UNIVERSITÉ DU QUÉBEC À MONTRÉAL

IDENTIFICATION DE MODÈLES APPROPRIÉS POUR L'INFÉRENCE CAUSALE À PARTIR DE DONNÉES D'OBSERVATION

THÈSE
PRÉSENTÉE
COMME EXIGENCE PARTIELLE
DU DOCTORAT EN MATHÉMATIQUES

PAR
DENIS TALBOT

JUILLET 2015

AVANT-PROPOS

Le 8 septembre 2010 commençait pour moi une longue, parfois difficile, mais très enrichissante, aventure. Je contactais alors Geneviève par courriel pour lui demander si elle était disponible et intéressée à superviser les études doctorales que j'allais débuter officiellement en septembre 2011, mais pour lesquelles je commençais déjà à me préparer en effectuant des demandes de bourse. Ce n'était que le début.

Durant les quatre années qui ont suivi, j'ai effectué de nombreuses lectures sur l'inférence causale, j'ai étudié comme un fou pour réussir les examens de synthèse en mathématiques, j'ai suivi plusieurs cours, j'ai rédigé trois articles scientifiques et collaboré à l'écriture de deux autres, j'ai participé à des congrès où j'ai effectué des présentations, ... J'approche maintenant la fin du périple doctoral : je finalise l'écriture de la thèse. J'ai beaucoup travaillé et je suis fier du travail que j'ai accompli, mais je ne pourrais passer sous silence toute l'aide que j'ai reçue pour mener à bien ce projet.

J'adresse mes premiers remerciements à Geneviève qui a accepté de me superviser et qui m'a fait découvrir ce domaine passionnant de la statistique qu'est l'inférence causale. Geneviève a énormément contribué à ma réussite, de par ses conseils, sa disponibilité et ses révisions détaillées. Je tiens également à remercier Juli, ma co-superviseure. Les conseils et les révisions de Juli ont grandement contribué à améliorer la qualité de la thèse et des articles qui y sont présentés. En particulier, Juli a un grand souci du détail et dispose d'une capacité impressionnante à repérer les éléments qui pourraient manquer de clarté pour les lecteurs.

Je veux aussi remercier les amis que j'ai côtoyés et qui m'ont offert un support moral pendant le doctorat : Sandra, Alexandre, Simon O., Simon J., Claudia, Anne-Sophie et Tanya. Un merci tout spécial à Sandra, Alexandre et Anne-Sophie

qui m'ont occasionnellement hébergé à Montréal (dans le cas de Sandra, plus qu'occasionnellement !).

Je remercie également mes parents pour leur support et pour l'éducation qu'ils m'ont donnée. Ils m'ont appris à persévérer pour atteindre mes objectifs.

Un merci tout spécial à Jimmy, qui partage ma vie depuis près de trois ans, pour son support, ses encouragements et sa capacité à relativiser les problèmes rencontrés.

Finalement, je veux remercier le Conseil de recherches en sciences naturelles et en génie du Canada ainsi que le Fonds de recherche du Québec - Nature et technologie pour leur support financier. Sans ce support, il aurait été beaucoup plus difficile, voire impossible, de me concentrer comme je l'ai fait à mes études et de compléter le doctorat aussi rapidement.

À Blake, mon filleul.
Tu es un petit garçon adorable,
j'aimerais tellement pouvoir te voir plus souvent.

«Entre
Ce que je pense
Ce que je veux dire
Ce que je crois dire
Ce que je dis
Ce que vous avez envie d'entendre
Ce que vous croyez entendre
Ce que vous entendez
Ce que vous avez envie de comprendre
Ce que vous croyez comprendre
Ce que vous comprenez
Il y a dix possibilités qu'on ait des difficultés à
communiquer.
Mais essayons quand même... »
Bernard Werber,
l'Encyclopédie du savoir relatif et absolu

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

RÉSUMÉ

L'inférence causale permet de tirer des conclusions sur des relations de cause à effet à l'aide d'observations empiriques. Alors que l'inférence causale peut généralement s'effectuer assez facilement à l'aide de données provenant d'études randomisées, il en va autrement lorsque des données d'études observationnelles sont utilisées. Toutefois, les études randomisées ne sont pas toujours possibles. Dans cette thèse, nous abordons la problématique de la sélection de modèles visant l'inférence causale à l'aide de données d'observation sous deux angles distincts.

Dans un premier temps, nous abordons le contexte où les connaissances substantielles (c'est-à-dire du domaine d'application) sont peu développées. Nous étudions la méthode de sélection de modèles pour l'inférence causale *Bayesian Adjustment for Confounding* (BAC) et présentons des développements récents par rapport à cette méthode. Notamment, une justification plus formelle à la validité de BAC pour l'estimation de l'effet causal d'exposition ainsi que des procédures visant à choisir une valeur de l'hyperparamètre ω de BAC minimisant l'erreur quadratique moyenne (EQM) sont étudiées. La performance de BAC en termes d'EQM se révélant décevante, nous élaborons une nouvelle approche de sélection de modèles pour l'inférence causale, *Bayesian Causal Effect Estimation* (BCEE). BCEE partage certaines similarités avec BAC ainsi qu'avec le modèle moyen bayésien. Cependant, contrairement à ces approches, BCEE est motivée par le paradigme graphique causal. Les études de simulations réalisées suggèrent que BCEE permet généralement de réduire au moins légèrement, et parfois notablement, l'EQM de l'estimateur de l'effet causal d'exposition par rapport à un modèle de réponse complet et par rapport à BAC. La performance de BCEE peut même approcher celle du vrai modèle d'exposition dans des conditions idéales.

Dans un deuxième temps, nous abordons la sélection de modèles pour l'inférence causale dans un contexte où les connaissances du domaine d'application sont plus développées. Une méthodologie statistique utilisant les modèles structuraux marginaux (MSMs) et ayant servi à analyser les données du *Honolulu Heart Program* est décrite. Cette analyse visait à estimer les relations causales entre l'activité physique, la tension artérielle et la mortalité. La méthodologie développée comporte plusieurs éléments novateurs, dont une utilisation élargie du paradigme graphique à l'inférence causale, la validation et l'amélioration de graphes acycliques orientés à l'aide de modèles d'équations structurelles ainsi que l'introduction de MSMs conditionnels à des variables variant dans le temps. Pour l'application étudiée, l'approche proposée facilite la sélection des variables utilisées pour la pondération des MSMs, étape charnière et complexe de l'implantation de ces modèles.

Finalement, nous étudions le choix d'une stabilisation des poids utilisés pour les MSMs en relation avec le modèle structurel choisi. Il est montré que la stabilisation des poids des MSMs peut non seulement avoir un impact sur la variance des estimateurs obtenus, mais également sur leur biais. À l'aide d'un exemple illustratif puis d'études de simulations, nous démontrons que les poids stabilisés usuels peuvent conduire à des estimateurs biaisés lorsque le modèle structurel implanté n'inclut pas l'ensemble de l'historique d'exposition alors que le modèle structurel réel dépend de l'ensemble de l'historique d'exposition. Par contre, les poids stabilisés simples se révèlent être robustes au biais introduit par une telle erreur de spécification du modèle.

Mots-clés : Inférence causale, sélection de modèles, confusion, variable confondante, modèles structuraux marginaux, pondération par probabilité inverse.

INTRODUCTION

L'inférence causale pourrait être définie comme étant le processus qui permet de tirer des conclusions sur des relations de cause à effet. Une définition formelle de l'inférence causale ne saurait toutefois transmettre jusqu'à quel point la pensée causale est naturelle chez l'être humain. Dès l'enfance, nous observons les conséquences de nos actions et effectuons ainsi, d'une certaine manière, de l'inférence causale. À titre d'exemple, on peut s'imaginer une situation similaire à celle présentée par Rothman & Greenland (2005) : un enfant jouant avec un interrupteur et découvrant qu'une lumière s'allume ou s'éteint selon la position de l'interrupteur. Alors que l'inférence causale peut se faire assez facilement dans une situation aussi simple où l'effet (lumière) de la cause (position de l'interrupteur) peut s'observer immédiatement, le processus menant à l'inférence causale est souvent bien plus compliqué.

Imaginons maintenant que nous voulons déterminer les effets à long terme de la consommation régulière de vin rouge sur notre santé. D'une part, les effets qui nous intéressent sont à long terme ; ils ne peuvent pas être observés immédiatement. Il est alors difficile d'être certain que les résultats observés sont bel et bien dus à la consommation régulière de vin rouge et non à d'autres facteurs. D'autre part, l'expérience ne pourrait s'effectuer individuellement. Afin de déterminer comment sa santé à long terme est affectée par le vin rouge, un même individu ne pourrait à la fois en consommer régulièrement et ne pas en consommer.

Bien que le processus d'inférence causale soit généralement complexe, il est d'une importance extrême. Dans le domaine de la santé, les techniques d'inférence causale permettent d'évaluer l'effet d'une intervention potentielle sur la santé des individus et ainsi contribuer à améliorer la santé publique. Par exemple, des études utilisant des techniques statistiques d'inférence causale suggèrent que le Zidovudine améliore

la survie des hommes atteints du VIH (Hernán *et al.*, 2000), que l'utilisation d'antirétroviraux hautement actifs augmente le temps avant de devenir atteint du SIDA ou de décéder chez les hommes atteints du VIH (Cole *et al.*, 2003) et que la consommation d'Aspirin favorise la survie des femmes atteintes du cancer du sein (Holmes *et al.*, 2010).

Il n'est donc pas surprenant que la recherche en inférence causale ainsi que ses applications soient en pleine croissance. Toutefois, malgré cet intérêt grandissant pour l'inférence causale, il existe encore de nombreux défis dans ce domaine. Ceux-ci sont particulièrement importants lorsqu'on désire effectuer de l'inférence causale avec des données d'observation, c'est-à-dire des données où l'exposition des sujets n'est pas décidée ou contrôlée par l'investigateur. Sommairement, afin d'effectuer de l'inférence causale sur la base de données d'observation, il est nécessaire de contrôler pour les variables dites confondantes. Ce contrôle peut se faire directement au moment de récolter les données, par exemple par appariement, ou au moment d'analyser les données, par exemple avec un modèle statistique.

Dans cette thèse, nous nous intéressons à l'estimation d'un effet causal à l'aide d'un modèle statistique avec des données d'observation. Nous abordons plus spécifiquement la problématique de la sélection de modèles visant l'inférence causale. Le premier chapitre effectue une brève introduction aux paradigmes contrefactuel et graphique à l'inférence causale et présente la problématique étudiée dans cette thèse. Le reste de la thèse comporte deux grandes parties.

La première partie de la thèse, constituée des chapitres 2, 3 et 4, porte sur des méthodes statistiques d'inférence causale où la sélection de modèles est dirigée par les données observées. Plus spécifiquement, le deuxième chapitre effectue une revue non exhaustive des écrits scientifiques portant sur ce genre de méthodes. Le troisième chapitre présente en plus grand détail l'algorithme *Bayesian Adjustment for Confounding* (BAC) récemment proposé par Wang *et al.* (2012a), ainsi que les développements auxquels j'ai contribué durant mon doctorat concernant cet algorithme. Le chapitre

4 comporte une version longue d'un article en langue anglaise à paraître dans *Journal of Causal Inference* et qui introduit un nouvel algorithme de sélection de modèles pour l'inférence causale : *Bayesian Causal Effect Estimation* (BCEE). Cet algorithme possède quelques points en commun avec BAC, mais vise à produire un estimateur de l'effet causal plus performant, en termes d'erreur quadratique moyenne, que l'estimateur obtenu avec BAC.

La deuxième partie de la thèse est formée des chapitres 5, 6 et 7 et aborde l'identification de modèles adéquats pour l'inférence causale dans un contexte où un ensemble de variables potentiellement confondantes de taille raisonnable peut être identifié sur la base de la littérature scientifique. Les travaux présentés se rapportent à une application pratique de techniques d'inférence causale et aux développements méthodologiques qui y sont associés. Plus précisément, le chapitre 5 introduit l'étude sur laquelle l'application a porté, le *Honolulu Heart Program* (HHP), ainsi que les modèles utilisés pour analyser les données, les modèles structuraux marginaux (MSMs). Le sixième chapitre est formé d'une version longue d'un article qui sera prochainement soumis à *Epidemiology* et décrivant l'approche méthodologique suivie pour analyser les données du HHP. Nous y décrivons, entre autres, comment les MSMs peuvent être imbriqués dans le paradigme graphique pour l'inférence causale et ainsi faciliter l'implantation des MSMs. Le septième chapitre contient un article scientifique publié dans *Statistics in Medicine*. Nous y exposons une problématique concernant l'utilisation des poids stabilisés dans les MSMs. Nous nous intéressons par le fait même au choix du type de poids à utiliser dans les MSMs et, ainsi, à la spécification d'un modèle approprié pour effectuer de l'inférence causale. La thèse se termine avec une discussion permettant de mettre en perspective les travaux de recherches présentés dans les chapitres précédents.

CHAPITRE I

LES PARADIGMES CONTREFACTUEL ET GRAPHIQUE À L'INFÉRENCE CAUSALE

Dans ce chapitre, nous présentons les deux principaux paradigmes à l'inférence causale, soit le paradigme contrefactuel et le paradigme graphique. Ce faisant, nous introduisons également la problématique étudiée dans cette thèse.

1.1 Le paradigme contrefactuel

Le paradigme contrefactuel, également appelé le paradigme des issues potentielles (*potential outcomes*), a d'abord été élaboré par Neyman (Neyman, 1923) pour étudier les expériences randomisées. Une généralisation permettant d'étudier les liens causaux avec des données d'observation a par la suite été réalisée par Rubin (Rubin, 1974).

1.1.1 Notation

Soit Y, la variable aléatoire correspondant à l'issue étudiée, par exemple le fait de développer ou non une maladie, et soit X, la variable aléatoire correspondant à l'exposition[1], par exemple la prise d'un certain médicament. Afin de simplifier la présentation, nous supposons que Y et X sont binaires $(0/1)$, bien qu'une telle hypothèse ne soit pas nécessaire. Nous notons également par l'indice $i = 1, ..., n$ les

1. Nous utilisons les termes « exposition » et « traitement » de façon interchangeable dans cette thèse.

sujets échantillonnés et utilisons les lettres minuscules y et x pour représenter les réalisations des variables aléatoires Y et X, respectivement.

1.1.2 Définition de l'effet causal

Le paradigme contrefactuel demande d'imaginer que, pour chaque individu, il existe deux issues potentielles, ou contrefactuelles : 1) l'issue potentielle correspondant à l'exposition, $Y(x = 1)$ et 2) l'issue potentielle correspondant à l'absence d'exposition, $Y(x = 0)$. Pour que de telles quantités aient un sens, il est nécessaire de faire l'hypothèse que chaque individu a des probabilités non nulles d'être exposé et d'être non exposé (hypothèse de positivité, *positivity*). Cette hypothèse ne serait pas respectée, par exemple, dans un contexte où l'on chercherait à comparer deux traitements, mais qu'un des deux est contrindiqué pour certains patients en raison d'autres problèmes de santé. De tels patients auraient alors une probabilité nulle d'avoir le médicament contrindiqué. Formellement, l'hypothèse de positivité se définit comme suit.

Définition 1.1. *Soit un traitement ou une exposition X et un ensemble de covariables $\boldsymbol{Z}$ (possiblement $\boldsymbol{Z} = \varnothing$). Il y a* ***positivité*** *conditionnellement à $\boldsymbol{Z}$ si $P(X = x|\boldsymbol{Z} = \boldsymbol{z}) > 0$ pour tout x et $\boldsymbol{z}$ tels que $f(\boldsymbol{z}) > 0$.*

En pratique, il est également commun de faire l'hypothèse que le niveau d'exposition reçu par un individu n'affecte pas l'issue des autres individus et qu'il n'existe pas plusieurs versions de chaque niveau d'exposition (hypothèse de stabilité de la valeur du traitement pour l'unité, *stable unit treatment value assumption - SUTVA*). Cette hypothèse est en fait implicitement nécessaire pour définir les deux issues potentielles, car, sans elle, l'écriture de $Y(x = 1)$ et de $Y(x = 0)$ n'a plus de sens. L'hypothèse SUTVA n'est pas raisonnable, par exemple, dans le contexte de maladies contagieuses. Dans ce contexte, le fait qu'une partie de la population reçoive un traitement pour prévenir la maladie peut non seulement réduire le risque d'être malade chez les traités,

mais également chez les non-traités. Le traitement a ainsi une valeur, ou une utilité, différente pour un sujet donné en fonction du traitement reçu par les autres sujets.

L'utilisation des deux issues potentielles, $Y(x = 1)$ et $Y(x = 0)$, permet de définir différentes quantités causales.

Définition 1.2.

- ***Effet causal pour le sujet*** i : $Y_i(x = 1) - Y_i(x = 0)$;
- ***Effet causal moyen*** : $\mathbb{E}[Y(x = 1) - Y(x = 0)]$;
- ***Effet causal moyen chez les traités*** : $\mathbb{E}[Y(x = 1) - Y(x = 0)|X = 1]$;
- ***Rapport de risque causal*** : $\mathbb{E}[Y(x = 1)]/\mathbb{E}[Y(x = 0)]$;
- ***Rapport de cotes causal*** : $\frac{\mathbb{E}[Y(x=1)]/(1-\mathbb{E}[Y(x=1)])}{\mathbb{E}[Y(x=0)]/(1-\mathbb{E}[Y(x=0)])}$.

Le problème fondamental de l'inférence causale est que les deux issues potentielles ne peuvent pas être observées simultanément pour un même individu, de sorte que les quantités causales définies précédemment ne peuvent pas être estimées directement à partir des données. Cependant, on suppose que l'issue factuelle, c'est-à-dire l'issue réellement observée, est la même que l'issue potentielle correspondant au niveau d'exposition observé (hypothèse de cohérence, *consistency*).

Définition 1.3. *Il y a cohérence si*

- $x_i = 0 \Rightarrow Y_i = Y_i(x = 0)$,
- $x_i = 1 \Rightarrow Y_i = Y_i(x = 1)$.

Sous cette hypothèse, on observe pour chaque individu l'une des deux issues potentielles. Il est ainsi possible d'estimer $\mathbb{E}[Y(x = 1)]$ et $\mathbb{E}[Y(x = 0)]$ avec les données d'une expérience randomisée idéale. Notons par n_0 et n_1 le nombre d'observations pour lesquelles $x_i = 0$ et $x_i = 1$, respectivement. Alors $\mathbb{E}[Y(x = 0)]$ et $\mathbb{E}[Y(x = 1)]$ peuvent être estimées sans biais par $\bar{y}_0 = \sum_{\{i \mid x_i=0\}} y_i/n_0$ et $\bar{y}_1 = \sum_{\{i \mid x_i=1\}} y_i/n_1$ respectivement (Rubin, 1974), où $\{i \mid x_i = a\}$ dénote l'ensemble des is tels que

$x_i = a$. Ce résultat découle du fait que, dans le cadre d'une expérience randomisée idéale, le niveau d'exposition est décidé de façon totalement aléatoire. Ainsi, toutes les caractéristiques pré-exposition des sujets, ainsi que les réponses potentielles sont, en moyenne, équilibrées entre les deux groupes.

Proposition 1.1.1. $\mathbb{E}(\bar{Y}_0) = \mathbb{E}[Y_i(x_i = 0)]$ *et* $\mathbb{E}(\bar{Y}_1) = \mathbb{E}[Y_i(x_i = 1)]$.

Démonstration.

$$\begin{aligned}
\mathbb{E}[\bar{Y}_0] &= \mathbb{E}\left[\sum_{\{i|x_i=0\}} Y_i/n_0\right] \\
&= \mathbb{E}[Y_i|X_i = 0] \\
&= \mathbb{E}[Y_i(x_i = 0)|X_i = 0] && \text{(hypothèse de cohérence)} \\
&= \mathbb{E}[Y_i(x_i = 0)] && (Y_i(x_i = 0) \perp\!\!\!\perp X_i \text{ en raison de la randomisation}),
\end{aligned}$$

où $\perp\!\!\!\perp$ dénote l'indépendance statistique. De la même façon, on peut montrer que $\mathbb{E}[\bar{Y}_1] = \mathbb{E}[Y_i(x_i = 1)]$. □

1.1.3 Confusion

Nous avons montré qu'il est simple d'estimer $\mathbb{E}[Y(x = 1)]$ et $\mathbb{E}[Y(x = 0)]$ à partir de données provenant d'une expérience randomisée idéale. Il est cependant plus difficile d'accomplir cette tâche à partir de données d'observation. En effet, en absence de randomisation, différents facteurs peuvent influencer à la fois le niveau d'exposition des sujets et leur réponse observée. Dans une telle situation, les réponses contrefactuelles ne sont pas nécessairement équilibrées entre le groupe exposé et le groupe non exposé. La différence naïve $\bar{Y}_1 - \bar{Y}_0$ peut alors estimer de façon biaisée l'effet causal moyen $\mathbb{E}[Y(x = 1) - Y(x = 0)]$. En effet, la dernière étape de la preuve précédente ne pourrait pas être effectuée, car $Y_i(x_i = 0) \not\perp\!\!\!\perp X_i$. On dira alors qu'il y a de la confusion.

Définition 1.4. *Il y a absence de* ***confusion*** *pour la relation causale entre X et Y si et seulement si $\{Y(0), Y(1)\} \perp\!\!\!\perp X$ (hypothèse forte d'ignorabilité).*

Afin d'estimer sans biais l'effet causal, une approche est de « contrôler » pour les variables confondantes. Ces variables confondantes sont souvent conceptualisées comme étant les causes communes de l'exposition et de la réponse (VanderWeele & Shpitser, 2011). Une condition suffisante pour estimer sans biais l'effet causal moyen est de contrôler pour un ensemble de variables $\boldsymbol{Z}$ satisfaisant l'hypothèse faible d'ignorabilité conditionnelle (Rosenbaum & Rubin, 1983) :

$$Y(x) \perp\!\!\!\perp X|\boldsymbol{Z}, \text{ pour } x \in \{0,1\}. \tag{1.1}$$

Démonstration.

$$\begin{aligned}\sum_{\boldsymbol{z}} \mathbb{E}[Y|X=0,\boldsymbol{Z}=\boldsymbol{z}]P(\boldsymbol{Z}=\boldsymbol{z}) &= \sum_{\boldsymbol{z}} \mathbb{E}[Y(x=0)|X=0,\boldsymbol{Z}=\boldsymbol{z}]P(\boldsymbol{Z}=\boldsymbol{z}) \\ &= \sum_{\boldsymbol{z}} \mathbb{E}[Y(x=0)|\boldsymbol{Z}=\boldsymbol{z}]P(\boldsymbol{Z}=\boldsymbol{z}) \\ &= \mathbb{E}[Y(x=0)],\end{aligned}$$

où $\sum_{\boldsymbol{z}}$ représente une somme (possiblement infinie ou une intégrale) sur toutes les valeurs possibles de $\boldsymbol{Z}$ (voir Pearl (2009)). La preuve est effectuée de la même façon pour le cas $X = 1$. □

Remarquons que cette hypothèse est respectée dans le cas d'un étude randomisée idéale, conditionnellement à $\boldsymbol{Z} = \varnothing$, justement en raison de la randomisation. L'hypothèse faible d'ignorabilité conditionnelle est toutefois invérifiable en pratique à partir des données puisque, pour chaque individu, une seule des issues potentielles est observée. On comprend ainsi facilement la difficulté d'identifier les variables confondantes $\boldsymbol{Z}$. Par ailleurs, même si l'on parvient à identifier un ensemble $\boldsymbol{Z}$

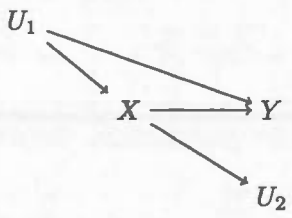
suffisant, il existe plusieurs options pour contrôler pour ces variables. Le choix d'une approche appropriée et efficace peut lui-même être ardu.

1.2 Le paradigme graphique

Le paradigme graphique à l'inférence causale a été, dans une large mesure, formalisé grâce aux travaux de Pearl (Pearl, 2009). Toutefois, l'idée d'utiliser des graphes pour représenter les liens entre différentes variables et d'en tirer des conclusions sur les relations causales entre ces variables proviendrait de Wright (Wright, 1921).

Afin de décrire le paradigme graphique à l'inférence causale, nous employons une notation similaire à celle utilisée dans la section précédente, c'est-à-dire que nous utilisons Y pour représenter la variable aléatoire correspondant à l'issue d'intérêt et X pour représenter la variable aléatoire correspondant à l'exposition. Nous nous servons de la figure 1.1 afin d'illustrer les différents concepts présentés.

Figure 1.1: Exemple de graphe acyclique orienté pour des données d'observation



1.2.1 Construction d'un graphe

Le paradigme graphique nécessite de tracer un graphe pour représenter les liens entre les variables pertinentes à l'étude de la relation causale entre X et Y (nous reviendrons sur le choix des variables à inclure dans le graphe). Dans le graphe, les variables sont représentées par des sommets (par exemple X, Y, U_1 et U_2 dans la figure 1.1) et les liens de cause à effet sont représentés par des flèches (par exemple $X \rightarrow Y$).

