \ldots, p$. Subséquemment, on définit une mesure de l'écart entre $\hat{\theta}_{\boldsymbol{c}}$ et $\hat{\theta}_{\boldsymbol{c}_{(-i)}}$, que l'on note $g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-i)}})$, et on trouve la valeur minimale de $g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-i)}})$ pour $i = 1, 2, \ldots, p$. Si cette distance est plus petite qu'un certain seuil, τ, on retire la variable C_j qui a engendré la valeur minimale de $g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-i)}})$, et on recommence le processus avec $\boldsymbol{C} \setminus C_j$. Le CIE se termine si la totalité des variables de l'ensemble initial $\boldsymbol{C}^*$ a été retirée, ou encore si la valeur minimale des $g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-i)}})$ $(i = 1, 2, \ldots, p)$ pour une itération donnée du CIE est supérieure ou égale à τ. Dans le cas de la régression linéaire, il est fréquent d'utiliser

$$g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-i)}}) = \left| \frac{\hat{\theta}_{\boldsymbol{c}_{(-i)}} - \hat{\theta}_{\boldsymbol{c}}}{\hat{\theta}_{\boldsymbol{c}}} \right| < \tau = .1. \tag{1.4.1}$$

En effet, puisque $\boldsymbol{C}$ est un ensemble suffisant, on a que $\mathbb{E}[\hat{\theta}_{\boldsymbol{c}}] = \theta$ [5], et on suppose que $\hat{\theta}_{\boldsymbol{c}} \approx \theta$. Par suite, la fonction $g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-i)}})$ proposée en 1.4.1 peut être conceptualisée comme un estimateur du biais relatif de l'effet total de l'exposition sur la réponse lorsque l'on omet une variable de l'ensemble suffisant $\boldsymbol{C}$. De ce fait, le CIE présenté en 1.4.1 tente de conserver les variables dont l'omission occasionne un biais supérieur à .1, ce qui est généralement le biais relatif maximal qui est jugé acceptable. La procédure du CIE est donnée par l'algorithme 1.

5. Ici, θ représente l'effet total causal de l'exposition sur la réponse.

Algorithme 1 : Changement d'estimation (CIE)

Résultat : $\boldsymbol{C}$
Définir $g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-i)}})$;
Définir τ ;
Définir A, Y et $\boldsymbol{C}^* = \{C_1, C_2, \ldots, C_p\}$;
$\boldsymbol{C} \leftarrow \boldsymbol{C}^*$;
$p \leftarrow card(\boldsymbol{C})$;
Condition $\leftarrow$ Vraie ;
tant que *Condition* **faire**
  Estimer $\hat{y} = \hat{\theta}_{\boldsymbol{c}} a + \hat{\theta}_0 + \hat{\theta}_1 c_1 + \hat{\theta}_2 c_2 + \cdots + \hat{\theta}_p c_p$;
  $i \leftarrow p$;
  tant que $i > 0$ **faire**
    $\hat{y} = \hat{\theta}_{\boldsymbol{c}_{(-i)}} a + \hat{\theta}^*_0 + \hat{\theta}^*_1 c_1 + \hat{\theta}^*_2 c_2 + \cdots + \hat{\theta}^*_{(i-1)} c_{(i-1)} + \hat{\theta}^*_{(i+1)} c_{(i+1)} + \cdots + \hat{\theta}^*_p c_p$;
    Calculer $g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-i)}})$;
    $i \leftarrow i - 1$;
  fin
  $M \leftarrow min(\{g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-1)}}), g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-2)}}), \ldots, g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-p)}})\})$;
  Déterminer C_j tel que $M = g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-j)}})$;
  $p \leftarrow p - 1$;
  si $M \geq \tau$ *ou* $p = 0$ **alors**
    Condition = Faux ;
  fin
  sinon
    $\boldsymbol{C} \leftarrow \boldsymbol{C} \setminus C_j$;
  fin
fin

1.4.3.2 Changement basé sur l'erreur quadratique moyenne

Une des problématiques du CIE est que cette méthode ne vise que l'estimation sans biais de l'effet total de la réponse, ce qui mène habituellement au rejet de prédicteurs purs de la réponse (Vansteelandt, Bekaert, & Claeskens, 2010). Ainsi, on peut affirmer que l'objectif du CIE est d'obtenir un sous-ensemble $\boldsymbol{C}$ qui est minimal, ce qui n'est pas toujours souhaitable lorsqu'un des objectifs de la sélec-

tion de variables est la réduction de la variance des estimateurs. Un exemple de ce phénomène est donné dans la sous-figure (a) de la figure 1.4. Dans ce contexte, ajuster sur l'ensemble suffisant $\{L, Z\}$ qui n'est pas minimal pourrait permettre d'obtenir une erreur standard du coefficient de l'exposition sur la réponse inférieure à ce qui serait obtenu si on ajustait sur l'ensemble minimal $\{L\}$. En effet, il est généralement reconnu, en régression linéaire, que l'inclusion de prédicteurs purs de la réponse réduit la variance de l'estimation du coefficient de l'exposition sur la réponse (Robinson & Jewell, 1991). Or, puisque $\{L\}$ est un ensemble suffisant, on ne devrait pas observer un changement important dans l'estimation de l'effet total de l'exposition lorsque Z est exclu. Ainsi, le CIE ne devrait pas conserver la variable Z.

Une alternative au CIE qui vise également à diminuer l'incertitude dans l'estimation de l'effet total est le ΔMSE proposé par Greenland, Daniel et Pearce (2016). Dans cette méthode, on définit le $MSE(\boldsymbol{c})$ comme l'erreur quadratique moyenne du modèle complet, et le $MSE(\boldsymbol{c}_{(-i)})$ comme l'erreur quadratique moyenne du modèle lorsque la $i^{ème}$ covariable (C_i) est omise. On définit ΔMSE_i comme la différence $MSE(\boldsymbol{c}_{(-i)}) - MSE(\boldsymbol{c})$. De plus, notons $V_{\boldsymbol{c}}$ la variance de l'estimateur du coefficient d'exposition du modèle complet, et $V_{\boldsymbol{c}_{(-i)}}$ la variance de l'estimateur du coefficient d'exposition du modèle lorsque la $i^{ème}$ covariable (C_i) est omise, avec $\widehat{V}_{\boldsymbol{c}}$ et $\widehat{V}_{\boldsymbol{c}_{(-i)}}$ les estimateurs respectifs de $V_{\boldsymbol{c}}$ et $V_{\boldsymbol{c}_{(-i)}}$. Ainsi, en supposant que $\boldsymbol{C}$ est un ensemble suffisant, on a que

$$\begin{aligned}
\Delta MSE_i =& MSE(\boldsymbol{c}_{(-i)}) - MSE(\boldsymbol{c}) \\
=& \left[(\mathbb{E}[\hat{\theta}_{\boldsymbol{c}_{(-i)}}] - \theta)^2 + V_{\boldsymbol{c}_{(-i)}}\right] - \left[(\mathbb{E}[\hat{\theta}_{\boldsymbol{c}}] - \theta)^2 + V_{\boldsymbol{c}}\right] \\
=& \left[Biais(\hat{\theta}_{\boldsymbol{c}_{(-i)}})^2 + V_{\boldsymbol{c}_{(-i)}}\right] - V_{\boldsymbol{c}} \\
=& Biais(\hat{\theta}_{\boldsymbol{c}_{(-i)}})^2 - (V_{\boldsymbol{c}} - V_{\boldsymbol{c}_{(-i)}}) \\
\approx& (\hat{\theta}_{\boldsymbol{c}_{(-i)}} - \hat{\theta}_{\boldsymbol{c}})^2 - (\widehat{V}_{\boldsymbol{c}} - \widehat{V}_{\boldsymbol{c}_{(-i)}}),
\end{aligned} \tag{1.4.2}$$

ce qui nous permet d'avoir une expression approximative de la différence en erreur quadratique moyenne. Un $\Delta MSE_i \geq 0$ suggère qu'il faut conserver la covariable, puisque son retrait a un effet délétère sur l'estimation du coefficient de l'exposition sur la réponse. La procédure complète du ΔMSE est présentée par l'algorithme 2. L'algorithme est similaire à celui proposé pour le CIE, mais on remplace $g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-i)}})$ par ΔMSE_i donné dans l'équation 1.4.2 et τ par 0.

Algorithme 2 : Différence en erreur quadratique moyenne (ΔMSE)

Résultat : $\boldsymbol{C}$
Définir A, Y et $\boldsymbol{C}^* = \{C_1, C_2, \ldots, C_p\}$;
$\boldsymbol{C} \leftarrow \boldsymbol{C}^*$;
$p \leftarrow card(\boldsymbol{C})$;
Condition $\leftarrow$ Vraie ;
tant que *Condition* **faire**
  Estimer $\hat{y} = \hat{\theta}_{\boldsymbol{c}} a + \hat{\theta}_0 + \hat{\theta}_1 c_1 + \hat{\theta}_2 c_2 + \cdots + \hat{\theta}_p c_p$;
  $i \leftarrow p$;
  tant que $i > 0$ **faire**
    $\hat{y} = \hat{\theta}_{\boldsymbol{c}_{(-i)}} a + \hat{\theta}^*_0 + \hat{\theta}^*_1 c_1 + \hat{\theta}^*_2 c_2 + \cdots + \hat{\theta}^*_{(i-1)} c_{(i-1)} + \hat{\theta}^*_{(i+1)} c_{(i+1)} + \cdots + \hat{\theta}^*_p c_p$;
    Calculer ΔMSE_i ;
    $i \leftarrow i - 1$;
  fin
  $M \leftarrow min(\{\Delta MSE_1, \Delta MSE_2, \ldots, \Delta MSE_p\})$;
  Déterminer C_j tel que $M = \Delta MSE_j$;
  $p \leftarrow p - 1$;
  si $M \geq 0$ *ou* $p = 0$ **alors**
    Condition = Faux ;
  fin
  sinon
    $\boldsymbol{C} \leftarrow \boldsymbol{C} \setminus C_j$;
  fin
fin

CHAPITRE II

SÉLECTION DE VARIABLES EN MÉDIATION CAUSALE

Ce chapitre présente les limites potentielles du CIE et du ΔMSE standard pour faire de la sélection de variables en médiation causale. De plus, quatre procédures adaptées du CIE (CIE_{effets} et CIE_{max}) et du ΔMSE (ΔMSE_{effets} et ΔMSE_{max}) sont proposées pour faire de la sélection de variables en médiation causale.

2.1 Méthodes de sélection de variables et médiation

Les méthodes de sélection de variables basées sur l'effet total de l'exposition sur la réponse ne sont pas nécessairement adéquates pour faire de la sélection de variables en médiation causale. En effet, un ensemble suffisant $\boldsymbol{C}$ pour permettre une interprétation causale de l'association entre A et Y ne permet pas assurément d'obtenir des effets naturels causaux direct et indirect. La discussion qui suit, qui illustre ce point, motive l'adaptation des méthodes CIE et ΔMSE pour l'estimation des effets de médiation.

La figure 2.1 présente un DAG qui exprime un modèle de médiation avec des variables confondantes ($\boldsymbol{C}_{ay}$, $\boldsymbol{C}_{am}$ et $\boldsymbol{C}_{my}$), des prédicteurs purs ($\boldsymbol{P}_a$, $\boldsymbol{P}_m$ et $\boldsymbol{P}_y$)

et des variables de bruit ($\boldsymbol{B}$). De manière cohérente avec ce qui a été explicité dans la sous-section 1.3.3, l'identification des effets naturels direct et indirect peut être réalisée dans la figure 2.1 si on ajuste adéquatement pour les variables des vecteurs $\boldsymbol{C}_{ay}$ (H_1), $\boldsymbol{C}_{am}$ (H_2) et $\boldsymbol{C}_{my}$ (H_3). L'objectif est d'identifier, à partir d'un algorithme basé sur les données, les variables mesurées ($\boldsymbol{C}_{ay}$, $\boldsymbol{C}_{am}$, $\boldsymbol{C}_{my}$, $\boldsymbol{P}_a$, $\boldsymbol{P}_m$, $\boldsymbol{P}_y$ et $\boldsymbol{B}$) qui doivent être considérées dans le modèle de médiation pour l'estimation sans biais [1] et efficace des effets naturels direct et indirect.

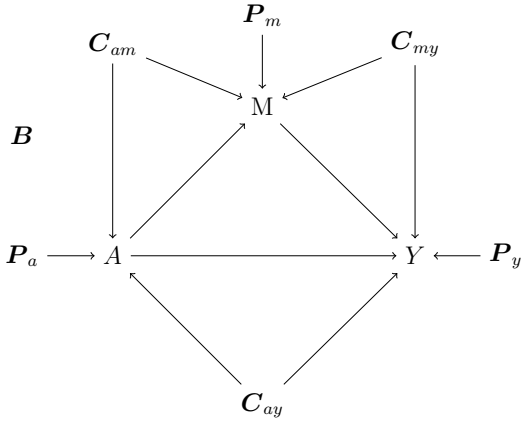


FIGURE 2.1 : Modèle de médiation et sélection de variables

Dans un premier temps, supposons que la sélection de variables s'effectue en visant l'estimation de l'exposition sur la réponse. En lien avec ce qui a été présenté dans la sous-section 1.4.2, on peut remarquer qu'un ensemble suffisant doit nécessairement inclure les covariables $\boldsymbol{C}_{am}$ et $\boldsymbol{C}_{ay}$. En effet, les vecteurs $\boldsymbol{C}_{am}$ et $\boldsymbol{C}_{ay}$ ne sont pas des descendants de l'exposition. De plus, afin de fermer les chemins porte arrière ouverts $A \leftarrow \boldsymbol{C}_{ay} \rightarrow Y$ et $A \leftarrow \boldsymbol{C}_{am} \rightarrow M \rightarrow Y$, il est nécessaire d'ajuster pour l'ensemble des covariables des vecteurs $\boldsymbol{C}_{ay}$ et $\boldsymbol{C}_{am}$, respectivement. Ainsi, en plus d'être un confondant de la relation entre A et M, le vecteur $\boldsymbol{C}_{am}$ est également un confondant de la relation entre A et Y, puisqu'il est une cause commune de ces

1. Dans le cas des procédures basées sur le ΔMSE, un certain biais peut être toléré si un gain important en variance est obtenu par l'exclusion d'une covariable dans le modèle de médiation.

deux variables. Les méthodes CIE et ΔMSE présentées dans les algorithmes 1 et 2 devraient donc permettre de conserver les variables des vecteurs $\boldsymbol{C}_{am}$ et $\boldsymbol{C}_{ay}$.

Pour le CIE standard, le problème émerge pour les variables confondantes $\boldsymbol{C}_{my}$. Dans ce contexte, il n'est pas nécessaire d'ajuster sur $\boldsymbol{C}_{my}$ pour permettre une interprétation causale de l'association entre l'exposition et la réponse. En effet, bien qu'un ensemble suffisant puisse inclure $\boldsymbol{C}_{my}$, l'ensemble minimal est $\{\boldsymbol{C}_{am}, \boldsymbol{C}_{ay}\}$. Ainsi, le CIE sur l'effet total de l'exposition sur la réponse ne devrait pas conserver les variables du vecteur $\boldsymbol{C}_{my}$, mais ce vecteur est nécessaire à l'identification des effets naturels direct et indirect (H_3). On peut illustrer ce phénomène à l'aide d'un jeu de données simulé. Supposons que l'on a les modèles suivants :

$$\begin{aligned}
\mathbb{E}[\boldsymbol{C}_{my}] &= \mathbf{0}, \\
\text{logit}(\mathbb{E}[A]) &= \gamma_0, \\
\mathbb{E}[M|A=a, \boldsymbol{C}_{my}=\boldsymbol{c}_{my}] &= \beta_0 + \beta_1 a + \boldsymbol{\beta}_2^T \boldsymbol{c}_{my}, \\
\mathbb{E}[Y|A=a, M=m, \boldsymbol{C}_{my}=\boldsymbol{c}_{my}] &= \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \boldsymbol{\theta}_4^T \boldsymbol{c}_{my},
\end{aligned}$$

avec $\mathbf{0}$ le vecteur nul. Pour simuler le jeu de données, on commence par générer $\boldsymbol{C}_{my} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, où $\boldsymbol{I}$ est la matrice identité et $\boldsymbol{C}_{my}$ est un vecteur composé de cinq variables aléatoires, soient C_1, C_2, C_3, C_4 et C_5. Par la suite, on génère $A \sim Bernoulli(expit(\gamma_0))$, avec $expit(x) = (1+\exp(-x))^{-1}$. Subséquemment, on génère $M \sim \mathcal{N}(\beta_0 + \beta_1 a + \beta_2^T \boldsymbol{c}_{my}, \sigma_m^2)$. Enfin, on génère $Y \sim \mathcal{N}(\theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \boldsymbol{\theta}_4^T \boldsymbol{c}_{my}, \sigma_y^2)$. Pour l'illustration, on prend un échantillon de taille $n = 500$. De plus, on utilise les paramètres suivants : $\gamma_0 = \beta_0 = \theta_0 = \theta_3 = 0$, $\sigma_m^2 = \sigma_y^2 = 1$, $\beta_1 = \theta_1 = 2$, $\theta_2 = .5$ et $\boldsymbol{\beta}_2^T = \boldsymbol{\theta}_4^T = (1,1,1,1,1)$. Avec ces paramètres, on a que $END(\boldsymbol{c}) = 2$, $ENI(\boldsymbol{c}) = 1$ et $ET(\boldsymbol{c}) = 3$.

Le tableau 2.1 présente un exemple de l'application du CIE sur l'effet total de

l'exposition sur la réponse, avec $\tau = 0.1$, en rapportant les métriques $g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-i)}})$ ($i = 1, \ldots, 5$; voir formule 1.4.1) et le minimum de ces mesures à chacune des itérations. Si la valeur minimale de $\left\{g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-i)}})\right\}_{i=1}^{5}$ est plus petite que τ, alors la variable qui est associée à ce minimum est exclue du processus de sélection.

TABLEAU 2.1 : Calcul du changement en estimation et exclusion des variables dans une application du CIE sur l'effet total de l'exposition sur la réponse

	C_1	C_2	C_3	C_4	C_5	$Minimum$
Itération 1	0.007	0.026	0.018	0.067	0.011	0.007
Itération 2	Exclue	0.027	0.018	0.069	0.011	0.011
Itération 3	Exclue	0.027	0.019	0.075	Exclue	0.019
Itération 4	Exclue	0.026	Exclue	0.071	Exclue	0.026
Itération 5	Exclue	Exclue	Exclue	0.072	Exclue	0.072
Itération 6	Exclue	Exclue	Exclue	Exclue	Exclue	Exclue

Dans cet exemple, on remarque que toutes les variables du vecteur $\boldsymbol{C}_{my}$ sont ultimement exclues par l'algorithme CIE standard. Or, le biais occasionné par l'omission de ces variables sur l'estimation des effets naturels direct et indirect peut être substantiel. Pour corroborer cette affirmation, on peut effectuer une simulation. Plus précisément, on génère 1000 jeux de données à l'aide de la même procédure que celle utilisée pour l'exemple précédent. Pour chacun des jeux de données, on applique un CIE sur l'effet total de l'exposition sur la réponse, et on calcule l'effet total, l'effet naturel direct et l'effet naturel indirect avec les variables sélectionnées par le CIE. Les résultats sont présentés dans le tableau 2.2.

Les cinq premières colonnes rapportent la proportion de sélection des variables du vecteur $\boldsymbol{C}_{my}$, alors que les colonnes subséquentes correspondent au biais relatif pour les différents effets. Les résultats montrent que la proportion de sélection pour les variables du vecteur $\boldsymbol{C}_{my}$ varie entre 2.70% et 3.90%, ce qui est cohérent

avec l'exemple présenté dans le tableau 2.1. Alors que le biais relatif de l'effet total est de -0.32%, le biais relatif pour les effets naturels direct et indirect est, respectivement, de -83.0% et 165%. Ainsi, les résultats de cette simulation illustrent la difficulté du CIE standard à conserver les variables de confusion entre M et Y, ce qui peut occasionner un biais relatif important pour les effets naturels direct et indirect.

TABLEAU 2.2 : Proportion de sélection des confondants de la relation entre M et Y et biais du CIE standard pour l'estimation des effets total, naturel direct et naturel indirect

Proportion de sélection (%)					Biais relatif (%)		
C_1	C_2	C_3	C_4	C_5	Total	Direct	Indirect
2.70	3.90	3.50	3.40	3.00	-0.32	-83.0	165

La situation est plus compliquée pour le ΔMSE qui, rappelons-le, tente de minimiser l'erreur quadratique moyenne pour l'estimation de l'effet total de l'exposition sur la réponse. Effectivement, puisque les variables du vecteur $\boldsymbol{C}_{my}$ sont des prédicteurs de M et de Y, l'exclusion de $\boldsymbol{C}_{my}$ devraient augmenter la variance de l'estimateur du coefficient de l'exposition sur la réponse. Ainsi, le ΔMSE ne devrait généralement pas exclure les variables du vecteur $\boldsymbol{C}_{my}$. Pour la sélection des variables qui ne font pas partie de l'ensemble minimal (c'est-à-dire $\boldsymbol{P}_a, \boldsymbol{P}_m, \boldsymbol{P}_y$ et $\boldsymbol{B}$), l'impact sur la variance de l'ajustement de variables diffère entre l'effet total, l'effet naturel direct et l'effet naturel indirect. Par exemple, alors que l'ajustement pour les prédicteurs purs du médiateur ($\boldsymbol{P_m}$) devrait diminuer la variance de l'effet total de l'exposition sur la réponse, une étude récente (Diop et al., 2021) a montré que la variance de l'estimation de l'effet naturel indirect augmentait lorsque les modèles ajustaient pour les variables $\boldsymbol{P}_m$, tandis que la variance de l'estimateur de l'effet naturel direct augmentait ou diminuait selon les paramètres de simulation. Ainsi, il est difficile d'établir dans quelle mesure le ΔMSE peut

faire de la sélection de variables de manière optimale en médiation causale. Dans ce contexte, il est important de proposer des méthodes alternatives au CIE et au ΔMSE pour permettre une sélection adéquate des variables dans le contexte de la médiation causale. Ces adaptations sont présentées dans la section suivante.