L'absence de flèche entre deux variables (par exemple entre Y et U_2) constitue donc une hypothèse d'absence de lien de cause à effet direct entre ces deux variables (on suppose que Y n'a pas d'effet causal direct sur U_2 et vice-versa). Tous les liens ont par ailleurs une et une seule direction (le graphe est dit **orienté**). Notons qu'on représente parfois la présence d'une cause commune non-observée entre deux variables par une flèche bidirectionnelle. Toutefois, cette notation n'est qu'un raccourci, puisque les liens causaux bidirectionnels ne sont pas permis dans le paradigme graphique causal.

En plus d'être orienté, le graphe doit également être **acyclique**, c'est-à-dire qu'en partant de n'importe quelle variable du graphe et en suivant le sens des flèches, il est impossible de retourner au point de départ. Cette hypothèse peut facilement se justifier par l'axiome causal qui veut que toute cause précède temporellement son effet. Il serait donc illogique qu'un graphe causal contienne un cycle. Puisque le graphe construit doit à la fois être acyclique et orienté, on dit qu'il s'agit d'un graphe acyclique orienté (*directed acyclic graph*, DAG).

Définition 1.5. *Un **graphe** est formé d'un ensemble $\boldsymbol{V}$ de sommets et d'un ensemble $\boldsymbol{E}$ d'arêtes permettant de relier un à un certains (possiblement aucun) sommets. Un graphe est dit **orienté** si toutes ses arêtes ont une et une seule direction (les arêtes sont des flèches unidirectionnelles). Un graphe orienté est **acyclique** si, pour chaque $v \in \boldsymbol{V}$, il est impossible de trouver un chemin (un ensemble contigu d'arêtes) menant de v à v en suivant le sens des arêtes.*

Nous avons mentionné précédemment que le graphe doit représenter les liens entre les variables pertinentes à l'étude de la relation causale entre X et Y, sans préciser la nature des variables pertinentes. En fait, pour pouvoir étudier la relation causale entre X et Y à partir du DAG, il faut que le DAG soit **causal**, c'est-à-dire qu'il doit inclure toutes les variables (observées ou non) qui sont des causes communes de toutes les paires de variables représentées sur le graphe (Hernán & Robins, 2015).

Définition 1.6. *Un* ***DAG causal*** *est un DAG pour lequel 1) une absence d'arête entre deux sommets v_1 et v_2 peut être interprétée comme une absence d'effet causal direct entre v_1 et v_2, 2) toutes les causes, observées ou non, de chaque paire d'éléments de $\mathbf{V}$ sont également sur le graphe, et 3) la présence d'un chemin menant de v_1 à v_2 en suivant le sens des arêtes implique que v_1 est une cause de v_2.*

1.2.2 Vocabulaire

Afin de simplifier le reste de la présentation, nous introduisons quelques mots de vocabulaire utilisés dans le paradigme graphique.

On dira qu'un ensemble de sommets et de flèches contigus constitue un **chemin** (par exemple $U_1 \rightarrow X \rightarrow Y$ ou $U_2 \leftarrow X \rightarrow Y$). Un chemin est **orienté** si les flèches vont toutes dans la même direction (par exemple $U_1 \rightarrow X \rightarrow Y$). Un chemin **porte-arrière** entre deux variables X et Y est un chemin entre X et Y ayant une flèche pointant vers X (par exemple $X \leftarrow U_1 \rightarrow Y$). Un **collisionneur** (*collider*) est une variable sur un chemin vers laquelle deux flèches pointent (par exemple, Y est un collisionneur dans $X \rightarrow Y \leftarrow U_1$).

Il est également commun d'utiliser un vocabulaire généalogique afin d'identifier les liens entre les variables dans le graphe. Dans la figure 1.1, X est un enfant de U_1 puisqu'il y a une flèche partant de U_1 pointant vers X. Inversement, U_1 est un parent de X. De plus, U_2 est un descendant de U_1 puisqu'il existe un chemin orienté menant de U_1 à U_2 ($U_1 \rightarrow X \rightarrow U_2$). À l'opposé, U_1 est un ancêtre de U_2.

Si un chemin inclut un collisionneur pour lequel on ne contrôle pas et pour lequel on ne contrôle pour aucun de ses descendants, alors ce chemin est **fermé** ou **bloqué**. Un chemin est également fermé si on contrôle pour une variable faisant partie du chemin qui n'est pas un collisionneur. Un chemin qui n'est pas fermé est un chemin **ouvert**. Un concept relié à celui de chemin ouvert et fermé est celui de d-séparation.

Définition 1.7. *Soit un DAG causal G. Deux variables X et Y sont* ***d-séparées*** *par un ensemble de variables* $\boldsymbol{Z}$ *dans G si et seulement si tous les chemins entre X et Y sont bloqués après avoir contrôlé pour* $\boldsymbol{Z}$.

1.2.3 DAGs et associations statistiques

Les DAGs peuvent être utilisés pour représenter les associations statistiques entre différentes variables. Par exemple, si P est la distribution conjointe des variables $X_1, X_2, ..., X_J$, alors un DAG G est compatible avec P s'il permet de représenter le processus aléatoire qui génère les données. Pour que G soit compatible avec P il est nécessaire que si les parents de X_j d-séparent X_j de ses autres ancêtres dans G, alors X_j est indépendant de ses autres ancêtres conditionnellement à ses parents (Pearl (2009), definition 1.2.2).

Définition 1.8. *Soit* $\boldsymbol{V} = \{X_1, ..., X_J\}$ *un ensemble de variables ordonnées, G un DAG et* pa_j *les parents de* X_j *dans G. Si la distribution conjointe de* $\boldsymbol{V}$, *P, permet la factorisation* $P(x_1, ..., x_J) = \prod_{j=1}^{J} p(x_j|pa_j)$, *alors G et P sont dits* ***compatibles***.

Lorsque G et P sont compatibles, il existe un fort lien entre le concept de d-séparation et d'association statistique. En effet, si X et Y sont d-séparées par $\boldsymbol{Z}$, alors $X \perp\!\!\!\perp Y|\boldsymbol{Z}$. La réciproque est d'ailleurs presque toujours vraie. En fait, elle n'est fausse que dans des cas très précis et improbables en pratique (Pearl (2009), sections 1.2.3, 2.4 et 2.9.1).

Conceptuellement, un chemin ouvert permet ainsi l'existence d'une association statistique entre les variables aux extrémités du chemin. Par exemple, le chemin porte-arrière $X \leftarrow U_1 \rightarrow Y$ de la figure 1.1 permet une association entre X et Y par le biais de U_1. Par opposition, un chemin bloqué n'engendre pas d'association entre deux variables, bien qu'une association pourrait provenir d'autres chemins. Ainsi, si on bloque le chemin $X \leftarrow U_1 \rightarrow Y$ en contrôlant pour U_1, les variables X et Y ne sont plus associées par l'intermédiaire de U_1, mais demeurent associées en raison du chemin

$X \rightarrow Y$. On peut déduire de cet exemple une approche permettant d'estimer l'effet causal de X sur Y : identifier et bloquer les chemins qui engendrent une association entre X et Y qui n'est pas une association de cause à effet et ne pas bloquer les chemins créant l'association causale. Cette approche est formalisée par le critère porte-arrière que nous présentons à la prochaine section.

1.2.4 Identifier et éviter la confusion à l'aide d'un DAG

En supposant qu'on dispose d'un DAG causal, il est relativement facile d'étudier la relation causale entre X et Y. Par la simple observation du DAG, on peut déterminer s'il existe de la confusion et les variables confondantes pour lesquelles on peut contrôler pour éliminer la confusion. Pour ce faire, une approche consiste à utiliser le critère porte-arrière (*back-door criterion*, Pearl (2009) section 3.3.1) :

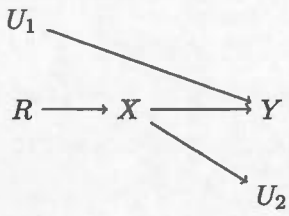
Définition 1.9. *Un ensemble de variables $\boldsymbol{Z}$ satisfait le critère porte-arrière pour une paire de variables ordonnée (X,Y) dans un DAG causal G si :*
(i) aucune variable dans $\boldsymbol{Z}$ n'est un descendant de X et
(ii) $\boldsymbol{Z}$ bloque tous les chemins porte-arrière entre X et Y.

Généralement, si l'ensemble $\boldsymbol{Z} = \varnothing$ n'est pas suffisant en vertu de la définition (1.9), il existe de la confusion pour la relation causale entre X et Y. Par contre, cette confusion peut être éliminée en contrôlant pour un ensemble $\boldsymbol{Z}$ suffisant. Dans la figure 1.1, il existe ainsi de la confusion engendrée par le chemin porte-arrière $X \leftarrow U_1 \rightarrow Y$. Toutefois, en contrôlant pour U_1, et en évitant de contrôler pour U_2, on obtient un ensemble $\boldsymbol{Z} = \{U_1\}$ qui permet d'éviter la confusion.

Remarquons que le paradigme graphique indique pourquoi une étude randomisée idéale permet l'estimation de l'effet causal moyen présenté à la définition 1.2. En effet, dans une expérience randomisée idéale, seul le hasard détermine le niveau d'exposition, ce qui élimine tous les chemins porte-arrière qui pourraient naturellement exister. La figure 1.2 est une version modifiée du DAG 1.1 qui

représente l'impact d'une randomisation idéale (R).

Figure 1.2: Exemple de graphe acyclique orienté pour une expérience randomisée idéale



En absence de randomisation, des problèmes similaires à ceux rencontrés avec le paradigme contrefactuel font surface. D'une part, afin d'utiliser le critère porte-arrière pour identifier un ensemble $\boldsymbol{Z}$ suffisant pour identifier l'effet causal de X sur Y, il faut d'abord tracer un DAG causal. Il est néanmoins difficile de déterminer quelles variables et flèches inclure dans le DAG, ainsi que la direction des flèches incluses. D'autre part, une façon appropriée et efficace de contrôler pour $\boldsymbol{Z}$ peut être difficile à déterminer.

1.3 La sélection de modèles pour l'inférence causale

À la fois pour le paradigme contrefactuel et pour le paradigme graphique, nous avons mis en évidence que l'identification des variables confondantes représente un défi de taille dans l'application des méthodes d'inférence causale avec des données d'observation. De plus, même lorsque ces variables sont identifiées, le choix d'une façon de contrôler pour ces variables peut elle aussi représenter un défi. Nous considérons que ces deux problèmes font en fait partie de la problématique plus générale de la sélection de modèles pour l'inférence causale avec des données d'observation.

Cette problématique est complexe et ne dispose pas d'une solution unique qui saurait s'appliquer à toute la diversité des problèmes rencontrés en pratique. Les solutions appropriées dépendront à coup sûr des particularités de chaque étude. Dans cette thèse, nous proposons des solutions à quelques situations.

Dans la première partie de la thèse, nous nous intéressons plus particulièrement à une situation où l'on dispose d'un grand ensemble de variables potentiellement confondantes qu'on suppose suffisant. Le problème étudié consiste à choisir un sous-ensemble réduit de variables potentiellement confondantes afin d'améliorer la précision de l'estimation de l'effet causal moyen. La deuxième partie de la thèse porte sur une application pratique des techniques d'inférence causale. Contrairement à la situation étudiée dans la première partie de la thèse, un nombre restreint de variables cliniquement pertinentes a pu être identifié *a priori* et un premier jet de DAG a dès lors pu être imaginé. Nous nous intéressons dans cette partie de la thèse à la validation et à l'amélioration de ce premier jet de DAG ainsi qu'à la spécification des MSMs utilisés pour analyser les données.

Première partie

Sélection de modèles guidée par les données pour l'inférence causale

CHAPITRE II

UNE REVUE DE QUELQUES MÉTHODES EXISTANTES

Une stratégie intéressante pour contourner le problème d'identification d'un ensemble suffisant consiste à mesurer un grand ensemble de variables potentiellement confondantes, $\boldsymbol{U} = \{U_1, ..., U_M\}$. Ces variables sont ainsi qualifiées sur la base des connaissances *a priori* du domaine d'application, de sorte qu'on soit raisonnablement certain que cet ensemble est suffisant. Une fois $\boldsymbol{U}$ identifié, le contrôle peut être effectué de différentes façons. L'effet causal moyen sur Y d'une augmentation de X d'une unité peut théoriquement être obtenu par

$$\sum_{\boldsymbol{u}} (\mathbb{E}[Y|\boldsymbol{U} = \boldsymbol{u}, X = x + 1] - \mathbb{E}[Y|\boldsymbol{U} = \boldsymbol{u}, X = x]) P(\boldsymbol{U} = \boldsymbol{u}). \tag{2.1}$$

Ce résultat découle de la démonstration effectuée à la section 1.1.3. En pratique, on peut estimer (2.1) à l'aide d'un modèle statistique approprié. Par exemple, l'estimation de (2.1) pourrait se faire par l'estimation du paramètre β dans le modèle de régression linéaire $\mathbb{E}[Y|X, \boldsymbol{U}] = \delta_0 + \beta X + \sum_{m=1}^{M} \delta_m U_m$.

Un inconvénient à l'utilisation d'un grand ensemble de variables potentiellement confondantes pour effectuer le contrôle est que celui-ci inclut probablement plus de variables que nécessaire. L'estimation de l'effet causal à l'aide de la totalité des variables dans $\boldsymbol{U}$ peut alors être fortement inefficace, c'est-à-dire que la variance de

l'estimateur de l'effet causal peut être très élevée. Ce problème est particulièrement bien connu pour la régression linéaire où l'utilisation d'un grand ensemble de variables de contrôle engendre non seulement une réduction du nombre de degrés de liberté, mais peut également conduire à une inflation de la variance de $\hat{\beta}$ en raison de la colinéarité entre X et $\boldsymbol{U}$ (voir par exemple O'brien (2007)). Une solution à ce problème est de tenter d'identifier, à l'aide des données, un sous-ensemble $\boldsymbol{Z} \subset \boldsymbol{U}$ également suffisant pour identifier l'effet causal de X sur Y. Il est permis d'espérer que l'estimation de l'effet causal à l'aide de $\boldsymbol{Z}$ soit plus efficace que l'estimation à l'aide de $\boldsymbol{U}$.

Nous effectuons dans ce chapitre une revue de quelques méthodes qui peuvent être utilisées pour mettre en application une telle stratégie en vue de l'estimation de l'effet causal de X sur Y. Nous présentons en plus grand détail des méthodes qui sont reliées à l'algorithme BCEE présenté au chapitre 4.

2.1 Le modèle moyen bayésien

Le modèle moyen bayésien (*Bayesian Model Averaging*, BMA) (Raftery *et al.*, 1997; Hoeting *et al.*, 1999) est une approche bayésienne de sélection de modèles qui a d'abord été suggérée dans un contexte pour effectuer de la prédiction. À quelques reprises, il a également été proposé d'utiliser le BMA afin d'estimer un effet (causal) et d'effectuer une sélection de variables confondantes (par exemple, Clyde (2000); Koop & Tole (2004)). Nous décrivons le BMA en faisant référence principalement à ce dernier contexte.

Soit $\boldsymbol{A} = \{A_1, ..., A_K\}$, un ensemble de modèles considérés pour estimer l'effet causal de X sur Y, et soit $\boldsymbol{U} = \{U_1, ..., U_M\}$ un ensemble de variables potentiellement confondantes. Les modèles $\boldsymbol{A}$ peuvent être, par exemple, les $K = 2^M$ modèles de régression linéaire de la forme $\mathbb{E}[Y|X, \boldsymbol{Z}] = \delta_0 + \beta X + \boldsymbol{\delta Z}$, où $\boldsymbol{Z} \subseteq \boldsymbol{U}$, bien que des modèles linéaires généralisés ou des modèles de survie peuvent être considérés, selon

le type de la variable Y. Les inférences par rapport à l'effet causal de X sur Y se font à partir de la loi *a posteriori* de β :

$$p(\beta|Y) = \sum_{k=1}^{K} p(\beta|A_k, Y)p(A_k|Y).$$

Ainsi, le BMA permet non seulement d'effectuer de la sélection de modèles basée sur les données observées, mais également de tenir compte de l'incertitude associée au choix du modèle, contrairement à des approches plus classiques de sélection de variables telle que la régression pas-à-pas (*stepwise*). Ces propriétés font en sorte que le BMA est particulièrement performant pour effectuer de la prédiction d'observations futures. En effet, Madigan & Raftery (1994) ont démontré qu'utiliser le BMA permet d'avoir une meilleure performance prédictive moyenne que de choisir un seul modèle A_k, quelque soit $A_k \in \boldsymbol{A}$. Le BMA produit aussi des intervalles de confiance (crédibilité) pour les valeurs prédites dont le niveau de couverture correspond approximativement au niveau de couverture désiré (Raftery & Zheng, 2003).

Néanmoins, des études de simulations ont montré que lorsque le BMA est utilisé pour l'estimation de l'effet causal moyen de X sur Y, l'estimateur obtenu peut être biaisé et les intervalles de confiance peuvent avoir un taux de couverture plus faible que celui attendu (Wang *et al.*, 2012a; Crainiceanu *et al.*, 2008). Ce résultat est peu surprenant puisque le BMA ne tient pas compte du contexte spécifique de l'inférence causale dans l'attribution des probabilités *a posteriori* des modèles; la probabilité *a posteriori* d'un modèle A_k ne dépend que de l'ajustement du modèle aux données et de sa probabilité *a priori*, $p(A_k|Y) \propto p(Y|A_k)p(A_k)$. Ainsi, les modèles incluant certaines variables confondantes importantes faiblement associées à la réponse, mais fortement associées à l'exposition, peuvent recevoir un poids *a posteriori* faible.

2.2 La méthode de Crainiceanu *et al.* (2008)

Réalisant que le BMA ne produit pas toujours des résultats satisfaisants pour l'inférence causale, Crainiceanu *et al.* (2008) (CDP) proposent une nouvelle méthode de sélection de modèles visant spécifiquement l'estimation de l'effet de X sur Y. L'approche de CDP cible l'estimation sans biais du paramètre de régression associé à l'exposition dans un modèle de réponse avec ajustement pour des facteurs potentiellement confondants, par exemple le modèle de régression linéaire :

$$\mathbb{E}[Y|X,\boldsymbol{U}] = \delta_0^{\boldsymbol{\alpha}} + \beta^{\boldsymbol{\alpha}} X + \sum_{m=1}^{M} \alpha_m \delta_m^{\boldsymbol{\alpha}} U_m, \tag{2.2}$$

où $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_M) \in \{0,1\}^M$ avec $\alpha_m = 1$ si et seulement si U_m est inclus dans le modèle. Toutefois, d'autres modèles peuvent être considérés, par exemple, des modèles linéaires généralisés. Tout comme CDP, afin de simplifier la présentation, nous considérons le vecteur $\boldsymbol{\alpha}$ comme un modèle et utilisons la notation $\boldsymbol{\alpha} \subset \boldsymbol{\alpha}'$ pour indiquer que le modèle $\boldsymbol{\alpha}$ est emboîté dans le modèle $\boldsymbol{\alpha}'$.

CDP remarquent que l'interprétation de $\beta^{\boldsymbol{\alpha}}$ peut varier selon le modèle. En supposant que $\boldsymbol{U}$ inclut un sous-ensemble suffisant pour identifier l'effet causal de X sur Y, ils postulent l'existence d'un ensemble **minimal** suffisant, c'est-à-dire un sous-ensemble de $\boldsymbol{U}$ suffisant et de taille minimale parmi tous les sous-ensembles de $\boldsymbol{U}$ suffisants. CDP dénotent par $\boldsymbol{\alpha}^*$ cet ensemble minimal suffisant et le qualifient comme étant le « vrai » modèle. Ils supposent par ailleurs que $\beta^{\boldsymbol{\alpha}} = \beta^{\boldsymbol{\alpha}^*}$ si $\boldsymbol{\alpha}^* \subseteq \boldsymbol{\alpha}$ et que, généralement, $\beta^{\boldsymbol{\alpha}} \neq \beta^{\boldsymbol{\alpha}^*}$ si $\boldsymbol{\alpha}^* \not\subseteq \boldsymbol{\alpha}$. Finalement, CDP notent que si $\boldsymbol{\alpha}^* \subset \boldsymbol{\alpha}$, $var(\hat{\beta}^{\boldsymbol{\alpha}}) \geq var(\hat{\beta}^{\boldsymbol{\alpha}^*})$. Ainsi, les intervalles de confiance basés sur $\boldsymbol{\alpha}$ seront plus larges que nécessaire. Il y a donc tout intérêt à tenter de sélectionner un ensemble suffisant pour estimer l'effet causal qui s'approche autant que possible de l'ensemble minimal suffisant $\boldsymbol{\alpha}^*$.

En utilisant la notation introduite, CDP expriment l'espérance *a posteriori* de β associée au BMA comme

$$\mathbb{E}[\beta|Y] = \sum_{\{\alpha|\alpha^* \subseteq \alpha\}} E[\beta^\alpha|\alpha, Y]p(\alpha|Y) + \sum_{\{\alpha|\alpha^* \not\subseteq \alpha\}} E[\beta^\alpha|\alpha, Y]p(\alpha|Y). \qquad (2.3)$$

Le biais du BMA pour l'estimation de l'effet causal moyen de X sur Y provient ainsi du poids *a posteriori* accordé aux modèles du deuxième terme de (2.3).

La méthode de CDP est similaire au BMA, mais cherche à concentrer les inférences sur le premier terme de (2.3). En bref, l'algorithme de CDP consiste d'abord à effectuer une modélisation de la variable d'exposition X dans le but d'identifier les forts prédicteurs de l'exposition, disons $\boldsymbol{U}_1$. Ensuite, une modélisation pour la variable réponse est effectuée, où les variables associées à l'exposition, $\boldsymbol{U}_1$, sont forcées dans le modèle. Les modèles $\boldsymbol{\alpha}$ pour lesquels $\hat{\beta}^\alpha$ est similaire à $\hat{\beta}^{(1,1,\ldots,1)}$, le coefficient de régression associé à X dans le modèle de réponse complet, sont conservés. Il est donc implicitement supposé que le modèle de réponse complet produit un estimateur non biaisé de l'effet causal de X sur Y. Les modèles donnant une estimation similaire de l'effet causal produisent donc vraisemblablement eux aussi un estimateur non biaisé. Une procédure d'inférence tenant compte de l'incertitude quant au choix du modèle, qui est similaire à celle du BMA, est finalement effectuée sur le sous-ensemble de modèles conservés. L'algorithme de CDP complet comporte 10 étapes.

La méthode de CDP constitue certainement une amélioration par rapport au BMA pour la sélection de modèles dans un contexte d'inférence causale. Cependant, cette méthode a encore des limites importantes. D'abord, certaines étapes menant à l'identification des modèles sur lesquels la procédure d'inférence est appliquée au final sont subjectives. De plus, la méthode de CDP ne tient compte de l'incertitude associée au choix des modèles que dans sa dernière étape et néglige ainsi l'incertitude associée à la sélection effectuée aux étapes précédentes. La justification même de la méthode de CDP a ses limites. Entre autres, même s'il existe un sous-ensemble de $\boldsymbol{U}$

qui soit suffisant pour identifier l'effet causal, cela n'assure pas qu'il existe un **unique** sous-ensemble minimal suffisant, le « vrai » modèle α^*. Par exemple, dans la figure 2.1, les ensembles $\{U_1\}$ et $\{U_2\}$ sont tous les deux suffisants et minimaux. De plus, s'il existe bien un unique ensemble suffisant minimal α^*, cela n'assure pas non plus que tous les modèles α tels que $\alpha^* \subseteq \alpha$, identifient également l'effet causal. Notamment, la figure 2.2 donne un exemple où l'ensemble vide $\varnothing$ est minimal suffisant, mais l'ensemble $\{U_2\}$, où $\varnothing \subseteq \{U_2\}$, n'identifie pas l'effet causal, car U_2 est un collisionneur.

Figure 2.1: Exemple de DAG avec deux ensembles minimaux suffisants ($\{U_1\}$ et $\{U_2\}$).

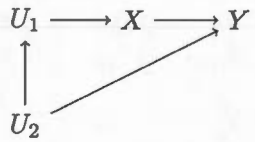
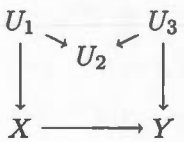


Figure 2.2: Exemple de DAG où il existe un α tel que $\alpha^* \subseteq \alpha$ qui n'identifie pas l'effet causal ($\varnothing$ est suffisant et $\varnothing \subset \{U_2\}$, mais $\{U_2\}$ n'identifie pas l'effet causal).



2.3 Les algorithmes BAC et TBAC

Tout comme la méthode de CDP, les algorithmes *Bayesian Adjustment for Confounding* (BAC) et *Two-Stage Bayesian Adjustment for Confounding* (TBAC) (Wang *et al.*, 2012a) visent à effectuer une estimation sans biais du paramètre β associé à l'exposition dans une régression de Y sur X en tenant compte des facteurs potentiellement confondants. Les méthodes BAC et TBAC répondent par contre à certaines limites de la méthode de CDP énoncées à la fin de la section 2.2. BAC et TBAC tiennent notamment compte de l'incertitude associée à l'ensemble du processus de sélection de modèles par le biais d'une approche complètement bayésienne. Tels que présentés par Wang *et al.* (2012a), BAC et TBAC ne s'appliquent toutefois que pour les situations où Y et X sont des variables continues.