2.2 Adaptation des méthodes de sélection de variables au contexte de médiation

2.2.1 Notation

Pour l'entièreté de cette présentation, on prend $\boldsymbol{C}^*$ un ensemble de covariables mesurées, avec $\boldsymbol{C} \subseteq \boldsymbol{C}^*$. Également, on note $\widehat{END}(\boldsymbol{c})$ et $\widehat{ENI}(\boldsymbol{c})$ les estimateurs des effets naturels direct et indirect, respectivement, lorsque l'on inclut toutes les variables d'un ensemble $\boldsymbol{C}$, c'est-à-dire que les effets sont estimés avec

$$\begin{aligned}\hat{m} &= \hat{\beta}_0 + \hat{\beta}_1 a + \hat{\beta}_2 c_1 + \hat{\beta}_3 c_2 + \cdots + \hat{\beta}_{(p+1)} c_p,\\ \hat{y} &= \hat{\theta}_0 + \hat{\theta}_1 a + \hat{\theta}_2 m + \hat{\theta}_3 am + \hat{\theta}_4 c_1 + \hat{\theta}_5 c_2 + \cdots + \hat{\theta}_{(p+3)} c_p,\end{aligned}$$

où $p = card(\boldsymbol{C})$. Similairement, on note $\widehat{END}(\boldsymbol{c}_{(-i)})$ et $\widehat{ENI}(\boldsymbol{c}_{(-i)})$ les estimateurs des effets naturels direct et indirect, respectivement, qui sont obtenus en ajustant sur les variables de l'ensemble $\boldsymbol{C} \setminus C_i$, c'est-à-dire que

$$\begin{aligned}\hat{m} &= \hat{\beta}^*_0 + \hat{\beta}^*_1 a + \hat{\beta}^*_2 c_1 + \hat{\beta}^*_3 c_2 + \cdots + \hat{\beta}^*_{(i)} c_{(i-1)} + \hat{\beta}^*_{(i+2)} c_{(i+1)} + \cdots + \hat{\beta}^*_{(p+1)} c_p,\\ \hat{y} &= \hat{\theta}^*_0 + \hat{\theta}^*_1 a + \hat{\theta}^*_2 m + \hat{\theta}^*_3 am + \hat{\theta}^*_4 c_1 + \hat{\theta}^*_5 c_2 + \cdots + \hat{\theta}^*_{(i+2)} c_{(i-1)}\\ &\quad + \hat{\theta}^*_{(i+4)} c_{(i+1)} + \cdots + \hat{\theta}^*_{(p+3)} c_p.\end{aligned}$$

De plus, on pose

$$g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c_{(-i)}}})_{direct} = \left| \frac{\widehat{END}(\boldsymbol{c}_{(-i)}) - \widehat{END}(\boldsymbol{c})}{\widehat{END}(\boldsymbol{c})} \right|$$

et

$$g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c_{(-i)}}})_{indirect} = \left| \frac{\widehat{ENI}(\boldsymbol{c}_{(-i)}) - \widehat{ENI}(\boldsymbol{c})}{\widehat{ENI}(\boldsymbol{c})} \right|.$$

Similairement, on note $\Delta MSE_i^{direct} = MSE(\boldsymbol{c}_{(-i)})^{direct} - MSE(\boldsymbol{c})^{direct}$, avec

$$MSE(\boldsymbol{c}_{(-i)})^{direct} = \mathbb{E}\left[\left(\widehat{END}(\boldsymbol{c}_{(-i)}) - END(\boldsymbol{c}) \right)^2 \right]$$

et

$$MSE(\boldsymbol{c})^{direct} = \mathbb{E}\left[\left(\widehat{END}(\boldsymbol{c}) - END(\boldsymbol{c}) \right)^2 \right].$$

Comme cela était le cas pour le ΔMSE_i sur l'effet total de l'exposition sur la réponse, on peut estimer ΔMSE_i^{direct} par

$$\left(\widehat{END}(\boldsymbol{c}_{(-i)}) - \widehat{END}(\boldsymbol{c}) \right)^2 - \left(\widehat{V}\left[\widehat{END}(\boldsymbol{c}) \right] - \widehat{V}\left[\widehat{END}(\boldsymbol{c}_{(-i)}) \right] \right);$$

on utilise une notation similaire pour le $\Delta MSE_i^{indirect}$.

2.2.2 Hypothèses des approches de sélection de variables en médiation causale

Un ensemble $\boldsymbol{C}$ est considéré comme *suffisant pour la médiation* si les critères suivants sont rencontrés. Premièrement, les variables du vecteur $\boldsymbol{C}$ ne doivent pas être causées par l'exposition. De plus, le vecteur $\boldsymbol{C}$ doit inclure l'ensemble des variables de confusion entre A et Y ($\boldsymbol{C}_{ay}$; H_1), entre A et M ($\boldsymbol{C}_{am}$; H_2)

et entre M et Y ($\boldsymbol{C}_{my}$; H_3). Également, on doit avoir que $Y(a,m) \perp\!\!\!\perp M(a^*)|\boldsymbol{C}$ (H_4). Ensuite, pour chaque valeur de $\boldsymbol{C}$, il est nécessaire que deux hypothèses de positivité soient respectées, soit que $P(A = a|\boldsymbol{C} = \boldsymbol{c}) > 0$ pour $a \in \{0,1\}$, et $g(m|A = a, \boldsymbol{C} = \boldsymbol{c}) > 0$ pour $a \in \{0,1\}$ et pour toutes les valeurs de m dans le support de $M|\boldsymbol{C} = \boldsymbol{c}$ (Nguyen, Schmid, Ogburn, & Stuart, 2020). Une des hypothèses fondamentales des quatre méthodes proposées pour faire de la sélection de variables en médiation causale est que l'ensemble initial $\boldsymbol{C}^*$ est suffisant pour la médiation. Enfin, une autre hypothèse est que pour tous les ensembles $\boldsymbol{C} \subseteq \boldsymbol{C}^*$, les modèles de régression 2.2.1 et 2.2.2 sont correctement spécifiés, et donc que les formules pour les effets naturels direct (formule 1.3.1) et indirect (formule 1.3.2) sont adéquates pour tous les ensembles suffisants pour la médiation :

$$\mathbb{E}[M|A = a, \boldsymbol{C} = \boldsymbol{c}] = \beta_0 + \beta_1 a + \boldsymbol{\beta}_2^T \boldsymbol{c}, \tag{2.2.1}$$

$$\mathbb{E}[Y|A = a, M = m, \boldsymbol{C} = \boldsymbol{c}] = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \boldsymbol{\theta}_4^T \boldsymbol{c}. \tag{2.2.2}$$

2.2.3 Sélection de variables sur les effets naturels marginaux

Une des difficultés pour faire de la sélection de variables avec l'approche par régression linéaire est que même lorsque $\boldsymbol{C}$ et $\boldsymbol{D}$ sont des ensembles de covariables suffisants pour la médiation, les quantités $\widehat{END}(\boldsymbol{c})$ et $\widehat{END}(\boldsymbol{d})$ ne sont pas nécessairement des estimateurs pour les mêmes estimands. Une solution à cette problématique est de faire la sélection de variables sur les effets naturels marginaux plutôt que sur les effets naturels conditionnels. Sous les hypothèses émises dans la sous-section 2.2.2, il est possible de proposer un estimateur pour les effets naturels marginaux en conditionnant sur l'espérance des covariables (voir annexe A). Conséquemment, les algorithmes de sélection de variables qui sont proposés estiment les effets naturels en conditionnant sur la valeur moyenne des covariables.

2.2.4 Adaptation du CIE au contexte de médiation

2.2.4.1 Proposition 1 : Sélection sur les effets naturels direct et indirect

Une première proposition, notée CIE_{effets}, est d'effectuer le CIE sur l'effet naturel direct et l'effet naturel indirect séparément, et d'inclure une variable dans l'ensemble $\boldsymbol{C}$ si cette dernière est sélectionnée dans le CIE sur l'effet naturel direct *ou* sur l'effet naturel indirect. Plus précisément, on débute avec l'ensemble suffisant pour la médiation $\boldsymbol{C}^*$, et on pose $\boldsymbol{C} = \boldsymbol{C}^*$. Par la suite, on estime $\widehat{END}(\boldsymbol{c})$ et $\left\{\widehat{END}(\boldsymbol{c}_{(-i)})\right\}_{i=1}^{p}$. Ensuite, on calcule $\left\{g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-i)}})_{direct}\right\}_{i=1}^{p}$, et on trouve la valeur minimale de $\left\{g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-i)}})_{direct}\right\}_{i=1}^{p}$. Si cette valeur est inférieure à τ, on rejette la variable C_j qui a engendré cette valeur minimale, et on recommence le processus avec $\boldsymbol{C} \setminus C_j$. Ce CIE sur l'effet naturel direct se conclut si $\boldsymbol{C} = \emptyset$ ou si la valeur minimale de $\left\{g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-i)}})_{direct}\right\}_{i=1}^{p}$ est supérieure ou égale à τ ; on note $\boldsymbol{C}_{Direct}$ les variables sélectionnées par ce CIE. Après avoir obtenu l'ensemble $\boldsymbol{C}_{Direct}$, on recommence cette même procédure, mais en utilisant l'effet naturel indirect, et on note $\boldsymbol{C}_{Indirect}$ les variables sélectionnées par ce CIE. Enfin, on prend $\boldsymbol{C} = \boldsymbol{C}_{Direct} \cup \boldsymbol{C}_{Indirect}$.

En effectuant le CIE séparément sur l'effet naturel direct et l'effet naturel indirect, il devrait être possible de contrôler le biais relatif à τ pour $\widehat{END}(\boldsymbol{c})$ et pour $\widehat{ENI}(\boldsymbol{c})$. De plus, cette méthode devrait être en mesure de retirer les prédicteurs purs ($\boldsymbol{P}_a, \boldsymbol{P}_m, \boldsymbol{P}_y$) et les variables de bruit ($\boldsymbol{B}$) de l'ensemble $\boldsymbol{C}^*$, puisque l'omission de ces covariables n'engendre pas de biais sur l'estimation des effets naturels direct et indirect. La procédure du CIE_{effets} est présentée dans l'algorithme 3.

Algorithme 3 : Changement d'estimation sur END et ENI (CIE_{effets})

Résultat : $\boldsymbol{C}$

Définir τ ;
Définir A, M, Y et $\boldsymbol{C}^* = \{C_1, C_2, \ldots, C_p\}$;
$k \leftarrow$ Direct ;
pour k *dans {Direct, Indirect}* **faire**
  si $k = Direct$ **alors**
    Définir $g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c_{(-i)}}})$ par $g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c_{(-i)}}})_{direct}$;
    Définir $\hat{E}_{\boldsymbol{c}}$ par $\widehat{END}(\boldsymbol{c})$;
    Définir $\hat{E}_{\boldsymbol{c}_{(-i)}}$ par $\widehat{END}(\boldsymbol{c}_{(-i)})$;
  fin
  sinon
    Définir $g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c_{(-i)}}})$ par $g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c_{(-i)}}})_{indirect}$;
    Définir $\hat{E}_{\boldsymbol{c}}$ par $\widehat{ENI}(\boldsymbol{c})$;
    Définir $\hat{E}_{\boldsymbol{c}_{(-i)}}$ par $\widehat{ENI}(\boldsymbol{c}_{(-i)})$;
  fin
  $\boldsymbol{C} \leftarrow \boldsymbol{C}^*$;
  $p \leftarrow card(\boldsymbol{C})$;
  Condition $\leftarrow$ Vraie ;
  tant que *Condition* **faire**
    Estimer $\hat{E}_{\boldsymbol{c}}$;
    $i \leftarrow p$;
    tant que $i > 0$ **faire**
      Estimer $\hat{E}_{\boldsymbol{c}_{(-i)}}$;
      Calculer $g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-i)}})$;
      $i \leftarrow i - 1$;
    fin
    $M \leftarrow min(\{g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-1)}}), g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-2)}}), \ldots, g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-p)}})\})$;
    Déterminer C_j tel que $M = g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-j)}})$;
    $p \leftarrow p - 1$;
    si $M \geq \tau$ *ou* $p = 0$ **alors**
      Condition = Faux ;
    fin
    sinon
      $\boldsymbol{C} \leftarrow \boldsymbol{C} \setminus C_j$;
    fin
  fin
  $\boldsymbol{C}_k = \boldsymbol{C}$;
fin
$\boldsymbol{C} \leftarrow \boldsymbol{C}_{Direct} \cup \boldsymbol{C}_{Indirect}$.

2.2.4.2 Proposition 2 : Sélection sur le maximum des effets naturels direct et indirect

Une limite du CIE_{effets} est que l'on doit effectuer deux procédures CIE, ce qui peut être computationnellement demandant lorsque la taille de l'échantillon ou le nombre de variables sont importants. Une solution alternative, notée CIE_{max}, consiste à effectuer un seul CIE. Toutefois, à chaque itération du processus, le CIE est calculé pour les effets naturels direct et indirect en prenant la valeur maximale comme mesure de distance $g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-i)}})$. Plus précisément, on débute avec l'ensemble suffisant pour la médiation $\boldsymbol{C}^*$, et on pose $\boldsymbol{C} = \boldsymbol{C}^*$. Par la suite, on obtient les estimations $\widehat{END}(\boldsymbol{c})$, $\widehat{ENI}(\boldsymbol{c})$, $\left\{\widehat{END}(\boldsymbol{c}_{(-i)})\right\}_{i=1}^{p}$ et $\left\{\widehat{ENI}(\boldsymbol{c}_{(-i)})\right\}_{i=1}^{p}$. De plus, pour $i = 1, 2, \ldots, p$, on calcule $g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-i)}})_{direct}$ et $g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-i)}})_{indirect}$, et on pose

$$g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-i)}}) = max\{g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-i)}})_{direct}, g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-i)}})_{indirect}\}.$$

Ensuite, on trouve la valeur minimale de $\left\{g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-i)}})\right\}_{i=1}^{p}$. Si cette valeur est inférieure à τ, on rejette la variable C_j qui a engendré cette valeur minimale, et on recommence le processus avec $\boldsymbol{C} \setminus C_j$. Le CIE_{max} se termine si $\boldsymbol{C} = \emptyset$ ou encore si la valeur minimale de $\left\{g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-i)}})\right\}_{i=1}^{p}$ est supérieure ou égale à τ. Ainsi, pour retirer une variable C_j de l'ensemble $\boldsymbol{C}$, il faut que $g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-j)}})_{direct}$ et $g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-j)}})_{indirect}$ soient inférieurs à τ. Comme cela était le cas pour CIE_{effets}, le CIE_{max} devrait rejeter les prédicteurs purs ($\boldsymbol{P}_a, \boldsymbol{P}_m, \boldsymbol{P}_y$) et les variables de bruit ($\boldsymbol{B}$). La procédure complète du CIE_{max} est présentée dans l'algorithme 4.

Algorithme 4 : Changement d'estimation maximal sur END et ENI (CIE_{max})

Résultat : $\boldsymbol{C}$
Définir τ ;
Définir A, M, Y et $\boldsymbol{C}^* = \{C_1, C_2, \ldots, C_p\}$;
$\boldsymbol{C} \leftarrow \boldsymbol{C}^*$;
$p \leftarrow card(\boldsymbol{C})$;
Condition $\leftarrow$ Vraie ;
tant que *Condition* **faire**
  Estimer $\widehat{END}(\boldsymbol{c})$;
  Estimer $\widehat{ENI}(\boldsymbol{c})$;
  $i \leftarrow p$;
  tant que $i > 0$ **faire**
    Estimer $\widehat{END}(\boldsymbol{c}_{(-i)})$;
    Estimer $\widehat{ENI}(\boldsymbol{c}_{(-i)})$;
    Calculer $g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-i)}})_{direct}$;
    Calculer $g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-i)}})_{indirect}$;
    $g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-i)}}) \leftarrow max\{g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-i)}})_{direct},\ g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-i)}})_{indirect}\}$;
    $i \leftarrow i - 1$;
  fin
  $M \leftarrow min(\{g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-1)}}), g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-2)}}), \ldots, g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-p)}})\})$;
  Déterminer C_j tel que $M = g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-j)}})$;
  $p \leftarrow p - 1$;
  si $M \geq \tau$ *ou* $p = 0$ **alors**
    Condition = Faux ;
  fin
  sinon
    $\boldsymbol{C} \leftarrow \boldsymbol{C} \setminus C_j$;
  fin
fin

2.2.5 Adaptation du ΔMSE au contexte de médiation

2.2.5.1 Proposition 1 : Sélection sur les effets naturels direct et indirect

Une première proposition, notée ΔMSE_{effets}, consiste à effectuer un ΔMSE séparément sur l'effet naturel direct et l'effet naturel indirect et à conserver une variable si cette dernière est sélectionnée dans le ΔMSE sur l'effet naturel direct *ou* dans le ΔMSE sur l'effet naturel indirect. La procédure est identique à celle proposée pour le CIE_{effets}, mais on remplace $g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-i)}})_{direct}$ par ΔMSE_i^{direct}, $g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-i)}})_{indirect}$ par $\Delta MSE_i^{indirect}$ et τ par 0. La procédure complète du ΔMSE_{effets} est présentée dans l'algorithme 5.

Comme cela a été mentionné antérieurement, l'omission des prédicteurs purs ($\boldsymbol{P}_a$, $\boldsymbol{P}_m$, $\boldsymbol{P}_y$) et des variables de bruit ($\boldsymbol{B}$) n'entraîne pas de biais dans l'estimation des effets de médiation. Toutefois, puisque le ΔMSE_{effets} tient également compte de l'impact de l'ajustement des variables sur la variance des estimateurs, certaines de ces variables pourraient ne pas être rejetées dans le processus de sélection. En règle générale, on peut émettre l'hypothèse que les variables $\boldsymbol{B}$ ne seront pas conservées, puisqu'il ne devrait pas y avoir de gain en variance lorsque l'on ajuste pour ces variables. Pour ce qui est des prédicteurs purs, une étude récente a montré que l'inclusion des prédicteurs purs de l'exposition augmentait la variance des estimateurs de $END(\boldsymbol{c})$ et de $ENI(\boldsymbol{c})$, alors que l'ajustement pour les prédicteurs purs de la réponse diminuait la variance de ces mêmes estimateurs (Diop et al., 2021). Conséquemment, on peut s'attendre à ce que le ΔMSE_{effets} rejette les variables de $\boldsymbol{P}_a$ mais conserve les variables de $\boldsymbol{P}_y$.

Algorithme 5 : Différence en erreur quadratique moyenne sur END et ENI (ΔMSE_{effets})

Résultat : $\boldsymbol{C}$
Définir A, M, Y et $\boldsymbol{C}^* = \{C_1, C_2, \ldots, C_p\}$;
$k \leftarrow Direct$;
pour k *dans {Direct, Indirect}* **faire**
  si $k = Direct$ **alors**
    Définir ΔMSE_i par ΔMSE_i^{direct} ;
    Définir $\hat{E}_{\boldsymbol{c}}$ par $\widehat{END}(\boldsymbol{c})$;
    Définir $\hat{E}_{\boldsymbol{c}_{(-i)}}$ par $\widehat{END}(\boldsymbol{c}_{(-i)})$;
  fin
  sinon
    Définir ΔMSE_i par $\Delta MSE_i^{indirect}$;
    Définir $\hat{E}_{\boldsymbol{c}}$ par $\widehat{ENI}(\boldsymbol{c})$;
    Définir $\hat{E}_{\boldsymbol{c}_{(-i)}}$ par $\widehat{ENI}(\boldsymbol{c}_{(-i)})$;
  fin
  $\boldsymbol{C} \leftarrow \boldsymbol{C}^*$;
  $p \leftarrow card(\boldsymbol{C})$;
  Condition $\leftarrow$ Vraie ;
  tant que *Condition* **faire**
    Estimer $\hat{E}_{\boldsymbol{c}}$;
    $i \leftarrow p$;
    tant que $i > 0$ **faire**
      Estimer $\hat{E}_{\boldsymbol{c}_{(-i)}}$;
      Calculer ΔMSE_i ;
      $i \leftarrow i - 1$;
    fin
    $M \leftarrow min(\{\Delta MSE_1, \Delta MSE_2, \ldots, \Delta MSE_p\})$;
    Déterminer C_j tel que $M = \Delta MSE_j$;
    $p \leftarrow p - 1$;
    si $M \geq 0$ *ou* $p = 0$ **alors**
      Condition = Faux ;
    fin
    sinon
      $\boldsymbol{C} \leftarrow \boldsymbol{C} \setminus C_j$;
    fin
  fin
  $\boldsymbol{C}_k = \boldsymbol{C}$;
fin
$\boldsymbol{C} \leftarrow \boldsymbol{C}_{Direct} \cup \boldsymbol{C}_{Indirect}$.

La situation est plus compliquée pour les prédicteurs purs du médiateur ($\boldsymbol{P}_m$). En effet, alors que l'inclusion des variables $\boldsymbol{P}_m$ augmente la variance de l'estimateur de $END(\boldsymbol{c})$, l'ajustement pour les variables $\boldsymbol{P}_m$ peut augmenter ou diminuer la variance de l'estimateur de $ENI(\boldsymbol{c})$ (Diop et al., 2021). Plus précisément, selon cette même étude, si le coefficient du lien entre le médiateur et la réponse (θ_2) était plus grand ou égal au coefficient du lien entre l'exposition et le médiateur (β_1), alors l'inclusion des prédicteurs purs du médiateur avait tendance à diminuer la variance de l'estimateur de $ENI(\boldsymbol{c})$; l'inverse était observé lorsque $\theta_2 < \beta_1$. Puisque le ΔMSE_{effets} inclut une variable dans $\boldsymbol{C}$ si cette dernière est sélectionnée dans $\boldsymbol{C}_{Direct}$ ou $\boldsymbol{C}_{Indirect}$, il est possible d'émettre l'hypothèse qu'un prédicteur pur du médiateur sera sélectionné si $\theta_2 \geq \beta_1$, avec une conséquence possiblement délétère pour l'effet naturel direct. Cela permet donc de penser que la sélection sur le maximum (voir plus bas) pourrait être plus appropriée pour le ΔMSE.

2.2.5.2 Proposition 2 : Sélection sur le maximum des effets naturels direct et indirect

Comme cela était le cas pour le CIE, il est possible de proposer un ΔMSE_{max} qui est basé sur un maximum entre le ΔMSE_i^{direct} et le $\Delta MSE_i^{indirect}$ à chaque itération de l'algorithme. La procédure est similaire à celle présentée pour le CIE_{max}, mais il est nécessaire de remplacer $g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-i)}})_{direct}$ par ΔMSE_i^{direct} et $g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-i)}})_{indirect}$ par $\Delta MSE_i^{indirect}$. De plus, on peut substituer $g(\hat{\theta}_{\boldsymbol{c}}, \hat{\theta}_{\boldsymbol{c}_{(-i)}})$ par

$$\Delta MSE_i = \begin{cases} \Delta MSE_i^{direct} & \text{si } |\Delta MSE_i^{direct}| > |\Delta MSE_i^{indirect}| \\ \Delta MSE_i^{indirect} & \text{sinon} \end{cases}.$$

On peut remarquer que cet algorithme propose de comparer les valeurs absolues de la différence en erreur quadratique moyenne des effets naturels direct et indirect. Si les valeurs absolues n'étaient pas utilisées, une variable serait rejetée uniquement si elle engendre un ΔMSE_i négatif (donc si le retrait de la variable diminue l'erreur quadratique moyenne) pour l'effet naturel direct et l'effet naturel indirect, ce qui serait similaire à la procédure ΔMSE_{effets}.

Les caractéristiques de sélection de cette procédure devraient être comparables à celle présentées pour le ΔMSE_{effets}. Toutefois, les variables $\boldsymbol{P}_m$ devraient être conservées dans l'ensemble $\boldsymbol{C}$ seulement si l'ajustement pour les prédicteurs purs du médiateur engendre une diminution de la variance de l'estimateur pour $ENI(\boldsymbol{c})$ qui est supérieure à l'augmentation de la variance de l'estimateur pour $END(\boldsymbol{c})$, ce qui est un avantage de cette approche comparativement au ΔMSE_{effets}. Toutefois, puisque le ΔMSE n'est pas une mesure relative ou standardisée, il est possible, dans certains contextes, que le ΔMSE_i^{direct} ou le $\Delta MSE_i^{indirect}$ ait une influence plus importante dans la sélection de variables, c'est-à-dire que le ΔMSE_i soit plus fréquemment égal à l'une ou l'autre de ces différences en erreur quadratique moyenne (par exemple, que le ΔMSE_i soit généralement égal au ΔMSE_i^{direct}). La procédure complète du ΔMSE_{max} est présentée dans l'algorithme 6.

Algorithme 6 : Différence en erreur quadratique moyenne maximale sur END et ENI (ΔMSE_{max})

Résultat : $\boldsymbol{C}$

Définir A, M, Y et $\boldsymbol{C}^* = \{C_1, C_2, \ldots, C_p\}$;

$\boldsymbol{C} \leftarrow \boldsymbol{C}^*$;

$p \leftarrow card(\boldsymbol{C})$;

Condition $\leftarrow$ Vraie ;

tant que *Condition* **faire**

  Estimer $\widehat{END}(\boldsymbol{c})$;

  Estimer $\widehat{ENI}(\boldsymbol{c})$;

  $i \leftarrow p$;

  tant que $i > 0$ **faire**

    Estimer $\widehat{END}(\boldsymbol{c}_{(-i)})$;

    Estimer $\widehat{ENI}(\boldsymbol{c}_{(-i)})$;

    Calculer ΔMSE_i^{direct} ;

    Calculer $\Delta MSE_i^{indirect}$;

    $$\Delta MSE_i = \begin{cases} \Delta MSE_i^{direct} & \text{si } |\Delta MSE_i^{direct}| > |\Delta MSE_i^{indirect}| \\ \Delta MSE_i^{indirect} & \text{sinon} \end{cases} ;$$

    $i \leftarrow i - 1$;

  fin

  $M \leftarrow min(\{\Delta MSE_1, \Delta MSE_2, \ldots, \Delta MSE_p\})$;

  Déterminer C_j tel que $M = \Delta MSE_j$;

  $p \leftarrow p - 1$;

  si $M \geq 0$ *ou* $p = 0$ **alors**

    Condition = Faux ;

  fin

  sinon

    $\boldsymbol{C} \leftarrow \boldsymbol{C} \setminus C_j$;

  fin

fin

CHAPITRE III

ÉTUDE DE SIMULATION

Ce chapitre présente les simulations qui sont effectuées pour évaluer la performance des différentes méthodes proposées pour faire la sélection de variables en médiation causale. La première section soumet le plan de simulation basé sur les recommandations de Morris, White et Crowther (2019), alors que la seconde section décrit les résultats de la simulation.