La notation utilisée pour présenter BAC et TBAC est très similaire à celle qui a servi à présenter la méthode de CDP. Tout comme l'approche de CDP, BAC et TBAC

considèrent un modèle pour la variable d'exposition X et un modèle pour la variable réponse Y. Soit $\boldsymbol{\alpha^X} = (\alpha_1^X, ..., \alpha_M^X) \in \{0,1\}^M$, où $\alpha_m^X = 1$ si et seulement si la variable U_m est inclue dans le modèle d'exposition. Le vecteur $\boldsymbol{\alpha^Y} = (\alpha_1^Y, ..., \alpha_M^Y) \in \{0,1\}^M$ est défini similairement pour le modèle de réponse. Les modèles d'exposition et de réponse considérés sont :

$$\begin{aligned} E[X] &= \delta_0^{\boldsymbol{\alpha^X}} + \sum_{m=1}^{M} \alpha_m^X \delta_m^{\boldsymbol{\alpha^X}} U_m, \\ E[Y|X] &= \delta_0^{\boldsymbol{\alpha^Y}} + \beta^{\boldsymbol{\alpha^Y}} X + \sum_{m=1}^{M} \alpha_m^Y \delta_m^{\boldsymbol{\alpha^Y}} U_m. \end{aligned} \tag{2.4}$$

Les processus de sélection des variables incluses dans chacun des deux modèles sont reliés par l'entremise d'une loi *a priori* informative :

$$\begin{aligned} P(\boldsymbol{\alpha^X}, \boldsymbol{\alpha^Y}) &= \prod_{m=1}^{M} P(\alpha_m^X, \alpha_m^Y), \\ P(\alpha_m^X = 0, \alpha_m^Y = 0) &= P(\alpha_m^X = 0, \alpha_m^Y = 1) \\ &= P(\alpha_m^X = 1, \alpha_m^Y = 1) = \omega/(3\omega + 1), \\ P(\alpha_m^X = 1, \alpha_m^Y = 0) &= 1/(3\omega + 1), \end{aligned} \tag{2.5}$$

où $1 \leq \omega \leq \infty$ [1] est un hyperparamètre choisi par l'utilisateur. Plus précisément, cette loi *a priori* relie les deux sélections de variables en favorisant l'inclusion des variables associées à l'exposition dans le modèle de réponse. Cependant, lorsque $\omega = 1$, BAC et TBAC sont identiques au BMA avec une loi *a priori* uniforme, car la sélection des variables pour le modèle de réponse se fait alors indépendamment des variables

1. On considère ici l'ensemble augmenté des nombres réels, $\mathbb{R}^*$, voir par exemple Apostol (1974) section 1.20. L'ensemble $\mathbb{R}^*$ inclut les éléments $+\infty$ et $-\infty$.

sélectionnées dans le modèle d'exposition. À l'opposé, lorsque $\omega = \infty$, il est impossible qu'une variable soit sélectionnée dans le modèle d'exposition sans également être sélectionnée dans le modèle de réponse. La loi *a priori* de BAC et TBAC peut être justifiée par le fait qu'une variable confondante doit nécessairement être à la fois associée à l'exposition et à la réponse, conditionnellement à l'exposition (Pearl (2009), section 6.5.2). Ainsi, si une variable est associée à l'exposition, il est plus « probable » que cette variable soit une variable confondante et devrait ainsi être incluse dans le modèle de réponse avec une plus grande probabilité qu'une variable qui ne serait pas associée à l'exposition.

Le processus de sélection des variables diffère légèrement entre les algorithmes BAC et TBAC.

2.3.1 TBAC

Pour l'algorithme TBAC, la sélection des variables se divise en deux étapes. D'abord, la distribution *a posteriori* du modèle d'exposition est obtenue :

$$P(\boldsymbol{\alpha}^{\boldsymbol{X}}|X) \propto P(X|\boldsymbol{\alpha}^{\boldsymbol{X}})P(\boldsymbol{\alpha}^{\boldsymbol{X}}).$$

Généralement, une loi *a priori* non-informative est considérée pour le modèle d'exposition, c'est-à-dire $P(\boldsymbol{\alpha}^{\boldsymbol{X}}) = 1/2^M \; \forall \, \boldsymbol{\alpha}^{\boldsymbol{X}}$. Les inférences concernant l'effet causal de l'exposition peuvent être faites à partir de la distribution *a posteriori* de β, qui peut être simplifiée en effectuant quelques hypothèses :

$$P(\beta|Y,X) \propto \sum_{\boldsymbol{\alpha}^{\boldsymbol{X}}} \sum_{\boldsymbol{\alpha}^{\boldsymbol{Y}}} P(\beta|\boldsymbol{\alpha}^{\boldsymbol{Y}},Y,X)P(Y|\boldsymbol{\alpha}^{\boldsymbol{Y}},X)P(\boldsymbol{\alpha}^{\boldsymbol{Y}}|\boldsymbol{\alpha}^{\boldsymbol{X}})P(\boldsymbol{\alpha}^{\boldsymbol{X}}|X),$$

où $P(\boldsymbol{\alpha}^{\boldsymbol{Y}}|\boldsymbol{\alpha}^{\boldsymbol{X}})$ peut facilement être déduite de (2.5) :

$$P(\boldsymbol{\alpha}^{\boldsymbol{Y}}|\boldsymbol{\alpha}^{\boldsymbol{X}}) = \prod_{m=1}^{M} P(\alpha_m^Y|\alpha_m^X),$$
$$P(\alpha_m^Y = 0|\alpha_m^X = 0) = P(\alpha_m^Y = 1|\alpha_m^X = 0) = 1/2,$$

$$P(\alpha_m^Y = 0|\alpha_m^X = 1) = 1/(\omega + 1),$$
$$P(\alpha_m^Y = 1|\alpha_m^X = 1) = \omega/(\omega + 1).$$

2.3.2 BAC

Contrairement à l'algorithme TBAC, l'algorithme BAC effectue conjointement la sélection de variables pour le modèle d'exposition et pour le modèle de réponse. Cette version de l'algorithme permet ainsi une rétroaction entre les deux sélections de variables. Encore une fois, les inférences sont basées sur la distribution *a posteriori* de β. Cette distribution a été obtenue explicitement par Lefebvre *et al.* (2014a) :

$$P(\beta|Y, X) \propto \sum_{\boldsymbol{\alpha}^{\boldsymbol{X}}} \sum_{\boldsymbol{\alpha}^{\boldsymbol{Y}}} P(\beta|\boldsymbol{\alpha}^{\boldsymbol{Y}}, Y, X) P(Y|\boldsymbol{\alpha}^{\boldsymbol{Y}}, X) P(X|\boldsymbol{\alpha}^{\boldsymbol{X}}) P(\boldsymbol{\alpha}^{\boldsymbol{Y}}, \boldsymbol{\alpha}^{\boldsymbol{X}}).$$

Les algorithmes BAC et TBAC peuvent être perçus comme une amélioration à la méthode de CDP pour effectuer de la sélection de modèles pour l'inférence causale, puisque l'incertitude par rapport au choix du modèle est pleinement incorporée dans ces procédures. De plus, l'application de ces approches est beaucoup moins subjective que l'application de la méthode de CDP.

Certaines limites demeurent néanmoins. Notamment, Wang *et al.* (2012a) ne fournissent pas de conditions sous lesquelles les inférences basées sur BAC et TBAC sont non biaisées. Les auteurs ne fournissent également pas de guide pour le choix

de l'hyperparamètre ω. De plus, certaines simulations suggèrent que BAC et TBAC sont peu performants en terme de la variance et de l'erreur quadratique moyenne de l'estimateur de l'effet causal en comparaison à celui obtenu par l'utilisation d'un modèle de réponse complet (Wang *et al.*, 2012a; Vansteelandt, 2012).

2.4 L'approche de VanderWeele & Shpitser (2011)

VanderWeele & Shpitser (2011) (VS) considèrent un contexte un peu différent de celui que nous étudions, mais abordent au passage la problématique de la sélection d'un ensemble suffisant pour l'estimation de l'effet causal de X sur Y sur la base des données.

Le principal objectif poursuivi par VS est de proposer une approche pratique pour l'identification de variables potentiellement confondantes sur la base des connaissances du domaine d'application. VS supposent que $\boldsymbol{U}$ est un ensemble de variables pré-exposition, ou, du moins, n'incluant pas de descendants de l'exposition. Ils démontrent que si $\boldsymbol{U}$ contient un sous-ensemble suffisant pour estimer l'effet causal, alors le sous-ensemble $\boldsymbol{Z} \subseteq \boldsymbol{U}$ formé de toutes les causes de l'exposition et de toutes les causes de la réponse dans $\boldsymbol{U}$ est suffisant. Afin d'utiliser cette approche, l'utilisateur doit être en mesure, à partir de ses connaissances du domaine, de déterminer si chaque variable dans $\boldsymbol{U}$ est une cause de l'exposition, une cause de la réponse, ni l'un ni l'autre ou les deux.

Deux algorithmes permettant de sélectionner un sous-ensemble de $\boldsymbol{Z}$ également suffisant sont aussi proposés par VS. Le premier algorithme est de type descendant (*backward*) et permet de retirer itérativement des variables qui ne sont pas associées avec la réponse. L'algorithme débute avec un ensemble $\boldsymbol{Z}^{(0)} = \boldsymbol{Z} = \{Z_1, Z_2, ..., Z_K\}$ suffisant. L'ensemble $\boldsymbol{Z}^{(1)} = \boldsymbol{Z}^{(0)} \backslash Z_{k_1}$ est également suffisant si $Y \perp\!\!\!\perp Z_{k_1} | X, \boldsymbol{Z}^{(1)}$. Cette réduction est répétée jusqu'à ce qu'il devienne impossible de retirer des variables tout en respectant la condition d'indépendance conditionnelle. Autrement dit, à l'étape j, $\boldsymbol{Z}^{(j)} = \boldsymbol{Z}^{(j-1)} \backslash Z_{k_j}$ est suffisant si $Y \perp\!\!\!\perp Z_{k_j} | X, \boldsymbol{Z}^{(j)}$, où Z_{k_j} est la variable retirée à

l'étape j, $k_j \neq k_{j'}\ \forall\ j \neq j'$. En pratique, la condition d'indépendance pourrait être vérifiée à l'aide de tests statistiques.

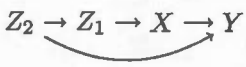
Le second algorithme est de type ascendant (*forward*) et permet d'ajouter itérativement des variables pré-exposition associées avec la réponse jusqu'à l'obtention d'un ensemble suffisant. Cet algorithme suppose une fois de plus qu'un ensemble $\boldsymbol{Z} = \{Z_1, Z_2, ..., Z_K\}$ suffisant a été identifié. Le point de départ de l'algorithme est $\boldsymbol{Z}^{(0)} = \varnothing$. L'ensemble $\boldsymbol{Z}^{(1)} = \boldsymbol{Z}^{(0)} \cup Z_{k_1}$ est construit de sorte que $Y \not\perp\!\!\!\perp Z_{k_1} | X, \boldsymbol{Z}^{(0)}$. À l'étape j, $\boldsymbol{Z}^{(j)} = \boldsymbol{Z}^{(j-1)} \cup Z_{k_j}$ est construit de sorte que $Y \not\perp\!\!\!\perp Z_{k_j} | X, \boldsymbol{Z}^{(j-1)}$, où Z_{k_j} est la variable ajoutée à l'étape j. Cet ajout de variables est poursuivi jusqu'à ce qu'il devienne impossible d'ajouter de nouvelles variables qui ne sont pas (conditionnellement) indépendantes de la réponse. L'ensemble obtenu au final est suffisant pour identifier l'effet causal de X sur Y. Encore une fois, un tel algorithme pourrait être implanté en vérifiant la condition d'indépendance à l'aide de tests statistiques.

Les algorithmes proposés par VS bénéficient d'une base théorique forte contrairement aux approches précédentes dont la justification était plutôt heuristique. Cependant, aucune implantation pratique de ces algorithmes n'est suggérée. D'ailleurs, les propriétés pour des échantillons finis de telles implantations restent à déterminer. En pratique, il est facilement vérifiable que l'ensemble suffisant final obtenu va non seulement dépendre de l'algorithme choisi, mais également de l'ordre dans lequel les variables sont sélectionnées pour être retirées/ajoutées.

Pour illustrer ceci, nous utilisons le DAG présenté à la figure 2.3. Considérons d'abord l'algorithme ascendant (on a donc $\boldsymbol{Z}^{(0)} = \varnothing$). Une première possibilité d'exécution de cet algorithme est de commencer par ajouter Z_1, car $Y \not\perp\!\!\!\perp Z_1 | X$, de sorte que $\boldsymbol{Z}^{(1)} = Z_1$. Ensuite, Z_2 est ajouté, car $Y \not\perp\!\!\!\perp Z_2 | X, Z_1$. L'autre possibilité d'exécution de ce même algorithme consiste à ajouter premièrement Z_2. Cependant, Z_1 n'est ensuite pas ajouté, car $Y \perp\!\!\!\perp Z_1 | X, Z_2$. L'ensemble final identifié, $\{Z_1, Z_2\}$ dans le premier cas et $\{Z_2\}$ dans le deuxième cas, n'est ainsi pas le même en fonction de l'ordre selon lequel les variables sont considérées pour être ajoutées. Notons que si l'algorithme descendant est exécuté,

la seule possibilité à la première étape est de retirer Z_1. De plus, Z_2 n'est ensuite pas retirée, de sorte que l'ensemble final obtenu est $\{Z_2\}$.

Figure 2.3: Exemple de DAG pour les algorithmes de VS.



2.5 Les méthodes de Persson *et al.* (2013)

Les méthodes proposées par Persson *et al.* (2013) (PHWD) sont basées sur des algorithmes de sélection de variables développés par de Luna *et al.* (2011). Tout comme les algorithmes de VS, leur base théorique est forte étant donné qu'il a été montré que les algorithmes proposés par de Luna *et al.* (2011) identifient, sous certaines conditions mentionnées plus bas, des sous-ensembles de $\boldsymbol{U}$ suffisants pour identifier l'effet causal de X sur Y.

Contrairement au BMA, à la méthode de CDP, à BAC et à TBAC, les méthodes de PHWD sont non paramétriques et se limitent à la situation où X est binaire $(0/1)$. Les observations pour lesquelles $X = 0$ sont appariées à des observations pour lesquelles $X = 1$ en fonction des valeurs d'un sous-ensemble $\boldsymbol{Z} \subseteq \boldsymbol{U}$ suffisant pour estimer l'effet causal (voir par exemple Stuart (2010) concernant les méthodes d'appariement en inférence causale). Dans ce contexte non paramétrique, la sélection d'un sous-ensemble $\boldsymbol{U}$ pour estimer l'effet causal est encore plus importante que dans un contexte paramétrique. En effet, bien que les approches d'appariement ne nécessitent pas d'hypothèses distributionnelles, c'est-à-dire sur la distribution conjointe des variables, ou sur la forme fonctionnelle des liens entre les variables, elles peuvent produire des estimateurs biaisés et dont la variance est très élevée si l'ensemble $\boldsymbol{Z}$ est grand (Persson *et al.*, 2013).

Les méthodes de PHWD sont imbriquées dans le paradigme contrefactuel à l'inférence causale. Elles supposent le respect des hypothèses de cohérence, de positivité ainsi

que de l'hypothèse faible d'ignorabilité conditionnelle à $\boldsymbol{U}$ (voir section 1.1). Sous ces hypothèses, de Luna *et al.* (2011) montrent que les sous-ensembles de $\boldsymbol{U}$ suivants existent et sont uniques :

1. $\boldsymbol{U}_X$ est l'ensemble $\boldsymbol{U}_X \subseteq \boldsymbol{U}$ de cardinalité minimale tel que $P(X|\boldsymbol{U}) = P(X|\boldsymbol{U}_X)$;
2. Pour $x = 0, 1$, $\boldsymbol{Q}_x$ est l'ensemble $\boldsymbol{Q}_x \subseteq \boldsymbol{U}_X$ de cardinalité minimale tel que $P(Y(x)|\boldsymbol{U}_X) = P(Y(x)|\boldsymbol{Q}_x)$;
3. Pour $x = 0, 1$, $\boldsymbol{U}_x$ est l'ensemble $\boldsymbol{U}_x \subseteq \boldsymbol{U}$ de cardinalité minimale tel que $P(Y(x)|\boldsymbol{U}) = P(Y(x)|\boldsymbol{U}_x)$;
4. Pour $x = 0, 1$, $\boldsymbol{Z}_x$ est l'ensemble $\boldsymbol{Z}_x \subseteq \boldsymbol{U}_x$ de cardinalité minimale tel que $P(X|\boldsymbol{U}_x) = P(X|\boldsymbol{Z}_x)$.

De plus, les ensembles $\boldsymbol{U}_X$, $\boldsymbol{Q} = \boldsymbol{Q}_0 \cup \boldsymbol{Q}_1$, $\boldsymbol{U}_Y = \boldsymbol{U}_0 \cup \boldsymbol{U}_1$ et $\boldsymbol{Z} = \boldsymbol{Z}_0 \cup \boldsymbol{Z}_1$ sont tous suffisants pour identifier l'effet causal de X sur Y, c'est-à-dire que l'hypothèse faible d'ignorabilité (1.1) est respectée conditionnellement à chacun de ces ensembles.

Afin d'identifier ces sous-ensembles, de Luna *et al.* (2011) proposent les algorithmes suivants :

Algorithme A - Sélectionner $\boldsymbol{Q}_0$ et $\boldsymbol{Q}_1$.

Étape 1. Choisir $\boldsymbol{U}_X$ tel que $(X \perp\!\!\!\perp \boldsymbol{U} \backslash \boldsymbol{U}_X | \boldsymbol{U}_X)$.

Étape 2. Pour $x = 0, 1$, choisir $\boldsymbol{Q}_x \subseteq \boldsymbol{U}_X$ tel que $(Y \perp\!\!\!\perp \boldsymbol{U}_X \backslash \boldsymbol{Q}_x | \boldsymbol{Q}_x, X = x)$.

Algorithme B - Sélectionner $\boldsymbol{Z}_0$ et $\boldsymbol{Z}_1$.

Étape 1. Pour $x = 0, 1$, choisir $\boldsymbol{U}_x$ tel que $(Y \perp\!\!\!\perp \boldsymbol{U} \backslash \boldsymbol{U}_x | \boldsymbol{U}_x, X = x)$.

Étape 2. Choisir $\boldsymbol{Z}_x \subseteq \boldsymbol{U}_x$ tel que $(X \perp\!\!\!\perp \boldsymbol{U}_x \backslash \boldsymbol{Z}_x | \boldsymbol{Z}_x)$.

Les méthodes de PHWD consistent à implanter ces algorithmes à l'aide d'une procédure descendante de sélection de variables. Pour chacun des deux algorithmes, la procédure débute avec $\boldsymbol{U}$, puis des variables sont retirées de $\boldsymbol{U}$ une à une si elles satisfont les critères d'indépendances conditionnelles énoncés dans l'algorithme considéré (algorithme A ou B). Ces indépendances conditionnelles sont testées à l'aide de tests statistiques non paramétriques. Lorsque la procédure se termine, des

sous-ensembles $\hat{\boldsymbol{U}}_X$ et $\hat{\boldsymbol{Q}} = \hat{\boldsymbol{Q}}_0 \cup \hat{\boldsymbol{Q}}_1$ sont identifiés pour l'algorithme A et des sous-ensembles $\hat{\boldsymbol{U}}_Y = \hat{\boldsymbol{U}}_0 \cup \hat{\boldsymbol{U}}_1$, et $\hat{\boldsymbol{Z}} = \hat{\boldsymbol{Z}}_0 \cup \hat{\boldsymbol{Z}}_1$ sont identifiés pour l'algorithme B. L'appariement des observations peut alors être effectué en fonction d'un de ces quatre ensembles $(\hat{\boldsymbol{U}}_X, \hat{\boldsymbol{Q}}, \hat{\boldsymbol{U}}_Y, \hat{\boldsymbol{Z}})$.

La principale qualité des méthodes proposées par PHWD est leur forte base théorique. Par ailleurs, l'approche non paramétrique peut être vue comme un avantage, en raison de sa flexibilité, mais également comme un inconvénient, puisque l'estimateur peut être biaisé ou instable lorsque l'appariement est effectué à partir d'un ensemble trop important de variables. Par ailleurs, bien que l'incertitude associée à l'appariement des observations est prise en compte dans les méthodes de PHWD, l'incertitude reliée à l'identification des sous-ensembles suffisants semble être négligée, ce qui pourrait conduire à une sous-estimation de l'erreur type de l'effet causal et à des intervalles de confiance trop étroits. Finalement, rien n'assure que les sous-ensembles identifiés sont optimaux pour réduire l'erreur quadratique moyenne de l'effet causal estimé ou même que la réduction de dimension est toujours souhaitable. Entre autres, les simulations présentées par PHWD suggèrent que l'estimation à partir de $\boldsymbol{Z}$ est souvent moins efficace que celle à partir de $\boldsymbol{U}_X$, bien que $\boldsymbol{Z} \subseteq \boldsymbol{U}_X$.

2.6 Autres méthodes

Il existe plusieurs autres méthodes de sélection de modèles guidées par les données pour l'inférence causale. Entre autres, une méthode similaire à BAC a été suggérée dans un cadre d'estimation de l'effet causal à l'aide de scores de propension (Zigler & Dominici, 2014). McCandless *et al.* (2009) proposent aussi une approche bayésienne de sélection de modèles pour l'inférence causale basée sur les scores de propension. Étonnamment, cette dernière approche semble moins efficace que celle consistant à négliger l'incertitude associée à l'estimation du score de propension. Une extension de l'algorithme BAC a été proposée pour des variables d'exposition binaires (Lefebvre *et al.*, 2014b). Une approche de modèle moyen doublement robuste a été introduite par

Cefalu *et al.* (2013). Tout comme la méthode de CDP ainsi que les algorithmes BAC et TBAC, cette approche effectue une modélisation bayésienne du modèle d'exposition et du modèle de réponse. Elle permet néanmoins une estimation non biaisée d'autant qu'un ou l'autre de ces deux modèles est correctement spécifié, alors que les méthodes de CDP, BAC et TBAC nécessitent que le modèle de réponse soit correctement spécifié, entre autres en ce qui a trait aux formes fonctionnelles des liens entre les variables $\boldsymbol{U}$ et Y.

Certaines méthodes de sélection de modèles pour l'inférence causale utilisent des techniques d'apprentissage machine (*machine learning*). Par exemple, Hill (2011) propose une approche non paramétrique basée sur les arbres de régression additifs bayésiens (*Bayesian Additive Regression Trees*, BART). Cette approche est simple d'utilisation, flexible et des intervalles de confiance (crédibilité) tenant compte du processus de sélection de modèles peuvent facilement être obtenus. Toutefois, puisque cette méthode ne se concentre que sur la modélisation de la variable réponse, elle est potentiellement sensible à des problèmes similaires à ceux du BMA : des variables confondantes faiblement associées à la réponse, mais fortement associées à l'exposition risquent de ne pas être identifiées par l'algorithme.

Le maximum de vraisemblance ciblé (*Targeted Maximum Likelihood*, TMLE) est une seconde approche qui est souvent implantée à l'aide de techniques d'apprentissage machine (van der Laan & Rubin, 2006; Gruber & van der Laan, 2009). Le maximum de vraisemblance ciblé vise l'estimation du maximum de vraisemblance du paramètre causal dans un modèle et considère les autres paramètres du modèle comme étant des paramètres de nuisance. L'estimation du paramètre ciblé est effectuée de façon à réduire le biais de l'estimateur de l'effet causal, au coût potentiel d'une augmentation de sa variance et d'une augmentation du biais et de la variance des estimateurs des paramètres de nuisance.

Plusieurs autres méthodes plus ou moins similaires à celles déjà présentées existent, mais en effectuer une revue exhaustive dépasse les objectifs de cette thèse.

CHAPITRE III

DÉVELOPPEMENTS CONCERNANT L'ALGORITHME BAC

À la section 2.3, nous avons brièvement présenté les algorithmes BAC et TBAC. Tel que mentionné, ces algorithmes sont basés sur une heuristique intéressante et ont plusieurs avantages. Par exemple, ils permettent de tenir compte de façon appropriée de l'incertitude associée au choix des modèles dans les inférences sur l'effet causal. Toutefois, nous avons également souligné quelques limites de ces algorithmes, dont l'absence de conditions théoriques sous lesquelles ils produisent des estimateurs non biaisés, la possible inefficacité des algorithmes en terme de la variance de l'estimateur de l'effet causal et la difficulté de choisir l'hyperparamètre ω. Étant donné les qualités de ces algorithmes, Lefebvre *et al.* (2014a) ont cru bon investiguer davantage les propriétés de l'algorithme BAC. Les résultats de cet article sont résumés dans ce chapitre, en soulignant au passage les éléments auxquels j'ai le plus contribué.