3.1 Plan de simulation

Le plan de simulation expose les objectifs de la simulation, les mécanismes de génération de données, les paramètres et autres cibles à estimer, les méthodes évaluées et les mesures de performance.

3.1.1 Objectifs

L'objectif principal de la simulation est d'évaluer et de comparer la performance des versions modifiées du CIE (CIE_{effets} et CIE_{max}) et du ΔMSE (ΔMSE_{effets} et ΔMSE_{max}) pour faire la sélection de variables en médiation causale. Un second

objectif est également d'illustrer l'incapacité du CIE standard à effectuer la sélection de variables pour estimer adéquatement les effets naturels direct et indirect. Enfin, comme cela a été mentionné dans la section 2.1, il est difficile d'établir si le ΔMSE standard peut effectuer la sélection de variables de manière optimale lorsque le but est d'estimer les effets naturels direct et indirect. Ainsi, un dernier objectif est de déterminer si, et dans quelles conditions, le ΔMSE sur total de l'exposition sur la réponse peut être utilisé pour faire de la sélection de variables dans le contexte de médiation.

3.1.2 Mécanismes de génération de données

Pour l'ensemble des scénarios de simulation, on utilise les modèles suivants :

$$\begin{aligned}
\text{logit}(\mathbb{E}[A|\boldsymbol{C}_{am}=\boldsymbol{c}_{am},\boldsymbol{C}_{ay}=\boldsymbol{c}_{ay},\boldsymbol{P}_a=\boldsymbol{p}_a]) =& \gamma_0+\boldsymbol{\gamma}_1^T\boldsymbol{c}_{am}+\boldsymbol{\gamma}_2^T\boldsymbol{c}_{ay}+\boldsymbol{\gamma}_3^T\boldsymbol{p}_a,\\
\mathbb{E}[M|A=a,\boldsymbol{C}_{am}=\boldsymbol{c}_{am},\boldsymbol{C}_{my}=\boldsymbol{c}_{my},\boldsymbol{P}_m=\boldsymbol{p}_m] =& \beta_0+\beta_1 a+\boldsymbol{\beta}_2^T\boldsymbol{c}_{am}+\boldsymbol{\beta}_3^T\boldsymbol{c}_{my}+\boldsymbol{\beta}_4^T\boldsymbol{p}_m,\\
\mathbb{E}[Y|A=a,M=m,\boldsymbol{C}_{ay}=\boldsymbol{c}_{ay},\boldsymbol{C}_{my}=\boldsymbol{c}_{my},\boldsymbol{P}_y=\boldsymbol{p}_y] =& \theta_0+\theta_1 a+\theta_2 m+\theta_3 am+\boldsymbol{\theta}_4^T\boldsymbol{c}_{ay}\\
&+\boldsymbol{\theta}_5^T\boldsymbol{c}_{my}+\boldsymbol{\theta}_6^T\boldsymbol{p}_y,
\end{aligned}$$

avec $\boldsymbol{P}_a$, $\boldsymbol{P}_m$, $\boldsymbol{P}_y$, $\boldsymbol{B}$, $\boldsymbol{C}_{ay}$, $\boldsymbol{C}_{am}$, $\boldsymbol{C}_{my}$ des vecteurs colonnes de dimension trois. Ainsi, il y a un total de 21 variables qui sont assujetties au processus de sélection. Pour effectuer la simulation, on commence par générer les variables $\boldsymbol{P}_a$, $\boldsymbol{P}_m$, $\boldsymbol{P}_y$, $\boldsymbol{B}$, $\boldsymbol{C}_{ay}$, $\boldsymbol{C}_{am}$, $\boldsymbol{C}_{my} \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{0},\boldsymbol{I}_3)$. Par la suite, on génère de manière successive les variables $A \sim Bernoulli(expit(\gamma_0+\boldsymbol{\gamma}_1^T\boldsymbol{c}_{am}+\boldsymbol{\gamma}_2^T\boldsymbol{c}_{ay}+\boldsymbol{\gamma}_3^T\boldsymbol{p}_a))$, $M \sim \mathcal{N}(\beta_0+\beta_1 a+\boldsymbol{\beta}_2^T\boldsymbol{c}_{am}+\boldsymbol{\beta}_3^T\boldsymbol{c}_{my}+\boldsymbol{\beta}_4^T\boldsymbol{p}_m,1)$ et $Y \sim \mathcal{N}(\theta_0+\theta_1 a+\theta_2 m+\theta_3 am+\boldsymbol{\theta}_4^T\boldsymbol{c}_{ay}+\boldsymbol{\theta}_5^T\boldsymbol{c}_{my}+\boldsymbol{\theta}_6^T\boldsymbol{p}_y,1)$, avec $expit(x)=(1+\exp(x))^{-1}$.

Les paramètres de simulation sont largement inspirés de Diop et ses collaborateurs

(2021). Au total, six scénarios sont proposés. Plus précisément, pour l'ensemble de la simulation, on prend $(\gamma_0, \beta_0, \theta_0, \theta_1, \theta_3) = (0, 0, 0, 0.4, 0)$. Le fait que $\theta_3 = 0$ signifie qu'il n'y a pas d'interaction entre l'exposition et le médiateur dans le modèle sur la réponse. Ce choix est justifié par le fait que les algorithmes de sélection de variables sur l'effet total (voir les procédures 1 et 2) sont basés sur des modèles où l'effet de l'exposition sur la réponse est invariant selon la valeur du médiateur. De plus, l'étude de Diop et ses collaborateurs (2021) n'incluait pas de terme d'interaction entre l'exposition et le médiateur dans ses analyses principales. En effet, l'omission du terme d'interaction permet de comparer plus aisément les conclusions de la présente simulation aux connaissances pré-existantes. Pour les trois premiers scénarios, les paramètres des vecteurs $\boldsymbol{\gamma}_1^T$, $\boldsymbol{\gamma}_2^T$, $\boldsymbol{\gamma}_3^T$, $\boldsymbol{\beta}_2^T$, $\boldsymbol{\beta}_3^T$, $\boldsymbol{\beta}_4^T$, $\boldsymbol{\theta}_4^T$, $\boldsymbol{\theta}_5^T$ et $\boldsymbol{\theta}_6^T$ sont fixes à (0.3, 0.6, 1.2). Également, on a que $(\theta_2, \beta_1) = (0.6, 0.6)$ dans le scénario 1, $(\theta_2, \beta_1) = (0.3, 1.2)$ dans le scénario 2 et $(\theta_2, \beta_1) = (1.2, 0.3)$ dans le scénario 3. Les scénarios 4, 5 et 6 sont identiques aux scénarios 1, 2 et 3, respectivement, à la différence que les valeurs des paramètres des vecteurs $\boldsymbol{\theta}_4^T$, $\boldsymbol{\theta}_5^T$ et $\boldsymbol{\beta}_2^T$ sont maintenant fixes à (1.2, 0.6, 0.3). Ainsi, alors que les trois premiers scénarios évaluent les propriétés de sélection lorsque les effets de confusion sont, respectivement, faibles, modérés et forts, les trois scénarios subséquents tentent de comparer les procédures lorsque les effets d'une cause commune entre deux variables peuvent différer. Les trois premiers scénarios sont dits *symétriques*, alors que les scénarios 4, 5 et 6 sont dits *asymétriques*. Pour ces six scénarios, l'effet naturel direct est de 0.4, l'effet naturel indirect est de 0.36 et l'effet total est de 0.76. Enfin, la taille des échantillons est de $n = 500$ et le nombre de réplications est de $m = 1000$. Le tableau 3.1 résume les coefficients utilisés pour les différents scénarios.

TABLEAU 3.1 : Paramètres pour les six scénarios de simulation

	Scénario 1	Scénario 2	Scénario 3	Scénario 4	Scénario 5	Scénario 6
γ_0	0.0	0.0	0.0	0.0	0.0	0.0
$\gamma_{(1,1)}$	0.3	0.3	0.3	0.3	0.3	0.3
$\gamma_{(1,2)}$	0.6	0.6	0.6	0.6	0.6	0.6
$\gamma_{(1,3)}$	1.2	1.2	1.2	1.2	1.2	1.2
$\gamma_{(2,1)}$	0.3	0.3	0.3	0.3	0.3	0.3
$\gamma_{(2,2)}$	0.6	0.6	0.6	0.6	0.6	0.6
$\gamma_{(2,3)}$	1.2	1.2	1.2	1.2	1.2	1.2
$\gamma_{(3,1)}$	0.3	0.3	0.3	0.3	0.3	0.3
$\gamma_{(3,2)}$	0.6	0.6	0.6	0.6	0.6	0.6
$\gamma_{(3,3)}$	1.2	1.2	1.2	1.2	1.2	1.2
β_0	0.0	0.0	0.0	0.0	0.0	0.0
β_1	0.6	1.2	0.3	0.6	1.2	0.6
$\beta_{(2,1)}$	0.3	0.3	0.3	1.2	1.2	1.2
$\beta_{(2,2)}$	0.6	0.6	0.6	0.6	0.6	0.6
$\beta_{(2,3)}$	1.2	1.2	1.2	0.3	0.3	0.2
$\beta_{(3,1)}$	0.3	0.3	0.3	0.3	0.3	0.3
$\beta_{(3,2)}$	0.6	0.6	0.6	0.6	0.6	0.6
$\beta_{(3,3)}$	1.2	1.2	1.2	1.2	1.2	1.2
$\beta_{(4,1)}$	0.3	0.3	0.3	0.3	0.3	0.3
$\beta_{(4,2)}$	0.6	0.6	0.6	0.6	0.6	0.6
$\beta_{(4,3)}$	1.2	1.2	1.2	1.2	1.2	1.2
θ_0	0.0	0.0	0.0	0.0	0.0	0.0
θ_1	0.4	0.4	0.4	0.4	0.4	0.4
θ_2	0.6	0.3	1.2	0.6	0.3	1.2
θ_3	0.0	0.0	0.0	0.0	0.0	0.0
$\theta_{(4,1)}$	0.3	0.3	0.3	1.2	1.2	1.2
$\theta_{(4,2)}$	0.6	0.6	0.6	0.6	0.6	0.6
$\theta_{(4,3)}$	1.2	1.2	1.2	0.3	0.3	0.3
$\theta_{(5,1)}$	0.3	0.3	0.3	1.2	1.2	1.2
$\theta_{(5,2)}$	0.6	0.6	0.6	0.6	0.6	0.6
$\theta_{(5,3)}$	1.2	1.2	1.2	0.3	0.3	0.3
$\theta_{(6,1)}$	0.3	0.3	0.3	0.3	0.3	0.3
$\theta_{(6,2)}$	0.6	0.6	0.6	0.6	0.6	0.6
$\theta_{(6,3)}$	1.2	1.2	1.2	1.2	1.2	1.2

Note : Les paramètres $\gamma_{(j,i)}$, $\beta_{(j,i)}$ et $\theta_{(j,i)}$ correspondent, respectivement, à la $i^{ème}$ valeur pour les vecteurs des paramètres γ_j, β_j et θ_j.

3.1.3 Paramètres et autres cibles à estimer

Les paramètres évalués sont l'effet naturel direct marginal (END) et l'effet naturel indirect marginal (ENI). Une seconde cible à estimer est les modèles choisis par

les différentes procédures de sélection de variables.

3.1.4 Méthodes évaluées

Au total, six procédures de sélection de variables sont évaluées, soit :

1. Les méthodes basées sur l'effet total [1] : CIE_{total} et ΔMSE_{total} ;
2. Les méthodes basées sur une sélection séparée pour les effets naturels direct et indirect : CIE_{effets} et ΔMSE_{effets} ;
3. Les méthodes basées sur le maximum des effets naturels direct et indirect : CIE_{max} et ΔMSE_{max}.

Pour l'ensemble des procédures, le processus de sélection débute avec la totalité des 21 variables décrites dans la sous-section 3.1.2 sur les mécanismes de génération de données. Pour les méthodes basées sur le CIE, on utilise $\tau = 0.1$. De plus, bien que la variable réponse soit simulée avec un modèle où le terme d'interaction entre l'exposition et la réponse est nul, le terme d'interaction est tout de même inclus pour estimer les effets de médiation. Cette décision est basée sur le fait qu'il est généralement recommandé d'inclure le terme d'interaction entre l'exposition et la réponse (VanderWeele, 2015). En pratique, la sélection de variables est donc réalisée avec un modèle sur la réponse qui incorpore le terme d'interaction entre le médiateur et l'exposition, même lorsque l'estimation de ce paramètre est proche de 0. Une conséquence de cette décision est que le CIE_{total} et le ΔMSE_{total} doivent

---

1. La notation CIE_{total} et ΔMSE_{total} est utilisée pour souligner le fait que ces procédures sont réalisées sur l'effet total de l'exposition sur la réponse estimé comme la somme des effets naturels direct et indirect.

être réalisés en estimant l'effet total [2] comme la somme de l'effet naturel direct et l'effet naturel indirect. Pour être en mesure d'estimer adéquatement les effets marginaux (voir sous-section 2.2.3 pour un rappel), on conditionne aux valeurs moyennes des variables qui sont sélectionnées par les différentes procédures, la valeur moyenne d'une variable étant une estimation de son espérance théorique.

Pour évaluer correctement les qualités des six procédures proposées, la simulation inclut également un modèle qui ajuste pour la totalité des 21 variables (méthode *Complète*) et un modèle où l'ensemble d'ajustement est l'ensemble vide (méthode *Brute*). Le premier modèle permet de déterminer la performance des méthodes de sélection (par exemple, on peut se demander s'il est possible d'observer une amélioration de l'erreur quadratique moyenne lorsqu'une sélection de variables est effectuée ou encore si le biais et le taux de couverture sont similaires entre le modèle complet et les modèles réduits). Différemment, le second modèle permet d'illustrer l'importance d'ajuster adéquatement pour les variables de confusion (hypothèses H_1-H_3).

3.1.5 Mesures de performance

Les mesures de performance (MP) principales pour les paramètres à estimer (END et ENI) sont le biais, le biais relatif (BR), l'erreur standard empirique (esEMP), l'erreur quadratique moyenne (EQM), le gain de précision relatif (GPR), l'erreur standard moyenne du modèle (esMod [3] moyenne) et le taux de couverture

---

2. Les méthodes CIE et MSE standards ne permettent pas l'ajout du terme d'interaction entre l'exposition et le médiateur dans le modèle de la réponse. En fait, puisque M est un descendant de A, M ne peut être présent dans l'ensemble suffisant initial $\boldsymbol{C}^*$.

3. Le $esMod_i$ est l'estimation de l'erreur standard de l'estimateur pour la $i^{\text{ème}}$ simulation, c'est-à-dire $\sqrt{\widehat{V}(\hat{\theta})_i}$.

(TC). Une attention particulière est accordée au taux de couverture, car la sélection de variables a généralement pour conséquence de réduire artificiellement la taille des intervalles de confiance (Fox, 2016), ce qui se traduit par un taux de couverture inférieur à la valeur nominale attendue. De plus, pour plusieurs de ces mesures, l'erreur standard Monte-Carlo des estimateurs est également présentée. Enfin, des intervalles de confiance approximatifs [4] pour l'estimation du taux de couverture et du gain de précision basés sur l'erreur standard Monte-Carlo sont calculés. La définition, l'estimation et l'erreur standard Monte-Carlo (le cas échéant) de ces mesures de performance sont formulées dans le tableau 3.2.

TABLEAU 3.2 : Mesures de performance : définition, estimation et erreur standard Monte-Carlo

Mesures de performance	Définition	Estimation	Erreur standard Monte-Carlo
Biais	$\mathbb{E}[\hat{\theta}] - \theta$	$\frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_i - \theta$	$\sqrt{\frac{1}{m(m-1)}\sum_{i=1}^{m}(\hat{\theta}_i - \bar{\theta})^2}$
BR	$\frac{\mathbb{E}[\hat{\theta}] - \theta}{\theta}$	$\frac{1}{m}\sum_{i=1}^{m}\frac{\hat{\theta}_i - \theta}{\theta}$	
esEMP	$\sqrt{V[\hat{\theta}]}$	$\sqrt{\frac{1}{m-1}\sum_{i=1}^{m}(\hat{\theta}_i - \bar{\theta})^2}$	$\frac{\widehat{esEMP}}{\sqrt{2(m-1)}}$
GPR	$100\left(\frac{V[\hat{\theta}_A]}{V[\hat{\theta}_B]} - 1\right)$	$100\left(\left(\frac{\widehat{esEMP}_A}{\widehat{esEMP}_B}\right)^2 - 1\right)$	$200\left(\frac{\widehat{esEMP}_A}{\widehat{esEMP}_B}\right)^2 \cdot b$
EQM	$\mathbb{E}[(\hat{\theta} - \theta)^2]$	$\frac{1}{m}\sum_{i=1}^{m}(\hat{\theta}_i - \theta)^2$	$\sqrt{\frac{\sum_{i=1}^{m}\left[(\hat{\theta}_i - \theta)^2 - \widehat{EQM}\right]^2}{m(m-1)}}$
esMod moyenne	$\sqrt{\mathbb{E}[\hat{V}(\hat{\theta})]}$	$\sqrt{\frac{1}{m}\sum_{i=1}^{m}\hat{V}(\hat{\theta}_i)}$	$\sqrt{\frac{\hat{V}[\hat{V}(\hat{\theta})]}{4m\left(\widehat{esMod}\right)^2}}$
TC	$Pr(\hat{\theta}_{inf} \leq \theta \leq \hat{\theta}_{sup})$	$\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}(\hat{\theta}_{inf,i} \leq \theta \leq \hat{\theta}_{sup,i})$	$\sqrt{\frac{\widehat{TC}(1-\widehat{TC})}{m}}$

BR=Biais relatif ; esEMP=Erreur standard empirique ; GPR=Gain de précision relatif ; EQM=Erreur quadratique moyenne ; esMod moyenne=Erreur standard moyenne du modèle ; TC=Taux de couverture ; Les erreurs standards Monte-Carlo pour le gain de précision relatif et le $esMod$ moyenne sont des approximations ; *Note :* $\bar{\theta} = \frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_i$; $\hat{V}[\hat{V}(\hat{\theta})] = \frac{1}{m-1}\sum_{i=1}^{m}\{\hat{V}[\hat{\theta}_i] - \frac{1}{m}\sum_{j=1}^{m}\hat{V}[\hat{\theta}_j]\}^2$; $\hat{\theta}_{inf}$ est la borne inférieure de l'intervalle de confiance ; $\hat{\theta}_{sup}$ est la borne supérieure de l'intervalle de confiance ; $\mathbb{1}(\hat{\theta}_{inf,i} \leq \theta \leq \hat{\theta}_{sup,i}) = 1$ si θ est compris dans l'intervalle de confiance de l'estimateur, 0 sinon ; $b = \sqrt{\frac{1 - Corr(\hat{\theta}_A, \hat{\theta}_B)^2}{m-1}}$.

4. Approximation normale. Soit $\hat{\theta}$ la quantité estimée (par exemple, le taux de couverture) et $\widehat{V}(\hat{\theta})$ la variance Monte-Carlo, alors les intervalles de confiance sont donnés par $\hat{\theta} \pm \sqrt{\widehat{V}(\hat{\theta})}Z_{1-\alpha/2}$, avec $Z_{1-\alpha/2}$ le quantile d'une loi $\mathcal{N}(0,1)$.

Pour évaluer les modèles choisis par les différentes procédures, trois mesures de performance sont utilisées, soit le nombre moyen de variables sélectionnées, la proportion de modèles où les variables sélectionnées forment un ensemble suffisant pour la médiation (inclut l'intégralité des covariables $\boldsymbol{C}_{am}$, $\boldsymbol{C}_{ay}$ et $\boldsymbol{C}_{my}$) et la proportion de sélection pour chacune des variables. Ces mesures sont principalement utilisées de manière descriptive pour aider dans l'interprétation et la compréhension des résultats de la simulation.

3.2 Résultats de la simulation

3.2.1 Visualisation des résultats de simulation

Les figures 3.1 et 3.2 présentent, respectivement, des boîtes à moustaches pour les résultats de l'estimation des effets naturels direct et indirect pour les six scénarios de simulation. On remarque que la méthode *Brute* surestime généralement l'effet naturel direct et l'effet naturel indirect. Similairement, la méthode CIE_{total} sous-estime l'effet naturel direct, mais surestime l'effet naturel indirect. En comparaison, très peu de différences majeures sont décelées entre les autres méthodes de sélection, bien que la méthode CIE_{effets} semble présenter des estimations légèrement plus grandes pour l'effet naturel indirect. Pour mieux comprendre les résultats de la simulation, on poursuit par une présentation des mesures de performance pour la sélection de variables, soit le nombre moyen de variables sélectionnées, la proportion de sélection d'un ensemble suffisant pour la médiation et la proportion de sélection pour chacune des 21 covariables incluses dans le modèle de médiation.