3.1 Loi *a posteriori* marginale de β

Afin d'améliorer la compréhension de l'effet de la loi *a priori* de l'algorithme BAC, Lefebvre *et al.* (2014a) proposent d'étudier la loi *a posteriori* marginale de β au lieu de sa loi *a posteriori* conditionnelle, tel qu'effectué par Wang *et al.* (2012a).

La spécification des modèles de réponse et d'exposition (2.4) implique les vraisemblances respectives suivantes pour ces variables :

$$P(Y|X,\boldsymbol{\theta_Y},\boldsymbol{\alpha^Y}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_Y^2}} \exp\left(-\frac{1}{2\sigma_Y^2}\left[Y_i - \left(\delta_0^{\boldsymbol{\alpha^Y}} + \beta^{\boldsymbol{\alpha^Y}} X_i + \sum_{m=1}^{M} \alpha_m^Y \delta_m^Y U_{i,m}\right)\right]^2\right),$$

$$P(X|\boldsymbol{\theta_X},\boldsymbol{\alpha^X}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp\left(-\frac{1}{2\sigma_X^2}\left[X_i - \left(\delta_0^{\boldsymbol{\alpha^X}} + \sum_{m=1}^{M} \alpha_m^X \delta_m^X U_{i,m}\right)\right]^2\right), \quad (3.1)$$

où $\boldsymbol{\theta_Y} = (\beta^{\boldsymbol{\alpha^Y}}, \delta_0^{\boldsymbol{\alpha^Y}}, \delta_1^{\boldsymbol{\alpha^Y}}, ..., \delta_M^{\boldsymbol{\alpha^Y}}, \sigma_Y^2)$, $\boldsymbol{\theta_X} = (\delta_0^{\boldsymbol{\alpha^X}}, \delta_1^{\boldsymbol{\alpha^X}}, ..., \delta_M^{\boldsymbol{\alpha^X}}, \sigma_X^2)$ et $i = 1, ..., n$ désigne les observations.

Les hypothèses ci-bas, qui sont implicitement effectuées par Wang *et al.* (2012a), sont utilisées :

$$P(\boldsymbol{\theta_Y},\boldsymbol{\theta_X}|\boldsymbol{\alpha^Y},\boldsymbol{\alpha^X}) = P(\boldsymbol{\theta_Y}|\boldsymbol{\alpha^Y})P(\boldsymbol{\theta_X}|\boldsymbol{\alpha^X});$$
$$P(Y|X,\boldsymbol{\theta_Y},\boldsymbol{\theta_X},\boldsymbol{\alpha^Y},\boldsymbol{\alpha^X}) = P(Y|X,\boldsymbol{\theta_Y},\boldsymbol{\alpha^Y});$$
$$P(X|\boldsymbol{\theta_Y},\boldsymbol{\theta_X},\boldsymbol{\alpha^Y},\boldsymbol{\alpha^X}) = P(X|\boldsymbol{\theta_X},\boldsymbol{\alpha^X}).$$

À l'aide de ces hypothèses et des vraisemblances, la loi *a posteriori* marginale de β s'exprime comme :

$$P(\beta^{\boldsymbol{\alpha}^{\boldsymbol{Y}}}|Y,X) = \sum_{\boldsymbol{\alpha}^{\boldsymbol{Y}}} P(\beta^{\boldsymbol{\alpha}^{\boldsymbol{Y}}}|\boldsymbol{\alpha}^{\boldsymbol{Y}},Y,X)P(\boldsymbol{\alpha}^{\boldsymbol{Y}}|Y,X), \text{ où} \tag{3.2}$$

$$P(\boldsymbol{\alpha}^{\boldsymbol{Y}}|Y,X) \propto P(Y|X,\boldsymbol{\alpha}^{\boldsymbol{Y}}) \underbrace{\sum_{\boldsymbol{\alpha}^{\boldsymbol{X}}} P(X|\boldsymbol{\alpha}^{\boldsymbol{X}})P(\boldsymbol{\alpha}^{\boldsymbol{Y}},\boldsymbol{\alpha}^{\boldsymbol{X}})}_{:=\text{Facteur multiplicatif } MF(\boldsymbol{\alpha}^{\boldsymbol{Y}})}.$$

Tel qu'espéré, l'expression (3.2) permet de plus facilement comprendre l'effet de l'hyperparamètre ω sur le fonctionnement de BAC. En effet, cette écriture met en évidence que c'est à travers le facteur multiplicatif $MF(\boldsymbol{\alpha}^{\boldsymbol{Y}})$ que la loi *a priori* de BAC influence la loi *a posteriori*. Par exemple, lorsque $\omega = 1$, $MF(\boldsymbol{\alpha}^{\boldsymbol{Y}}) = \sum_{\boldsymbol{\alpha}^{\boldsymbol{X}}} P(X|\boldsymbol{\alpha}^{\boldsymbol{X}})/4$. Il s'agit donc d'une constante par rapport à $\boldsymbol{\alpha}^{\boldsymbol{Y}}$. On obtient ainsi que $P(\boldsymbol{\alpha}^{\boldsymbol{Y}}|Y,X) \propto P(Y|X,\boldsymbol{\alpha}^{\boldsymbol{Y}})$, tel que dans le BMA avec la loi *a priori* uniforme $P(\alpha_m^Y) = 1/2 \ \forall \ m$. Lorsque $\omega = \infty$, le facteur multiplicatif est proportionnel à la somme de la vraisemblance marginale des modèles d'exposition imbriqués dans le modèle de réponse $\boldsymbol{\alpha}^{\boldsymbol{Y}}$. En effet

$$MF(\boldsymbol{\alpha}^{\boldsymbol{Y}}) = \sum_{\{\boldsymbol{\alpha}^{\boldsymbol{X}}|\boldsymbol{\alpha}^{\boldsymbol{X}} \subseteq \boldsymbol{\alpha}^{\boldsymbol{Y}}\}} P(X|\boldsymbol{\alpha}^{\boldsymbol{X}}), \tag{3.3}$$

où $\{\boldsymbol{\alpha}^{\boldsymbol{X}}|\boldsymbol{\alpha}^{\boldsymbol{X}} \subseteq \boldsymbol{\alpha}^{\boldsymbol{Y}}\}$ désigne l'ensemble des $\boldsymbol{\alpha}^{\boldsymbol{X}}$ tels que $\alpha_m^X \leq \alpha_m^Y \ \forall \ m$.

Par ailleurs, à l'aide de l'expression (3.2), j'ai développé un *package* R, `BACprior`, permettant d'approximer les inférences *a posteriori* produites par l'algorithme BAC sans recourir à des méthodes de Monte-Carlo par chaînes de Markov (Talbot *et al.*, 2014). Cette implantation permet ainsi d'obtenir facilement et rapidement des inférences *a posteriori*. Le *package* `BACprior` est décrit davantage à la section 3.4.

3.2 Justification de BAC par le paradigme graphique causal

Ma principale contribution aux travaux de Lefebvre *et al.* (2014a) a été apportée dans l'élaboration de la justification causale de l'algorithme BAC à partir du paradigme graphique à l'inférence causale. Cette justification se résume comme suit.

Soient G, un DAG causal compatible avec la loi conjointe des variables $\{Y, X, \boldsymbol{U}\}$, et $\boldsymbol{PA_X}$, l'ensemble des parents (des causes directes) de X dans G. Supposons que $\boldsymbol{PA_X} \subseteq \boldsymbol{U}$ et que $\boldsymbol{U}$ n'inclut aucun descendant de X (voir la condition (ii) du critère porte-arrière (1.9)). On a que X et $\boldsymbol{U} \backslash \boldsymbol{PA_X}$ sont d-séparés par $\boldsymbol{PA_X}$ et donc que $X \perp\!\!\!\perp \boldsymbol{U} | \boldsymbol{PA_X}$. De plus, en supposant les vraisemblances (3.1), les coefficients de régression associés à $\boldsymbol{U} \backslash \boldsymbol{PA_X}$ dans le modèle d'exposition sont tous nuls. Sous de telles conditions, on peut s'attendre à ce que la vraisemblance marginale du modèle d'exposition ne contenant que les variables $\boldsymbol{PA_X}$ (le modèle d'exposition structurel) domine la vraisemblance marginale des autres modèles d'exposition lorsque la taille d'échantillon, n, est assez grande (Wasserman, 2000). Par ailleurs, lorsque $\omega = \infty$, étant donné la forme du facteur multiplicatif (3.3), on peut également s'attendre à ce que seuls les modèles de réponse emboîtant le modèle d'exposition structurel reçoivent un fort poids *a posteriori*. Or, lorsqu'un modèle de réponse contient toutes les causes directes de l'exposition ($\boldsymbol{PA_X}$), son estimateur de l'effet causal de l'exposition est non biaisé. En effet, le théorème 3.2.5 de (Pearl, 2009) assure que l'ajustement pour $\boldsymbol{PA_X}$ est une condition suffisante pour l'estimation de l'effet causal de l'exposition sur la réponse. L'ajustement pour d'autres covariables qui ne sont pas des descendants de l'exposition, en plus de $\boldsymbol{PA_X}$, produit également un estimateur non biaisé de l'effet causal.

Bien que ce fonctionnement de BAC avec $\omega = \infty$ favorise une estimation sans biais de l'effet causal, il contribue possiblement aussi à produire un estimateur dont la variance est élevée. Pour illustrer ce fait, Lefebvre *et al.* (2014a) considèrent la variance estimée de l'estimateur de l'effet de l'exposition dans un modèle de régression linéaire ordinaire fréquentiste :

$$\widehat{Var}(\hat{\beta}) = \frac{\left[\frac{(1-R_Y^2)\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-p-1}\right]}{(1-R_X^2)\sum_{i=1}^{n}(x_i-\bar{x})^2}, \tag{3.4}$$

où R_Y^2 désigne le coefficient de détermination du modèle de réponse, R_X^2 le coefficient de détermination du modèle de régression linéaire de la variable d'exposition sur les variables potentiellement confondantes inclues dans le modèle de réponse et p est le nombre de covariables dans le modèle (O'brien, 2007). L'équation (3.4) montre que $\widehat{Var}(\hat{\beta})$ est inversement proportionnelle à R_X^2. Ainsi, inclure de forts prédicteurs de l'exposition (par exemple, les causes directes de l'exposition) dans le modèle de réponse contribue à accroître $\widehat{Var}(\hat{\beta})$.

3.3 Choix de ω

Le choix $\omega = \infty$ apparaît dès lors conservateur puisqu'il favorise la production d'un estimateur sans biais, au coût potentiel d'une augmentation de la variance. En pratique, il est parfois plus intéressant de choisir une valeur de ω qui minimise l'erreur quadratique moyenne. Lefebvre *et al.* (2014a) suggèrent deux approches par validation-croisée et une par bootstrap visant cet objectif. L'implantation de ces trois méthodes par rééchantillonnage est facilitée par le *package* R `BACprior` (Talbot *et al.*, 2014). Les études de simulations que nous avons réalisées suggèrent que la performance de ces trois approches est très variable et dépend fortement, entre autres, du processus générant les données.

3.4 Le *package* BACprior

Le *package* `BACprior` comporte trois fonctions. La fonction `BACprior.lm` permet d'obtenir des inférences *a posteriori* approximatives rapidement pour plusieurs valeurs de ω. Les fonctions `BACprior.CV` et `BACprior.boot` implantent respectivement les approches par validation-croisée et par bootstrap proposées par Lefebvre *et al.* (2014a)

et visant à guider l'utilisateur dans un choix de valeur ω qui minimiserait l'erreur quadratique moyenne.

3.4.1 La fonction `BACprior.lm`

À l'aide de la fonction `regsubset` du *package* `leaps` (Lumley, 2009), la fonction `BACprior.lm` identifie d'abord pour chaque taille de modèles, de 1 à M, les meilleurs modèles d'exposition (ceux ayant la plus grande vraisemblance). Le nombre exact de modèles de chaque taille retenus peut être décidé par l'utilisateur. Plus ce nombre est élevé, meilleure est l'approximation effectuée, puisque l'algorithme BAC considère théoriquement l'ensemble de tous les modèles. Cependant, l'utilisation d'un plus grand nombre de modèles augmente le temps de calcul. La vraisemblance marginale de chacun des modèles retenus est par la suite approximée à l'aide de $\exp(-0.5BIC)$. (Clyde, 2003). De la même façon, les meilleurs modèles de réponse de chaque taille sont identifiés et leur vraisemblance marginale est calculée.

Avec ces éléments, $MF(\boldsymbol{\alpha}^{\boldsymbol{Y}})$ et la loi *a posteriori* du modèle de réponse $P(\boldsymbol{\alpha}^{\boldsymbol{Y}}|Y,X)$ peuvent facilement être calculés pour chacune des valeurs de ω choisies par l'utilisateur. Les modèles disposant d'une probabilité *a posteriori* non négligeable, selon un seuil choisi par l'utilisateur, sont conservés pour les calculs subséquents. L'espérance et la variance *a posteriori* de β sont finalement approximés à l'aide de formules suggérées par Hoeting *et al.* (1999) dans le cadre du BMA :

$$E[\beta|Y,X] \approx \sum_{\boldsymbol{\alpha}^{\boldsymbol{Y}}} \hat{\beta}^{\boldsymbol{\alpha}^{\boldsymbol{Y}}} P(\boldsymbol{\alpha}^{\boldsymbol{Y}}|Y,X)$$

$$Var[\beta|Y,X] \approx \left\{ \sum_{\boldsymbol{\alpha}^{\boldsymbol{Y}}} \left[\widehat{Var}(\hat{\beta}^{\boldsymbol{\alpha}^{\boldsymbol{Y}}}) + (\hat{\beta}^{\boldsymbol{\alpha}^{\boldsymbol{Y}}})^2 \right] P(\boldsymbol{\alpha}^{\boldsymbol{Y}}|Y,X) \right\} - E[\beta|Y,X]^2,$$

où $\hat{\beta}^{\boldsymbol{\alpha}^{\boldsymbol{Y}}}$ est l'estimateur du maximum de vraisemblance de $\beta^{\boldsymbol{\alpha}^{\boldsymbol{Y}}}$ et $\widehat{Var}(\hat{\beta}^{\boldsymbol{\alpha}^{\boldsymbol{Y}}})$ est sa variance estimée.

En sortie, la fonction `BACprior.lm` fournit une approximation de l'espérance et de l'écart-type *a posteriori* de β pour chaque valeur de ω sélectionnée par l'utilisateur. Optionnellement, la fonction fournit également les modèles de réponses considérés dans les calculs finaux avec leur probabilité *a posteriori*.

3.4.2 Les fonctions `BACprior.CV` et `BACprior.boot`

La fonction `BACprior.CV` divise aléatoirement l'échantillon en deux parties égales V fois. À chaque fois, l'effet de l'exposition est estimé avec $\omega = \infty$ sur une partie de l'échantillon. Sur l'autre partie, l'effet de l'exposition est estimé avec chacune des valeurs de ω sélectionnées par l'utilisateur. Un critère à minimiser choisi par l'utilisateur est alors calculé. Ces critères, inspirés de celui proposé par Brookhart & van der Laan (2006), sont détaillés dans Lefebvre *et al.* (2014a).

Le fonctionnement de `BACprior.boot` consiste à d'abord former B échantillons bootstrap à partir de l'échantillon original. L'erreur quadratique moyenne est ensuite estimée en considérant que la véritable valeur de l'effet de l'exposition correspond à l'effet de l'exposition estimé sur l'échantillon original avec $\omega = \infty$.

3.5 Discussion sur l'algorithme BAC

Les travaux de Lefebvre *et al.* (2014a) ont permis de mettre en évidence que l'algorithme BAC peut être justifié plus formellement à l'aide du paradigme graphique à l'inférence causale. Ils confirment également l'inquiétude par rapport à la capacité de BAC d'estimer l'effet causal de l'exposition avec précision accrue. L'élaboration d'une procédure performante permettant de choisir ω afin de minimiser l'erreur quadratique moyenne semble par ailleurs très complexe.

Au lieu de poursuivre dans cette voie, le prochain chapitre est formé d'un article scientifique en langue anglaise et décrivant un nouvel algorithme de sélection de modèles pour l'inférence causale que j'ai développé, *Bayesian Causal Effect Estimation*.

Cet algorithme vise, entre autres, à produire un estimateur de l'effet causal de l'exposition plus efficace que celui obtenu avec BAC ou TBAC avec $\omega = \infty$.

CHAPITRE IV

PREMIER ARTICLE : THE BAYESIAN CAUSAL EFFECT ESTIMATION ALGORITHM

Denis Talbot, Geneviève Lefebvre, Juli Atherton

Abstract: Estimating causal exposure effects in observational studies ideally requires the analyst to have a vast knowledge of the domain of application. Investigators often bypass difficulties related to the identification and selection of confounders through the use of fully adjusted outcome regression models. However, since such models likely contain more covariates than required, the variance of the regression coefficient for exposure may be unnecessarily large. Instead of using a fully adjusted model, model selection can be attempted. Most classical statistical model selection approaches, such as Bayesian model averaging, do not readily address causal effect estimation. We present a new model averaged approach to causal inference, Bayesian causal effect estimation (BCEE), which is motivated by the graphical framework for causal inference. BCEE aims to unbiasedly estimate the causal effect of a continuous exposure on a continuous outcome while being more efficient than a fully adjusted approach.

Keywords: model selection, causal diagrams, exposure effect estimation, variance reduction.

4.1 Introduction

Estimating causal exposure effects in observational studies demands a vast knowledge of the domain of application. For instance, to estimate the causal effect of an exposure on an outcome, the graphical framework to causality usually involves postulating a causal graph to identify an appropriate set of confounding variables (Pearl, 2009). Specifying such a graph can be difficult, especially in subject areas where prior knowledge is scarce or limited.

Investigators often bypass difficulties related to the identification and selection of confounders through the use of fully adjusted outcome regression models. Such models express the outcome variable as a function of the exposure variable and all available potential confounding variables. A fully adjusted outcome regression model is commonly assumed to yield an unbiased estimator of the true effect of the exposure. However, since such models likely contain more covariates than required, the variance of the regression coefficient for exposure may be unnecessarily large. Instead of using a fully adjusted model, model selection can be attempted.

Most classical statistical model selection approaches do not readily address causal effect estimation. One such approach is Bayesian model averaging (BMA) (Raftery *et al.*, 1997; Hoeting *et al.*, 1999). BMA averages quantities of interest (e.g. a regression coefficient or the value of a future observation) over all possible models under consideration: in the average, each estimate is weighted by the posterior probability attributed to the corresponding model. When the goal is prediction, BMA accounts for the uncertainty associated with model choice and produces confidence intervals that have adequate coverage probabilities (Raftery & Zheng, 2003). Unfortunately, BMA can perform poorly when used to estimate a causal effect of exposure (Crainiceanu *et al.*, 2008; Wang *et al.*, 2012a).

Wang *et al.* (2012a) suggested two novel approaches that modify BMA to specifically target causal effect estimation: Bayesian adjustment for confounding (BAC) and

two-stage Bayesian adjustment for confounding (TBAC). Graph-based simulations presented in Wang *et al.* (2012a) show that the causal effect estimators of BAC and TBAC are unbiased in a variety of scenarios, hence confirming their adequacy for causal inference. Moreover, a theoretical justification for the use of BAC for causal inference purposes is further discussed in Lefebvre *et al.* (2014a). However, some simulations comparing BAC and TBAC to fully adjusted models show little difference in the variance of the causal effect estimators of each method (Wang *et al.*, 2012a; Vansteelandt, 2012). Moreover, the choice of BAC's hyperparameter ω has been recognized as challenging (Wang *et al.*, 2012b). The value $\omega = \infty$ has been recommended if one seeks an unbiased causal exposure effect estimator (Lefebvre *et al.*, 2014a). Lefebvre *et al.* (2014a) proposed using cross-validation and bootstrap for selecting an ω value that aims to minimize the mean-square-error (MSE) of the BAC's causal effect of exposure estimator. These results suggest that the optimal ω value not only depends on the data-generating scenario, but also on sample size, thus making it very hard in practice to select an appropriate ω value.

In this paper we propose a new model averaging approach to causal inference: Bayesian causal effect estimation (BCEE). BCEE aims to unbiasedly estimate the causal effect of a continuous exposure on a continuous outcome, while being more efficient than a fully adjusted approach. Our method has some similarities with TBAC unlike TBAC however, the motivation for our method lies in the graphical framework for causal inference (e.g. Pearl (2009)).

The paper is structured as follows. In Section 4.2, we present the BCEE algorithm and discuss, in Section 4.3, a number of aspects of its practical implementation. We compare BCEE to some existing approaches for causal effect estimation in Section 4.4. In Section 4.5, we apply BCEE to a real dataset where we estimate the causal effect of mathematical perceived competence on the self-reported average in mathematics for highschool students in the province of Quebec. We conclude in Section 4.6 with a discussion of our results and provide suggestions for further research.

4.2 Bayesian causal effect estimation (BCEE)

Before presenting BCEE in Section 4.2.3, we first describe the modeling framework in Section 4.2.1 and provide a proposition and corollary concerning directed acyclic graphs (DAGs) in Section 4.2.2. The description of how the proposition and the corollary are used to develop BCEE is presented in Section 4.2.4. We conclude, in Section 4.2.5, with a toy example that sheds light on BCEE's properties. Note that although we refer to BCEE as a Bayesian algorithm, strictly speaking, it is approximately Bayesian since it requires specifying prior distributions only for a subset of the parameters. To simplify the discussion, we motivate BCEE from a frequentist perspective.

4.2.1 Modeling framework

We consider estimating the causal effect of a continuous exposure on a continuous outcome. Let X be the random exposure variable, Y be the random outcome variable and $\boldsymbol{U} = \{U_1, U_2, ..., U_M\}$ be a set of M available, pre-exposure, potentially confounding random covariates. Let i index the units of observations, $i = 1, ..., n$. Our goal is to estimate the causal effect of exposure using a linear regression model for the outcome with normal, independent and identically distributed errors. Assuming the set $\boldsymbol{U}$ is sufficient to identify the average causal effect and the model is correctly specified, a fully adjusted linear regression model can be used to estimate the causal effect. Under such assumptions, parameter β encodes the average causal effect of a unit increase in X on Y in the linear model

$$\mathbb{E}\left(Y_i | X_i, \boldsymbol{U_i}\right) = \delta_0 + \beta X_i + \sum_{m=1}^{M} \delta_m U_{im}, \tag{4.1}$$

where δ_0 is the intercept and δ_m is the regression coefficient associated with covariate U_m. A disadvantage to using a fully adjusted outcome model is that the variance of the exposure effect estimator $\hat{\beta}$ can be large. Therefore, one might want to include a

reduced number of covariates in the outcome model (4.1), that is, to adjust for a strict subset of $\boldsymbol{U}$ also sufficient to estimate the causal effect of X on Y.

Consider G an assumed causal directed acyclic graph (DAG) compatible with the distribution of the observed covariates in G, $\{Y, X, \boldsymbol{U}\}$. Let $\boldsymbol{D} = \{D_1, D_2, ..., D_J\} \subset \boldsymbol{U}$ be the set of parents (direct causes) of X in G. Then using Pearl's back-door criterion (Pearl, 2009), it is straightforward to show that adjusting for the set $\boldsymbol{D}$ is sufficient to avoid confounding. In other words, the parameter β in the linear model

$$\mathbb{E}\left(Y_i|X_i, \boldsymbol{D_i}\right) = \delta_0 + \beta X_i + \sum_{j=1}^{J} \delta_j D_{ij} \tag{4.2}$$

can also be interpreted as the average causal effect of X on Y. It can also be shown that outcome models adjusting for sets of pre-exposure covariates that *at least* include the direct causes of exposure are unbiased; BAC may be seen to be exploiting this feature (Lefebvre *et al.*, 2014a). Adjusting for the set of direct causes of X in the outcome model thus seems appealing since $\boldsymbol{D}$ is generally smaller than the full set $\boldsymbol{U}$. However, this approach can also yield an estimator of β, $\hat{\beta}$, whose variance is large unless those direct causes of X are also strong predictors of Y (e.g. Lefebvre *et al.* (2014a)).

BAC, TBAC and BCEE all rely on the fact that the set of direct causes of X is sufficient for estimating the causal effect and that this set of covariates can be identified from the data. A differentiating feature of BCEE however, is that it aims to disfavor outcome models which include one or more direct causes of X that are unnecessary to eliminate confounding. This is viewed as desirable since these variables generally increase the variance of $\hat{\beta}$. By doing so, BCEE targets sufficient models

$$\mathbb{E}\left(Y_i|X_i, \boldsymbol{Z_i}\right) = \delta_0 + \beta X_i + \sum_{k=1}^{K} \delta_k Z_{ik} \tag{4.3}$$

for which the variance of $\hat{\beta}$ is smaller than the variance of $\hat{\beta}$ in model (4.1) and the variance of $\hat{\beta}$ obtained using BAC or TBAC. In Section 4.2.2 we present a proposition and a corollary that underlie the functioning of BCEE.

4.2.2 A motivation based on directed acyclic graphs

The results presented in this section are based on Pearl's back-door criterion and are thus obtained from a graphical perspective to causality using directed acyclic graphs (DAGs). For a brief review of this framework, we refer the reader to the appendix of VanderWeele & Shpitser (2011).

Proposition 4.2.1 presented below gives a sufficient condition to identify a set $\boldsymbol{Z}$ that yields an unbiased estimator $\hat{\beta}$ of the causal effect of X in (4.3). Corollary 4.2.1 starts with such a sufficient set $\boldsymbol{Z}$ and provides conditions under which a direct cause of X included in $\boldsymbol{Z}$ can be excluded so that the resulting set $\boldsymbol{Z}'$ is also sufficient. Remark that this corollary is akin to Proposition 1 from VanderWeele & Shpitser (2011). In the sequel, the concept of d-separation is used to entail notions of conditional independence between variables. Moreover, in Appendix 4.7.1, we show how the distribution-free adjustment defined in Pearl (2009) relates to the adjustment in the linear model setting introduced in Section 4.2.1.

Proposition 4.2.1. *Consider data compatible with a causal DAG G. Let $\boldsymbol{D} = \{D_1, D_2, ..., D_J\}$ be the set of direct causes of X and let $\boldsymbol{Z}$ be a set of covariates that we consider adjusting for. Adjusting for $\boldsymbol{Z}$ is sufficient to identify the average causal effect of X on Y if*

1) no descendants of X are in $\boldsymbol{Z}$ and

2) if for each $D_j \in \boldsymbol{D}$, either

(a) $D_j \in \boldsymbol{Z}$ or

(b) if $D_j \notin \boldsymbol{Z}$ then Y and D_j are d-separated by $\{X \cup \boldsymbol{Z}\}$.

Proof: see Appendix 4.7.2.1.

Corollary 4.2.1. *Consider a $D_j \in \boldsymbol{Z}$ and let $\boldsymbol{Z}' = \boldsymbol{Z} \backslash D_j$.*

1. *If D_j and Y are d-separated by $\{X \cup \boldsymbol{Z}'\}$ then all back-door paths $X \leftarrow D_j \cdots \rightarrow Y$ are blocked by $\boldsymbol{Z}'$.*
2. *If in addition to 1., $\boldsymbol{Z}$ is sufficient to identify the average causal effect according to Proposition 4.2.1, then $\boldsymbol{Z}'$ is also sufficient to identify the average causal effect of X on Y.*

Proof: see Appendix 4.7.2.2.

We now address how the proposition and the corollary are used in the linear regression setting presented in Section 4.2.1. First, Theorem 1.2.4 from Pearl (2009) states the quasi-equivalence between d-separation and conditional independence. That is, unless a very precise tuning of parameters occurs, d-separation of Y and D_j by $\{X \cup \boldsymbol{Z}\}$ is equivalent to conditional independence between Y and D_j given $\{X \cup \boldsymbol{Z}\}$. Hence, we can replace d-separation by conditional independence in Proposition 4.2.1 and in Corollary 4.2.1. Under the assumption that all variables in the graph G are multivariate normal, we have that conditional independence is equivalent to zero partial correlation and thus to zero regression parameter in the linear model (Baba *et al.*, 2004). More specifically, if Y and D_j are conditionally independent given $\{X \cup \boldsymbol{Z}\}$, then the regression parameter associated to D_j in the linear regression of Y on D_j, X and $\boldsymbol{Z}$ is 0; and this parameter is 0 only if Y and D_j are conditionally independent given $\{X \cup \boldsymbol{Z}\}$. The assumption of multivariate normality is quite stringent; a weaker assumption is that model (4.1) is correctly specified (see Appendix 4.7.3).

4.2.3 The BCEE algorithm

BCEE is viewed as a BMA procedure where the prior distribution of the outcome model is informative and constructed by using estimates from earlier steps of the algorithm, including the exposure model. In this section, we introduce BCEE and define the aforementioned prior distribution. The connections between

Proposition 4.2.1, Corollary 4.2.1 and BCEE's prior distribution are discussed in Section 4.2.4.

We now define the outcome model using the same model averaging notation as in BAC and TBAC. Let $\boldsymbol{\alpha}^{\boldsymbol{Y}} = (\alpha_1^Y, ..., \alpha_M^Y)$ be an M-dimensional vector for the inclusion of the covariates $\boldsymbol{U}$ in the outcome model, where component α_m^Y equals 1 if covariate U_m is included in the model and α_m^Y equals 0 if covariate U_m is not included, $m = 1, ..., M$. Letting i index the units of observation, $i = 1, ..., n$, the outcome model is the following normal linear model

$$Y_i = \delta_0^{\boldsymbol{\alpha}^{\boldsymbol{Y}}} + \beta^{\boldsymbol{\alpha}^{\boldsymbol{Y}}} X_i + \sum_{m=1}^{M} \alpha_m^Y \delta_m^{\boldsymbol{\alpha}^{\boldsymbol{Y}}} U_{im} + \varepsilon_i^{\boldsymbol{\alpha}^{\boldsymbol{Y}}}, \tag{4.4}$$

where $\delta_m^{\boldsymbol{\alpha}^{\boldsymbol{Y}}}$ and $\beta^{\boldsymbol{\alpha}^{\boldsymbol{Y}}}$ denote respectively the unknown regression coefficients associated with U_m and X in the outcome model specified by $\boldsymbol{\alpha}^{\boldsymbol{Y}}$. The parameter $\delta_0^{\boldsymbol{\alpha}^{\boldsymbol{Y}}}$ denotes the unknown intercept in model $\boldsymbol{\alpha}^{\boldsymbol{Y}}$ and the distribution of the error terms is given by $\varepsilon_i^{\boldsymbol{\alpha}^{\boldsymbol{Y}}} \overset{iid}{\sim} N(0, \sigma^2_{\boldsymbol{\alpha}^{\boldsymbol{Y}}})$.

Given model (4.4) and a prior distribution $P(\boldsymbol{\alpha}^{\boldsymbol{Y}})$, the use of BMA for the estimation of the exposure effect requires first obtaining the posterior distribution of the outcome model $P(\boldsymbol{\alpha}^{\boldsymbol{Y}}|Y) \propto P(Y|\boldsymbol{\alpha}^{\boldsymbol{Y}})P(\boldsymbol{\alpha}^{\boldsymbol{Y}})$. Standard implementation of BMA often involves selecting a uniform prior distribution $P(\boldsymbol{\alpha}^{\boldsymbol{Y}}) = 1/2^M \ \forall \ \boldsymbol{\alpha}^{\boldsymbol{Y}}$, in which case $P(\boldsymbol{\alpha}^{\boldsymbol{Y}}|Y) \propto P(Y|\boldsymbol{\alpha}^{\boldsymbol{Y}})$. The model-averaged exposure effect is then given by

$$\mathbb{E}[\beta|Y] = \sum_{\boldsymbol{\alpha}^{\boldsymbol{Y}}} \left[\int_{-\infty}^{\infty} \beta^{\boldsymbol{\alpha}^{\boldsymbol{Y}}} P\left(\beta^{\boldsymbol{\alpha}^{\boldsymbol{Y}}}|\boldsymbol{\alpha}^{\boldsymbol{Y}}, Y\right) d\beta^{\boldsymbol{\alpha}^{\boldsymbol{Y}}} \right] P\left(\boldsymbol{\alpha}^{\boldsymbol{Y}}|Y\right). \tag{4.5}$$

In BCEE, we utilize an informative prior distribution rather than the usual non-informative one. This distribution is such that the outcome models in which $\beta^{\boldsymbol{\alpha}^{\boldsymbol{Y}}}$ has a causal interpretation according to Proposition 4.2.1, and that cannot be reduced

according to Corollary 4.2.1, receive the bulk of the prior probability. As will be seen, this prior distribution is constructed by borrowing information from the data.

The first step in the construction of BCEE's prior distribution $P^B(\boldsymbol{\alpha}^{\boldsymbol{Y}})$ is to compute the posterior distribution of the exposure model. This step is also present in TBAC and is performed in BCEE to identify possible causal exposure models and thus likely direct causes of the exposure. Recall that direct causes of exposure play a pivotal role in both Proposition 4.2.1 and Corollary 4.2.1. We now introduce the notation for the exposure model. Let $\boldsymbol{\alpha}^{\boldsymbol{X}} = (\alpha_1^X, ..., \alpha_M^X)$ be an M-dimensional vector for the inclusion of the covariates $\boldsymbol{U}$ in the exposure model. The exposure model is the following normal linear model

$$X_i = \delta_0^{\boldsymbol{\alpha}^{\boldsymbol{X}}} + \sum_{m=1}^{M} \alpha_m^X \delta_m^{\boldsymbol{\alpha}^{\boldsymbol{X}}} U_{im} + \varepsilon_i^{\boldsymbol{\alpha}^{\boldsymbol{X}}}, \tag{4.6}$$

where $\delta_m^{\boldsymbol{\alpha}^{\boldsymbol{X}}}$ denotes the unknown regression coefficient of U_m, $m = 1, ..., M$, in the exposure model specified by $\boldsymbol{\alpha}^{\boldsymbol{X}}$. The parameter $\delta_0^{\boldsymbol{\alpha}^{\boldsymbol{X}}}$ denotes the unknown intercept in $\boldsymbol{\alpha}^{\boldsymbol{X}}$ and $\varepsilon_i^{\boldsymbol{\alpha}^{\boldsymbol{X}}} \overset{iid}{\sim} N(0, \sigma^2_{\boldsymbol{\alpha}^{\boldsymbol{X}}})$. In this step, each model $\boldsymbol{\alpha}^{\boldsymbol{X}}$ is attributed a weight corresponding to its posterior probability, $P(\boldsymbol{\alpha}^{\boldsymbol{X}}|X) \propto P(X|\boldsymbol{\alpha}^{\boldsymbol{X}})P(\boldsymbol{\alpha}^{\boldsymbol{X}})$. For simplification, $P(\boldsymbol{\alpha}^{\boldsymbol{X}})$ is taken to be uniform (that is, $P(\boldsymbol{\alpha}^{\boldsymbol{X}}) = 1/2^M \ \forall \ \boldsymbol{\alpha}^{\boldsymbol{X}}$), although other prior distributions could be considered.

We are now ready to define $P^B(\boldsymbol{\alpha}^{\boldsymbol{Y}})$, which depends not only on $P(\boldsymbol{\alpha}^{\boldsymbol{X}}|X)$, but also on the regression coefficients $\delta_m^{\boldsymbol{\alpha}^{\boldsymbol{Y}}}$. Remember that Proposition 4.2.1 and Corollary 4.2.1 both require verifying conditional independences. This can be achieved through the examination of the outcome model regression coefficients (see the final remarks of Section 4.2.4 and Appendix 4.7.3). To simplify the presentation, we assume for now that the true values of the regression coefficients are provided by an oracle. The BCEE prior distribution is as follows:

$$P^B(\boldsymbol{\alpha}^{\boldsymbol{Y}}) = \sum_{\boldsymbol{\alpha}^{\boldsymbol{X}}} P^B\left(\boldsymbol{\alpha}^{\boldsymbol{Y}}|\boldsymbol{\alpha}^{\boldsymbol{X}}\right) P(\boldsymbol{\alpha}^{\boldsymbol{X}}|X), \text{ where}$$

$$P^B\left(\boldsymbol{\alpha}^{\boldsymbol{Y}}|\boldsymbol{\alpha}^{\boldsymbol{X}}\right) \propto \prod_{m=1}^{M} Q_{\boldsymbol{\alpha}^{\boldsymbol{Y}}}\left(\alpha_m^Y|\alpha_m^X\right).$$

For vectors $\boldsymbol{\alpha}^{\boldsymbol{Y}}$ and $\boldsymbol{\alpha}^{\boldsymbol{X}}$, $Q_{\boldsymbol{\alpha}^{\boldsymbol{Y}}}\left(\alpha_m^Y|\alpha_m^X\right)$ is given by one of the following:

$$\begin{aligned} Q_{\boldsymbol{\alpha}^{\boldsymbol{Y}}}\left(\alpha_m^Y = 1|\alpha_m^X = 1\right) &= \frac{\omega_m^{\boldsymbol{\alpha}^{\boldsymbol{Y}}}}{\omega_m^{\boldsymbol{\alpha}^{\boldsymbol{Y}}} + 1}, \quad Q_{\boldsymbol{\alpha}^{\boldsymbol{Y}}}\left(\alpha_m^Y = 0|\alpha_m^X = 1\right) = \frac{1}{\omega_m^{\boldsymbol{\alpha}^{\boldsymbol{Y}}} + 1}, \\ Q_{\boldsymbol{\alpha}^{\boldsymbol{Y}}}\left(\alpha_m^Y = 1|\alpha_m^X = 0\right) &= \frac{1}{2}, \quad Q_{\boldsymbol{\alpha}^{\boldsymbol{Y}}}\left(\alpha_m^Y = 0|\alpha_m^X = 0\right) = \frac{1}{2}, \end{aligned} \tag{4.7}$$

where $\omega_m^{\boldsymbol{\alpha}^{\boldsymbol{Y}}}$ is defined in (4.8). To properly define $\omega_m^{\boldsymbol{\alpha}^{\boldsymbol{Y}}}$ we must first define the notion of an m-nearest neighbor outcome model. For a given model $\boldsymbol{\alpha}^{\boldsymbol{Y}}$ where $\alpha_m^Y = 0$, the m-nearest neighbor model to $\boldsymbol{\alpha}^{\boldsymbol{Y}}$, $\boldsymbol{\alpha}^{\boldsymbol{Y}}(m)$, has exactly the same covariates as $\boldsymbol{\alpha}^{\boldsymbol{Y}}$ except with $\alpha_m^Y = 1$ instead of $\alpha_m^Y = 0$. In the case where $\alpha_m^Y = 1$, there is no need to define an m-nearest neighbor model. We now define a new set of regression parameters:

$$\tilde{\delta}_m^{\boldsymbol{\alpha}^{\boldsymbol{Y}}} = \begin{cases} \delta_m^{\boldsymbol{\alpha}^{\boldsymbol{Y}}} & \text{if } \alpha_m^Y = 1 \\ \delta_m^{\boldsymbol{\alpha}^{\boldsymbol{Y}}(m)} & \text{if } \alpha_m^Y = 0. \end{cases}$$

For example, if $\boldsymbol{U} = \{U_1, U_2\}$ and $\boldsymbol{\alpha}^{\boldsymbol{Y}} = (1,0)$ then $\tilde{\delta}_1^{\boldsymbol{\alpha}^{\boldsymbol{Y}}} = \delta_1^{\boldsymbol{\alpha}^{\boldsymbol{Y}}}$ can be directly taken from model $\boldsymbol{\alpha}^{\boldsymbol{Y}}$, whereas $\tilde{\delta}_2^{\boldsymbol{\alpha}^{\boldsymbol{Y}}} = \delta_2^{\boldsymbol{\alpha}^{\boldsymbol{Y}}(2)}$ needs to be taken from model $\boldsymbol{\alpha}^{\boldsymbol{Y}}(2) = (1,1)$.

With this additional notation, we define the hyperparameter $\omega_m^{\boldsymbol{\alpha}^{\boldsymbol{Y}}}$ as:

$$\omega_m^{\boldsymbol{\alpha}^{\boldsymbol{Y}}} = \omega \times \left(\tilde{\delta}_m^{\boldsymbol{\alpha}^{\boldsymbol{Y}}} \frac{\sigma_{U_m}}{\sigma_Y}\right)^2, \tag{4.8}$$

where $0 \leq \omega \leq \infty$ is a user-defined hyperparameter, σ_{U_m} and σ_Y are respectively the (true) standard deviations of U_m and Y. Note that $\tilde{\delta}_m^{\boldsymbol{\alpha}^Y} \sigma_{U_m}/\sigma_Y$ is a standardization of $\tilde{\delta}_m^{\boldsymbol{\alpha}^Y}$ which makes it insensitive to the measurement units of both Y and U_m. In practice, we cannot rely on an oracle to provide $\tilde{\delta}_m^{\boldsymbol{\alpha}^Y}$; in the sequel, we use the maximum likelihood estimator of $\tilde{\delta}_m^{\boldsymbol{\alpha}^Y}$ instead. Also, the true values of σ_{U_m} and σ_Y are not known and are estimated by s_{U_m} and s_Y. The prior distribution $P^B(\boldsymbol{\alpha}^Y)$ thus has an empirical Bayes flavor. Once $P^B(\boldsymbol{\alpha}^Y)$ is obtained, the posterior distribution of the outcome model $P(\boldsymbol{\alpha}^Y|Y)$ is computed and the posterior exposure effect calculated according to (4.5). In Section 4.3.2, we discuss how one can account for using the data for the specification of $P^B(\boldsymbol{\alpha}^Y)$ to obtain appropriate inferences.

4.2.4 The rationale behind BCEE

In this section, we explain in detail how BCEE's prior distribution $P^B(\boldsymbol{\alpha}^Y)$ is motivated by causal graphs through Proposition 4.2.1 and Corollary 4.2.1. To do so, we rely on the relationship between d-separation and linear regression established in Appendix 4.7.1.

To begin, recall that the first step of BCEE serves to identify likely exposure models. Classical properties of Bayesian model selection ensure that the true (structural) exposure model, the one including only and all direct causes of X ($\boldsymbol{D} = \{D_1, ..., D_J\}$), is asymptotically attributed all the posterior probability by the first step of BCEE (e.g. Haughton (1988); Wasserman (2000)). This result follows from assuming that the set of potential confounding covariates $\boldsymbol{U}$ includes all direct causes of X and no descendants of X and that the specification of the model is correct: that is, the true exposure model is indeed a normal linear model of the form $X_i = \delta_0 + \sum_{j=1}^{J} \delta_j D_{ij} + \varepsilon_i^X$, with $\varepsilon_i^X \overset{iid}{\sim} N(0, \sigma_X^2)$.

The algorithm BCEE aims to give the bulk of the posterior weight to outcome models in which $\beta^{\boldsymbol{\alpha}^Y}$ has a causal interpretation according to Proposition 4.2.1 and that cannot be reduced according to Corollary 4.2.1. In such outcome models, $\boldsymbol{\alpha}^Y$ includes any

given direct cause (identified in the first step) only if the inclusion of this direct cause of exposure is necessary for β^{α^Y} to have a causal interpretation in α^Y. To do so, $P^B(\alpha^Y)$ places small prior weight on outcome models that do not respect condition 2 of Proposition 4.2.1. In such models, some direct causes of X are excluded (condition point 2a from Proposition 4.2.1) and Y is dependent on those excluded direct causes of X given X and the potential confounding covariates already included (condition point 2b from Proposition 4.2.1). Moreover, $P^B(\alpha^Y)$ limits the prior weight attributed to outcome models which could be reduced according to Corollary 4.2.1. In such models, some direct causes of X are included, but these are not associated with Y conditionally on X and the other covariates included.

To illustrate how Proposition 4.2.1 and Corollary 4.2.1 motivate the formulation of $P^B(\alpha^Y)$ we provide the following thought experiment. To simplify our presentation, we assume that the direct causes of exposure are known and that the outcome model (4.1) is correctly specified. Moreover, we order the elements of $\boldsymbol{U}$ so that the first J elements are $\boldsymbol{D}$, that is $\{U_1, ..., U_M\} = \{D_1, ..., D_J, U_{J+1}, ..., U_M\}$. For ease of interpretation, we also assume that the covariates $\boldsymbol{U}$ are standardized, although, due to the way $\omega_m^{\alpha^Y}$ is defined, this is not necessary in practice. We consider four different situations to illustrate how BCEE functions. In each situation, a direct cause of exposure $D_j = U_j$ is either included or excluded from the outcome model α^Y and the maximum likelihood estimate $|\hat{\tilde{\delta}}_j^{\alpha^Y}|$ is either close to 0 or large. The anticipated magnitudes of $Q_{\alpha^Y}(\alpha_j^Y|\alpha_j^X)$ and of $P^B(\alpha^Y|\alpha^X)$ for each situation are presented in Table 4.1. Considering jointly those four situations, we see that only outcome models that both correctly identify the average causal effect of exposure and that solely include necessary direct causes of exposure receive non-negligible prior probabilities. In the next paragraph, we describe in detail the first situation, which supposes that direct cause D_j is omitted from α^Y and its associated estimated parameter $|\hat{\tilde{\delta}}_j^{\alpha^Y}|$ is large.

Suppose α^Y does not include D_j. Note that $Q_{\alpha^Y}(\alpha_j^Y = 0|\alpha_j^X = 1)$ depends on $\hat{\tilde{\delta}}_j^{\alpha^Y}$ through $\omega_j^{\alpha^Y}$. Therefore, $P^B(\alpha^Y|\alpha^X)$ also depends on $\hat{\tilde{\delta}}_j^{\alpha^Y}$. If $|\hat{\tilde{\delta}}_j^{\alpha^Y}|$ is large, then Y is likely not independent of D_j conditionally on X and the potential

confounding covariates included in $\boldsymbol{\alpha}^{\boldsymbol{Y}}$ (see Appendix 4.7.3). It follows that $\boldsymbol{\alpha}^{\boldsymbol{Y}}$ does not respect condition 2b from Proposition 4.2.1. Since the value of $\omega_j^{\boldsymbol{\alpha}^{\boldsymbol{Y}}}$ is large, $Q_{\boldsymbol{\alpha}^{\boldsymbol{Y}}}(\alpha_j^Y = 0|\alpha_j^X = 1)$ is small and so is $P^B(\boldsymbol{\alpha}^{\boldsymbol{Y}}|\boldsymbol{\alpha}^{\boldsymbol{X}})$. In this situation, $P^B(\boldsymbol{\alpha}^{\boldsymbol{Y}})$ is well behaved: the model $\boldsymbol{\alpha}^{\boldsymbol{Y}}$ is not sufficient to identify the average causal effect of exposure and hence it receives little prior probability. A similar reasoning can be applied for situation 4 of Table 4.1. The reasoning for situations 2 and 3 is also quite similar, but requires invoking Corollary 4.2.1 to determine whether the inclusion of D_j is necessary or not.

Table 4.1: Magnitudes of $Q_{\boldsymbol{\alpha}^{\boldsymbol{Y}}}(\alpha_j^Y|\alpha_j^X)$ and $P^B(\boldsymbol{\alpha}^{\boldsymbol{Y}}|\boldsymbol{\alpha}^{\boldsymbol{X}})$ for four situations defined by the inclusion of a direct cause of exposure D_j and the magnitude of $|\hat{\tilde{\delta}}_j^{\boldsymbol{\alpha}^{\boldsymbol{Y}}}|$.

Situation	D_j	$\lvert\hat{\tilde{\delta}}_j^{\boldsymbol{\alpha}^{\boldsymbol{Y}}}\rvert$	$Y \perp\!\!\!\perp D_j \lvert X, \boldsymbol{Z}'$	$\omega_j^{\boldsymbol{\alpha}^{\boldsymbol{Y}}}$	$Q_{\boldsymbol{\alpha}^{\boldsymbol{Y}}}(\alpha_j^Y\lvert\alpha_j^X)$	$P^B(\boldsymbol{\alpha}^{\boldsymbol{Y}}\lvert\boldsymbol{\alpha}^{\boldsymbol{X}})$
(1)	Excl.	Large	Not likely	Large	Close to 0	Close to 0
(2)	Incl.	Close to 0	Likely	Close to 0	Close to 0	Close to 0
(3)	Incl.	Large	Not likely	Large	Close to 1	Depends
(4)	Excl.	Close to 0	Likely	Close to 0	Close to 1	Depends

LEGEND: $\boldsymbol{Z}'$ denotes the potential confounding covariates included in $\boldsymbol{\alpha}^{\boldsymbol{Y}}$ excluding D_j, Excl. = Excluded, Incl. = Included, Depends = Depends on other D_js.

Remark in Table 4.1 that in situations 3 and 4, where $Q_{\boldsymbol{\alpha}^{\boldsymbol{Y}}}(\alpha_j^Y|\alpha_j^X)$ is close to 1, $P^B(\boldsymbol{\alpha}^{\boldsymbol{Y}}|\boldsymbol{\alpha}^{\boldsymbol{X}})$ depends in a large part on the $Q_{\boldsymbol{\alpha}^{\boldsymbol{Y}}}$ associated with the other direct causes of exposure. If none of the $Q_{\boldsymbol{\alpha}^{\boldsymbol{Y}}}$ are close to 0, then $P^B(\boldsymbol{\alpha}^{\boldsymbol{Y}}|\boldsymbol{\alpha}^{\boldsymbol{X}})$ is non-negligible and hence favors models that identify the causal effect according to Proposition 4.2.1 and Corollary 4.2.1. However, if any of the $Q_{\boldsymbol{\alpha}^{\boldsymbol{Y}}}$ is close to 0, then $P^B(\boldsymbol{\alpha}^{\boldsymbol{Y}}|\boldsymbol{\alpha}^{\boldsymbol{X}})$ is close to 0.

4.2.5 A toy example

We consider a toy example to gain preliminary insights on the finite sample properties of BCEE. We generated a sample of size $n = 500$ satisfying the following relationships:

$$X = U_1 + U_2 + \varepsilon_X$$
$$Y = X + 0.1U_1 + \varepsilon_Y,$$

with $U_1, U_2 \sim N(0,1)$ and $\varepsilon_X, \varepsilon_Y \sim N(0,1)$, all independent.