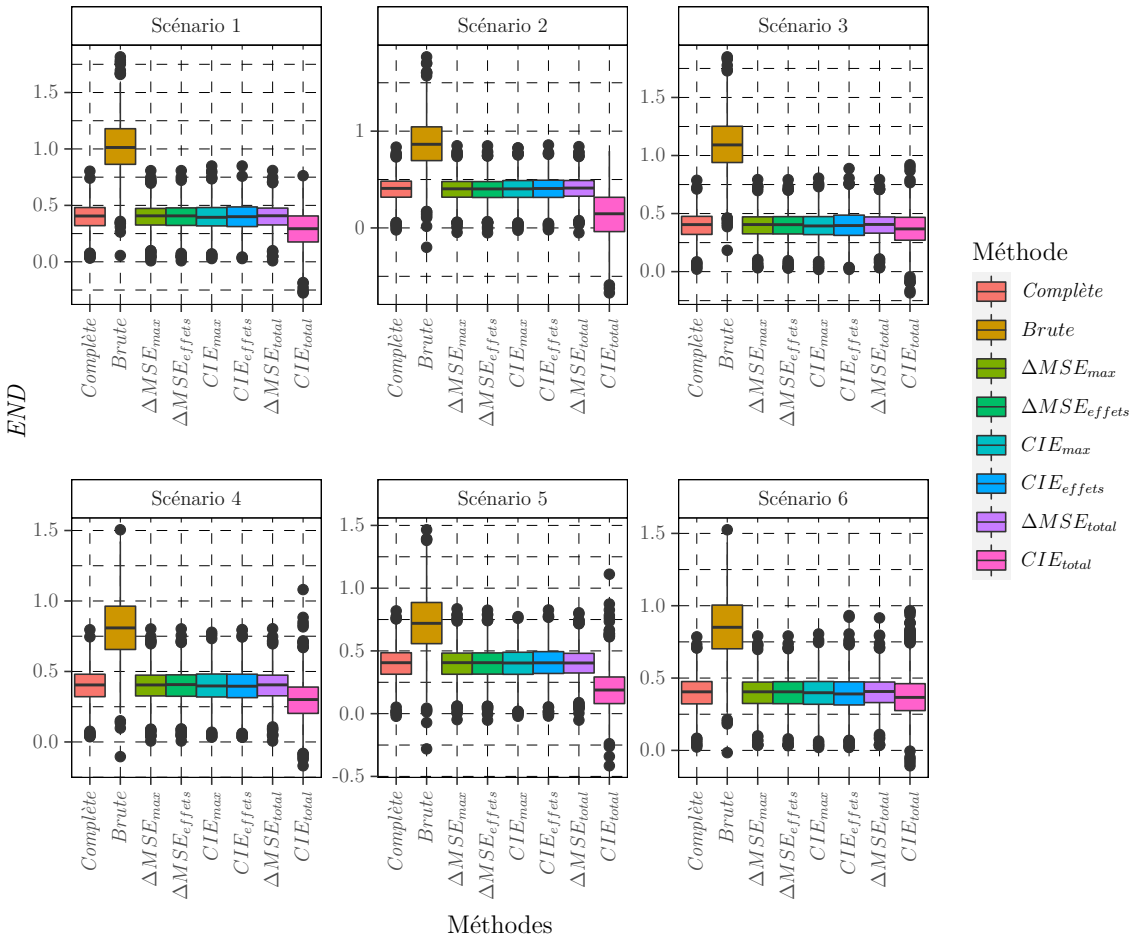


FIGURE 3.1 : Boîtes à moustaches pour les effets naturels directs (END=0.4)

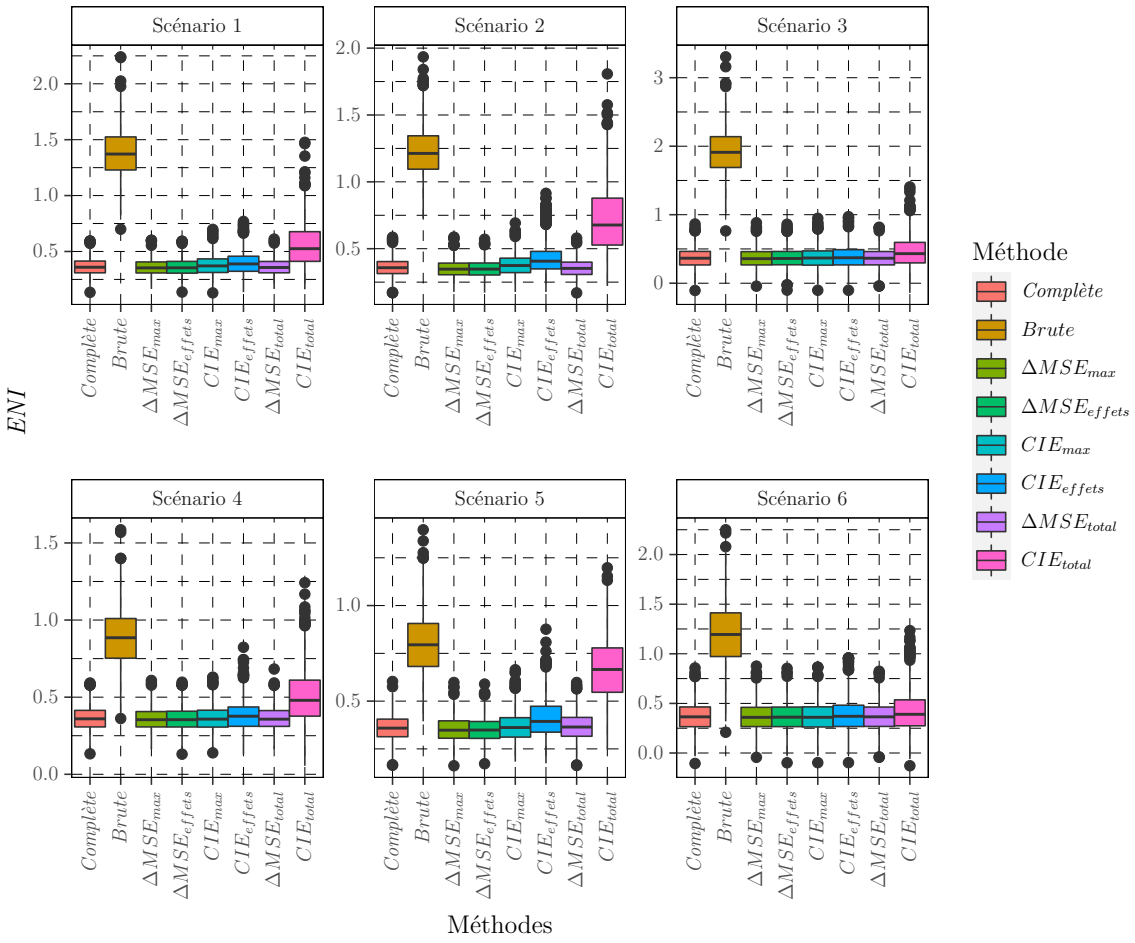


FIGURE 3.2 : Boîtes à moustaches pour les effets naturels indirects (ENI=0.36)

3.2.2 Sélection de variables

Le tableau 3.3 présente le nombre moyen de variables sélectionnées par chacune des six méthodes appliquées sur l'ensemble des jeux de données simulés dans les différents scénarios. Les résultats montrent que les méthodes *total* (ΔMSE et CIE) conservent généralement moins de variables, respectivement, que les méthodes ΔMSE et CIE qui ont été développées spécifiquement pour faire de la sélection dans un contexte de médiation. Ce phénomène est particulièrement prononcé pour le CIE_{total} qui, en moyenne, ne maintient jamais plus de dix variables dans le modèle de médiation à l'arrêt de l'algorithme. On note toutefois que le ΔMSE_{total} rejette, en moyenne, un nombre inférieur de variables que le ΔMSE_{max} dans les conditions où $\beta_1 > \theta_2$ (scénarios 2 et 5). On rappelle que β_1 correspond à l'effet de l'exposition sur le médiateur, alors que θ_2 traduit l'impact du médiateur sur la réponse. Un second constat est que les méthodes ΔMSE semblent sélectionner davantage de variables que les méthodes basées sur le CIE. Enfin, alors que le ΔMSE_{max} préserve moins de variables que la méthode ΔMSE_{effets}, l'inverse est observé lorsque l'on compare le CIE_{max} et CIE_{effets}.

TABLEAU 3.3 : Nombre moyen de variables sélectionnées selon les différentes méthodes proposées

Scénario	ΔMSE_{max}	ΔMSE_{effets}	CIE_{max}	CIE_{effets}	ΔMSE_{total}	CIE_{total}
1	16.93	18.13	11.57	11.19	16.84	6.83
2	15.71	17.28	11.02	10.52	16.41	5.62
3	17.00	18.20	12.96	12.59	16.89	8.53
4	16.94	18.10	12.40	11.96	16.84	8.04
5	15.74	17.32	11.96	11.17	16.40	5.76
6	17.01	18.23	13.51	13.08	16.89	9.76

Le tableau 3.4 présente la proportion de sélection d'un ensemble suffisant pour

la médiation pour les six méthodes comparées. On rappelle que l'ensemble suffisant doit nécessairement contenir les variables des vecteurs $\boldsymbol{C}_{ay}$, $\boldsymbol{C}_{am}$ et $\boldsymbol{C}_{my}$. Les résultats montrent que les méthodes ΔMSE_{max} et ΔMSE_{effets} préservent généralement un ensemble suffisant pour la médiation, avec des proportions qui varient de 95.2% (scénario 5) à 100.0% (scénarios 3, 4 et 6) pour le ΔMSE_{max}, et de 97.6% (scénario 2) à 100% (scénarios 1, 3, 4 et 6) pour le ΔMSE_{effet}. La proportion de sélection d'un ensemble suffisant varie de 76.6% (scénario 5) à 100% (scénario 3) pour le ΔMSE_{total}, avec les scénarios 2 et 5 ($\beta_1 > \theta_2$) qui présentent des proportions inférieures à 80%.

TABLEAU 3.4 : Proportion de sélection d'un ensemble suffisant pour la médiation (%)

Scénario	ΔMSE_{max}	ΔMSE_{effets}	CIE_{max}	CIE_{effets}	ΔMSE_{total}	CIE_{total}
1	99.9	100.0	37.4	25.3	98.9	0.7
2	96.4	97.6	12.7	6.1	78.5	0.0
3	100.0	100.0	47.5	38.9	100.0	4.6
4	100.0	100.0	91.8	75.9	98.6	8.4
5	95.2	98.7	77.8	38.5	76.6	1.8
6	100.0	100.0	87.9	74.5	99.6	21.2

Les méthodes basées sur le CIE affichent des proportions plus faibles que les méthodes basées sur le ΔMSE. En effet, la proportion de sélection d'un ensemble suffisant pour la médiation varie de 0% (scénario 2) à 21.2% (scénario 6) pour le CIE_{total}, de 6.1% (scénario 2) à 75.9% (scénario 4) pour le CIE_{effets} et de 12.7% (scénario 2) à 91.8% (scénario 4) pour le CIE_{max}. Toutefois, il est intéressant de noter qu'un ensemble qui n'est pas suffisant de manière théorique peut malgré tout mener à des résultats acceptables empiriquement. En effet, bien que certaines procédures pourraient ne pas parvenir à conserver un ensemble suffisant dans certains scénarios, le biais relatif des estimateurs qui résulte de ces méthodes de

sélection pourrait être inférieur à 10%.

Pour discerner davantage la qualité des six algorithmes, le tableau 3.5 illustre la proportion de sélection pour les variables qui sont dans l'ensemble minimal. Les résultats montrent que les méthodes ΔMSE_{effets} et ΔMSE_{max} sélectionnent généralement la totalité des neuf variables de confusion. Similairement, à l'exception de la variable C_{am1} dans le scénario 2 (78.8%) et de la variable C_{am3} dans le scénario 5 (76.8%), le ΔMSE_{total} conserve les variables de l'ensemble minimal avec des proportions supérieures à 95%. Ces résultats suggèrent donc que les variables $\boldsymbol{C}_{am}$ qui possèdent une association relativement faible avec le médiateur sont davantage rejetées par le ΔMSE_{total}.

Pour le CIE_{max}, à l'exception des facteurs de confusion qui ont des effets faibles (C_{ay1} et C_{my1}) dans les scénarios symétriques (scénarios 1, 2 et 3) et de la variable C_{am1} dans les scénarios 1, 2 et 5, les facteurs de confusion sont sélectionnés avec des proportions supérieures à 90%. Il est possible de faire deux constats pour le CIE_{effets}. D'une part, les proportions de sélection sont généralement inférieures à celles rapportées par la méthode CIE_{max}. D'autre part, en plus des variables précédemment mentionnées pour le CIE_{max}, le CIE_{effets} conserve avec des proportions inférieures à 90% le facteur de confusion C_{am1} dans les scénarios 3 et 4, C_{am2} lorsque $\beta_1 > \theta_2$ (scénarios 2 et 5), C_{am3} dans le scénario 5 et C_{my3} dans le scénario 6. Enfin, de manière cohérente avec ce qui a été présenté antérieurement, le CIE_{total} a généralement de la difficulté à maintenir les variables de confusion entre le médiateur et la réponse ($\boldsymbol{C}_{my}$). De plus, la proportion de sélection est inférieure à 90% pour la variable C_{ay1} dans les scénarios symétriques (scénarios 1, 2 et 3) et le scénario 5, pour la variable C_{ay3} dans le scénario 6, la variable C_{am1} dans toutes les conditions sauf le scénario 6, la variable C_{am2} lorsque $\theta_2 \leq \beta_1$ (scénarios 1, 2, 4 et 5) et pour la variable C_{am3} dans la condition asymétrique lorsque $\theta_2 \leq \beta_1$ (scénarios 4 et 5).

TABLEAU 3.5 : Proportion de sélection des variables appartenant à l'ensemble minimal (%)

Scénario	Méthode	C_{ay_1}	C_{ay_2}	C_{ay_3}	C_{am_1}	C_{am_2}	C_{am_3}	C_{my_1}	C_{my_2}	C_{my_3}
1	ΔMSE_{max}	100.0	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0
	ΔMSE_{effets}	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	CIE_{max}	80.5	100.0	100.0	59.8	99.9	100.0	70.3	99.8	100.0
	CIE_{effets}	75.9	99.8	100.0	47.7	97.9	100.0	66.6	99.8	100.0
	ΔMSE_{total}	100.0	100.0	100.0	98.9	100.0	100.0	100.0	100.0	100.0
	CIE_{total}	38.2	99.0	100.0	12.8	86.8	100.0	10.9	40.2	62.3
2	ΔMSE_{max}	100.0	100.0	100.0	96.6	99.8	100.0	100.0	100.0	100.0
	ΔMSE_{effets}	100.0	100.0	100.0	97.6	100.0	100.0	100.0	100.0	100.0
	CIE_{max}	78.9	100.0	100.0	18.7	90.5	100.0	87.9	100.0	100.0
	CIE_{effets}	77.1	99.9	100.0	12.4	60.9	94.3	83.9	100.0	100.0
	ΔMSE_{total}	100.0	100.0	100.0	78.8	99.7	100.0	100.0	100.0	100.0
	CIE_{total}	34.8	99.3	100.0	2.0	42.4	98.5	4.3	30.1	55.4
3	ΔMSE_{max}	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	ΔMSE_{effets}	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	CIE_{max}	83.2	100.0	100.0	90.1	100.0	100.0	63.1	98.6	99.9
	CIE_{effets}	76.0	99.8	100.0	83.8	99.9	100.0	59.7	96.3	99.7
	ΔMSE_{total}	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	CIE_{total}	38.2	98.0	100.0	50.4	99.0	100.0	25.8	55.2	72.2
4	ΔMSE_{max}	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	ΔMSE_{effets}	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	CIE_{max}	97.3	100.0	100.0	96.6	99.9	99.9	99.6	99.9	97.2
	CIE_{effets}	95.8	99.8	99.9	89.9	98.5	99.2	96.3	98.8	91.6
	ΔMSE_{total}	100.0	100.0	100.0	100.0	100.0	98.6	100.0	100.0	100.0
	CIE_{total}	91.0	98.3	91.9	83.2	87.1	74.9	52.6	40.8	44.2
5	ΔMSE_{max}	100.0	100.0	100.0	95.8	99.9	99.4	100.0	100.0	100.0
	ΔMSE_{effets}	100.0	100.0	100.0	98.7	100.0	100.0	100.0	100.0	100.0
	CIE_{max}	97.8	100.0	100.0	86.6	91.8	95.1	99.7	100.0	99.8
	CIE_{effets}	96.2	99.9	99.7	71.2	66.9	73.4	98.7	100.0	99.5
	ΔMSE_{total}	100.0	100.0	100.0	100.0	99.7	76.8	100.0	100.0	100.0
	CIE_{total}	89.7	97.8	90.5	40.9	36.0	38.3	44.4	26.6	21.4
6	ΔMSE_{max}	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	ΔMSE_{effets}	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	CIE_{max}	97.3	100.0	100.0	98.9	100.0	100.0	97.4	97.9	92.1
	CIE_{effets}	95.6	99.9	99.9	96.6	100.0	100.0	92.8	95.0	88.7
	ΔMSE_{total}	100.0	100.0	99.6	100.0	100.0	100.0	100.0	100.0	100.0
	CIE_{total}	91.9	97.5	89.4	95.7	98.6	95.0	59.8	57.6	62.8

Le tableau 3.6 présente les proportions de sélection pour les variables qui n'appartiennent pas à l'ensemble minimal. Les résultats montrent que les méthodes basées sur le CIE rejettent dans des proportions importantes les variables de bruit ($\boldsymbol{B}$). De plus, la proportion de sélection est généralement peu élevée pour les prédicteurs purs P_{a1}, P_{m1} et P_{y1}, mais augmente graduellement avec l'amplification des effets des variables des vecteurs $\boldsymbol{P}_a$, $\boldsymbol{P}_m$ et $\boldsymbol{P}_y$.

Les procédures basées sur le ΔMSE rejettent généralement les variables de bruit. Toutefois, les proportions de sélection sont plus élevées que pour les méthodes qui reposent sur le CIE, et ce phénomène est particulièrement notable pour le ΔMSE_{effets}. La proportion de sélection pour les prédicteurs purs de l'exposition ($\boldsymbol{P}_a$) est similaire à celle observée pour les variables de bruit. De plus, on note que les prédicteurs purs de la réponse ($\boldsymbol{P}_y$), d'une part, et les prédicteurs purs du médiateur ($\boldsymbol{P}_m$) lorsque $\theta_2 \geq \beta_1$ (scénarios 1, 3, 4 et 6), d'autre part, sont régulièrement maintenus pour le ΔMSE_{total}, le ΔMSE_{max} et le ΔMSE_{effets}. Lorsque $\theta_2 < \beta_1$ (scénarios 2 et 5), la procédure ΔMSE_{total} sélectionne généralement les prédicteurs purs P_{m2} et P_{m3}, mais P_{m1} est conservée dans des proportions plus faibles (environ 80%). Différemment, la méthode ΔMSE_{max} rejette fréquemment les variables du prédicteur purs $\boldsymbol{P}_m$, avec des proportions de sélection qui varient de 53.2% (variable P_{m1} du scénario 2) à 64% (variable P_{m3} du scénario 5). Similairement, la proportion de sélection des variables du vecteur $\boldsymbol{P}_m$ varie de 62% (variable P_{m1} du scénario 2) à 90.5% (variable P_{m3} du scénario 5) pour la procédure ΔMSE_{effets}. Le fait que les variables du vecteur $\boldsymbol{P}_m$ sont rejetées dans des proportions plus importantes par les méthodes ΔMSE_{max} et ΔMSE_{effets} dans certains scénarios n'est pas surprenant, car l'étude de Diop et ses collaborateurs (2021) a montré que l'inclusion des prédicteurs purs du médiateur avait un effet délétère sur l'estimation de la variance pour les effets naturels direct et indirect lorsque $\theta_2 < \beta_1$.

TABLEAU 3.6 : Proportion de sélection des variables qui n'appartenant pas à l'ensemble minimal (%)

Scénario	Méthode	P_{a_1}	P_{a_2}	P_{a_3}	P_{m_1}	P_{m_2}	P_{m_3}	P_{y_1}	P_{y_2}	P_{y_3}	B_1	B_2	B_3
1	ΔMSE_{max}	34.1	32.9	30.2	99.0	99.9	100.0	100.0	100.0	100.0	31.2	33.3	32.2
	ΔMSE_{effets}	52.9	51.9	51.5	100.0	100.0	100.0	100.0	100.0	100.0	50.6	53.1	53.3
	CIE_{max}	5.4	19.8	52.0	10.4	37.8	62.7	26.1	55.6	75.4	0.7	0.6	0.3
	CIE_{effets}	5.2	20.5	49.5	10.3	33.4	60.4	25.5	53.2	71.2	0.8	0.8	0.6
	ΔMSE_{total}	30.4	31.4	32.8	99.1	100.0	100.0	100.0	100.0	100.0	31.1	30.5	29.5
	CIE_{total}	0.8	5.6	22.0	0.6	4.5	28.8	4.1	20.6	46.1	0.1	0.0	0.0
2	ΔMSE_{max}	36.4	31.8	30.4	53.2	60.9	62.3	100.0	100.0	100.0	32.4	33.9	33.5
	ΔMSE_{effets}	57.6	54.7	52.2	62.0	65.7	90.0	100.0	100.0	100.0	49.5	49.5	49.6
	CIE_{max}	4.2	13.8	39.2	5.5	26.0	57.7	28.7	65.1	83.7	0.9	0.5	0.6
	CIE_{effets}	4.8	16.3	38.7	6.9	28.4	55.1	30.2	61.9	78.7	0.9	0.9	1.1
	ΔMSE_{total}	30.4	29.5	31.7	81.5	99.8	100.0	100.0	100.0	100.0	29.7	30.1	29.3
	CIE_{total}	0.2	3.4	18.0	0.2	0.7	5.5	3.4	19.6	44.2	0.0	0.0	0.0
3	ΔMSE_{max}	35.4	34.5	31.4	100.0	100.0	100.0	100.0	100.0	100.0	32.7	34.4	32.0
	ΔMSE_{effets}	52.8	51.6	50.5	99.8	100.0	100.0	100.0	100.0	100.0	54.1	54.6	56.1
	CIE_{max}	13.4	40.5	70.4	34.3	60.7	79.6	28.0	55.5	72.2	2.7	2.2	2.0
	CIE_{effets}	14.1	39.9	66.9	34.3	58.4	77.3	26.1	52.4	66.3	2.9	2.9	2.4
	ΔMSE_{total}	31.7	31.8	31.5	99.9	100.0	100.0	100.0	100.0	100.0	31.7	31.3	30.8
	CIE_{total}	1.9	11.6	30.6	6.5	28.6	57.1	5.5	24.1	47.6	0.3	0.2	0.0
4	ΔMSE_{max}	34.3	32.8	30.9	99.0	99.9	100.0	100.0	100.0	100.0	31.8	33.1	31.8
	ΔMSE_{effets}	53.3	51.3	49.7	100.0	100.0	100.0	100.0	100.0	100.0	50.4	52.3	52.7
	CIE_{max}	4.9	20.0	53.0	10.7	37.2	63.8	26.6	55.2	76.9	0.8	0.6	0.2
	CIE_{effets}	5.0	20.0	49.7	10.8	33.4	59.3	24.8	50.9	70.5	0.9	0.7	0.6
	ΔMSE_{total}	30.3	31.3	32.7	99.0	100.0	100.0	100.0	100.0	100.0	31.5	30.8	29.4
	CIE_{total}	0.7	6.5	23.8	0.8	4.8	29.8	4.0	22.5	47.2	0.0	0.1	0.0
5	ΔMSE_{max}	36.3	32.2	30.3	54.2	61.0	64.0	100.0	100.0	100.0	32.7	34.4	33.4
	ΔMSE_{effets}	56.8	55.2	51.3	63.5	67.2	90.5	100.0	100.0	100.0	50.6	49.9	48.8
	CIE_{max}	4.4	13.1	39.7	5.3	26.0	56.3	29.1	64.1	84.6	1.1	0.6	0.6
	CIE_{effets}	4.4	16.1	38.5	6.6	25.2	52.3	28.8	58.7	77.8	1.3	0.8	0.7
	ΔMSE_{total}	30.5	29.9	32.1	80.9	99.8	100.0	100.0	100.0	100.0	30.3	30.6	29.3
	CIE_{total}	0.0	4.2	17.6	0.1	0.4	4.6	2.8	17.6	42.7	0.0	0.0	0.0
6	ΔMSE_{max}	35.5	34.2	31.5	100.0	100.0	100.0	99.9	100.0	100.0	33.5	34.5	31.8
	ΔMSE_{effets}	54.0	52.5	50.8	99.8	100.0	100.0	100.0	100.0	100.0	53.3	55.3	56.8
	CIE_{max}	14.7	41.4	72.0	34.0	61.3	80.3	28.4	55.8	72.2	2.9	2.4	2.1
	CIE_{effets}	14.2	39.8	68.6	33.7	58.0	75.1	25.7	51.0	65.0	3.3	2.5	2.8
	ΔMSE_{total}	31.7	31.8	31.6	99.9	100.0	100.0	100.0	100.0	100.0	32.1	31.7	30.9
	CIE_{total}	2.0	12.6	32.9	7.9	31.1	56.8	7.0	26.6	49.9	0.3	0.1	0.0

Maintenant que la description des variables sélectionnées a été effectuée, il est opportun de présenter les mesures de performance (biais, EQM, etc.) des six procédures proposées, de la méthode *Complète* et de la méthode *Brute*. La présentation est réalisée séparément pour l'effet naturel direct et l'effet naturel indirect.

3.2.3 Résultats pour l'effet naturel direct

3.2.3.1 Biais, erreur standard et erreur quadratique moyenne

Le tableau 3.7 présente les résultats pour le biais, le biais relatif, l'erreur standard empirique, l'erreur standard moyenne du modèle et la racine carrée de l'erreur quadratique moyenne (rEQM) pour l'effet naturel direct. L'erreur standard Monte-Carlo pour plusieurs de ces mesures est également indiquée entre parenthèses. La simulation montre que la méthode *Brute* surestime l'effet naturel direct, avec un biais relatif qui varie de 79.17% (scénario 5) à 173.3% (scénario 3). Le CIE_{total} possède un biais relatif qui varie de -6.349% (scénario 6) à -67.63% (scénario 2), ce qui suggère une sous-estimation de l'effet naturel direct. En effet, il est possible de remarquer que la valeur absolue du biais relatif pour le CIE_{total} est supérieure à 10% pour les scénarios où $\beta_1 \geq \theta_2$ (scénarios 1, 2, 4 et 5), mais entre 5% et 10% dans les conditions où $\beta_1 < \theta_2$ (scénarios 3 et 6). Pour les autres procédures, la valeur absolue du biais relatif est inférieure à 2.5% pour l'ensemble des situations évaluées.