The first step of BCEE is to calculate the posterior distribution of the exposure model $P(\boldsymbol{\alpha}^{\boldsymbol{X}}|X)$. The four possible exposure models in this example are:

$$\boldsymbol{\alpha}_{\boldsymbol{1}}^{\boldsymbol{X}} : X| \rightarrow (\alpha_1^X = 0, \alpha_2^X = 0),$$
$$\boldsymbol{\alpha}_{\boldsymbol{2}}^{\boldsymbol{X}} : X|U_1 \rightarrow (\alpha_1^X = 1, \alpha_2^X = 0),$$
$$\boldsymbol{\alpha}_{\boldsymbol{3}}^{\boldsymbol{X}} : X|U_2 \rightarrow (\alpha_1^X = 0, \alpha_2^X = 1),$$
$$\boldsymbol{\alpha}_{\boldsymbol{4}}^{\boldsymbol{X}} : X|U_1, U_2 \rightarrow (\alpha_1^X = 1, \alpha_2^X = 1).$$

We approximate $P(X|\alpha^{\boldsymbol{X}})$ using $\exp[-0.5BIC(\boldsymbol{\alpha}^{\boldsymbol{X}})]$ (Clyde, 2003), where $BIC(\boldsymbol{\alpha}^{\boldsymbol{X}})$ is the Bayesian information criterion for exposure model $\boldsymbol{\alpha}^{\boldsymbol{X}}$. In our example, model $\boldsymbol{\alpha}_{\boldsymbol{4}}^{\boldsymbol{X}}$ receives all posterior weight, that is $P(\boldsymbol{\alpha}^{\boldsymbol{X}} = (1,1)|X) = 1$.

Next, we compute the posterior distribution of the outcome model using $P^B(\boldsymbol{\alpha}^{\boldsymbol{Y}})$. We take $\omega = 100\sqrt{n}$, a choice that is subsequently discussed in Section 4.3.1. The four possible outcome models are:

$$\boldsymbol{\alpha_1^Y} : Y|X \rightarrow (\alpha_1^Y = 0, \alpha_2^Y = 0),$$
$$\boldsymbol{\alpha_2^Y} : Y|X, U_1 \rightarrow (\alpha_1^Y = 1, \alpha_2^Y = 0),$$
$$\boldsymbol{\alpha_3^Y} : Y|X, U_2 \rightarrow (\alpha_1^Y = 0, \alpha_2^Y = 1),$$
$$\boldsymbol{\alpha_4^Y} : Y|X, U_1, U_2 \rightarrow (\alpha_1^Y = 1, \alpha_2^Y = 1).$$

Note that only models $\boldsymbol{\alpha_2^Y}$ and $\boldsymbol{\alpha_4^Y}$ correctly identify the causal effect of exposure. We present the calculation of $P^B(\boldsymbol{\alpha^Y}|\boldsymbol{\alpha^X})$ for model $\boldsymbol{\alpha_2^Y}$. Since we obtained $P(\boldsymbol{\alpha^X} = (1,1)|X) = 1$, we only need to calculate $P^B(\boldsymbol{\alpha^Y} = (1,0)|\boldsymbol{\alpha^X} = (1,1)) \propto Q_{\boldsymbol{\alpha_2^Y}}(\alpha_1^Y = 1|\alpha_1^X = 1) Q_{\boldsymbol{\alpha_2^Y}}(\alpha_2^Y = 0|\alpha_2^X = 1)$. We have:

$$Q_{\boldsymbol{\alpha_2^Y}}(\alpha_1^Y = 1|\alpha_1^X = 1) = \frac{\omega_1^{\boldsymbol{\alpha_2^Y}}}{\omega_1^{\boldsymbol{\alpha_2^Y}} + 1},$$
$$Q_{\boldsymbol{\alpha_2^Y}}(\alpha_2^Y = 0|\alpha_2^X = 1) = \frac{1}{\omega_2^{\boldsymbol{\alpha_2^Y}} + 1}.$$

We get $\omega_1^{\boldsymbol{\alpha_2^Y}} = \omega(\hat{\delta}_1^{\boldsymbol{\alpha_2^Y}} \times s_{U_1}/s_Y)^2 = 100\sqrt{500}(0.14 \times 1.00/2.01)^2 = 9.75$. Note that because U_1 is included in $\boldsymbol{\alpha_2^Y}$, $\hat{\delta}_1^{\boldsymbol{\alpha_2^Y}} = \hat{\delta}_1^{\boldsymbol{\alpha_2^Y}}$. Also, we have $\omega_2^{\boldsymbol{\alpha_2^Y}} = \omega(\hat{\delta}_2^{\boldsymbol{\alpha_2^Y}} \times s_{U_2}/s_Y)^2 = 100\sqrt{500}(-0.01 \times 1.04/2.01)^2 = 0.05$. Because U_2 is not in $\boldsymbol{\alpha_2^Y}$, we get the regression parameter estimate for U_2 from its 2-nearest neighbor model, that is $\hat{\delta}_2^{\boldsymbol{\alpha_2^Y}} = \hat{\delta}_2^{\boldsymbol{\alpha_4^Y}}$. Finally, the value of the (unnormalized) prior probability of model $\boldsymbol{\alpha_2^Y}$ is 0.8658.

Following the same process for the three other outcome models, we calculate the prior probabilities. From there, we calculate the posterior distribution of the outcome model using the relationship $P(\boldsymbol{\alpha^Y}|Y) \propto P(Y|\boldsymbol{\alpha^Y}) P^B(\boldsymbol{\alpha^Y})$. Again, we use $\exp[-0.5BIC(\boldsymbol{\alpha^Y})]$ to approximate $P(Y|\boldsymbol{\alpha^Y})$. Table 4.2 provides the results with the details of the intermediate steps.

Table 4.2: Calculation of the BCEE outcome model posterior distribution with intermediate steps.

model	$U.P^B(\boldsymbol{\alpha}^Y)$	$P^B(\boldsymbol{\alpha}^Y)$	BIC	BMA $P(\boldsymbol{\alpha}^Y \mid Y)$	$P(\boldsymbol{\alpha}^Y \mid Y)$
$\boldsymbol{\alpha}_1^Y$	0.0230	0.0229	1435.82	0.4602	0.0254
$\boldsymbol{\alpha}_2^Y$	0.8658	0.8618	1435.81	0.4629	0.9625
$\boldsymbol{\alpha}_3^Y$	0.0749	0.0746	1440.04	0.0560	0.0101
$\boldsymbol{\alpha}_4^Y$	0.0409	0.0407	1442.00	0.0209	0.0021

LEGEND: $U.P^B(\boldsymbol{\alpha}^Y)$ is the unnormalized prior probability, $P^B(\boldsymbol{\alpha}^Y)$ is the prior probability, BIC is the Bayesian information criterion, BMA $P(\boldsymbol{\alpha}^Y \mid Y)$ is the posterior probability the model resulting from a BMA procedure with a non-informative prior distribution, and $P(\boldsymbol{\alpha}^Y \mid Y)$ is the posterior probability using BCEE.

We see from these results how BCEE, as compared to BMA, shifts the posterior weight toward models that identify the causal effect of exposure. In fact, in this toy example, BCEE puts almost all the posterior weight on the true outcome model. BCEE accomplishes this by using an informative prior distribution for the outcome model that borrows information both from the exposure selection step and from neighboring regression coefficient estimates in the outcome models.

4.3 Practical considerations regarding BCEE

In this section we discuss practical considerations regarding the usage of the BCEE algorithm. First, we discuss the choice of the hyperparameter ω value in (4.8), then we suggest two alternative ways of implementing BCEE.

4.3.1 Choice of ω

Recall that BCEE's prior distribution $P^B(\boldsymbol{\alpha}^Y)$ depends on a user-selected hyperparameter ω. In what follows, we suggest making ω proportional to $\sqrt{n}$ on the basis of asymptotic results related to the quantities $Q_{\boldsymbol{\alpha}^Y}$ in Equation (4.7). Without loss of generality, we only discuss the case $Q_{\boldsymbol{\alpha}^Y}(\alpha_m^Y = 1 | \alpha_m^X = 1)$. Indeed, the cases $Q_{\boldsymbol{\alpha}^Y}(\alpha_m^Y = 1 | \alpha_m^X = 0)$ and $Q_{\boldsymbol{\alpha}^Y}(\alpha_m^Y = 0 | \alpha_m^X = 0)$ are trivial because the two quantities

are both equal to 1/2. Moreover, the case $Q_{\boldsymbol{\alpha}^Y}(\alpha_m^Y = 0|\alpha_m^X = 1)$ is essentially equivalent to the case $Q_{\boldsymbol{\alpha}^Y}(\alpha_m^Y = 1|\alpha_m^X = 1)$ since these quantities are closely (and negatively) associated. Remark that because we consider the case where $\alpha_m^Y = 1$, $\tilde{\delta}_m^{\boldsymbol{\alpha}^Y} = \delta_m^{\boldsymbol{\alpha}^Y}$. However, we present the reasoning in terms of $\tilde{\delta}_m^{\boldsymbol{\alpha}^Y}$ to allow a direct generalization to the case where $\alpha_m^Y = 0$.

Assume that the true outcome model is a normal linear model of the form (4.1) and first consider the case $\tilde{\delta}_m^{\boldsymbol{\alpha}^Y} = 0$ for a given model $\boldsymbol{\alpha}^Y$. Then covariate U_m is conditionally independent of Y given the (other) covariates included in model $\boldsymbol{\alpha}^Y$ (see Appendix 4.7.3). Hence U_m should be left out of $\boldsymbol{\alpha}^Y$ on the basis of Corollary 4.2.1. It is thus desirable that $Q_{\boldsymbol{\alpha}^Y}(\alpha_m^Y = 1|\alpha_m^X = 1) \to 0$ as $n \to \infty$, which happens if $\hat{\omega}^{\boldsymbol{\alpha}_m^Y} = \omega \times \left(\hat{\tilde{\delta}}_m^{\boldsymbol{\alpha}^Y} s_{U_m}/s_Y\right)^2 \to 0$ as $n \to \infty$.

Consider the case $\tilde{\delta}_m^{\boldsymbol{\alpha}^Y} \neq 0$. According to Proposition 4.2.1, it is now desirable that $Q_{\boldsymbol{\alpha}^Y}(\alpha_m^Y = 1|\alpha_m^X = 1) \to 1$ as $n \to \infty$, since this would allow for covariates causing less confounding to be forced in the outcome model as n grows. Thus, we need $\hat{\omega}_m^{\boldsymbol{\alpha}^Y} \to \infty$ as $n \to \infty$ if $\tilde{\delta}_m^{\boldsymbol{\alpha}^Y} \neq 0$.

If $\tilde{\delta}_m^{\boldsymbol{\alpha}^Y} = 0$ then $\hat{\tilde{\delta}}_m^{\boldsymbol{\alpha}^Y} \xrightarrow{P} 0$ and thus, for any finite constant value of ω, $\hat{\omega}_m^{\boldsymbol{\alpha}^Y} \xrightarrow{P} 0$, where $\xrightarrow{P}$ means convergence in probability. However, if $\tilde{\delta}_m^{\boldsymbol{\alpha}^Y} \neq 0$, we need to choose ω as a function of sample size n in order to ensure that $\hat{\omega}_m^{\boldsymbol{\alpha}^Y} \to \infty$ as $n \to \infty$. We consider rates of convergence to find an appropriate function of n.

Recall that $\hat{\omega}_m^{\boldsymbol{\alpha}^Y}$ is a function of the MLE $\hat{\tilde{\delta}}_m^{\boldsymbol{\alpha}^Y}$ (Section 4.2.3). Under mild regularity conditions, it follows from the results in Yuan & Chan (2011) that $\hat{\tilde{\delta}}_m^{\boldsymbol{\alpha}^Y} s_{U_m}/s_Y \xrightarrow{P} \tilde{\delta}_m^{\boldsymbol{\alpha}^Y} \sigma_{U_m}/\sigma_Y$ at rate $O_p(1/\sqrt{n})$, where O_p is the usual big-O_p notation (Agresti (2013), p.588). Thus $\left(\hat{\tilde{\delta}}_m^{\boldsymbol{\alpha}^Y} s_{U_m}/s_Y\right)^2 \xrightarrow{P} \left(\tilde{\delta}_m^{\boldsymbol{\alpha}^Y} \sigma_{U_m}/\sigma_Y\right)^2$ at rate $O_p(1/n)$.

By taking $\omega = cn^b$, with $0 < b < 1$, where c is a user-fixed constant that does not depend on sample size, we obtain $\hat{\omega}_m^{\boldsymbol{\alpha}^Y} \to \infty$ (at rate n^b) if $\tilde{\delta}_m^{\boldsymbol{\alpha}^Y} \neq 0$ and $\hat{\omega}_m^{\boldsymbol{\alpha}^Y} \xrightarrow{P} 0$ (with convergence rate $O_p(1/n^{1-b})$) if $\tilde{\delta}_m^{\boldsymbol{\alpha}^Y} = 0$, as desired. The value $b = 1/2$ appears to make a good compromise between the two desired convergence behaviors. The simulation

study presented in Section 4 shows that BCEE performs well for $\omega = c\sqrt{n}$ with $100 \leq c \leq 1000$. We also see that larger values of c yield less bias and more variance in the estimator of the causal effect, and conversely for smaller values of c. Appendix 4.7.4 illustrates how $Q_{\boldsymbol{\alpha}^{\boldsymbol{Y}}}(\alpha_m^Y = 1 | \alpha_m^X = 1)$ behaves for different values of c in some simple settings.

4.3.2 Implementing BCEE

In this section, we first consider a naive implementation of BCEE that closely follows our presentation of the algorithm in Section 4.2.3. Then we describe a modified implementation that accounts for using the MLE $\hat{\tilde{\delta}}_m^{\boldsymbol{\alpha}^{\boldsymbol{Y}}}$ in $P^B(\boldsymbol{\alpha}^{\boldsymbol{Y}})$.

We perform three steps to sample one draw from the posterior distribution of the average causal exposure effect $P(\beta|Y)$. Several such draws are taken to obtain approximations to quantities of interest, such as the posterior mean and variance of β. The steps of the sampling procedure are:

S1. Draw $\boldsymbol{\alpha}^{\boldsymbol{X}}$ from the posterior distribution of the exposure model $P(\boldsymbol{\alpha}^{\boldsymbol{X}}|X) \propto P(X|\boldsymbol{\alpha}^{\boldsymbol{X}})$, using $\exp[-0.5BIC(\boldsymbol{\alpha}^{\boldsymbol{X}})]$ to approximate $P(X|\boldsymbol{\alpha}^{\boldsymbol{X}})$;

S2. Draw $\boldsymbol{\alpha}^{\boldsymbol{Y}}$ from the conditional posterior distribution $P(\boldsymbol{\alpha}^{\boldsymbol{Y}}|\boldsymbol{\alpha}^{\boldsymbol{X}}, Y) \propto P^B(\boldsymbol{\alpha}^{\boldsymbol{Y}}|\boldsymbol{\alpha}^{\boldsymbol{X}})P(Y|\boldsymbol{\alpha}^{\boldsymbol{Y}})$, where the regression coefficients $\tilde{\delta}_m^{\boldsymbol{\alpha}^{\boldsymbol{Y}}}$ are estimated by their MLEs and $P(Y|\boldsymbol{\alpha}^{\boldsymbol{Y}})$ is approximated by $\exp[-0.5BIC(\boldsymbol{\alpha}^{\boldsymbol{Y}})]$;

S3. Draw β from the conditional posterior distribution $P(\beta^{\boldsymbol{\alpha}^{\boldsymbol{Y}}}|\boldsymbol{\alpha}^{\boldsymbol{Y}}, Y)$, which we approximate by its limit normal distribution $N\left(\hat{\beta}^{\boldsymbol{\alpha}^{\boldsymbol{Y}}}, \widehat{SE}(\hat{\beta}^{\boldsymbol{\alpha}^{\boldsymbol{Y}}})\right)$ (Dawid, 1970; Walker, 1969), where $\hat{\beta}^{\boldsymbol{\alpha}^{\boldsymbol{Y}}}$ is the maximum likelihood estimator of $\beta^{\boldsymbol{\alpha}^{\boldsymbol{Y}}}$ and $\widehat{SE}(\hat{\beta}^{\boldsymbol{\alpha}^{\boldsymbol{Y}}})$ is its estimated standard error.

The sampling for the first two steps is done using Markov chain Monte Carlo model composition (MC^3) (Madigan *et al.*, 1995). We refer to this naive implementation of BCEE as N-BCEE.

Because N-BCEE does not take into account the uncertainty related to the estimation of the regression coefficients $\tilde{\delta}_m^{\alpha^Y}$ in $P^B(\boldsymbol{\alpha}^{\boldsymbol{Y}})$, we anticipate that the confidence (credible) interval for β will be too narrow. Our insight relies on the empirical Bayes literature, where it has been extensively shown that data-dependent prior distributions lead to confidence intervals that tend to be "too short, inappropriately centered, or both" (Carlin & Gelfand, 1990). Also, narrow confidence intervals for β are observed in simulations presented in Section 4.4. Although many solutions to this problem have been proposed (see Carlin & Louis (2000) for a short discussion), most cannot be realistically applied to BCEE due to the complexity of the algorithm. Therefore, we propose the following simple ad hoc solution, which happens to be notably faster than N-BCEE. We refer to this modified implementation of BCEE as A-BCEE.

A-BCEE is the same as N-BCEE except for step S2. Recall that this step is directed at sampling from the conditional posterior distribution $P(\boldsymbol{\alpha}^{\boldsymbol{Y}}|\boldsymbol{\alpha}^{\boldsymbol{X}}, Y)$ using MC^3. This MC^3 scheme requires calculating a Metropolis-Hasting ratio which involves the ratio of the (conditional) prior probabilities of the proposed outcome model, $\boldsymbol{\alpha}_1^{\boldsymbol{Y}}$, to the current outcome model, $\boldsymbol{\alpha}_2^{\boldsymbol{Y}}$:

$$RP = \frac{P^B\left(\boldsymbol{\alpha}_1^{\boldsymbol{Y}}|\boldsymbol{\alpha}^{\boldsymbol{X}}\right)}{P^B\left(\boldsymbol{\alpha}_2^{\boldsymbol{Y}}|\boldsymbol{\alpha}^{\boldsymbol{X}}\right)} = \frac{\prod_{m=1}^{M} Q_{\boldsymbol{\alpha}_1^{\boldsymbol{Y}}}\left(\alpha_m^Y|\alpha_m^X\right)/C}{\prod_{m=1}^{M} Q_{\boldsymbol{\alpha}_2^{\boldsymbol{Y}}}\left(\alpha_m^Y|\alpha_m^X\right)/C} = \prod_{m=1}^{M} \frac{Q_{\boldsymbol{\alpha}_1^{\boldsymbol{Y}}}\left(\alpha_m^Y|\alpha_m^X\right)}{Q_{\boldsymbol{\alpha}_2^{\boldsymbol{Y}}}\left(\alpha_m^Y|\alpha_m^X\right)}, \tag{4.9}$$

where C is a normalizing constant such that $P^B(\boldsymbol{\alpha}^{\boldsymbol{Y}}|\boldsymbol{\alpha}^{\boldsymbol{X}}) = \prod_{m=1}^{M} Q_{\boldsymbol{\alpha}^{\boldsymbol{Y}}}\left(\alpha_m^Y|\alpha_m^X\right)/C$. In RP, $\boldsymbol{\alpha}_1^{\boldsymbol{Y}}$ and $\boldsymbol{\alpha}_2^{\boldsymbol{Y}}$ are two neighbor outcome models that differ only by their inclusion of a single covariate $U_{m'}$. A-BCEE utilizes the following simplification for RP:

$$RP \approx \frac{Q_{\boldsymbol{\alpha}_1^{\boldsymbol{Y}}}\left(\alpha_{m'}^Y|\alpha_{m'}^X\right)}{Q_{\boldsymbol{\alpha}_2^{\boldsymbol{Y}}}\left(\alpha_{m'}^Y|\alpha_{m'}^X\right)}. \tag{4.10}$$

The heuristic for suggesting this approximation is that the individual ratio that is the most likely to significantly differ from 1 in (4.9) is the one associated to covariate $U_{m'}$, that is $Q_{\boldsymbol{\alpha}_1^{\boldsymbol{Y}}}\left(\alpha_{m'}^Y|\alpha_{m'}^X\right)/Q_{\boldsymbol{\alpha}_2^{\boldsymbol{Y}}}\left(\alpha_{m'}^Y|\alpha_{m'}^X\right)$. In fact, unless the covariates

U are very strongly correlated with each other, we expect the $\hat{\tilde{\delta}}^{\alpha^Y}_m$s ($m \neq m'$) to be of the same magnitude between two neighboring models. Note that we also expect many terms in the RP product to be exactly equal to 1 since an individual ratio equals 1 when its corresponding covariate is not included in the exposure model $\left(Q_{\alpha^Y_1}\left(\alpha^Y_m|\alpha^X_m=0\right)/Q_{\alpha^Y_2}\left(\alpha^Y_m|\alpha^X_m=0\right)=1\right)$. Simulations were performed to verify the validity of approximation (4.10) (results not presented).

Using simplified RP (4.10), it becomes an easy task to incorporate the variability associated with the estimation of the $\tilde{\delta}^{\alpha^Y}$s. We assume that $\tilde{\delta}^{\alpha^Y}_{m'} \sim N\left(\hat{\tilde{\delta}}^{\alpha^Y}_{m'}, \widehat{SE}\left(\hat{\tilde{\delta}}^{\alpha^Y}_{m'}\right)\right)$, where $\widehat{SE}\left(\hat{\tilde{\delta}}^{\alpha^Y}_{m'}\right)$ is the estimated standard error of $\hat{\tilde{\delta}}^{\alpha^Y}_{m'}$. In summary, in step S2 of the sampling procedure of A-BCEE we simply draw $\tilde{\delta}^{\alpha^Y}_{m'}$ from $N\left(\hat{\tilde{\delta}}^{\alpha^Y}_{m'}, \widehat{SE}\left(\hat{\tilde{\delta}}^{\alpha^Y}_{m'}\right)\right)$ and use it in approximation (4.10). We remark that this strategy is akin to specifying an Empirical Bayes type of hyperprior for $\tilde{\delta}^{\alpha^Y}$.

The finite sample properties of N-BCEE and A-BCEE are studied and compared in some simulation scenarios presented in the next section. We also consider nonparametric bootstrap (Laird & Louis, 1987) in a few simple and small scale simulations as an alternative to A-BCEE to correct confidence intervals. Note that, due to computing time, this bootstrapped BCEE (B-BCEE) approach is considerably less practical than A-BCEE to evaluate in simulations and to apply to real data sets of moderate to large sizes.

4.4 Simulation studies

In this section, we study the finite sample properties of BCEE in various simulation scenarios. The first primary objective of the simulations is to compare BCEE to standard or related methods that are used to estimate total average causal effects of exposure. The second primary objective is to study the sensitivity of BCEE to the choice of its user-selected hyperparameter ω. The three secondary objectives relate to the large, whilst finite, properties of BCEE, to the performance of the B-BCEE and to

BCEE robustness to non normality of errors (model mispecification). The primary and secondary objectives are examined in separate subsections.

4.4.1 Main simulations

To achieve the two main objectives, we examine 24 different simulation scenarios obtained by considering three factors: data-generating process (DGP1, DGP2, DGP3 and DGP4), sample size (200, 600 and 1000) and true causal effect of exposure ($\beta = 0.1$ or $\beta = 0$). The four data-generating processes are described below.

The first data-generating process (DGP1) satisfies the following relationships:

$$\begin{aligned} U_3 &= U_2 + \varepsilon_3 \\ U_5 &= U_4 + \varepsilon_5 \\ X &= U_1 + U_2 + U_4 + \varepsilon_X \\ Y &= U_3 + 0.1U_4 + U_5 + \beta X + \varepsilon_Y, \end{aligned}$$

with $U_1, U_2, U_4, \varepsilon_3, \varepsilon_5, \varepsilon_X, \varepsilon_Y \sim N(0,1)$ all independent. The set of available covariates is $\boldsymbol{U} = \{U_1, U_2, \ldots, U_5\}$.

The second data-generating process (DGP2) involves a larger number of covariates than DGP1 and features an indirect effect of X on Y:

$$\begin{aligned} U_1 &= U_4 + \varepsilon_1 & U_2 &= U_4 + \varepsilon_2 \\ U_3 &= U_4 + \varepsilon_3 & U_5 &= U_1 + \varepsilon_5 \\ X &= U_1 + U_2 + U_3 + \varepsilon_X \\ U_6 &= 0.5X + U_3 + \varepsilon_6 \\ Y &= 0.1U_4 + 0.1U_5 + \beta U_6 + 0.5\beta X + \varepsilon_Y, \end{aligned}$$

where $U_4, \varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_5, \varepsilon_X, \varepsilon_6, \varepsilon_Y \sim N(0,1)$ all independent. The set of available covariates is $\boldsymbol{U} = \{U_1, U_2, \ldots, U_5, U_7, \ldots, U_{15}\}$, where $U_7, \ldots, U_{15}$ are all independent $N(0,1)$. We exclude U_6 from the set of potential confounding covariates since one must not adjust for descendants of the exposure X to identify the total average causal effect. Here the total effect of X on Y (direct effect plus indirect effect through U_6) is $0.5\beta + 0.5\beta = \beta$. For simulation purposes, we consider the model $\boldsymbol{\alpha^Y} = (0,0,1,1,1,0,...,0)$ as the "true" outcome model.

The third data-generating process (DGP3) is similar to the first simulation example in Wang *et al.* (2012a) but includes only 18 additional (noise) covariates (instead of 49):

$$X = 0.7U_1 + \sqrt{(1-0.7^2)}\varepsilon_X$$
$$Y = 0.1U_1 + 0.1U_2 + \beta X + \varepsilon_Y,$$

where $U_1, U_2, \varepsilon_X, \varepsilon_Y \sim N(0,1)$ all independent. The set of available covariates is $\boldsymbol{U} = \{U_1, U_2, \ldots, U_{20}\}$, where $U_3, \ldots, U_{20}$ are also independent $N(0,1)$.

The fourth data-generating process (DGP4) is inspired by a DAG presented in Morgan & Winship (2007), Figure 1.1, page 25:

$$X = 0.1U_1 + 0.1U_2 + 0.1U_3 + \varepsilon_X$$
$$U_6 = U_3 + \varepsilon_6$$
$$Y = 0.1U_4 + 0.5U_5 + 0.5U_6 + \beta X + \varepsilon_Y,$$

where $\varepsilon_X, \varepsilon_6, \varepsilon_Y \sim N(0,1)$ all independent. Covariates U_1, U_2, U_3, U_4, U_5 are also $N(0,1)$ and are all independent except U_1, U_2 and U_1, U_4 for which we have $Cov(U_1, U_2) = 0.7$ and $Cov(U_1, U_4) = 0.7$. Notice that U_1 is a collider between U_2 and U_4 and thus $Cov(U_2, U_4) = 0$.

For each of the 24 simulation scenarios, we randomly generated 500 datasets. We estimated the average causal effect of exposure using 8 different procedures: 1) the

true outcome model, 2) the fully adjusted model, 3) Bayesian model averaging (BMA) with a uniform prior distribution on the outcome model, 4) Bayesian adjustment for confounding (BAC) with ω chosen with cross-validation criterion $C_V^m(\omega)$ proposed in Lefebvre *et al.* (2014a), 5) BAC with $\omega = \infty$, 6) Two-stage Bayesian adjustment for confounding (TBAC) with $\omega = \infty$, 7) N-BCEE and 8) A-BCEE. For both N-BCEE and A-BCEE, we used $\omega = c\sqrt{n}$ and considered $c = 100$, $c = 500$ and $c = 1000$. For each scenario and each method of estimation, we computed the average causal effect estimate ($Mean$), the average standard error estimate ($\overline{SEE}$), the standard deviation of the estimates (SDE), the root mean squared error ($\sqrt{MSE}$) and the coverage probability of 95% confidence intervals (CP). All 95% confidence intervals were computed using the normal approximation $\hat{\beta} \pm 1.96SEE$. Tables 4.3, 4.4, 4.5 and 4.6 summarize the results for $\beta = 0.1$. The marginal posterior probability of inclusion of each potential confounding covariate can be found in Tables 4.11 to 4.14 in Appendix 4.7.5. The results for $\beta = 0$ are similar and are presented in Appendix 4.7.6, Tables 4.16-4.19.

Table 4.3: Comparison of estimates of β obtained from the true outcome model, the fully adjusted model, BMA, BAC, TBAC, N-BCEE and A-BCEE for 500 Monte Carlo replicates of the first data-generating process (DGP1).

n	*Method*	*Mean*	$\overline{SEE}$	SDE	$\sqrt{MSE}$	CP
200	True model	0.100	0.045	0.047	0.047	94
200	Fully adjusted model	0.098	0.072	0.074	0.074	94
200	BMA	0.113	0.047	0.047	0.048	95
200	BAC ($C_V^m(\omega)$)	0.104	0.055	0.064	0.064	92
200	BAC ($\omega = \infty$)	0.098	0.072	0.074	0.074	94
200	TBAC ($\omega = \infty$)	0.098	0.072	0.074	0.074	94
200	N-BCEE (c = 100)	0.108	0.051	0.055	0.056	93
200	N-BCEE (c = 500)	0.104	0.055	0.062	0.062	92
200	N-BCEE (c = 1000)	0.102	0.057	0.065	0.065	93
200	A-BCEE (c = 100)	0.107	0.055	0.054	0.054	95
200	A-BCEE (c = 500)	0.104	0.061	0.060	0.060	96
200	A-BCEE (c = 1000)	0.103	0.063	0.063	0.063	96
600	True model	0.100	0.026	0.025	0.025	96
600	Fully adjusted model	0.100	0.041	0.039	0.039	97
600	BMA	0.111	0.027	0.027	0.029	94
600	BAC ($C_V^m(\omega)$)	0.105	0.031	0.035	0.035	95
600	BAC ($\omega = \infty$)	0.100	0.041	0.039	0.039	97
600	TBAC ($\omega = \infty$)	0.100	0.041	0.039	0.039	96
600	N-BCEE (c = 100)	0.108	0.029	0.031	0.031	93
600	N-BCEE (c = 500)	0.106	0.030	0.033	0.034	93
600	N-BCEE (c = 1000)	0.105	0.031	0.034	0.034	93
600	A-BCEE (c = 100)	0.108	0.030	0.030	0.031	95
600	A-BCEE (c = 500)	0.105	0.033	0.032	0.032	97
600	A-BCEE (c = 1000)	0.105	0.035	0.033	0.033	97
1000	True model	0.101	0.020	0.020	0.020	95
1000	Fully adjusted model	0.100	0.032	0.033	0.033	94
1000	BMA	0.111	0.021	0.022	0.025	92
1000	BAC ($C_V^m(\omega)$)	0.102	0.026	0.030	0.030	93
1000	BAC ($\omega = \infty$)	0.100	0.032	0.033	0.033	94
1000	TBAC ($\omega = \infty$)	0.100	0.032	0.033	0.033	94
1000	N-BCEE (c = 100)	0.107	0.022	0.024	0.025	94
1000	N-BCEE (c = 500)	0.105	0.023	0.026	0.026	94
1000	N-BCEE (c = 1000)	0.104	0.024	0.026	0.027	94
1000	A-BCEE (c = 100)	0.107	0.023	0.024	0.025	95
1000	A-BCEE (c = 500)	0.105	0.026	0.026	0.026	96
1000	A-BCEE (c = 1000)	0.104	0.027	0.027	0.027	96

LEGEND: $Mean$ is the mean estimated value of β where the true value is 0.1, $\overline{SEE}$ is the mean standard error estimate, SDE is the standard deviation of the estimates of β, $\sqrt{MSE}$ is the squared-root of the mean squared error, CP is the coverage probability in % of 95% confidence intervals.

Table 4.4: Comparison of estimates of β obtained from the true outcome model, the fully adjusted model, BMA, BAC, TBAC, N-BCEE and A-BCEE for 500 Monte Carlo replicates of the second data-generating process (DGP2).

n	*Method*	*Mean*	$\overline{SEE}$	*SDE*	$\sqrt{MSE}$	*CP*
200	True model	0.102	0.046	0.045	0.045	96
200	Fully adjusted model	0.104	0.075	0.078	0.078	94
200	BMA	0.148	0.044	0.046	0.067	68
200	BAC ($C_V^m(\omega)$)	0.118	0.052	0.075	0.077	76
200	BAC ($\omega = \infty$)	0.103	0.073	0.077	0.077	95
200	TBAC ($\omega = \infty$)	0.103	0.073	0.076	0.076	95
200	N-BCEE (c = 100)	0.120	0.053	0.068	0.071	83
200	N-BCEE (c = 500)	0.110	0.058	0.073	0.074	86
200	N-BCEE (c = 1000)	0.107	0.060	0.074	0.074	88
200	A-BCEE (c = 100)	0.120	0.062	0.066	0.069	92
200	A-BCEE (c = 500)	0.112	0.067	0.071	0.072	95
200	A-BCEE (c = 1000)	0.110	0.068	0.072	0.073	95
600	True model	0.100	0.026	0.026	0.026	96
600	Fully adjusted model	0.102	0.042	0.041	0.041	95
600	BMA	0.133	0.030	0.032	0.046	70
600	BAC ($C_V^m(\omega)$)	0.106	0.036	0.042	0.042	85
600	BAC ($\omega = \infty$)	0.102	0.041	0.041	0.041	96
600	TBAC ($\omega = \infty$)	0.102	0.041	0.041	0.041	96
600	N-BCEE (c = 100)	0.114	0.032	0.037	0.040	86
600	N-BCEE (c = 500)	0.108	0.034	0.039	0.039	91
600	N-BCEE (c = 1000)	0.106	0.035	0.039	0.040	91
600	A-BCEE (c = 100)	0.114	0.036	0.037	0.039	92
600	A-BCEE (c = 500)	0.109	0.038	0.038	0.039	94
600	A-BCEE (c = 1000)	0.107	0.039	0.039	0.039	94
1000	True model	0.100	0.020	0.021	0.021	95
1000	Fully adjusted model	0.100	0.032	0.032	0.032	95
1000	BMA	0.121	0.024	0.027	0.034	80
1000	BAC ($C_V^m(\omega)$)	0.100	0.029	0.031	0.031	92
1000	BAC ($\omega = \infty$)	0.099	0.032	0.032	0.031	95
1000	TBAC ($\omega = \infty$)	0.099	0.032	0.032	0.031	95
1000	N-BCEE (c = 100)	0.107	0.025	0.029	0.029	90
1000	N-BCEE (c = 500)	0.103	0.026	0.029	0.029	90
1000	N-BCEE (c = 1000)	0.102	0.026	0.030	0.030	91
1000	A-BCEE (c = 100)	0.108	0.028	0.028	0.029	93
1000	A-BCEE (c = 500)	0.104	0.029	0.029	0.029	94
1000	A-BCEE (c = 1000)	0.103	0.030	0.029	0.030	95

LEGEND: See Table 4.3.

Table 4.5: Comparison of estimates of β obtained from the true outcome model, the fully adjusted model, BMA, BAC, TBAC, N-BCEE and A-BCEE for 500 Monte Carlo replicates of the third data-generating process (DGP3).

n	*Method*	*Mean*	$\overline{SEE}$	*SDE*	$\sqrt{MSE}$	*CP*
200	True model	0.103	0.100	0.100	0.100	95
200	Fully adjusted model	0.101	0.105	0.104	0.104	96
200	BMA	0.149	0.085	0.086	0.099	89
200	BAC ($C_V^m(\omega)$)	0.116	0.087	0.103	0.104	90
200	BAC ($\omega = \infty$)	0.101	0.100	0.101	0.101	95
200	TBAC ($\omega = \infty$)	0.102	0.101	0.101	0.101	95
200	N-BCEE (c = 100)	0.113	0.093	0.100	0.101	93
200	N-BCEE (c = 500)	0.106	0.096	0.101	0.101	94
200	N-BCEE (c = 1000)	0.104	0.097	0.101	0.101	94
200	A-BCEE (c = 100)	0.116	0.096	0.098	0.099	95
200	A-BCEE (c = 500)	0.109	0.098	0.099	0.100	95
200	A-BCEE (c = 1000)	0.108	0.099	0.100	0.100	95
600	True model	0.098	0.057	0.060	0.060	96
600	Fully adjusted model	0.098	0.058	0.061	0.061	96
600	BMA	0.138	0.054	0.061	0.072	80
600	BAC ($C_V^m(\omega)$)	0.104	0.054	0.065	0.065	87
600	BAC ($\omega = \infty$)	0.097	0.058	0.060	0.060	96
600	TBAC ($\omega = \infty$)	0.097	0.057	0.060	0.060	95
600	N-BCEE (c = 100)	0.108	0.056	0.064	0.064	88
600	N-BCEE (c = 500)	0.101	0.056	0.062	0.062	92
600	N-BCEE (c = 1000)	0.100	0.056	0.061	0.061	92
600	A-BCEE (c = 100)	0.111	0.057	0.063	0.064	90
600	A-BCEE (c = 500)	0.104	0.057	0.062	0.062	92
600	A-BCEE (c = 1000)	0.103	0.057	0.062	0.062	94
1000	True model	0.098	0.044	0.043	0.043	96
1000	Fully adjusted model	0.098	0.045	0.043	0.043	95
1000	BMA	0.130	0.045	0.050	0.058	79
1000	BAC ($C_V^m(\omega)$)	0.102	0.043	0.046	0.046	91
1000	BAC ($\omega = \infty$)	0.098	0.045	0.043	0.043	96
1000	TBAC ($\omega = \infty$)	0.098	0.044	0.043	0.043	96
1000	N-BCEE (c = 100)	0.106	0.044	0.048	0.048	91
1000	N-BCEE (c = 500)	0.101	0.044	0.045	0.045	93
1000	N-BCEE (c = 1000)	0.100	0.044	0.044	0.044	94
1000	A-BCEE (c = 100)	0.108	0.045	0.048	0.048	92
1000	A-BCEE (c = 500)	0.103	0.045	0.046	0.046	94
1000	A-BCEE (c = 1000)	0.102	0.045	0.045	0.045	94

LEGEND: See Table 4.3.

Table 4.6: Comparison of estimates of β obtained from the true outcome model, the fully adjusted model, BMA, BAC, TBAC, N-BCEE and A-BCEE for 500 Monte Carlo replicates of the fourth data-generating process (DGP4).

n	*Method*	*Mean*	$\overline{SEE}$	*SDE*	$\sqrt{MSE}$	*CP*
200	True model	0.103	0.054	0.052	0.052	96
200	Fully adjusted model	0.105	0.072	0.068	0.068	95
200	BMA	0.119	0.060	0.054	0.057	96
200	BAC ($C_V^m(\omega)$)	0.110	0.061	0.062	0.063	95
200	BAC ($\omega = \infty$)	0.103	0.072	0.068	0.068	95
200	TBAC ($\omega = \infty$)	0.105	0.071	0.067	0.067	96
200	N-BCEE (c = 100)	0.108	0.061	0.063	0.064	93
200	N-BCEE (c = 500)	0.106	0.064	0.066	0.066	94
200	N-BCEE (c = 1000)	0.105	0.065	0.066	0.066	95
200	A-BCEE (c = 100)	0.110	0.066	0.062	0.062	96
200	A-BCEE (c = 500)	0.107	0.068	0.064	0.064	96
200	A-BCEE (c = 1000)	0.107	0.068	0.065	0.065	96
600	True model	0.099	0.031	0.031	0.031	95
600	Fully adjusted model	0.097	0.041	0.043	0.043	95
600	BMA	0.110	0.036	0.036	0.038	92
600	BAC ($C_V^m(\omega)$)	0.100	0.037	0.042	0.042	92
600	BAC ($\omega = \infty$)	0.096	0.041	0.043	0.043	95
600	TBAC ($\omega = \infty$)	0.096	0.041	0.042	0.043	94
600	N-BCEE (c = 100)	0.102	0.036	0.040	0.040	92
600	N-BCEE (c = 500)	0.099	0.037	0.041	0.041	92
600	N-BCEE (c = 1000)	0.098	0.037	0.042	0.042	92
600	A-BCEE (c = 100)	0.102	0.038	0.040	0.040	94
600	A-BCEE (c = 500)	0.100	0.039	0.041	0.041	94
600	A-BCEE (c = 1000)	0.099	0.040	0.041	0.041	94
1000	True model	0.099	0.024	0.024	0.024	96
1000	Fully adjusted model	0.099	0.032	0.032	0.032	95
1000	BMA	0.107	0.028	0.029	0.030	92
1000	BAC ($C_V^m(\omega)$)	0.100	0.029	0.032	0.032	91
1000	BAC ($\omega = \infty$)	0.098	0.032	0.032	0.032	94
1000	TBAC ($\omega = \infty$)	0.098	0.032	0.032	0.032	93
1000	N-BCEE (c = 100)	0.102	0.028	0.030	0.030	92
1000	N-BCEE (c = 500)	0.100	0.028	0.031	0.031	92
1000	N-BCEE (c = 1000)	0.100	0.029	0.032	0.031	92
1000	A-BCEE (c = 100)	0.102	0.030	0.031	0.031	94
1000	A-BCEE (c = 500)	0.101	0.030	0.031	0.031	94
1000	A-BCEE (c = 1000)	0.100	0.031	0.032	0.032	94

LEGEND: See Table 4.3.

We start by discussing the results pertaining to non-BCEE methods for estimating the average causal effect of exposure. Then, we discuss the results for BCEE and contrast them to the former results.

As expected, Bayesian model averaging (BMA) can perform very poorly to estimate the average causal effect. More precisely, the simulation results show that the bias can be substantial when the most important confounding covariates are only slightly associated with the outcome (DGP2 and DGP3). For instance, in DGP2, U_3 and U_4 are important confounding covariates often excluded by BMA (see Table 4.12 in Appendix 4.7.5). This situation also yields confidence intervals with poor coverage probabilities. Although increasing sample size seems to reduce the bias, the coverage probability remains mostly unchanged. In situations where the most important confounding covariates are strongly associated with the outcome (DGP1 and DGP4), BMA performs very well both in terms of mean squared error (MSE) and coverage probability.

The simulation results also support the claim that BAC and TBAC with $\omega = \infty$ do not yield a notable reduction in the variance of the estimated causal effect as compared to the fully adjusted model. This is partly due to the fact that BAC and TBAC tend to include more covariates than needed to achieve unbiasedness (see Appendix 4.7.5). Moreover, using BAC with cross-validation criterion $C_V^m(\omega)$ gives relatively poor results. Even though this method sometimes gives smaller MSE than BAC with $\omega = \infty$, the estimated standard error remarkably underestimate the true standard error (the standard deviation of the estimates of β). One possible explanation for this underestimation is that BAC with $C_V^m(\omega)$ neglects the uncertainty associated with the choice of the hyperparameter ω.

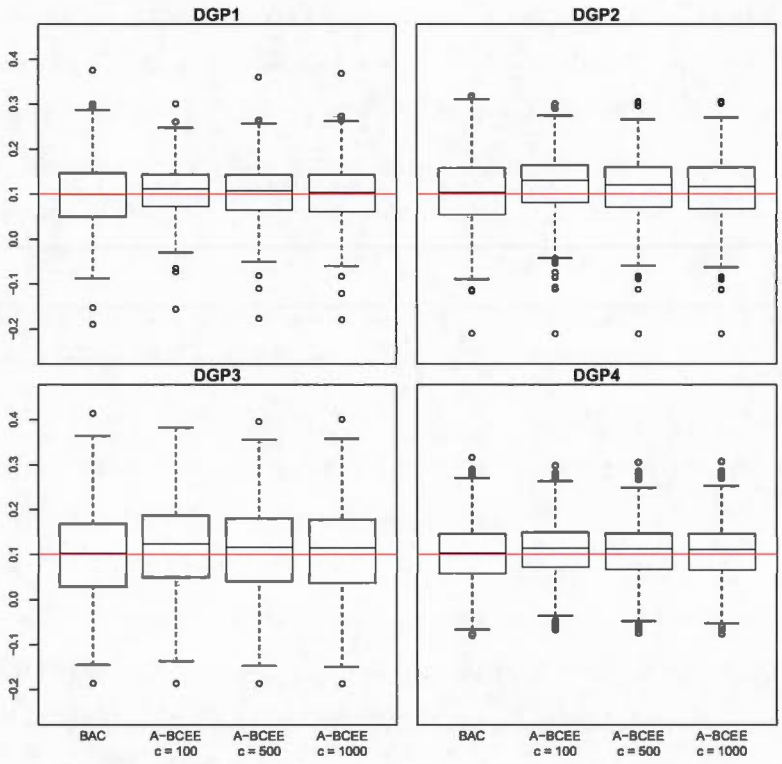
The simulation results show that the choice of using $\omega = c\sqrt{n}$, $c \in [100, 1000]$, for A-BCEE and N-BCEE is reasonable. The results do not appear too sensitive to the choice of c in this interval. The simulation results also confirm that N-BCEE can

yield lower than expected coverage probabilities. This seems to be particularly true in complex scenarios that contain many covariates, such as DGP2, DGP3 and DGP4.

Despite sometimes producing slightly biased estimates, A-BCEE performs at least as well as BAC and TBAC with $\omega = \infty$ in terms of MSE. The bias is small enough that in all simulation scenarios we considered, A-BCEE (with any c) yields appropriate coverage probability. In general, A-BCEE gives less weight to variables only associated with the exposure than BAC and TBAC (see Appendix 4.7.5). In DGP1, A-BCEE outperforms BAC and TBAC with $\omega = \infty$ in terms of MSE. In DGP2 and DGP4, A-BCEE has smaller MSE than BAC and TBAC although comparatively to a lesser extent. Results are quite similar between BAC, TBAC and A-BCEE in DGP3. Note that in DGP3, the true model and the fully adjusted model have the same MSE. There is thus no possible gain in using another model than the fully adjusted model. Figure 4.1 illustrates the distribution of $\hat{\beta}$ obtained by using A-BCEE and BAC with $\omega = \infty$ for all four data-generating processes with $n = 200$ (analogous figures are displayed in Appendix 4.7.8 with $n = 600$ and $n = 1000$). This figure shows how estimates obtained with A-BCEE, despite being slightly biased, are more concentrated around the true value β than estimates obtained with BAC. Moreover, Figure 4.1 illustrates the bias-variance tradeoff associated with the choice of c in A-BCEE: smaller values of c, as compared to larger values of c, favor a reduced variance in the estimator of the causal effect at the cost of an increase in bias.

On the basis of these results, we hypothesized that BCEE would perform best when 1) there are some direct causes of the exposure that are strongly associated with the exposure, and 2) there exists variables that can d-separate those direct causes from the outcome. In such situations, we expect BCEE to favor models excluding those direct causes and including the d-separating variables. To verify this, we simulated data according to a fifth data-generating scenario (DGP5) which meets these two conditions.

Figure 4.1: Comparison of the distribution of $\hat{\beta}$ obtained from A-BCEE and BAC ($\omega = \infty$) for all four data-generating processes and a sample size $n = 200$. The red line corresponds to the true value $\beta = 0.1$



The equations for DGP5 are:

$$U_5 = U_1 + U_2 + U_3 + U_4 + \varepsilon_5$$
$$X = U_1 + U_2 + U_3 + U_4 + \varepsilon_X$$
$$Y = U_5 + \beta X + \varepsilon_Y,$$

where $\varepsilon_5, \varepsilon_X, \varepsilon_Y \sim$ N(0,1), all independent. In this example, BCEE's prior distribution, $P^B(\boldsymbol{\alpha}^{\boldsymbol{Y}})$, is devised to give non negligible prior weight to the two following sufficient outcome models : (i) the one including $\{U_1, U_2, U_3, U_4\}$, and (ii) the one including only

$\{U_5\}$. However, because the marginal likelihood of the model (ii) should dominate the one of model (i) for large n, we expect the second outcome model to receive increased posterior weight as n grows. To reduce computational burden, we only considered $\beta = 0.1$ and did not estimate β with N-BCEE. The results are presented in Table 4.7. Those results show how under such ideal conditions, the MSE obtained by using A-BCEE is much smaller than the one obtained using the fully adjusted outcome model, BAC or TBAC. In fact, A-BCEE's MSE is similar to the MSE of the true outcome model. Moreover, Table 4.15 in Appendix 4.7.5 reveals that models including U_5, but excluding U_1, U_2, U_3 and U_4 are favored by A-BCEE, particularly for the larger sample sizes. Indeed, the marginal posterior probabilities of covariates U_1 to U_4 decrease with sample size while the posterior probability of U_5 remains at 1 for all sample sizes considered. This is as opposed to BAC and TBAC where the full model (including U_1 to U_5) receives a posterior probability of 1 at all sample sizes.