Les méthodes CIE_{total} et *Brute* présentent une rEQM qui est supérieure à la méthode *Complète*. Similairement, les résultats révèlent que la méthode CIE_{effets} a une rEQM qui est supérieure à la méthode *Complète* pour l'ensemble des scénarios, mais cette différence est moins importante que celle observée pour les méthodes CIE_{total} et *Brute*. La rEQM pour la sélection basée sur le CIE_{max} est similaire à celle de la méthode *Complète* dans l'ensemble des conditions. On note toutefois que les valeurs de la rEQM sont plus grandes pour la méthode CIE_{max} que la méthode *Complète* dans les scénarios symétriques (scénarios 1, 2 et 3), alors que l'inverse est observé dans les conditions asymétriques (scénarios 4, 5 et 6). Enfin, les procédures basées sur le ΔMSE affichent des rEQMs qui sont similaires, voir

inférieures à celles obtenues par la méthode *Complète*.

TABLEAU 3.7 : Mesures de performance (Erreur standard Monte-Carlo entre parenthèses) pour l'effet naturel direct (END=0.4)

Scénario	Méthode	Estimation	BR (%)	Biais	esEMP	esMod	rEQM
1	*Complète*	0.401	0.336	0.001 (0.004)	0.118 (0.003)	0.124 (0.000)	0.118 (0.001)
	Brute	1.018	154.4	0.618 (0.008)	0.245 (0.005)	0.242 (0.000)	0.664 (0.010)
	ΔMSE_{max}	0.401	0.255	0.001 (0.004)	0.116 (0.003)	0.115 (0.000)	0.116 (0.001)
	ΔMSE_{effets}	0.401	0.328	0.001 (0.004)	0.117 (0.003)	0.117 (0.000)	0.117 (0.001)
	CIE_{max}	0.398	-0.528	-0.002 (0.004)	0.120 (0.003)	0.143 (0.001)	0.120 (0.001)
	CIE_{effets}	0.400	-0.087	-0.000 (0.004)	0.125 (0.003)	0.146 (0.001)	0.125 (0.001)
	ΔMSE_{total}	0.403	0.795	0.003 (0.004)	0.114 (0.003)	0.115 (0.000)	0.114 (0.001)
	CIE_{total}	0.283	-29.15	-0.117 (0.006)	0.177 (0.004)	0.179 (0.001)	0.212 (0.002)
2	*Complète*	0.402	0.495	0.002 (0.004)	0.129 (0.003)	0.134 (0.000)	0.129 (0.001)
	Brute	0.867	116.8	0.467 (0.008)	0.264 (0.006)	0.259 (0.000)	0.536 (0.008)
	ΔMSE_{max}	0.401	0.243	0.001 (0.004)	0.126 (0.003)	0.121 (0.000)	0.126 (0.001)
	ΔMSE_{effets}	0.400	-0.013	-0.000 (0.004)	0.127 (0.003)	0.125 (0.000)	0.127 (0.001)
	CIE_{max}	0.405	1.186	0.005 (0.004)	0.132 (0.003)	0.143 (0.001)	0.132 (0.001)
	CIE_{effets}	0.405	1.371	0.005 (0.004)	0.134 (0.003)	0.147 (0.001)	0.134 (0.001)
	ΔMSE_{total}	0.410	2.405	0.010 (0.004)	0.127 (0.003)	0.126 (0.000)	0.127 (0.001)
	CIE_{total}	0.129	-67.63	-0.271 (0.008)	0.241 (0.005)	0.192 (0.001)	0.362 (0.005)
3	*Complète*	0.401	0.256	0.001 (0.004)	0.115 (0.003)	0.121 (0.000)	0.115 (0.001)
	Brute	1.093	173.3	0.693 (0.007)	0.236 (0.005)	0.235 (0.000)	0.732 (0.011)
	ΔMSE_{max}	0.401	0.202	0.001 (0.004)	0.112 (0.002)	0.112 (0.000)	0.112 (0.001)
	ΔMSE_{effets}	0.401	0.254	0.001 (0.004)	0.114 (0.003)	0.114 (0.000)	0.114 (0.001)
	CIE_{max}	0.397	-0.751	-0.003 (0.004)	0.117 (0.003)	0.146 (0.001)	0.117 (0.001)
	CIE_{effets}	0.399	-0.230	-0.001 (0.004)	0.126 (0.003)	0.150 (0.001)	0.126 (0.001)
	ΔMSE_{total}	0.402	0.545	0.002 (0.003)	0.110 (0.002)	0.112 (0.000)	0.110 (0.001)
	CIE_{total}	0.368	-8.029	-0.032 (0.005)	0.152 (0.003)	0.173 (0.001)	0.156 (0.001)
4	*Complète*	0.401	0.277	0.001 (0.004)	0.118 (0.003)	0.124 (0.000)	0.118 (0.001)
	Brute	0.806	101.5	0.406 (0.007)	0.235 (0.005)	0.240 (0.000)	0.469 (0.006)
	ΔMSE_{max}	0.401	0.216	0.001 (0.004)	0.116 (0.003)	0.115 (0.000)	0.116 (0.001)
	ΔMSE_{effets}	0.401	0.352	0.001 (0.004)	0.117 (0.003)	0.117 (0.000)	0.117 (0.001)
	CIE_{max}	0.398	-0.443	-0.002 (0.004)	0.118 (0.003)	0.142 (0.001)	0.118 (0.001)
	CIE_{effets}	0.396	-1.021	-0.004 (0.004)	0.121 (0.003)	0.149 (0.001)	0.121 (0.001)
	ΔMSE_{total}	0.402	0.562	0.002 (0.004)	0.114 (0.003)	0.115 (0.000)	0.114 (0.001)
	CIE_{total}	0.300	-25.07	-0.100 (0.005)	0.150 (0.003)	0.186 (0.001)	0.180 (0.001)
5	*Complète*	0.402	0.386	0.002 (0.004)	0.129 (0.003)	0.134 (0.000)	0.129 (0.001)
	Brute	0.717	79.17	0.317 (0.008)	0.252 (0.006)	0.254 (0.000)	0.404 (0.006)
	ΔMSE_{max}	0.401	0.299	0.001 (0.004)	0.127 (0.003)	0.121 (0.000)	0.127 (0.001)
	ΔMSE_{effets}	0.400	0.035	0.000 (0.004)	0.127 (0.003)	0.125 (0.000)	0.127 (0.001)
	CIE_{max}	0.401	0.300	0.001 (0.004)	0.128 (0.003)	0.142 (0.001)	0.128 (0.001)
	CIE_{effets}	0.405	1.200	0.005 (0.004)	0.134 (0.003)	0.148 (0.001)	0.134 (0.001)
	ΔMSE_{total}	0.401	0.372	0.001 (0.004)	0.125 (0.003)	0.125 (0.000)	0.125 (0.001)
	CIE_{total}	0.189	-52.68	-0.211 (0.005)	0.172 (0.004)	0.199 (0.001)	0.272 (0.003)
6	*Complète*	0.401	0.222	0.001 (0.004)	0.115 (0.003)	0.121 (0.000)	0.115 (0.001)
	Brute	0.851	112.7	0.451 (0.007)	0.229 (0.005)	0.235 (0.000)	0.505 (0.007)
	ΔMSE_{max}	0.401	0.194	0.001 (0.004)	0.112 (0.002)	0.112 (0.000)	0.112 (0.001)
	ΔMSE_{effets}	0.401	0.332	0.001 (0.004)	0.113 (0.003)	0.115 (0.000)	0.113 (0.001)
	CIE_{max}	0.398	-0.591	-0.002 (0.004)	0.115 (0.003)	0.147 (0.001)	0.115 (0.001)
	CIE_{effets}	0.392	-1.987	-0.008 (0.004)	0.123 (0.003)	0.156 (0.001)	0.124 (0.001)
	ΔMSE_{total}	0.403	0.761	0.003 (0.004)	0.112 (0.002)	0.112 (0.000)	0.112 (0.001)
	CIE_{total}	0.375	-6.349	-0.025 (0.005)	0.156 (0.003)	0.180 (0.001)	0.158 (0.001)

Note : Estimation=La moyenne des estimateurs ; BR=Biais relatif ; esEMP=Erreur standard empirique ; rEQM=Racine carrée de l'erreur quadratique moyenne. L'Erreur standard Monte-Carlo associée au rEQM est celle de l'erreur quadratique moyenne.

3.2.3.2 Gain de précision relatif

Le tableau 3.8 présente une estimation ponctuelle et par intervalle du gain de précision relatif (GPR) pour les méthodes qui ne présentaient pas de biais relatif supérieur à 10% pour l'effet naturel direct comparativement au modèle qui inclut toutes les variables. Si on se réfère au tableau 3.2, $\hat{\theta}_A$ correspond à la méthode *Complète* et $\hat{\theta}_B$ représente les autres procédures comparées.

TABLEAU 3.8 : Gain de précision relatif (%) pour les méthodes étudiées comparativement au modèle complet pour l'effet naturel direct (Erreur standard Monte-Carlo entre parenthèses)

Scénario	Méthode	Gain de précision relatif	Intervalle de confiance (95%)
1	ΔMSE_{max}	3.71 (1.49)	[0.79, 6.62]
	ΔMSE_{effets}	2.12 (1.18)	[-0.18, 4.43]
	ΔMSE_{total}	7.11 (1.78)	[3.63, 10.6]
	CIE_{max}	-2.78 (2.25)	[-7.19, 1.62]
	CIE_{effets}	-10.5 (2.42)	[-15.2, -5.72]
2	ΔMSE_{max}	3.81 (1.54)	[0.80, 6.82]
	ΔMSE_{effets}	2.90 (1.27)	[0.41, 5.39]
	ΔMSE_{total}	3.32 (1.46)	[0.46, 6.18]
	CIE_{max}	-5.45 (2.17)	[-9.70, -1.19]
	CIE_{effets}	-8.22 (2.22)	[-12.6, -3.86]
3	ΔMSE_{max}	6.59 (1.85)	[2.97, 10.2]
	ΔMSE_{effets}	2.06 (1.20)	[-0.30, 4.42]
	ΔMSE_{total}	9.47 (2.21)	[5.13, 13.8]
	CIE_{max}	-2.84 (2.04)	[-6.84, 1.15]
	CIE_{effets}	-16.8 (2.38)	[-21.4, -12.1]
4	ΔMSE_{max}	3.64 (1.48)	[0.74, 6.54]
	ΔMSE_{effets}	2.83 (1.21)	[0.46, 5.19]
	ΔMSE_{total}	7.67 (1.85)	[4.04, 11.3]
	CIE_{max}	0.66 (2.20)	[-3.65, 4.97]
	CIE_{effets}	-4.49 (2.86)	[-10.1, 1.12]
5	ΔMSE_{max}	3.74 (1.53)	[0.74, 6.74]
	ΔMSE_{effets}	3.44 (1.28)	[0.94, 5.94]
	ΔMSE_{total}	7.16 (1.70)	[3.83, 10.5]
	CIE_{max}	1.19 (2.21)	[-3.14, 5.52]
	CIE_{effets}	-6.78 (2.44)	[-11.6, -2.01]
6	ΔMSE_{max}	6.43 (1.84)	[2.81, 10.0]
	ΔMSE_{effets}	3.29 (1.23)	[0.88, 5.70]
	ΔMSE_{total}	6.54 (2.26)	[2.11, 11.0]
	CIE_{max}	0.34 (2.08)	[-3.73, 4.41]
	CIE_{effets}	-12.7 (2.77)	[-18.1, -7.24]

Relativement à la méthode *Complète*, il y a un gain (perte) de précision si le GPR est supérieur (inférieur) à 0. Les résultats montrent que l'utilisation des algorithmes basés sur le ΔMSE permet d'obtenir un gain de précision statistiquement significatif comparativement à la méthode *Complète*, à l'exception du ΔMSE_{effets} dans les scénarios symétriques lorsque $\beta_1 \leq \theta_2$ (scénarios 1 et 3). Ce gain de précision est généralement plus important pour le ΔMSE_{total}, alors qu'il semble être moins accentué pour le ΔMSE_{effets}. Inversement, hormis le scénario 4 où l'intervalle de confiance inclut la valeur 0, le CIE_{effets} présente un GPR qui est négatif, ce qui suggère que la variance de l'estimateur de l'effet naturel direct pour cette procédure est plus importante qu'un modèle qui n'opère aucune sélection de variables. Enfin, en dehors du scénario 2 où il y a une perte de précision relative, la variance associée au CIE_{max} semble être similaire à celle obtenue par la méthode *Complète*.

3.2.3.3 Taux de couverture

Le tableau 3.9 affiche le taux de couverture pour l'effet naturel direct des différentes méthodes. Un taux de couverture est dit *à 95%* si l'intervalle de confiance de ce taux inclut 0.95. Pour mieux comprendre les résultats, il est également possible de se référer au tableau 3.7 pour obtenir l'erreur standard moyenne du modèle et l'erreur standard empirique. Les résultats montrent que les taux de couverture sont à 95% pour les méthodes *Complète*, ΔMSE_{effets} et ΔMSE_{total} pour l'ensemble des scénarios. Similairement, hormis pour le scénario 2, le ΔMSE_{max} produit également des taux de couverture à 95%. À l'opposé, les taux de couverture des méthodes CIE_{total} et *Brute* s'éloignent du taux de couverture nominal à 95%, à l'exception du scénario 4 pour le CIE_{total}. La méthode *Brute* démontre un taux de couverture qui est nettement inférieur à 95%. Différemment, le taux

de couverture du CIE_{total} est supérieur à 95% lorsque $\theta_2 > \beta_1$ (scénarios 3 et 6), mais il est inférieur à 95% lorsque $\beta_1 \geq \theta_2$ (scénarios 1, 2 et 5).

TABLEAU 3.9 : Taux de couverture pour l'effet naturel direct (Erreur standard Monte-Carlo entre parenthèse)

Scénario	Méthode	Taux de couverture	Intervalle de confiance (95%)
1	*Complète*	0.955 (0.007)	[0.942, 0.968]
	Brute	0.267 (0.014)	[0.240, 0.294]
	ΔMSE_{max}	0.947 (0.007)	[0.933, 0.961]
	ΔMSE_{effets}	0.946 (0.007)	[0.932, 0.960]
	CIE_{max}	0.971 (0.005)	[0.961, 0.981]
	CIE_{effets}	0.968 (0.006)	[0.957, 0.979]
	ΔMSE_{total}	0.955 (0.007)	[0.942, 0.968]
	CIE_{total}	0.907 (0.009)	[0.889, 0.925]
2	*Complète*	0.955 (0.007)	[0.942, 0.968]
	Brute	0.561 (0.016)	[0.530, 0.592]
	ΔMSE_{max}	0.934 (0.008)	[0.919, 0.949]
	ΔMSE_{effets}	0.947 (0.007)	[0.933, 0.961]
	CIE_{max}	0.964 (0.006)	[0.952, 0.976]
	CIE_{effets}	0.957 (0.006)	[0.944, 0.970]
	ΔMSE_{total}	0.947 (0.007)	[0.933, 0.961]
	CIE_{total}	0.679 (0.015)	[0.650, 0.708]
3	*Complète*	0.959 (0.006)	[0.947, 0.971]
	Brute	0.160 (0.012)	[0.137, 0.183]
	ΔMSE_{max}	0.954 (0.007)	[0.941, 0.967]
	ΔMSE_{effets}	0.950 (0.007)	[0.936, 0.964]
	CIE_{max}	0.976 (0.005)	[0.967, 0.985]
	CIE_{effets}	0.969 (0.005)	[0.958, 0.980]
	ΔMSE_{total}	0.956 (0.006)	[0.943, 0.969]
	CIE_{total}	0.968 (0.006)	[0.957, 0.979]
4	*Complète*	0.955 (0.007)	[0.942, 0.968]
	Brute	0.605 (0.015)	[0.575, 0.635]
	ΔMSE_{max}	0.948 (0.007)	[0.934, 0.962]
	ΔMSE_{effets}	0.949 (0.007)	[0.935, 0.963]
	CIE_{max}	0.971 (0.005)	[0.961, 0.981]
	CIE_{effets}	0.973 (0.005)	[0.963, 0.983]
	ΔMSE_{total}	0.954 (0.007)	[0.941, 0.967]
	CIE_{total}	0.953 (0.007)	[0.940, 0.966]
5	*Complète*	0.954 (0.007)	[0.941, 0.967]
	Brute	0.763 (0.013)	[0.737, 0.789]
	ΔMSE_{max}	0.940 (0.008)	[0.925, 0.955]
	ΔMSE_{effets}	0.952 (0.007)	[0.939, 0.965]
	CIE_{max}	0.963 (0.006)	[0.951, 0.975]
	CIE_{effets}	0.956 (0.006)	[0.943, 0.969]
	ΔMSE_{total}	0.952 (0.007)	[0.939, 0.965]
	CIE_{total}	0.858 (0.011)	[0.836, 0.880]
6	*Complète*	0.962 (0.006)	[0.950, 0.974]
	Brute	0.517 (0.016)	[0.486, 0.548]
	ΔMSE_{max}	0.955 (0.007)	[0.942, 0.968]
	ΔMSE_{effets}	0.952 (0.007)	[0.939, 0.965]
	CIE_{max}	0.980 (0.004)	[0.971, 0.989]
	CIE_{effets}	0.973 (0.005)	[0.963, 0.983]
	ΔMSE_{total}	0.953 (0.007)	[0.940, 0.966]
	CIE_{total}	0.970 (0.005)	[0.959, 0.981]

Pour l'ensemble des scénarios, le CIE_{max} présente un taux de couverture qui est légèrement supérieur à 95%. Lorsque $\beta_1 > \theta_2$ (scénarios 2 et 5), le taux de couverture de la procédure CIE_{effets} est à 95%, mais il est légèrement supérieur à cette quantité lorsque $\theta_2 \leq \beta_1$ (scénarios 1, 3, 4 et 6). Pour les méthodes CIE_{max} et CIE_{effets}, l'erreur standard empirique est plus petite que l'erreur standard moyenne du modèle, ce qui suggère que la variance est surestimée lorsque l'on réduit le nombre de variables à l'aide de ces algorithmes. Cette observation explique probablement la sur-couverture observée pour ces méthodes dans certains scénarios.

Principaux points saillants

–> Biais

(a) Les méthodes CIE_{total} (4/6 scénarios) et *Brute* (6/6) possèdent généralement un biais dans l'estimation de l'effet naturel direct.

(b) Les méthodes CIE_{max} (6/6), CIE_{effets} (6/6), ΔMSE_{total} (6/6), ΔMSE_{max} (6/6) et ΔMSE_{effets} (6/6) présentent des biais relatifs inférieurs à 10%.

–> Gain de précision des méthodes sans biais

(a) Les méthodes ΔMSE_{total} (6/6), ΔMSE_{max} (6/6) et ΔMSE_{effets} (4/6) offrent généralement un gain de précision comparativement à la méthode *Complète*.

(b) La méthode CIE_{effets} (5/6) mène habituellement à une perte de précision relativement à la méthode *Complète*.

(c) Le plus souvent, le CIE_{max} (5/6) ne produit ni gain ni perte de précision comparativement à la méthode *Complète*.

–> Taux de couverture

(a) Seules les méthodes ΔMSE_{total} (6/6), ΔMSE_{max} (5/6) et ΔMSE_{effets} (6/6) obtiennent communément des taux de couverture à 95%.

3.2.4 Résultats pour l'effet naturel indirect

3.2.4.1 Biais, erreur standard et erreur quadratique moyenne

Le tableau 3.10 présente les résultats pour le biais, le biais relatif, l'erreur standard empirique, l'erreur standard moyenne du modèle et la rEQM pour l'effet naturel indirect. Pour la majorité de ces mesures de performance, une erreur standard Monte-Carlo est également affichée entre parenthèses. La simulation montre que la méthode *Brute* surestime l'effet naturel indirect, avec un biais relatif qui varie de 122.4% (scénario 5) à 432.3% (scénario 3). Le CIE_{total} surestime également l'effet naturel indirect, avec un biais relatif qui varie de 14.34% (scénario 6) à 98.54% (scénario 2). Similairement aux résultats concernant l'effet naturel direct, les méthodes ΔMSE possèdent un biais relatif (en valeur absolue) qui est inférieur à 3% pour l'ensemble des scénarios. Le CIE_{max} possède d'un biais relatif qui est légèrement plus élevé comparativement aux méthodes ΔMSE, mais demeure inférieur à 10%. En effet, le biais relatif pour le CIE_{max} varie de 0.707% (scénario 4) à 5.102% (scénario 2). Enfin, le biais relatif du CIE_{effets} varie de 5.521% (scénario 4) à 17.28% (scénario 2), et il est supérieur à 10% dans les scénarios où $\beta_1 > \theta_2$ (scénarios 2 et 5).

Les méthodes CIE_{total} et *Brute* présentent une rEQM qui est largement supérieure à la méthode *Complète*. Les résultats suggèrent que les méthodes CIE_{max} et CIE_{effets} ont une rEQM qui est supérieure à la méthode *Complète*, mais cette différence est moins importante que celle observée pour les méthodes CIE_{total} et *Brute*. En règle générale, dans l'ensemble des scénarios, il semble que le CIE_{max} présente une rEQM qui est inférieure à celle obtenue par la méthode CIE_{effets}. Enfin, les procédures basées sur le ΔMSE affichent des rEQMs qui sont similaires, voir inférieures à celles obtenues par la méthode *Complète*, à l'exception du

ΔMSE_{total} dans le scénario 5.

TABLEAU 3.10 : Mesures de performance (Erreur standard Monte-Carlo entre parenthèses) pour l'effet naturel indirect (ENI=0.36)

Scénario	Méthode	Estimation	BR (%)	Biais	esEMP	esMod	rEQM
1	*Complète*	0.362	0.624	0.002 (0.002)	0.079 (0.002)	0.078 (0.000)	0.079 (0.000)
	Brute	1.379	283.2	1.019 (0.007)	0.224 (0.005)	0.218 (0.000)	1.044 (0.015)
	ΔMSE_{max}	0.358	-0.503	-0.002 (0.002)	0.074 (0.002)	0.073 (0.000)	0.074 (0.000)
	ΔMSE_{effets}	0.358	-0.434	-0.002 (0.002)	0.077 (0.002)	0.074 (0.000)	0.077 (0.000)
	CIE_{max}	0.377	4.793	0.017 (0.003)	0.093 (0.002)	0.097 (0.001)	0.095 (0.000)
	CIE_{effets}	0.395	9.643	0.035 (0.003)	0.106 (0.002)	0.100 (0.001)	0.111 (0.001)
	ΔMSE_{total}	0.360	0.102	0.000 (0.002)	0.076 (0.002)	0.073 (0.000)	0.076 (0.000)
	CIE_{total}	0.555	54.16	0.195 (0.006)	0.195 (0.004)	0.165 (0.002)	0.276 (0.004)
2	*Complète*	0.360	0.014	0.000 (0.002)	0.068 (0.002)	0.070 (0.000)	0.068 (0.000)
	Brute	1.224	240.0	0.864 (0.006)	0.192 (0.004)	0.186 (0.000)	0.885 (0.011)
	ΔMSE_{max}	0.351	-2.614	-0.009 (0.002)	0.067 (0.001)	0.065 (0.000)	0.067 (0.000)
	ΔMSE_{effets}	0.350	-2.843	-0.010 (0.002)	0.067 (0.001)	0.065 (0.000)	0.068 (0.000)
	CIE_{max}	0.378	5.102	0.018 (0.003)	0.080 (0.002)	0.075 (0.000)	0.082 (0.000)
	CIE_{effets}	0.422	17.28	0.062 (0.004)	0.113 (0.003)	0.080 (0.001)	0.129 (0.001)
	ΔMSE_{total}	0.354	-1.555	-0.006 (0.002)	0.067 (0.002)	0.068 (0.000)	0.067 (0.000)
	CIE_{total}	0.715	98.54	0.355 (0.008)	0.248 (0.006)	0.132 (0.001)	0.433 (0.007)
3	*Complète*	0.366	1.578	0.006 (0.005)	0.148 (0.003)	0.144 (0.000)	0.149 (0.001)
	Brute	1.916	432.3	1.556 (0.011)	0.346 (0.008)	0.339 (0.001)	1.594 (0.035)
	ΔMSE_{max}	0.364	1.112	0.004 (0.005)	0.144 (0.003)	0.133 (0.000)	0.144 (0.001)
	ΔMSE_{effets}	0.365	1.435	0.005 (0.005)	0.145 (0.003)	0.136 (0.000)	0.145 (0.001)
	CIE_{max}	0.374	3.766	0.014 (0.005)	0.159 (0.004)	0.170 (0.001)	0.160 (0.001)
	CIE_{effets}	0.385	7.054	0.025 (0.005)	0.171 (0.004)	0.174 (0.001)	0.173 (0.001)
	ΔMSE_{total}	0.365	1.275	0.005 (0.005)	0.143 (0.003)	0.133 (0.000)	0.143 (0.001)
	CIE_{total}	0.458	27.16	0.098 (0.007)	0.227 (0.005)	0.244 (0.002)	0.247 (0.003)
4	*Complète*	0.362	0.678	0.002 (0.002)	0.079 (0.002)	0.078 (0.000)	0.079 (0.000)
	Brute	0.886	146.2	0.526 (0.006)	0.195 (0.004)	0.188 (0.000)	0.561 (0.007)
	ΔMSE_{max}	0.358	-0.452	-0.002 (0.002)	0.075 (0.002)	0.073 (0.000)	0.075 (0.000)
	ΔMSE_{effets}	0.359	-0.409	-0.001 (0.002)	0.077 (0.002)	0.074 (0.000)	0.077 (0.000)
	CIE_{max}	0.363	0.707	0.003 (0.003)	0.084 (0.002)	0.096 (0.001)	0.084 (0.000)
	CIE_{effets}	0.380	5.521	0.020 (0.003)	0.099 (0.002)	0.104 (0.001)	0.101 (0.001)
	ΔMSE_{total}	0.362	0.532	0.002 (0.002)	0.077 (0.002)	0.073 (0.000)	0.077 (0.000)
	CIE_{total}	0.506	40.64	0.146 (0.006)	0.179 (0.004)	0.155 (0.001)	0.231 (0.003)
5	*Complète*	0.360	0.124	0.000 (0.002)	0.069 (0.002)	0.070 (0.000)	0.069 (0.000)
	Brute	0.801	122.4	0.441 (0.005)	0.162 (0.004)	0.153 (0.000)	0.469 (0.005)
	ΔMSE_{max}	0.352	-2.277	-0.008 (0.002)	0.067 (0.002)	0.065 (0.000)	0.068 (0.000)
	ΔMSE_{effets}	0.350	-2.678	-0.010 (0.002)	0.067 (0.002)	0.065 (0.000)	0.068 (0.000)
	CIE_{max}	0.366	1.794	0.006 (0.002)	0.077 (0.002)	0.074 (0.000)	0.077 (0.000)
	CIE_{effets}	0.407	13.14	0.047 (0.003)	0.103 (0.002)	0.080 (0.000)	0.113 (0.001)
	ΔMSE_{total}	0.366	1.577	0.006 (0.002)	0.072 (0.002)	0.069 (0.000)	0.073 (0.000)
	CIE_{total}	0.669	85.73	0.309 (0.005)	0.173 (0.004)	0.126 (0.001)	0.354 (0.004)
6	*Complète*	0.366	1.605	0.006 (0.005)	0.149 (0.003)	0.144 (0.000)	0.149 (0.001)
	Brute	1.191	230.9	0.831 (0.010)	0.323 (0.007)	0.315 (0.001)	0.892 (0.018)
	ΔMSE_{max}	0.364	1.074	0.004 (0.005)	0.144 (0.003)	0.133 (0.000)	0.144 (0.001)
	ΔMSE_{effets}	0.364	1.185	0.004 (0.005)	0.145 (0.003)	0.136 (0.000)	0.145 (0.001)
	CIE_{max}	0.365	1.296	0.005 (0.005)	0.151 (0.003)	0.173 (0.002)	0.151 (0.001)
	CIE_{effets}	0.379	5.385	0.019 (0.005)	0.168 (0.004)	0.184 (0.002)	0.169 (0.001)
	ΔMSE_{total}	0.364	1.213	0.004 (0.005)	0.143 (0.003)	0.133 (0.000)	0.143 (0.001)
	CIE_{total}	0.412	14.34	0.052 (0.006)	0.200 (0.004)	0.236 (0.002)	0.206 (0.002)

Note : Estimation=La moyenne des estimateurs ; BR=Biais relatif ; esEMP=Erreur standard empirique ; rEQM=Racine carrée de l'erreur quadratique moyenne. L'Erreur standard Monte-Carlo associée au rEQM est celle de l'erreur quadratique moyenne.

3.2.4.2 Gain de précision relatif

Le tableau 3.11 présente le GPR avec un intervalle de confiance pour les méthodes qui ne présentaient pas de biais relatif supérieur à 10% pour l'effet naturel indirect comparativement au modèle où l'ensemble des variables est inclus.

TABLEAU 3.11 : Gain de précision relatif (%) pour les méthodes étudiées comparativement au modèle complet pour l'effet naturel indirect (Erreur standard Monte-Carlo entre parenthèses)

Scénario	Méthode	Gain de précision relatif	Intervalle de confiance (95%)
1	ΔMSE_{max}	12.4 (1.98)	[8.52, 16.3]
	ΔMSE_{effets}	4.42 (1.10)	[2.26, 6.57]
	ΔMSE_{total}	8.97 (2.10)	[4.86, 13.1]
	CIE_{max}	-28.8 (2.13)	[-33.0, -24.7]
2	ΔMSE_{max}	4.34 (2.24)	[-0.05, 8.73]
	ΔMSE_{effets}	4.62 (1.28)	[2.11, 7.12]
	ΔMSE_{total}	3.22 (1.70)	[-0.11, 6.54]
	CIE_{max}	-27.7 (2.32)	[-32.3, -23.2]
3	ΔMSE_{max}	6.59 (1.53)	[3.60, 9.59]
	ΔMSE_{effets}	4.73 (1.18)	[2.43, 7.04]
	ΔMSE_{total}	7.87 (1.83)	[4.28, 11.5]
	CIE_{max}	-12.8 (1.54)	[-15.9, -9.81]
4	ΔMSE_{max}	11.2 (1.94)	[7.45, 15.0]
	ΔMSE_{effets}	4.24 (1.15)	[1.99, 6.48]
	ΔMSE_{total}	4.20 (2.25)	[-0.21, 8.60]
	CIE_{max}	-10.8 (2.50)	[-15.7, -5.93]
5	ΔMSE_{max}	3.68 (2.22)	[-0.67, 8.04]
	ΔMSE_{effets}	4.71 (1.31)	[2.14, 7.28]
	ΔMSE_{total}	-9.87 (2.22)	[-14.2, -5.51]
	CIE_{max}	-20.2 (2.49)	[-25.1, -15.3]
6	ΔMSE_{max}	6.60 (1.53)	[3.60, 9.60]
	ΔMSE_{effets}	4.27 (1.13)	[2.06, 6.48]
	ΔMSE_{total}	8.24 (1.87)	[4.57, 11.9]
	CIE_{max}	-3.22 (1.59)	[-6.33, -0.10]

Pour l'effet naturel indirect, le ΔMSE_{effets} affiche un gain de précision comparativement à la méthode *Complète* dans l'ensemble des scénarios, alors que l'inverse est observé pour le CIE_{max}. Le ΔMSE_{max} mène à un gain de précision comparativement au modèle *Complet* dans les scénarios où $\theta_2 \geq \beta_1$ (scénarios 1, 3, 4 et 6), alors que la variance des deux approches est similaire dans les autres condi-

tions. Le ΔMSE_{total} démontre un gain de précision relatif comparativement à la méthode *Complète* dans les scénarios 1, 3 et 6. Toutefois, dans le scénario 5, on remarque que le gain de précision relatif est négatif (-9.87%) et différent de 0. Le gain de précision est particulièrement élevé pour la méthode ΔMSE_{max} dans les scénarios où $\beta_1 = \theta_2$ (scénarios 1 et 4).

3.2.4.3 Taux de couverture

Le tableau 3.12 affiche le taux de couverture pour l'effet naturel indirect des différentes méthodes. Il est à noter que les bornes de l'intervalle de confiance pour le taux couverture peuvent être des valeurs inférieures à 0 ou supérieures à 1. Dans tous les scénarios, la méthode *Complète* possède des taux de couverture à 95%, alors que l'inverse est observé pour la méthode *Brute* où la totalité des taux de couverture est nettement inférieure à 95%. La méthode CIE_{total} produit un taux de couverture à 95% lorsque $\theta_2 > \beta_1$ (scénarios 3 et 6), mais affiche généralement une forte sous-couverture dans les autres situations.

Les approches basées sur le ΔMSE ne présentent pas la même performance concernant le taux de couverture pour l'effet naturel indirect comparativement à ce qui a été observé pour l'effet naturel direct. En effet, on observe une légère sous-couverture dans certains scénarios. Néanmoins, le ΔMSE_{total} obtient un taux de couverture à 95% lorsque $\beta_1 \geq \theta_2$ (scénarios 1, 2, 4 et 5), alors que le ΔMSE_{max} présente un taux de couverture à 95% lorsque $\beta_1 = \theta_2$ (scénarios 1 et 4). De plus, le taux de couverture du ΔMSE_{effets} est à 95% pour les scénarios 1, 2 et 4. Si on observe le tableau 3.10, on remarque que l'erreur standard moyenne du modèle pour les méthodes basées sur le ΔMSE est généralement similaire ou inférieure à l'erreur standard empirique, ce qui explique probablement la sous-couverture observée dans plusieurs scénarios.

TABLEAU 3.12 : Taux de couverture pour l'effet naturel indirect (Erreur standard Monte-Carlo entre parenthèse)

Scénario	Méthode	Taux de couverture	Intervalle de confiance (95%)
1	*Complète*	0.939 (0.008)	[0.924, 0.954]
	Brute	0.001 (0.001)	[-0.001, 0.003]
	ΔMSE_{max}	0.940 (0.008)	[0.925, 0.955]
	ΔMSE_{effets}	0.937 (0.008)	[0.922, 0.952]
	CIE_{max}	0.956 (0.006)	[0.943, 0.969]
	CIE_{effets}	0.931 (0.008)	[0.915, 0.947]
	ΔMSE_{total}	0.936 (0.008)	[0.921, 0.951]
	CIE_{total}	0.803 (0.013)	[0.778, 0.828]
2	*Complète*	0.960 (0.006)	[0.948, 0.972]
	Brute	0.000 (0.000)	[0.000, 0.000]
	ΔMSE_{max}	0.932 (0.008)	[0.916, 0.948]
	ΔMSE_{effets}	0.942 (0.007)	[0.928, 0.956]
	CIE_{max}	0.923 (0.008)	[0.906, 0.940]
	CIE_{effets}	0.841 (0.012)	[0.818, 0.864]
	ΔMSE_{total}	0.947 (0.007)	[0.933, 0.961]
	CIE_{total}	0.363 (0.015)	[0.333, 0.393]
3	*Complète*	0.937 (0.008)	[0.922, 0.952]
	Brute	0.001 (0.001)	[-0.001, 0.003]
	ΔMSE_{max}	0.930 (0.008)	[0.914, 0.946]
	ΔMSE_{effets}	0.931 (0.008)	[0.915, 0.947]
	CIE_{max}	0.950 (0.007)	[0.936, 0.964]
	CIE_{effets}	0.932 (0.008)	[0.916, 0.948]
	ΔMSE_{total}	0.932 (0.008)	[0.916, 0.948]
	CIE_{total}	0.951 (0.007)	[0.938, 0.964]
4	*Complète*	0.943 (0.007)	[0.929, 0.957]
	Brute	0.191 (0.012)	[0.167, 0.215]
	ΔMSE_{max}	0.942 (0.007)	[0.928, 0.956]
	ΔMSE_{effets}	0.936 (0.008)	[0.921, 0.951]
	CIE_{max}	0.968 (0.006)	[0.957, 0.979]
	CIE_{effets}	0.953 (0.007)	[0.940, 0.966]
	ΔMSE_{total}	0.937 (0.008)	[0.922, 0.952]
	CIE_{total}	0.831 (0.012)	[0.808, 0.854]
5	*Complète*	0.953 (0.007)	[0.940, 0.966]
	Brute	0.143 (0.011)	[0.121, 0.165]
	ΔMSE_{max}	0.934 (0.008)	[0.919, 0.949]
	ΔMSE_{effets}	0.934 (0.008)	[0.919, 0.949]
	CIE_{max}	0.940 (0.008)	[0.925, 0.955]
	CIE_{effets}	0.864 (0.011)	[0.843, 0.885]
	ΔMSE_{total}	0.946 (0.007)	[0.932, 0.960]
	CIE_{total}	0.350 (0.015)	[0.320, 0.380]
6	*Complète*	0.936 (0.008)	[0.921, 0.951]
	Brute	0.248 (0.014)	[0.221, 0.275]
	ΔMSE_{max}	0.930 (0.008)	[0.914, 0.946]
	ΔMSE_{effets}	0.932 (0.008)	[0.916, 0.948]
	CIE_{max}	0.955 (0.007)	[0.942, 0.968]
	CIE_{effets}	0.953 (0.007)	[0.940, 0.966]
	ΔMSE_{total}	0.933 (0.008)	[0.918, 0.948]
	CIE_{total}	0.962 (0.006)	[0.950, 0.974]

La procédure CIE_{max} possède un taux de couverture à 95% dans les scénarios 1, 3, 5 et 6, mais elle présente une légère sous-couverture dans le scénario 2 et une légère sur-couverture dans le scénario 4. De manière cohérente avec ces conclusions, les résultats illustrés dans le tableau 3.10 montrent que l'erreur standard moyenne du modèle est inférieure à l'erreur standard empirique pour le scénario 2, alors que l'inverse est remarqué dans le scénario 4. La méthode CIE_{effets} obtient un taux de couverture à 95% dans les scénarios asymétriques où $\beta_1 \leq \theta_2$ (scénarios 4 et 6), alors qu'une sous-couverture est observée dans les autres contextes. Cette sous-couverture résulte probablement de la présence d'un certain biais dans les estimations et d'une erreur standard du modèle (à l'exception du scénario 3) qui est inférieure à l'erreur standard empirique.

Principaux points saillants

–> Biais

(a) Les méthodes CIE_{total} (6/6 scénarios) et *Brute* (6/6) possèdent un biais dans l'estimation de l'effet naturel indirect.

(b) Les méthodes CIE_{max} (6/6), ΔMSE_{total} (6/6), ΔMSE_{max} (6/6) et ΔMSE_{effets} (6/6) présentent des biais relatifs inférieurs à 10%. La méthode CIE_{effets} offre un biais relatif supérieur à 10% dans certaines conditions (2/6).

–> Gain de précision des méthodes sans biais

(a) Le ΔMSE_{max} (4/6) et le ΔMSE_{effets} (6/6) offrent généralement un gain de précision comparativement à la méthode *Complète*.

(b) Le CIE_{max} (6/6) mène à une perte de précision relativement à la méthode *Complète*.

(c) Le ΔMSE_{total} produit un gain de précision comparativement à la méthode *Complète* dans certaines conditions (3/6), mais une perte de précision dans un scénario (1/6).

–> Taux de couverture

(a) Les méthodes CIE_{total} (2/6), ΔMSE_{max} (2/6), CIE_{effets} (2/6) et ΔMSE_{effets} (3/6) ont des taux de couverture à 95% dans 50% ou moins des scénarios, alors que

(b) Les méthodes CIE_{max} (4/6) et ΔMSE_{total} (4/6) obtiennent des taux de couverture à 95% dans plus de 50% des conditions.

CHAPITRE IV

APPLICATION

Ce chapitre propose un exemple d'application des méthodes de sélection de variables sur un jeu de données réelles provenant de la vague 1999 de l'étude sur l'alcool du *Harvard School of Public Health College*. Après une brève mise en contexte du modèle de médiation évalué, la méthodologie utilisée et les résultats des analyses de médiation qui découlent des différentes méthodes de sélection de variables sont présentés.

4.1 Mise en contexte

La consommation d'alcool (CA) et la consommation excessive d'alcool sur une courte période de temps (*Binge drinking* ; CEA) ont été associées à une diminution de la performance académique chez les étudiants universitaires (Edkins, Edgerton, & Roberts, 2017; Williams, Powell, & Wechsler, 2003). Selon Williams et ses collaborateurs (2003), cet effet est partiellement attribuable au fait que la consommation d'alcool réduit le temps accordé aux études à l'extérieur des classes, ce qui réduit la performance académique. Ainsi, on peut tenter de déterminer si la consommation régulière d'alcool (variable d'exposition) est associée à une diminution de la performance académique (variable réponse) directement et

indirectement à travers une réduction du temps accordé aux études à l'extérieur des classes (variable de médiation) (voir figure 4.1 pour un DAG du modèle de médiation excluant les covariables). L'application est largement inspirée des travaux de Williams et ses collaborateurs (2003). Puisque les méthodes de sélection de variables évaluées sont particulièrement utiles pour les petits échantillons, cette illustration est réalisée sur la base d'un sous-échantillon d'étudiants sexuellement actifs qui ont mentionné avoir déjà eu un rapport sexuel avec un partenaire de même sexe.

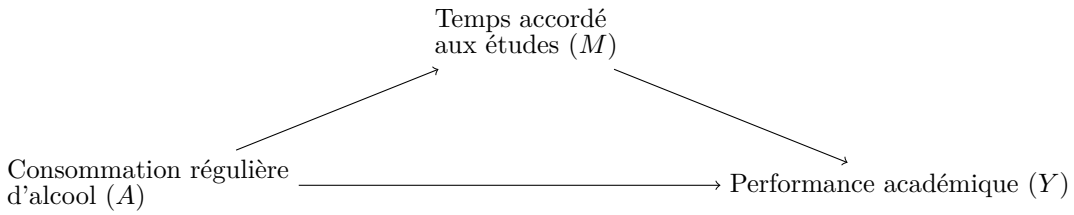


FIGURE 4.1 : DAG du modèle de médiation

4.2 Méthodologie

Les données proviennent de la vague 1999 de l'étude sur l'alcool du *Harvard School of Public Health College*. Puisque le terme *college* au Canada n'a pas la même signification qu'aux États-Unis, le terme *établissement post-secondaire* est utilisé dans la section qui suit pour faire référence aux collèges et aux universités américaines. Cette étude a débuté en 1993 comme une enquête nationalement représentative des établissements post-secondaires sélectionnés à partir de la liste des établissements post-secondaires accrédités de l'*American Council on Education*. Les participants ont été sélectionnés de manière aléatoire avec une probabilité qui était proportionnelle à la taille d'échantillonnage. Comme cela est recommandé par certains analystes lorsque l'on utilise une approche par modèle (voir Heeringa,

West et Berglund (2020), p. 207-208, pour une discussion), le poids d'échantillonnage est omis dans les modèles de régression subséquents. Initialement, 13 954 participants étaient inclus dans la version 1999 de l'étude sur l'alcool du *Harvard School of Public Health College*. Un total de 2996 étudiants ont été exclus parce qu'ils possédaient au moins une valeur manquante sur une des variables utilisées dans l'analyse de médiation. Parmi les 10 958 étudiants restants, 472 ont indiqué avoir eu au moins un rapport sexuel avec une personne de même sexe au cours de leur vie, ce qui correspond à la taille finale de l'échantillon.

4.2.1 Mesures

Un indicateur binaire (oui/non) de consommation régulière d'alcool est obtenu à l'aide de deux questions. Dans un premier temps, les étudiants devaient mentionner le moment où ils avaient consommé un verre pour la dernière fois (Q1). Ceux qui avaient consommé un verre au cours des 30 derniers jours devaient répondre à la question suivante : «À combien de reprises avez-vous bu de l'alcool au cours des 30 derniers jours ?» [1] (Q2). Les réponses possibles étaient «N'a pas bu au cours des 30 derniers jours (1)», «1 à 2 occasions (2)», «3 à 5 occasions (3)», «6 à 9 occasions (4)», «10 à 19 occasions (5)», «20 à 39 occasions (6)» et «40 occasions ou plus (7)». Les étudiants qui n'ont pas bu au cours des 30 derniers jours (Q1) ou qui ont consommé un verre à moins de 10 occasions au cours des 30 derniers jours (Q2) sont considérés comme n'ayant pas une consommation régulière d'alcool (non), alors que les étudiants qui ont rapporté avoir consommé un verre à 10 occasions ou plus au cours des 30 derniers jours (Q2) sont considérés comme ayant une consommation régulière d'alcool (oui). Le médiateur est évalué à l'aide

1. Pour l'ensemble des questions et choix de réponses, il s'agit de traductions qui ne sont pas officielles.

de la question suivante : «Au cours des 30 derniers jours, combien d'heures par jour en moyenne avez-vous consacrées à chacune des activités suivantes ?», avec une des activités qui était «Étudier en dehors des cours». Les choix de réponses étaient 0 heure (0), 1 heure (1), 2 heures (2), 3 heures (3), 4 heures (4) ou 5 heures ou plus (5). La performance académique est mesurée à l'aide du « *Grade Point Average* (GPA)». Plus précisément, les étudiants devaient indiquer «[l]equel des énoncés suivants décrit le mieux votre moyenne pondérée cumulative cette année ?». Comme cela a été effectué par Williams et ses collaborateurs (2003), les réponses ont été codifiées en neuf catégories, soit A (4), A- (3.7), B+ (3.3), B (3), B- (2.7), C+(2.3), C(2), C- (1.7) et D (1). Les participants qui n'avaient pas encore obtenu de note ou qui ne connaissaient pas la réponse à cette question ont reçu une valeur manquante pour cet item.

Les covariables utilisées pour la sélection de variables peuvent être conceptualisées en huit catégories. Les variables *sociodémographiques* (six items) incluent l'âge (15 ans à 25 ans et plus), le sexe biologique (Homme/Femme), la religiosité (Importance de la religion : «Très importante (1)», «Importante (2)», «Plutôt importante (3)», «Pas du tout importante (4)»), le fait de n'avoir jamais été marié (Oui/-Non), le fait d'être caucasien (Oui/Non) et l'éducation parentale (Éducation plus élevée que le secondaire : Oui/Non). Comme cela a été effectué par Halpern et ses collaborateurs (2004), l'éducation parentale est obtenue en observant l'éducation la plus élevée entre le père et la mère, le cas échéant.

Les variables en lien avec l'*établissement post-secondaire* (six items) incluent le fait d'être membre d'une fraternité ou d'une sororité (Oui/Non), la satisfaction avec l'éducation obtenue («Très satisfait (1)», «Plutôt satisfait (2)», «Plutôt insatisfait (3)», «Très insatisfait (4)»), l'acceptation de la gestion de l'alcool par l'établissement post-secondaire («Fortement d'accord (1)», «En accord (2)», «En désaccord (3)», «Fortement en désaccord (4)»), le fait d'avoir porté plainte à

l'établissement post-secondaire au cours de la dernière année à propos de comportements d'étudiants qui présentaient un niveau élevé d'intoxication (Oui/Non), le fait d'avoir une résidence sur le campus (Oui/Non) et le fait d'être un étudiant gradué (Oui/Non).

Les variables en lien avec la *consommation de substance* (deux items) incluent le fait d'avoir fumé au moins une cigarette au cours des 30 derniers jours (Oui/Non) et le fait d'avoir consommé un autre type de drogue (par exemple, marijuana, cocaïne, etc.) au cours des 30 derniers jours (Oui/Non).

Les variables en lien avec la *consommation d'alcool des autres* (trois items) incluent le fait que les parents consomment de l'alcool (Oui/Non), le fait que plus de 70% des amis à l'université sont des buveurs excessifs (*Binge drinker* : Oui/Non) et le fait de croire que la consommation d'alcool est un problème pour les étudiants sur le campus («Un problème majeur (1)», «Un problème (2)», «Un problème mineur (3)», «Pas un problème (4)»).

Les variables concernant les *conséquences de la consommation d'alcool des autres* (trois items) incluent le fait d'avoir subi à cause de la consommation d'alcool des autres étudiants au moins un épisode au cours de l'année scolaire d'humiliation ou d'insulte (Oui/Non), de préjudice physique (A été poussé, frappé ou agressé : Oui/Non) ou de préjudice sexuel (A été victime d'agression sexuelle : Oui/Non).

Deux items portent sur la santé des répondants. Plus précisément, on évalue la santé générale subjective («Excellente (1)», «Très bonne (2)», «Bonne (3)», «Passable (4)», «Mauvaise (5)») et la vitalité au cours des 30 derniers jours. La vitalité au cours des 30 derniers jours est mesurée à l'aide de quatre questions comme «Combien de temps au cours des 30 derniers jours avez-vous eu beaucoup d'énergie». Les choix de réponses étaient «Tout le temps (1)», «La plupart du temps (2)», «Une bonne partie du temps(3)», «Une partie du temps (4)», «Un peu de

temps (5)» et «Jamais (6)». Initialement, les questions ont été inversées pour qu'un score élevé indique un haut niveau de vitalité. Le score de vitalité a ensuite été construit en prenant la somme des quatre questions. Par la suite, on a soustrait cette somme par 4 (le minimum possible de la somme) et divisée par 21 (le nombre de réponses possibles pour la somme des quatre questions). Enfin, pour obtenir un score entre 0 et 100, on a multiplié le résultat par 100 (voir McDowell (2006), p. 650-654, pour la codification de ce concept). La cohérence interne [2], telle que mesurée par l'alpha de Cronbach, est acceptable ($\alpha = 0.772$).

Une autre série de questions (six items) portent sur le *temps moyen par jour consacré à certaines activités* au cours des 30 derniers jours (0 heure (0), 1 heure (1), 2 heures (2), 3 heures (3), 4 heures (4) ou 5 heures ou plus (5)), comme regarder la télévision ou des vidéos, travailler, socialiser avec des amis, participer à une organisation étudiante, jouer ou pratiquer un sport interuniversitaire, avoir d'autres activités physiques et faire du bénévolat.

Enfin, deux questions portent sur *l'école secondaire*. Plus particulièrement, on évalue si les participants étaient des buveurs excessifs au secondaire (Oui/Non) et s'ils faisaient de l'athlétisme au secondaire (Oui/Non). Le lecteur peut se référer au livre de code (*codebook* ; `https://www.icpsr.umich.edu/web/HMCA/studies/3818/datadocumentation#`) pour obtenir plus d'information sur les questions et les choix de réponses. Un code SAS pour la construction des variables qui combinent différents indicateurs y est également disponible. Le tableau 4.1 présente les statistiques descriptives (sans le poids d'échantillonnage) pour les différentes variables, selon la consommation régulière d'alcool.

---

2. La cohérence interne permet de déterminer si les items qui mesurent le même construit obtiennent des valeurs similaires. L'alpha de Cronbach varie de l'infini négatif à 1. Un alpha de Cronbach qui est proche de 1 est indicatif d'une cohérence interne élevée.

TABLEAU 4.1 : Statistiques descriptives selon la consommation régulière d'alcool

		Consommation régulière d'alcool	
	Total (n=472)	Oui (n=78)	Non (n=394)
Variable	Moyenne (ES)	Moyenne (ES)	Moyenne (ES)
GPA	3.265 (0.546)	3.265 (0.468)	3.265 (0.561)
Étude en dehors des cours (Heure)	2.871 (1.415)	2.577 (1.264)	2.929 (1.437)
Age	21.57 (2.275)	21.28 (1.999)	21.63 (2.324)
Religiosité	3.112 (1.054)	3.269 (0.949)	3.081 (1.072)
Satisfaction avec l'éducation obtenue	1.589 (0.696)	1.641 (0.720)	1.579 (0.692)
Gestion de l'alcool par l'établissement	2.373 (0.723)	2.436 (0.695)	2.360 (0.729)
Problème d'alcool parmi les étudiants	2.587 (0.920)	2.846 (0.955)	2.536 (0.905)
Santé générale	2.263 (0.869)	2.308 (0.708)	2.254 (0.898)
Vitalité	49.34 (16.75)	51.40 (16.98)	48.94 (16.70)
Télévision (temps)	1.856 (1.405)	2.115 (1.339)	1.805 (1.414)
Travail (temps)	2.331 (2.118)	2.256 (2.153)	2.345 (2.113)
Organisation étudiante (temps)	0.809 (1.291)	0.731 (1.326)	0.825 (1.285)
Sport (temps)	0.333 (0.998)	0.397 (1.155)	0.320 (0.965)
Autres activités physiques (temps)	1.085 (1.223)	1.090 (1.311)	1.084 (1.207)
Bénévolat (temps)	0.506 (1.106)	0.564 (1.169)	0.495 (1.094)
	Effectif (%)	Effectif (%)	Effectif (%)
Sexe (Homme)	158 (33.47)	33 (42.31)	125 (31.73)
Jamais marié (Oui)	411 (87.08)	72 (92.31)	339 (86.04)
Caucasien (Oui)	350 (74.15)	66 (84.62)	284 (72.08)
Éducation parentale (Plus élevée que le secondaire)	383 (81.14)	69 (88.46)	314 (79.70)
Fraternité/Sororité (Oui)	51 (10.81)	14 (17.95)	37 (9.391)
Porter plainte (Oui)	28 (5.932)	3 (3.846)	25 (6.345)
Résidence sur le campus (Oui)	149 (31.57)	25 (32.05)	124 (31.47)
Étudiant gradué (Oui)	7 (1.483)	3 (3.846)	4 (1.015)
Consommation de cigarette(s) (Oui)	193 (40.89)	47 (60.26)	146 (37.06)
Consommation d'autre(s) drogue(s) (Oui)	136 (28.81)	38 (48.72)	98 (24.87)
CA des parents (Oui)	315 (66.74)	59 (75.64)	256 (64.97)
70% des amis sont des buveurs excessifs (Oui)	60 (12.71)	26 (33.33)	34 (8.629)
Victime d'humiliation (Oui)	156 (33.05)	38 (48.72)	118 (29.95)
Préjudice physique (Oui)	53 (11.23)	21 (26.92)	32 (8.122)
Préjudice sexuel (Oui)	13 (2.754)	3 (3.846)	10 (2.538)
Buveur excessif au secondaire (Oui)	159 (33.69)	39 (50.00)	120 (30.46)
Athlétisme au secondaire (Oui)	265 (56.14)	52 (66.67)	213 (54.06)

Note : ES=Erreur standard

4.2.2 Analyses

Une sélection de variables est réalisée à l'aide des six procédures discutées (CIE_{total}, ΔMSE_{total}, CIE_{max}, ΔMSE_{max}, CIE_{effets}, ΔMSE_{effets}). Comme cela a été effectué pour la simulation, on conditionne sur la valeur moyenne des covariables. Pour chacune des méthodes de sélection, le terme d'interaction entre l'exposition et le médiateur dans le modèle sur la réponse est inclus. Ce choix est effectué pour être cohérent avec la recommandation de VanderWeele (2015) d'inclure, à

moins d'indication contraire, le terme d'interaction entre l'exposition et le médiateur. Dans la méthode *Complète*, le terme d'interaction n'est pas statistiquement différent de 0 ($\hat{\beta} = -0.045$, $p = 0.373$). Les réponses sont présentées pour les six procédures proposées, la méthode *Complète* et la méthode *Brute*.

4.3 Résultats

Le tableau 4.2 illustre les covariables sélectionnées par les différentes procédures. Globalement, comme cela était le cas pour la simulation, il est intéressant de remarquer que la méthode CIE_{total}, et dans une moindre mesure le CIE_{max}, rejette davantage de variables que les autres procédures proposées, alors que la méthode ΔMSE_{effets} conserve plus de variables que les autres algorithmes ; les méthodes CIE_{effets}, ΔMSE_{max} et ΔMSE_{total} conservent un nombre similaire de covariables. Il est également possible de noter quelques similarités avec les résultats de simulation dans le patron des variables sélectionnées. En effet, de manière cohérente avec les résultats de l'étude de simulation pour les variables de confusion entre l'exposition et le médiateur ou entre l'exposition et la variable réponse, on peut constater que certaines covariables sont maintenues dans l'analyse de médiation par l'ensemble des méthodes (par exemple, la consommation d'autres drogues, le fait que 70% des amis qui sont des buveurs excessifs et le temps consacré à la télévision). Également, à l'instar des confondants entre le médiateur et la réponse dans l'étude de simulation, certaines covariables (par exemple, la satisfaction avec l'éducation obtenue, le sexe des participants et la gestion de l'alcool par l'établissement post-secondaire) sont sélectionnées par toutes les procédures à l'exception du CIE_{total}. De plus, les variables *temps consacré au bénévolat* et *être victime d'humiliation à cause de la consommation d'alcool des autres* sont conservées uniquement par les procédures basées sur le ΔMSE, ce qui est similaire aux

conclusions de l'étude de simulation pour les prédicteurs purs du médiateur et de la réponse. Enfin, les items *temps consacré au travail* et *temps consacré au sport* sont rejetés par la totalité des procédures, ce qui rappelle les résultats de l'étude de simulation pour les prédicteurs purs de l'exposition et les variables de bruit.

TABLEAU 4.2 : Variables sélectionnées par les différentes méthodes

Variables	CIE_{max}	CIE_{effets}	ΔMSE_{max}	ΔMSE_{effets}	CIE_{total}	ΔMSE_{total}
Sociodémographique						
Age	✗	✓	✗	✓	✗	✓
Sexe	✓	✓	✓	✓	✗	✓
Religiosité	✓	✓	✓	✓	✓	✓
Jamais marié	✓	✓	✓	✓	✓	✓
Caucasien	✓	✓	✓	✓	✓	✓
Éducation parentale	✗	✗	✓	✓	✓	✓
Établissement post-secondaire						
Fraternité/Sorority	✓	✓	✓	✓	✓	✓
Satisfaction avec l'éducation obtenue	✓	✓	✓	✓	✗	✓
Gestion de l'alcool par l'établissement	✓	✓	✓	✓	✗	✓
Porter plainte	✓	✗	✓	✓	✗	✗
Résidence sur le campus	✗	✗	✗	✗	✗	✓
Étudiant gradué	✓	✓	✓	✓	✓	✓
Consommation de substance						
Consommation de cigarette(s)	✗	✗	✗	✗	✗	✗
Consommation d'autre(s) drogue(s)	✓	✓	✓	✓	✓	✓
CA des autres						
CA des parents	✗	✗	✗	✗	✗	✗
70% des amis sont des buveurs excessifs	✓	✓	✓	✓	✓	✓
Problème d'alcool parmi les étudiants	✗	✗	✗	✗	✗	✗
Conséquence de la CA des autres						
Victime d'humiliation	✗	✗	✓	✓	✗	✗
Préjudice physique	✓	✓	✗	✗	✓	✗
Préjudice sexuel	✗	✗	✗	✓	✗	✗
Santé						
Santé générale	✗	✓	✓	✓	✗	✓
Vitalité	✗	✗	✗	✗	✗	✗
Temps consacré à divers activités						
Télévision (temps)	✓	✓	✓	✓	✓	✓
Travail (temps)	✗	✗	✗	✗	✗	✗
Organisation étudiante (temps)	✗	✗	✗	✓	✗	✗
Sport (temps)	✗	✗	✗	✗	✗	✗
Autres activités physiques (temps)	✗	✓	✓	✓	✗	✗
Bénévolat (temps)	✗	✗	✓	✓	✗	✓
École secondaire						
Buveur excessif au secondaire	✗	✗	✗	✓	✗	✗
Athlétisme au secondaire	✗	✓	✗	✗	✗	✓
Total de variables sélectionnées	13	16	17	21	10	17

Note : ✓=La variable est sélectionnée ; ✗=La variable n'est pas sélectionnée

Le tableau 4.3 présente les résultats des estimations pour l'ensemble des méthodes. Ce tableau illustre également la différence relative de l'estimation (DRE) et le gain de précision relatif estimé du modèle (GPR) entre les méthodes proposées

et le modèle avec l'ensemble complet de covariables. Pour la méthode *Complète*, les résultats montrent que l'effet naturel direct ($\widehat{END}(\bar{\boldsymbol{c}}) = 0.081$, Intervalle de confiance[IC]=[-0.057, 0.219]) et l'effet naturel indirect ($\widehat{ENI}(\bar{\boldsymbol{c}}) = -0.007$, IC=[-0.035, 0.021]) ne sont pas statistiquement significatifs.

TABLEAU 4.3 : Résultats selon les méthodes de sélection

Effet	Méthode	Estimation	DRE (%)	ES	GPR (%)	Intervalle de confiance (95%)
Direct	*Complète*	0.081	0.000	0.070	0.000	[-0.057, 0.219]
	Brute	0.019	-76.60	0.069	2.081	[-0.116, 0.154]
	CIE_{max}	0.075	-7.201	0.069	1.510	[-0.061, 0.211]
	CIE_{effets}	0.077	-5.087	0.069	2.455	[-0.058, 0.211]
	ΔMSE_{max}	0.068	-16.31	0.068	3.788	[-0.065, 0.200]
	ΔMSE_{effets}	0.067	-16.91	0.068	3.142	[-0.067, 0.201]
	CIE_{total}	0.061	-23.88	0.069	1.569	[-0.074, 0.197]
	ΔMSE_{total}	0.063	-22.46	0.068	3.835	[-0.070, 0.195]
Indirect	*Complète*	-0.007	0.000	0.014	0.000	[-0.035, 0.021]
	Brute	-0.018	173.3	0.019	-25.80	[-0.056, 0.020]
	CIE_{max}	-0.006	-14.98	0.013	11.09	[-0.031, 0.020]
	CIE_{effets}	-0.007	2.021	0.014	3.719	[-0.034, 0.020]
	ΔMSE_{max}	-0.008	11.58	0.013	14.58	[-0.032, 0.017]
	ΔMSE_{effets}	-0.007	-0.731	0.012	18.22	[-0.031, 0.017]
	CIE_{total}	-0.007	10.91	0.015	-6.258	[-0.038, 0.023]
	ΔMSE_{total}	-0.006	-6.714	0.014	0.526	[-0.034, 0.022]

Note : DRE=Différence relative dans l'estimation ; ES=Erreur standard ; GPR=Gain de précision relative

De manière similaire aux résultats obtenus dans la simulation, la méthode *Brute* mène à des estimations qui sont nettement différentes de celles obtenues par la méthode *Complète*, avec une différence relative de -76.60% pour l'effet naturel direct et de 173.3% pour l'effet naturel indirect. Pour l'effet naturel direct, à l'exception de la méthode *Brute*, les différences relatives avec l'approche *Complète* sont les plus petites (en valeur absolue) pour le CIE_{max} (-7.201%) et le CIE_{effets} (-5.087%), alors que les différences relatives les plus élevées sont obtenues par les algorithmes ΔMSE_{total} (-22.46%) et CIE_{total} (-23.88%). Pour l'effet naturel indirect, à l'exception de la méthode *Brute*, les différences relatives avec l'approche *Complète* sont les plus petites (en valeur absolue) pour le CIE_{effets} (2.021%) et le ΔMSE_{effets} (-0.731%), alors que les différences relatives les plus élevées sont obtenues par les algorithmes CIE_{total} (10.91%) et CIE_{max} (-14.98%). Dans tous les cas, les effets naturels direct et indirect ne sont pas statistiquement significatifs.

Hormis le CIE_{total} pour l'effet naturel indirect, l'ensemble des méthodes de sélection permet d'obtenir un gain de précision relativement à la méthode *Complète*. Pour l'effet naturel direct, les gains de précision relatifs sont les plus importants pour le ΔMSE_{max} (3.788%), le ΔMSE_{effets} (3.142%) et le ΔMSE_{total} (3.835%). Pour l'effet naturel indirect, ce sont principalement les méthodes ΔMSE_{max} (14.58%) et ΔMSE_{effets} (18.22%) qui présentent des gains de précision relatifs importants. Notons également que le GPR pour le ΔMSE_{total} (0.526) est proche de 0. Ce résultat est similaire à ce qui est observé dans certains scénarios de simulation où le GPR du ΔMSE_{total} était important pour l'effet naturel direct, mais indiscernable voir inférieur à 0 pour l'effet naturel indirect (scénarios 2, 4 et 5).

Les conclusions de cet exemple diffèrent des résultats de la simulation à certains égards. Évidemment, il est difficile d'établir si ces distinctions résultent d'une variation aléatoire normale dans l'application des méthodes de sélection, ou s'il s'agit d'une caractéristique propre aux données utilisées. Néanmoins, deux dissemblances méritent d'être mentionnées. D'une part, le CIE_{total} permet d'obtenir une estimation qui est relativement proche de la méthode *Complète* et des autres méthodes de sélection. Une explication potentielle est que peu de variables de confusion importantes entre le médiateur et la variable réponse étaient présentes. Une autre distinction est le fait que, pour l'effet naturel indirect, la DRE pour le CIE_{effets} est plus petite (en valeur absolue) que la DRE observée suite à l'application du CIE_{max}. Or, rappelons que le biais relatif du CIE_{effets} était systématiquement plus élevé que celui du CIE_{max} pour l'effet naturel indirect. Certainement, ce constat suppose que l'estimation de l'effet naturel direct de la méthode *Complète* est comparable au paramètre d'intérêt, ce qui ne peut pas être vérifié.

Malgré ces différences, cet exemple permet d'illustrer que les méthodes qui performaient de manière adéquate dans les scénarios de simulation sont en mesure de réduire grandement le nombre de covariables dans une analyse de médiation, tout

en proposant des valeurs estimées comparables à celles obtenues par la méthode *Complète* et une variance qui est similaire voir inférieure à la variance produite par cette dernière.

CHAPITRE V

CONCLUSION

5.1 Rappel des objectifs et faits saillants de l'étude

L'analyse de médiation causale est une méthode statistique largement employée dans plusieurs domaines de recherche pour aider à comprendre comment et par quels mécanismes une variable d'exposition transmet son effet sur une variable réponse. Une des difficultés particulières de la médiation causale est la nécessité d'ajuster sur un ensemble de covariables suffisant pour permettre une interprétation causale de l'effet total et des effets naturels direct et indirect. Une fois qu'un ensemble suffisant pour la médiation est identifié, il peut être pertinent de réduire la dimension de cet ensemble pour obtenir un modèle plus parcimonieux ou encore pour réduire la variance des estimateurs. Bien que certaines méthodes pour faire de la sélection de variables sur l'effet total (le CIE et le ΔMSE) aient été proposées dans le cadre de l'inférence causale, ces techniques n'ont pas été adaptées au contexte de médiation. Ainsi, l'objectif principal de cette étude était de réviser et d'évaluer par simulation Monte-Carlo le CIE et le ΔMSE pour faire la sélection de variables dans le cadre de la médiation par inférence causale lorsque la variable de médiation et la réponse sont continues. Au total, quatre algorithmes ont été développés spécifiquement pour faire de la sélection de variables en médiation

causale, soit le CIE_{max}, le CIE_{effets}, le ΔMSE_{max} et le ΔMSE_{effets}.

Les analyses que nous avons effectuées suggèrent que les méthodes basées sur le CIE rejettent davantage de covariables que celles fondées sur le ΔMSE. De plus, la simulation confirme que le CIE_{total} n'est pas adéquat pour faire de la sélection de variables dans le contexte de médiation. Parmi les cinq autres algorithmes évalués, seul le CIE_{effets} ne maintient pas un biais relatif acceptable dans la totalité des conditions. Ainsi, ces résultats suggèrent que les méthodes CIE_{max}, ΔMSE_{total}, ΔMSE_{max} et ΔMSE_{effets} sont à privilégier pour faire la sélection de variables en médiation causale.

Un avantage des méthodes qui s'appuient sur le ΔMSE comparativement aux méthodes basées sur le CIE est qu'elles tiennent compte de la variabilité des estimations. Cette propriété se traduit par le fait que les méthodes fondées sur le ΔMSE mènent généralement à une amélioration de la précision des estimateurs comparativement à la méthode *Complète*, alors que le CIE_{max} offre une précision qui est similaire ou inférieure à celle obtenue par la méthode *Complète*. Parmi les méthodes qui reposent sur le ΔMSE, le ΔMSE_{total} est la seule procédure qui est associée à une augmentation de la variance comparativement à la méthode *Complète*, bien que ce phénomène n'a été observé que dans une seule condition évaluée.

Il est généralement reconnu que les intervalles de confiance ne sont pas toujours adéquats à la suite d'une sélection de variables (Fox, 2016; Harrell, 2015). Toutefois, les résultats de la présente étude montrent que le ΔMSE_{total} atteint un taux de couverture nominal à 95% dans la majorité des conditions. De plus, les méthodes avec peu de biais (CIE_{max}, ΔMSE_{total}, ΔMSE_{max}, ΔMSE_{effets}) qui n'atteignent pas le taux de couverture attendu dans certains contextes procurent généralement qu'une légère sous-couverture ou sur-couverture. Ainsi, il serait mal-

avisé d'exclure une de ces procédures sur la base du taux de couverture.

En regard de l'ensemble des résultats obtenus dans ce mémoire, il est possible d'émettre les recommandations suivantes. Premièrement, si l'objectif est de réduire de manière maximale la dimension du vecteur de covariables, alors le CIE_{max} est probablement à privilégier. Deuxièmement, si le but est d'améliorer la précision des estimateurs, alors les chercheurs devraient prioriser les méthodes fondées sur le ΔMSE. Par contre, il est difficile de suggérer le recours à une procédure spécifique basée sur le ΔMSE pour améliorer la précision des estimateurs, car elles semblent toutes bénéficier de quelques avantages dans certaines conditions. Enfin, si l'intention est de réduire le nombre de covariables tout en maintenant un taux de couverture nominal le plus près de la valeur attendue, alors le ΔMSE_{total} est la procédure à favoriser.

5.2 Limites et forces de l'étude

Cette étude présente certaines limites. Premièrement, les simulations ont été réalisées en fixant le terme d'interaction entre l'exposition et le médiateur dans le modèle de réponse à 0. Or, l'impact de l'inclusion de certaines covariables sur la variance des estimateurs des effets naturels direct et indirect n'est pas nécessairement la même en présence d'une telle interaction. Par exemple, dans leurs analyses complémentaires, Diop et ses collaborateurs (2021) ont montré que l'inclusion d'un prédicteur pur du médiateur dans le cas où la relation entre le médiateur et la réponse est plus forte que l'association entre l'exposition et le médiateur permet de réduire la variance de l'estimateur pour l'effet naturel indirect, alors que l'inverse est observé lorsqu'il n'y a pas d'interaction.

Deuxièmement, les covariables ont été générées de manière indépendante comme

des variables normalement distribuées. Conséquemment, la simulation Monte-Carlo ne reflète pas nécessairement la complexité de la structure de dépendance des covariables qui peut être observée dans certains domaines d'investigation, comme la recherche en santé (voir Franklin, Schneeweiss, Polinski, & Rassen, 2014).

Une troisième limite est que la simulation porte uniquement sur les approches par régression dans le cas où le médiateur et la réponse sont continus. Il est important de se rappeler que les approches par régression permettent d'obtenir des estimateurs conditionnels pour les effets naturels direct et indirect. Ainsi, les estimateurs des effets naturels d'un modèle qui inclut une covariable, d'une part, et d'un modèle qui omet cette même covariable, d'autre part, ne visent pas toujours l'estimation des mêmes paramètres. Conséquemment, les méthodes basées sur un changement d'estimation comme le ΔMSE et le CIE ne sont pas nécessairement adéquates pour faire de la sélection de variables lorsque les effets naturels sont conditionnels, car un changement d'estimation peut simplement indiquer qu'un estimand différent est visé par la procédure. Dans la présente étude, ce problème a été contourné en conditionnant sur la valeur moyenne des covariables [1] pour permettre une estimation marginale (voir annexe A pour un rappel). Or, les analyses de médiation basées sur les approches par régression peuvent se réaliser avec des modèles non linéaires, comme les modèles où la paire médiateur/réponse est évaluée à l'aide de régressions logistique/linéaire (Valeri & VanderWeele, 2013; VanderWeele, 2015), de régressions linéaire/log-binomiale (VanderWeele, 2015; VanderWeele & Vansteelandt, 2010), de régressions linéaire/logistique (approximation proposée par Gaynor, Schwartz, & Lin, 2019) et de régressions logistique/logistique (Samoilenko & Lefebvre, 2021). Dans ce contexte,

1. On rappelle que lorsque le paramètre d'interaction entre l'exposition et le médiateur est nul, l'effet naturel direct demeure inchangé en fonction de la valeur de conditionnement des covariables. Toutefois, puisque les estimations du terme d'interaction ne sont pas égales à 0, la décision d'effectuer le conditionnement à l'espérance des covariables demeure pertinente.

en règle générale, conditionner sur les valeurs moyennes des covariables ne permet pas d'obtenir un estimateur marginal. Une solution à cette difficulté est d'adopter des estimateurs par pseudopopulation, comme les approches par pondération (Lange et al., 2012; Steen et al., 2017) et les approches par imputation (Steen et al., 2017; Vansteelandt, Bekaert, & Lange, 2012), qui permettent d'obtenir des estimations marginales pour les effets naturels direct et indirect de l'ensemble des paires médiateur/réponse précédemment mentionnées.

Enfin, certains chercheurs (par exemple, Hayes, 2013) suggèrent généralement d'utiliser les techniques de bootstrap comme méthode d'inférence dans les analyses de médiation. Or, étant donné les demandes computationnelles élevées des méthodes par bootstrap, il a été décidé de calculer les intervalles de confiance à l'aide de la méthode delta. Puisque les taux de couverture de la méthode *Complète* sont acceptables, les conclusions importantes de la présente étude devraient demeurer inchangées. Néanmoins, pour se rapprocher de l'application réelle des méthodes de médiation, l'évaluation des techniques par bootstrap pour former les intervalles de confiance devrait être considérée dans les études futures.

Malgré ces limites, la présente étude possède également plusieurs forces qui méritent d'être mentionnées. Premièrement, la simulation inclut six mécanismes de génération de données, ce qui permet d'évaluer la performance des méthodes proposées dans de multiples contextes. Deuxièmement, divers types de covariables ont été incorporés dans les simulations, comme les prédicteurs purs (exposition, médiateur et réponse), les facteurs de confusion qui correspondent aux différentes hypothèses de l'analyse de médiation (variables de confusion entre l'exposition et la réponse, entre l'exposition et le médiateur et entre le médiateur et la réponse) et des variables de bruit. Conséquemment, cette recherche couvre plusieurs situations qui peuvent être rencontrées en pratique. Troisièmement, cette étude a généralisé et évalué la performance de méthodes basées sur le CIE et le ΔMSE

pour faire de la sélection de variables dans le contexte de médiation, plutôt que de privilégier un des ces algorithmes. Comme cela est indiqué dans la conclusion, ces deux approches semblent être pourvues de caractéristiques distinctes ; il a donc été possible de proposer différentes recommandations selon les objectifs de la sélection de variables.

5.3 Remarques finales

En somme, la sélection de variables est un problème important dans l'analyse de médiation causale, et les méthodes basées sur les données comme le CIE et le ΔMSE sont des techniques largement employées en pratique dans certains domaines de recherche comme l'épidémiologie. Bien que plusieurs études novatrices aient été réalisées dans les dernières années pour améliorer les caractéristiques de ces approches (par exemple Dunkler & Heinze, 2014; Loh & Vansteelandt, 2021; Vansteelandt et al., 2010), l'application des méthodes de sélection de variables basées sur les données à la médiation causale a généralement été ignorée dans les écrits scientifiques. Les travaux réalisés dans le cadre de ce mémoire innovent en proposant des modifications aux méthodes CIE et ΔMSE pour permettre une sélection de variables dans le cadre des analyses de médiation, et, malgré les limites susmentionnées, confirment l'efficacité du CIE_{max}, du ΔMSE_{total}, du ΔMSE_{max} et du ΔMSE_{effets} dans le contexte de la médiation. Dans les recherches futures, des modifications aux algorithmes évalués ou encore le développement de nouvelles propositions pour combiner les changements d'estimation sur l'effet naturel direct et l'effet naturel indirect devraient être examinés. Ultimement, ces procédures de sélection de variables peuvent être utiles pour les chercheurs de domaines appliqués, afin de diminuer la complexité des modèles proposés ou encore d'améliorer la précision des estimateurs dans les analyses de médiation causale.

ANNEXE A

EFFETS MARGINAUX ET ESTIMATION PAR RÉGRESSION

Supposons qu'on a les modèles suivants :

$$\mathbb{E}[Y|A=a,M=m,\boldsymbol{C}=\boldsymbol{c}] = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \boldsymbol{\theta}_4^T \boldsymbol{c}, \tag{1.0.1}$$

$$\mathbb{E}[M|A=a,\boldsymbol{C}=\boldsymbol{c}] = \beta_0 + \beta_1 a + \boldsymbol{\beta}_2^T \boldsymbol{c}, \tag{1.0.2}$$

$$\text{logit}(\mathbb{E}[A|\boldsymbol{C}=\boldsymbol{c}]) = \gamma_0 + \boldsymbol{\gamma}_1^T \boldsymbol{c}.$$

Dans ce contexte, $\boldsymbol{C}$ est un ensemble suffisant pour la médiation. Ainsi, on a que

$$\begin{aligned}\mathbb{E}[Y(a,M(a^*))] &= \int \mathbb{E}[Y(a,M(a^*))|\boldsymbol{C}=\boldsymbol{c}]g(\boldsymbol{c})d\boldsymbol{c} \\ &= \int\int f(m|a^*,\boldsymbol{c})\,\mathbb{E}[Y|A=a,M=m,\boldsymbol{C}=\boldsymbol{c}]g(\boldsymbol{c})dmd\boldsymbol{c},\end{aligned}$$

avec $g(\cdot)$ la densité des variables du vecteur $\boldsymbol{C}$ et $f(\cdot)$ la densité de la variable de médiation. La dernière égalité découle du fait que $\boldsymbol{C}$ est un ensemble suffisant pour la médiation. De plus, on note END l'effet naturel direct marginal et $END(\boldsymbol{c})$ l'effet naturel direct conditionnel lorsque $\boldsymbol{C}=\boldsymbol{c}$. En particulier, on a que $END(\bar{\boldsymbol{c}})$ est l'effet naturel direct conditionnel lorsque $\boldsymbol{C}=\mathbb{E}[\boldsymbol{C}]$.

Proposition A.0.1. L'effet naturel direct marginal est donné par

$$END = END(\bar{\boldsymbol{c}}).$$

Démonstration.

$$\begin{aligned}
END &= \int\int f(m|a^*, \boldsymbol{c})\, \mathbb{E}[Y|A = a, M = m, \boldsymbol{C} = \boldsymbol{c}] g(\boldsymbol{c}) dm d\boldsymbol{c} \\
&\quad - \int\int f(m|a^*, \boldsymbol{c})\, \mathbb{E}[Y|A = a^*, M = m, \boldsymbol{C} = \boldsymbol{c}] g(\boldsymbol{c}) dm d\boldsymbol{c} \\
&= \int g(\boldsymbol{c}) \Big\{ \int f(m|a^*, \boldsymbol{c})\, \mathbb{E}[Y|A = a, M = m, \boldsymbol{C} = \boldsymbol{c}] \\
&\quad - f(m|a^*, \boldsymbol{c})\, \mathbb{E}[Y|A = a^*, M = m, \boldsymbol{C} = \boldsymbol{c}] \Big\} dm d\boldsymbol{c} \\
&= \int g(\boldsymbol{c}) END(\boldsymbol{c}) dc \\
&= \int (\theta_1 + \theta_3\beta_0 + \theta_3\beta_1 a^* + \theta_3 \boldsymbol{\beta}_2^T \boldsymbol{c})(a - a^*) g(\boldsymbol{c}) d\boldsymbol{c} \\
&= (\theta_1 + \theta_3\beta_0 + \theta_3\beta_1 a^* + \theta_3 \boldsymbol{\beta}_2^T\, \mathbb{E}[\boldsymbol{C}])(a - a^*) \\
&= END(\bar{\boldsymbol{c}}).
\end{aligned}$$

□

La démonstration de la proposition A.0.1 illustre également l'importance de modéliser adéquatement la relation entre les covariables et le médiateur, car le résultat établissant la correspondance entre l'effet naturel direct marginal et conditionnel à la moyenne suppose que les covariables sont entrent de façon linéaire dans le modèle du médiateur.

Corollaire A.0.1. L'effet total marginal est donné par

$$ET = ET(\bar{\boldsymbol{c}}).$$

Démonstration. Rappelons que l'effet naturel indirect ne dépend pas de $\boldsymbol{C}$. Ainsi,

on a automatiquement que $ENI = ENI(\bar{\boldsymbol{c}})$. De plus, on a que

$$\begin{aligned} ET &= END + ENI \\ &= END(\bar{\boldsymbol{c}}) + ENI(\bar{\boldsymbol{c}}) \\ &= ET(\bar{\boldsymbol{c}}). \end{aligned}$$

□

Corollaire A.0.2. Soit $\boldsymbol{C}$ et $\boldsymbol{D}$ des ensembles suffisants pour la médiation où $\boldsymbol{D} \neq \boldsymbol{C}$. De plus, supposons que les modèles suivants sont correctement spécifiés :

$$\mathbb{E}[Y|A = a, M = m, \boldsymbol{D} = \boldsymbol{d}] = \theta_0^* + \theta_1^* a + \theta_2^* m + \theta_3^* am + \boldsymbol{\theta}_4^{*^T} \boldsymbol{d}, \tag{1.0.3}$$

$$\mathbb{E}[M|A = a, \boldsymbol{D} = \boldsymbol{d}] = \beta_0^* + \beta_1^* a + \boldsymbol{\beta}_2^{*^T} \boldsymbol{d}, \tag{1.0.4}$$

$$\text{logit}(\mathbb{E}[A|\boldsymbol{D} = \boldsymbol{d}]) = \gamma_0^* + \boldsymbol{\gamma}_1^{*^T} \boldsymbol{d}.$$

Alors, on a que

$$END(\bar{\boldsymbol{c}}) = END(\bar{\boldsymbol{d}}).$$

Démonstration.

$$\begin{aligned} END(\bar{\boldsymbol{c}}) &= END \\ &= END(\bar{\boldsymbol{d}}). \end{aligned}$$

□

Il est possible d'illustrer la conclusion du corollaire A.0.2 avec un exemple par simulation, à l'aide des modèles suivants :

$$
\begin{aligned}
\mathbb{E}[U] &= \delta_0 \\
\mathbb{E}[C|U=u] &= \psi_0 + \psi_1 u \\
\mathbb{E}[P_m|U=u] &= \lambda_0 + \lambda_1 u \\
\mathbb{E}[P_y] &= \omega_0 \\
\mathbb{E}[Y|A=a, M=m, C=c, P_y=p_y] &= \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4 c + \theta_5 p_y, \\
\mathbb{E}[M|A=a, C=c, P_m=p_m] &= \beta_0 + \beta_1 a + \beta_2 c + \beta_3 p_m, \\
\text{logit}(\mathbb{E}[A|C=c]) &= \gamma_0 + \gamma_1 c.
\end{aligned}
$$

Ce modèle possède une covariable entre l'exposition, le médiateur et la réponse (C), un prédicteur pur du médiateur (P_m), un prédicteur pur de la réponse (P_y) et une cause commune qui n'est pas mesurée entre C et P_m (U). L'ensemble minimal est $\{C\}$. Pour simuler le jeu de données, on commence par générer $U \sim \mathcal{N}(\delta_0, 1)$ et $P_y \sim \mathcal{N}(\omega_0, 1)$. Par la suite, on génère $C \sim \mathcal{N}(\psi_0 + \psi_1 u, 1)$ et $P_m \sim \mathcal{N}(\lambda_0 + \lambda_1 u, 1)$. Ensuite, on génère $A \sim Bernoulli(expit(\gamma_0 + \gamma_1 c))$, avec $expit(x) = (1+\exp(-x))^{-1}$. Enfin, on génère successivement $M \sim \mathcal{N}(\beta_0 + \beta_1 a + \beta_2 c + \beta_3 p_m, 1)$ et $Y \sim \mathcal{N}(\theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4 c + \theta_5 p_y, 1)$. Pour l'ensemble des 1000 réplications, on prend un échantillon de taille $n = 1000$. De plus, on utilise les paramètres suivants : $\delta_0 = \psi_0 = \lambda_0 = \omega_0 = \theta_0 = \beta_0 = \gamma_0 = 0$, $\theta_3 = 0.5$, $\psi_1 = \lambda_1 = \beta_1 = \theta_1 = \theta_5 = \beta_3 = 2$, $\theta_2 = .5$, $\beta_2 = \theta_4 = \gamma_1 = 1$. Pour cette simulation, on compare quatre modèles de médiation différents qui incluent la variable C, soit : un modèle qui inclut P_m et P_y (*Modèle Complet*), un modèle qui inclut P_m (*Modèle P_m*), un modèle qui inclut P_y (*Modèle P_y*) et un modèle qui n'inclut ni P_m ni P_y (*Modèle Minimal*). Le tableau A.1 présente les résultats de cette simulation pour l'effet naturel direct lorsque le conditionnement est effectué à l'espérance des variables qui sont incorporées dans les différents modèles.

TABLEAU A.1 : Effet naturel direct conditionnel à la moyenne pour différents ensembles de conditionnement suffisants pour la médiation

	$Complet$	P_m	P_y	$Minimal$
END	2.005	2.014	2.004	2.008

Les résultats montrent que les quatre modèles procurent les mêmes conclusions lorsque le conditionnement s'effectue à l'espérance des covariables, ce qui est cohérent avec le corollaire A.0.2.

BIBLIOGRAPHIE

Andrews, R. M., & Didelez, V. (2021). Insights into the cross-world independence assumption of causal mediation analysis. *Epidemiology*, *32*(2), 209–219. doi: 10.1097/EDE.0000000000001313

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research. Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173–1182. doi: 10.1037/0022-3514.51.6.1173

Berzenski, S. R., & Yates, T. M. (2010). A developmental process analysis of the contribution of childhood emotional abuse to relationship violence. *Journal of Aggression, Maltreatment and Trauma*, *19*(2), 180–203. doi: 10.1080/10926770903539474

Dardis, C. M., Dixon, K. J., Edwards, K. M., & Turchik, J. A. (2015). An examination of the factors related to dating violence perpetration among young men and women and associated theoretical explanations : A review of the literature. *Trauma, Violence, and Abuse*, *16*(2), 136–152. doi: 10.1177/1524838013517559

Diop, A., Lefebvre, G., Duchaine, C. S., Laurin, D., & Talbot, D. (2021). The impact of adjusting for pure predictors of exposure, mediator, and outcome on the variance of natural direct and indirect effect estimators. *Statistics in Medicine*, *40*(10), 2339–2354. doi: 10.1002/sim.8906

Dunkler, D., & Heinze, G. (2014). *A SAS macro for augmented backward elimination* (Rapport technique). Consulté sur `https://cemsiis.meduniwien.ac.at/en/kb/science-research/software/statistical-software/abe/`

Edkins, T., Edgerton, J. D., & Roberts, L. W. (2017). Correlates of binge drinking in a sample of canadian university students. *International Journal of Child, Youth and Family Studies*, *8*(1), 112–144. doi: 10.18357/ijcyfs81201716944

Etminan, M., Collins, G. S., & Mansournia, M. A. (2020). Using causal diagrams to improve the design and interpretation of medical research. *Chest*, *158*(1), S21–S28. doi: 10.1016/j.chest.2020.03.011

Fox, J. (2016). *Applied regression analysis and generalized linear models* (2e éd.). Sage publications.

Franklin, J. M., Schneeweiss, S., Polinski, J. M., & Rassen, J. A. (2014). Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. *Computational Statistics and Data Analysis*, *72*, 219–226. doi: 10.1016/j.csda.2013.10.018

Gaynor, S. M., Schwartz, J., & Lin, X. (2019). Mediation analysis for common binary outcomes. *Statistics in Medicine*, *38*(4), 512–529. doi: 10.1002/sim.7945

Greenland, S., Daniel, R., & Pearce, N. (2016). Outcome modelling strategies in epidemiology : Traditional methods and basic alternatives. *International Journal of Epidemiology*, *45*(2), 565–575. doi: 10.1093/ije/dyw040

Greenland, S., Pearl, J., & Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, *10*(1), 37–48. doi: 10.1097/00001648-199901000-00008

Hair, J. F., Jr., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. (2016). *A primer on partial least squares structural equation modeling (PLS-SEM)* (2e éd.). Sage publications.

Halpern, C. T., Young, M. L., Waller, M. W., Martin, S. L., & Kupper, L. L. (2004). Prevalence of partner violence in same-sex romantic and sexual relationships in a national sample of adolescents. *Journal of Adolescent Health*, *35*(2), 124–131. doi: 10.1016/j.jadohealth.2003.09.003

Harrell, F. E., Jr. (2015). *Regression modeling strategies : With Applications to linear models, logistic and ordinal regression, and survival analysis* (2e éd.). Springer.

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis : A regression-based approcah* (1re éd.). Guilford Press.

Heeringa, S. G., West, B. T., & Berglund, P. A. (2020). *Applied survey data analysis* (2e éd.). Chapman and Hall/CRC.

Heinze, G., Wallisch, C., & Dunkler, D. (2018). Variable selection – A review and recommendations for the practicing statistician. *Biometrical Journal*, *60*(3), 431–449. doi: 10.1002/bimj.201700067

Hosmer, D. W., Jr., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression.* (3e éd.). John Wiley & Sons.

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, *15*(4), 309–334. doi: 10.1037/a0020761

Kline, R. B. (2015). *Principles and practices of structural equation modelling* (4e éd.). Guilford Press.

Lange, T., Vansteelandt, S., & Bekaert, M. (2012). A simple unified approach for estimating natural direct and indirect effects. *American Journal of Epidemiology*, *176*(3), 190–195. doi: 10.1093/aje/kwr525

Lefebvre, G., Delaney, J. A. C., & Platt, R. W. (2008). Impact of mis-specification of the treatment model on estimates from a marginal structural model. *Statistics in Medicine*, *27*(18), 3629–3642. doi: 10.1002/sim.3200

Lewis, M., & Kuerbis, A. (2016). An overview of causal directed acyclic graphs for substance abuse researchers. *Journal of Drug and Alcohol Research*, *5*, 1–8. doi: 10.4303/jdar/235992

Loh, W. W., & Vansteelandt, S. (2021). Confounder selection strategies targeting stable treatment effect estimators. *Statistics in Medicine*, *40*(3), 607–630. doi: 10.1002/sim.8792

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis* (1re éd.). Routledge.

MacKinnon, D. P., Warsi, G., & Dwyer, J. H. (1995). A simulation study of mediated effect measures. *Multivariate Behavioral Research*, *30*(1), 41–62. doi: doi:10.1207/s15327906mbr3001_3

Maldonado, G., & Greenland, S. (1993). Simulation study of confounder-selection strategies. *American Journal of Epidemiology*, *138*(11), 923–936.

McDowell, I. (2006). *Measuring Health : A guide to rating scales and questionnaires* (3e éd.). Oxford University Press.

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. doi: 10.1002/sim.8086

Muthén, L. K., & Muthén, B. O. (2017). *MPlus user' guide*. Muthén & Muthén.

Nguyen, T. Q., Schmid, I., Ogburn, E. L., & Stuart, E. A. (2020). Clarifying causal mediation analysis for the applied researcher : Effect identification via three assumptions and five potential outcomes. *arXiv*.

Nguyen, T. Q., Schmid, I., & Stuart, E. A. (2020). Clarifying causal mediation analysis for the applied researcher : Defining effects based on what we want to learn. *Psychological Methods*. doi: 10.1037/met0000299

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics : A primer*. John Wiley & Sons.

Robinson, L. D., & Jewell, N. P. (1991). Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review / Revue Internationale de Statistique*, *59*(2), 227–240. doi: 10.2307/1403444

Rosseel, Y. (2012). Lavaan : An R package for structural equation modeling and more. *Journal of Statistical Software*, *48*(2), 1–36.

Rothman, E. F., McNaughton Reyes, L., Johnson, R. M., & LaValley, M. (2012). Does the alcohol make them do it ? Dating violence perpetration and drinking among youth. *Epidemiologic Reviews*, *34*(1), 103–119. doi: 10.1093/epirev/mxr027

Samoilenko, M., & Lefebvre, G. (2021). Parametric regression-based causal mediation analysis of binary outcomes and binary mediators : Moving beyond the rareness or commonness of the outcome. *American Journal of Epidemiology*. doi: 10.1093/aje/kwab055

Shorey, R. C., Stuart, G. L., & Cornelius, T. L. (2011). Dating violence and substance use in college students : A review of the literature. *Aggression and Violent Behavior*, *16*(6), 541–550. doi: 10.1016/j.avb.2011.08.003

Shrier, I., & Platt, R. W. (2008). Reducing bias through directed acyclic graphs. *BMC Medical Research Methodology*, *8*(1), 1–15. doi: 10.1186/1471-2288-8-70

StataCorp. (2019). *Stata statistical software : Release 16*. StataCorp LLC.

Steen, J., Loeys, T., Moerkerke, B., & Vansteelandt, S. (2017). Medflex : An R package for flexible mediation analysis using natural effect models. *Journal of Statistical Software*, *76*(11), 1–45. doi: 10.18637/jss.v076.i11

Talbot, D., & Massamba, V. K. (2019). A descriptive review of variable selection methods in four epidemiologic journals : There is still room for improvement. *European Journal of E3pidemiology*, *34*(8), 725–730.

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). Mediation : R package for causal mediation analysis. *Journal of Statistical Software*, *59*(5), 1–38. doi: 10.18637/jss.v059.i05

Tonmyr, L., Thornton, T., Draca, J., & Wekerle, C. (2010). A review of childhood maltreatment and adolescent substance use relationship. *Current Psychiatry Reviews*, *6*(3), 223–234. doi: 10.2174/157340010791792581

Valeri, L., & VanderWeele, T. J. (2013). Mediation analysis allowing for exposure-mediator interactions and causal interpretation : Theoretical assumptions and implementation with SAS and SPSS macros. *Psychological Methods*, *18*(2), 137–150. doi: 10.1037/a0031034

VanderWeele, T. J. (2015). *Explanation in causal inference : Methods for mediation and interaction*. Oxford University Press.

VanderWeele, T. J., & Vansteelandt, S. (2010). Odds ratios for mediation analysis for a dichotomous outcome. *American Journal of Epidemiology*, *172*(12), 1339–1348. doi: 10.1093/aje/kwq332

Vansteelandt, S., Bekaert, M., & Claeskens, G. (2010). On model selection and model misspecification in causal inference. *Statistical Methods in Medical Research*, *21*(1), 7–30. doi: 10.1177/0962280210387717

Vansteelandt, S., Bekaert, M., & Lange, T. (2012). Imputation strategies for the estimation of natural direct and indirect effects. *Epidemiologic Methods*, *1*(1), 129–158. doi: 10.1515/2161-962X.1014

Wang, Z. (2007). Two postestimation commands for assessing confounding effects in epidemiological studies. *The Stata Journal*, *7*(2), 183–196.

Weng, H. Y., Hsueh, Y. H., Messam, L. L., & Hertz-Picciotto, I. (2009). Methods of covariate selection : Directed acyclic graphs and the change-in-estimate procedure. *American Journal of Epidemiology*, *169*(10), 1182–1190. doi: 10.1093/aje/kwp035

Williams, J., Powell, L. M., & Wechsler, H. (2003). Does alcohol consumption reduce human capital accumulation ? Evidence from the College Alcohol Study. *Applied Economics*, *35*(10), 1227–1239. doi: 10.1080/0003684032000090735

Witte, J., & Didelez, V. (2019). Covariate selection strategies for causal inference : Classification and comparison. *Biometrical Journal*, *61*(5), 1270–1289. doi: 10.1002/bimj.201700294

Wright, S. (1934). The method of path coefficients. *The Annals of Mathematical Statistics*, *5*(3), 161–215. doi: 10.1214/aoms/1177732676