Table 4.7: Comparison of estimates of β obtained from the true outcome model, the fully adjusted model, BMA, BAC, TBAC and A-BCEE for 500 Monte Carlo replicates of the fifth data-generating process (DGP5).

n	Method	Estimate	$\overline{SEE}$	SDE	$\sqrt{MSE}$	CP
200	True model	0.103	0.053	0.054	0.054	92
200	Fully adjusted model	0.102	0.072	0.076	0.076	94
200	BMA	0.103	0.054	0.055	0.055	93
200	BAC ($C_V^m(\omega)$)	0.102	0.059	0.066	0.066	92
200	BAC ($\omega = \infty$)	0.102	0.072	0.076	0.076	94
200	TBAC ($\omega = \infty$)	0.102	0.072	0.076	0.076	95
200	A-BCEE (c = 100)	0.103	0.055	0.056	0.056	93
200	A-BCEE (c = 500)	0.103	0.059	0.059	0.059	94
200	A-BCEE (c = 1000)	0.102	0.061	0.061	0.061	95
600	True model	0.099	0.031	0.029	0.029	96
600	Fully adjusted model	0.097	0.041	0.040	0.040	96
600	BMA	0.099	0.031	0.029	0.029	96
600	BAC ($C_V^m(\omega)$)	0.097	0.034	0.036	0.036	95
600	BAC ($\omega = \infty$)	0.097	0.041	0.040	0.040	96
600	TBAC ($\omega = \infty$)	0.097	0.041	0.040	0.040	96
600	A-BCEE (c = 100)	0.098	0.031	0.030	0.030	96
600	A-BCEE (c = 500)	0.098	0.033	0.030	0.030	97
600	A-BCEE (c = 1000)	0.098	0.034	0.031	0.031	97
1000	True model	0.100	0.024	0.023	0.023	95
1000	Fully adjusted model	0.100	0.032	0.031	0.031	94
1000	BMA	0.100	0.024	0.023	0.023	96
1000	BAC ($C_V^m(\omega)$)	0.101	0.027	0.027	0.027	95
1000	BAC ($\omega = \infty$)	0.100	0.032	0.031	0.031	94
1000	TBAC ($\omega = \infty$)	0.100	0.032	0.031	0.031	95
1000	A-BCEE (c = 100)	0.100	0.024	0.023	0.023	96
1000	A-BCEE (c = 500)	0.100	0.025	0.023	0.023	96
1000	A-BCEE (c = 1000)	0.100	0.025	0.023	0.023	96

LEGEND: See Table 4.3.

4.4.2 Additional simulations

We now address the secondary objectives with additional simulations. The first secondary goal is to study the large, whilst finite, sample properties of BCEE. To do this, we examine four different simulation scenarios obtained by considering the four data-generating processes (DGP1, DGP2, DGP3 and DGP4) with a sample size of 10,000. Once again, for each scenario, we randomly generated 500 datasets. We estimated the average causal effect of exposure using A-BCEE and N-BCEE with $\omega = c\sqrt{n}$. Because the sample size is large and the computational burden is heavy, we considered only one value of c ($c = 500$). The results are shown in Table 4.8. These simulations suggest that A-BCEE and N-BCEE with $\omega = c\sqrt{n}$ unbiasedly estimate the causal effect of exposure when n is large and when BCEE's working assumptions hold (i.e., $\boldsymbol{U}$ includes all direct causes of X and the normal linear model is a correct specification for both X and Y).

Table 4.8: Estimates of β for N-BCEE and A-BCEE with a sample size of n = 10,000 for 500 Monte Carlo replicates of each data-generating process.

DGP	$Method$	$Mean$	$\overline{SEE}$	SDE	$\sqrt{MSE}$	CP
1	N-BCEE (c = 500)	0.100	0.0068	0.0067	0.0067	97
1	A-BCEE (c = 500)	0.100	0.0069	0.0066	0.0066	97
2	N-BCEE (c = 500)	0.100	0.0074	0.0075	0.0075	96
2	A-BCEE (c = 500)	0.100	0.0079	0.0075	0.0075	98
3	N-BCEE (c = 500)	0.100	0.0140	0.0141	0.0141	95
3	A-BCEE (c = 500)	0.100	0.0140	0.0141	0.0141	95
4	N-BCEE (c = 500)	0.099	0.0083	0.0084	0.0084	96
4	A-BCEE (c = 500)	0.099	0.0089	0.0086	0.0086	97

LEGEND: DGP is the data-generating process, $Mean$ is the mean estimated value of β where the true value is 0.1, $\overline{SEE}$ is the mean standard error estimate, SDE is the standard deviation of the estimates of β, $\sqrt{MSE}$ is the squared-root of the mean squared error, CP is the coverage probability in % of 95% confidence intervals.

The second secondary objective is to study the performance of the bootstrapped BCEE (B-BCEE) to correct the confidence intervals of N-BCEE. Since this bootstrapped implementation is very computationally intensive, we only considered two simulation

scenarios: DGP1 and DGP4 with a sample size of 200. In this case, only 100 datasets were generated for each scenario. We estimated the causal effect of exposure using the fully adjusted model, the true outcome model, BMA, BAC, TBAC, A-BCEE, N-BCEE and B-BCEE. For A-BCEE, N-BCEE and B-BCEE we took $\omega = c\sqrt{n}$ with $c = 500$. For B-BCEE we performed 200 bootstrap resamplings and considered an estimate with and without a bias correction. The results are presented in Table 4.9. We find that the non-parametric bootstrap implementation of BCEE yields correct estimates of the standard error of estimate and correct coverage probabilities. However, B-BCEE does not seem to be as efficient nor as practical as A-BCEE.

Table 4.9: Comparison of estimates of β obtained from the true model, the fully adjusted model, BMA, BAC, TBAC, N-BCEE, A-BCEE and B-BCEE for the first and fourth data-generating processes (DGP1 and DGP4). Sample size is $n = 200$, 100 datasets were generated for each data-generating process. For B-BCEE, 200 bootstrap resamplings were performed.

DGP	*Method*	*Mean*	$\overline{SEE}$	*SDE*	$\sqrt{MSE}$	*CP*
1	True model	0.105	0.045	0.053	0.053	93
1	Fully adjusted model	0.104	0.072	0.075	0.075	94
1	BMA	0.121	0.048	0.050	0.054	93
1	BAC ($\omega = \infty$)	0.104	0.072	0.075	0.075	94
1	TBAC ($\omega = \infty$)	0.104	0.072	0.075	0.074	93
1	N-BCEE (c = 500)	0.112	0.056	0.063	0.064	92
1	A-BCEE (c = 500)	0.111	0.062	0.061	0.062	96
1	B-BCEE (c = 500, no bias corr.)	0.112	0.067	0.063	0.064	96
1	B-BCEE (c = 500, w/ bias corr.)	0.107	0.067	0.066	0.066	96
4	True model	0.111	0.063	0.063	0.063	96
4	Fully adjusted model	0.106	0.072	0.064	0.064	96
4	BMA	0.120	0.060	0.051	0.055	96
4	BAC ($\omega = \infty$)	0.105	0.072	0.064	0.064	96
4	TBAC ($\omega = \infty$)	0.106	0.071	0.064	0.063	96
4	N-BCEE (c = 500)	0.108	0.064	0.061	0.061	97
4	A-BCEE (c = 500)	0.109	0.068	0.060	0.060	96
4	B-BCEE (c = 500, no bias corr.)	0.108	0.071	0.061	0.061	97
4	B-BCEE (c = 500, w/ bias corr.)	0.107	0.071	0.062	0.062	98

LEGEND: See Table 4.8

The last secondary goal is to verify if BCEE is robust to non normality of errors. To do so, we considered 6 simulation scenarios built by considering 2 factors: data-generating

process (DGP6, DGP7 and DGP8) and sample size (200, 600, 1000). The data-generating processes DPG6, DGP7 and DGP8 mimic DGP1, DGP2 and DGP4 respectively but all the ε error terms have the following shifted exponential distribution $f(\varepsilon) = \exp(\varepsilon + 1), \varepsilon \geq -1$, instead of a $N(0, 1)$ distribution. The results from these additional simulations are presented in Appendix 4.7.7. The results are very similar to the ones presented in Tables 4.3, 4.4 and 4.6 and suggest that BCEE is very robust to deviations from normality.

4.5 Application: Estimation of the causal effect of perceived mathematical competence on grades in mathematics

In this section we use A-BCEE to estimate the causal effect of perceived competence in mathematics (measured on a scale from 1 to 7) on self-reported grades (in %) in mathematics. We consider longitudinal data obtained from 1430 students during their first three years of highschool. Participants lived in various regions throughout Quebec, Canada. The data were collected by postal questionnaires every year for a period of three years (time 1, time 2 and time 3). Further details can be found in Guay *et al.* (2011).

We used measures of perceived competence in mathematics at time 2 as the exposure and grades in mathematics at time 3 as the outcome to estimate the causal effect of interest. Recall that A-BCEE requires specifying a set of potential confounding covariates that includes all direct causes of the exposure and none of its descendants. Moreover, it is beneficial that this set also includes strong predictors of the outcome. We took advantage of the longitudinal feature of the data to build the set of potential confounding covariates. Because a cause always precedes its effect in time, we constructed the set of potential confounding covariates by including variables at time 1 that were potential direct causes of perceived-competence at time 2. We also included variables at time 2 that were thought to be strong predictors of grades in mathematics at time 3.

We selected the following 26 covariates: gender, highest level of education reached by the mother, highest level of education reached by the father, perceived competence in mathematics (at time 1), perceived autonomy support from the mother, perceived autonomy support from the father, perceived autonomy support from the mathematics teacher, perceived autonomy support from friends at school, self-reported mathematics' grades, intrinsic motivation in mathematics, identified motivation in mathematics, introjected motivation in mathematics, externally regulated motivation in mathematics, victimization and sense of belonging to school. All variables except the first four were considered both at times 1 and 2.

Before applying A-BCEE on these data, we obtained some descriptive statistics. We drew scatter plots of the outcome versus the exposure and versus each potential confounding covariate to roughly verify the linearity assumption and to check for outliers. For the same reasons, we drew scatter plots of the exposure versus each potential confounding covariate. We also noticed that only 46.5% of the participants have complete information for all the selected covariates. The variables measured at time 1 have generally few missing cases (between 1.8% and 8.3%), but the variables measured at time 2 and 3 have a larger degree of missingness (between 26.4% and 36.4%). We performed multiple imputation (Little & Rubin, 2002) to account for the missing data, using 50 imputed datasets to ensure the power falloff is negligible (Graham *et al.*, 2007).

We estimated the causal effect of perceived competence on grades in mathematics using the fully adjusted outcome model, A-BCEE with $\omega = c\sqrt{n}$ (c = 100, 500, 1000), BAC and TBAC (with $\omega = \infty$). Results are summarized in Table 4.10. The computational burden of BCEE on these data is manageable and comparable to the one of TBAC, although quite heavier than the one of BAC when using the `BACprior` package (Talbot *et al.*, 2014). The approximate running times of A-BCEE, BAC, and TBAC on one imputed dataset are respectively 22.5 minutes, 1.2 minutes, and 21.2 minutes on a PC with 2.4GHz and 8 Gb RAM.

Because Step S1 of A-BCEE aims to find the direct causes of the exposure, it is reasonable to only allow covariates measured before the exposure to be selected in this step. Hence, we ran the A-BCEE algorithm a second time, but this time excluding the possibility that covariates measured at time 2 enter the exposure model. We denote this implementation of A-BCEE as A-BCEE* in Table 4.10.

Table 4.10: Comparison of the estimated causal effect of perceived mathematical competence in mathematics on self-reported mathematics' grades.

Method	*Estimate*	*SEE*	*CI*
Fully adjusted model	0.693	0.460	(-0.208, 1.594)
BAC ($\omega = \infty$)	0.729	0.462	(-0.178, 1.635)
TBAC ($\omega = \infty$)	0.778	0.465	(-0.133, 1.690)
A-BCEE (c = 100)	0.807	0.451	(-0.076, 1.691)
A-BCEE (c = 500)	0.790	0.456	(-0.105, 1.685)
A-BCEE (c = 1000)	0.786	0.459	(-0.113, 1.685)
A-BCEE* (c = 100)	0.823	0.445	(-0.049, 1.696)
A-BCEE* (c = 500)	0.808	0.444	(-0.062, 1.679)
A-BCEE* (c = 1000)	0.803	0.444	(-0.066, 1.673)

LEGEND: *Estimate* is the estimated causal effect, SEE is the standard error estimate, CI is a 95% confidence interval for the causal effect.

Table 4.10 shows that the results from A-BCEE and A-BCEE* are very similar. This is not surprising since the marginal posterior probability of inclusion of covariates do not differ much between A-BCEE and A-BCEE* (not shown). Using A-BCEE instead of the fully adjusted model slightly decreases the standard error of estimate, between 0.3% and 3.5%, which translates in a small decrease of the 95% confidence intervals' width. Moreover the standard errors of estimate for BAC and TBAC are slightly larger than the one for the fully adjusted model in this illustration. Although the point estimates appear to vary substantially between methods, the differences are small relative to the magnitude of the estimated standard errors. We conclude that perceived competence in mathematics at one point in time likely has little or no causal effect on self-reported grades in mathematics a year later.

4.6 Discussion

We have introduced the Bayesian causal effect estimation (BCEE) algorithm to estimate causal exposure effects in observational studies. This novel data-driven approach avoids the need to rely on the specification of a causal graph and aims to control the variability of the estimator of the exposure effect. BCEE employs a prior distribution that is motivated by a theoretical proposition embedded the graphical framework to causal inference. We also proposed a practical implementation of BCEE, A-BCEE, that accounts for the fact that this prior distribution uses information from the data. Using simulation studies, we found that A-BCEE generally achieves at least some reduction of the MSE of the causal effect estimator as compared to the one generated by a fully-adjusted model approach or by other data-driven approaches to causal inference, such as BAC and TBAC, thus resulting in estimates that are overall closer to the true value. In some circumstances, the reduction of the MSE can be substantial. Moreover, confidence intervals with appropriate coverage probabilities were obtained. Hence, we believe that BCEE is a promising algorithm to perform causal inference.

Some current limitations of BCEE could be addressed in future research. The generalization to non continuous exposure variable (e.g. binary) is straightforward. Recall that the first step of BCEE aims at identifying the direct causes of the exposure. As in the normal case we have considered, classical Bayesian procedures asymptotically select the true exposure model with probability 1 when assuming X belongs to an exponential family (e.g. Bernoulli) and that an adequate parametric model is considered (Haughton, 1988). The generalization of BCEE to other types of outcome variables is less straightforward. One could specify a generalized linear model for the outcome of the form $g(\mathbb{E}[Y_i|X_i, \boldsymbol{U_i}]) = \delta_0 + \beta X_i + \sum_{m=1}^{M} \delta_m U_{im}$. However, unless g is the identity or the log link, such models are generally not collapsible for β over covariate U_m (Greenland *et al.*, 1999). In other words, the true value of β, and thus its interpretation, depends on whether U_m is included or not in the outcome model,

even when U_m is a not confounding covariate. In such circumstances, averaging the estimated value of β over different outcome models would not be advisable.

We think that BCEE can be particularly helpful to those working in fields where current subject-matter knowledge is sparse. To facilitate usage of the BCEE algorithm, we provide an R package named `BCEE` (available at http://cran.r-project.org).

4.7 Appendix

4.7.1 Back-door adjustment and linear regression adjustment

We describe how the distribution-free back-door adjustment is related to the linear regression adjustment. Assume $\boldsymbol{Z} = \{Z_1, ..., Z_K\}$ is a sufficient set to identify the causal effect of X on Y according to the back-door criterion. We consider the average causal effect of a unit increase of the exposure on the outcome using Pearl (2009)'s do-calculus:

$$
\begin{aligned}
\beta &= \mathbb{E}[Y|do(X = x+1)] - \mathbb{E}[Y|do(X = x)] && (4.11)\\
&= \sum_y yP(Y = y|do(X = x+1)) - \sum_y yP(Y = y|do(X = x))\\
&= \sum_y y \sum_{\boldsymbol{z}} P(Y = y|X = x+1, \boldsymbol{Z} = \boldsymbol{z})P(\boldsymbol{Z} = \boldsymbol{z})\\
&\quad - \sum_y y \sum_{\boldsymbol{z}} P(Y = y|X = x, \boldsymbol{Z} = \boldsymbol{z})P(\boldsymbol{Z} = \boldsymbol{z}) && \text{(Back-door adjustment)}\\
&= \sum_{\boldsymbol{z}} \left[\sum_y yP(Y = y|X = x+1, \boldsymbol{Z} = \boldsymbol{z}) - \sum_y yP(Y = y|X = x, \boldsymbol{Z} = \boldsymbol{z}) \right] P(\boldsymbol{Z} = \boldsymbol{z})\\
&= \mathbb{E}_{\boldsymbol{z}} \left[\mathbb{E}[Y|X = x+1, \boldsymbol{Z} = \boldsymbol{z}] - \mathbb{E}[Y|X = x, \boldsymbol{Z} = \boldsymbol{z}] \right]. && (4.12)
\end{aligned}
$$

We now show that the regression coefficient associated with the exposure in the linear regression model of Y on X and $\boldsymbol{Z}$ corresponds to β in (4.11) if the model is correctly specified. Since we assumed the model is correct, $\mathbb{E}[Y|X = x, \boldsymbol{Z} = \boldsymbol{z}] = \delta_0 + \beta^* x + \sum_{k=1}^{K} \delta_k z_k$, where the symbolic β^* is used to reflect the fact that the exposure effect may not be causal. Inserting this in (4.12) yields:

$$\beta = \mathbb{E}_z \left[\left(\delta_0 + \beta^*(x+1) + \sum_{k=1}^{K} \delta_k z_k \right) - \left(\delta_0 + \beta^* x + \sum_{k=1}^{K} \delta_k z_k \right) \right]$$
$$= \mathbb{E}_z [\beta^*] = \beta^*.$$

We find that the regression coefficient associated with the exposure in the linear regression model of Y on X and $\boldsymbol{Z}$ is indeed the average causal effect. Therefore, if the postulated linear regression model holds, adjusting in the nonparametric manner proposed in the back-door criterion is the same as adjusting in the linear regression.

4.7.2 Proofs

4.7.2.1 Proof of Proposition 4.2.1

Proof First, we know from Pearl (2009) Section 3.3.1 that a set $\boldsymbol{Z}$ is sufficient for identifying the causal effect of an exposure X on an outcome Y if (i) no descendants of X are in $\boldsymbol{Z}$ and (ii) $\boldsymbol{Z}$ blocks all back-door paths between X and Y. According to condition 1 we assume that there are no descendants of X in $\boldsymbol{Z}$. Suppose that G admits some back-door paths. All back-door paths are such that the second variable appearing in the path is a direct cause of X; the back-door paths thus have the form $X \leftarrow D_j \cdots \rightarrow Y$.

Suppose that a direct cause D_j is included in $\boldsymbol{Z}$. Then D_j (and therefore the set $\boldsymbol{Z}$) blocks all back-door paths of the form $X \leftarrow D_j \cdots \rightarrow Y$. Indeed no variable in $\boldsymbol{Z} \backslash D_j$ can reopen a path $X \leftarrow D_j \cdots \rightarrow Y$ once closed by D_j. Therefore, all back-door paths admitting a direct cause in $\boldsymbol{Z}$ are blocked by $\boldsymbol{Z}$.

It remains to show that all back-door paths for which the second variable in the path is not a direct cause included in $\boldsymbol{Z}$ are closed when condition 2b in the proposition holds. Consider $D_j \notin \boldsymbol{Z}$. Now assume that Y and D_j are d-separated by $\{X \cup \boldsymbol{Z}\}$. By the definition of d-separation, this means that every path connecting D_j to Y is blocked by $\{X \cup \boldsymbol{Z}\}$. Recall that all back-door paths associated with this D_j are of the form $X \leftarrow D_j \cdots \rightarrow Y$. Because by (2b) D_j and Y are d-separated by $\{X \cup \boldsymbol{Z}\}$ and since each subpath $D_j \cdots \rightarrow Y$ in these back-door paths does not contain the variable X, these subpaths are blocked by $\boldsymbol{Z}$. This reasoning is applied to each $D_j \notin \boldsymbol{Z}$ separately.

The proof is complete by the back-door criterion as we realize that all back-door paths, whether their D_j is contained in $\boldsymbol{Z}$ or not, are blocked by $\boldsymbol{Z}$. □

4.7.2.2 Proof of Corollary 4.2.1

Proof

1. Suppose that G admits some back-door paths of the form $X \leftarrow D_j \cdots \rightarrow Y$. If D_j and Y are d-separated by $\{X \cup \boldsymbol{Z}'\}$, then by definition of d-separation all paths between D_j and Y are blocked by $\{X \cup \boldsymbol{Z}'\}$. Using the same argument as the one used in the third paragraph of the proof of Proposition 4.2.1, it follows that all back-door paths $X \leftarrow D_j \cdots \rightarrow Y$ are blocked by $\boldsymbol{Z}'$.

2. To prove that $\boldsymbol{Z}'$ is sufficient for estimating the causal effect of X on Y, we show that all back-door paths between X and Y are blocked by $\boldsymbol{Z}'$.

First, we consider the back-door paths that admit D_j as second variable. From point 1. of the corollary, we already know that these back-door paths are blocked.

Next, we divide the back-door paths that do not admit D_j as second variable into two categories: 1) the paths whose second variable is a $D_{j'} \in \boldsymbol{Z}'$, $j' \neq j$, and 2) the paths whose second variable is a $D_{j'} \notin \boldsymbol{Z}'$. For 1), following the same argument as in the second paragraph of the proof of Proposition 4.2.1, we know that all back-door paths whose second variable is a $D_{j'} \in \boldsymbol{Z}'$ are blocked.

The case where the second variable is a $D_{j'} \notin \boldsymbol{Z}'$ is more involved. Here, note that $D_{j'}$ is not in $\boldsymbol{Z}$ either since $\boldsymbol{Z}' = \boldsymbol{Z} \backslash D_j$. The fact that $\boldsymbol{Z}$ is sufficient to identify the average causal effect according to Proposition 4.2.1 implies that $D_{j'}$ and Y are d-separated by $\{X \cup \boldsymbol{Z}\}$. Therefore, every path between $D_{j'}$ and Y is blocked by $\{X \cup \boldsymbol{Z}\}$. For those paths that do not include D_j, it is easy to see that they are also blocked by $\{X \cup \boldsymbol{Z}'\}$. For those paths that include D_j, that is, paths of the form $D_{j'} \cdots D_j \cdots \rightarrow Y$, we know from point 1. that they are blocked in the subpaths $D_j \cdots \rightarrow Y$ by $\{X \cup \boldsymbol{Z}'\}$. Thus, every path between $D_{j'}$ and Y is blocked by $\{X \cup \boldsymbol{Z}'\}$, whether or not it includes D_j. Using the same arguments as the ones used in the third paragraph of the proof of Proposition

4.2.1, it follows that all back-door paths $X \leftarrow D_{j'} \cdots \rightarrow Y$ are blocked by $\{X \cup \boldsymbol{Z}'\}$. The whole reasoning is applied for each possible $D_{j'}$, according to their inclusion or exclusion in $\boldsymbol{Z}'$.

Hence, all back-door paths between X and Y in G are blocked by $\{X \cup \boldsymbol{Z}'\}$. Also, because $\boldsymbol{Z}$ is sufficient to identify the average causal effect according to Proposition 4.2.1, $\boldsymbol{Z}$ does not include any descendants of X and therefore $\boldsymbol{Z}'$ does not either. According to the back-door criterion, $\boldsymbol{Z}'$ is thus sufficient to identify the average causal effect and the proof is complete.

□

4.7.3 General conditions for the equivalence of zero regression coefficient and conditional independence

We show that the independence of Y and U_k conditional on X and $U_1, ..., U_{k-1}, U_{k+1}, ..., U_M$ is equivalent to having regression parameter δ_k associated to U_k in the linear regression of Y on X and $\boldsymbol{U}$ equal to zero under less stringent assumptions than multivariate normality for the covariates X and $\boldsymbol{U}$.

Consider the same normal linear model as in (4.1)

$$Y_i = \delta_0 + \beta X_i + \sum_{m=1}^{M} \delta_m U_{im} + \varepsilon_i,$$

where $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. We assume that this model is correctly specified, that is, the data for Y is generated according to (4.1) with possibly some regression coefficients set to 0. However, we make no assumptions about the distribution of variables X and $\boldsymbol{U}$. To simplify the notation, we denote $\{U_1, ..., U_{k-1}, U_{k+1}, ...U_M\}$ by $\boldsymbol{U}\backslash U_k$. We consider the case where U_k is a continuous variable. Similar arguments can be used when U_k is discrete or has a mixture distribution. Using a conditional normal distribution for Y, we have

$$f_{Y|X,\boldsymbol{U}}(y|x,\boldsymbol{u}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2} \left[y - \left(\delta_0 + \beta x + \sum_{m=1}^{M} \delta_m u_m \right) \right]^2 \right\}$$

and the conditional distribution of $Y|X, \boldsymbol{U}\backslash U_k$ can be calculated as

$$f_{Y|X,\boldsymbol{U}\backslash U_k}(y|x,\boldsymbol{u}\backslash u_k) = \int_{-\infty}^{\infty} f_{U_k|X,\boldsymbol{U}\backslash U_k}(u_k|x,\boldsymbol{u}\backslash u_k) f_{Y|X,\boldsymbol{U}}(y|x,\boldsymbol{u}) du_k. \qquad (4.13)$$

If $\delta_k = 0$

$$f_{Y|X,U}(y|x,\boldsymbol{u}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\left[y - \left(\delta_0 + \beta x + \sum_{m\neq k}\delta_m u_m\right)\right]^2\right\},$$

and the expression (4.13) for $f_{Y|X,U\backslash U_k}(y|x,\boldsymbol{u}\backslash u_k)$ becomes

$$\int_{-\infty}^{\infty} f_{U_k|X,U\backslash U_k}(u_k|x,\boldsymbol{u}\backslash u_k)\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\left[y - \left(\delta_0 + \beta x + \sum_{m\neq k}\delta_m u_m\right)\right]^2\right\} du_k$$
$$= f_{Y|X,U}(y|x,\boldsymbol{u}) \int_{-\infty}^{\infty} f_{U_k|X,U\backslash U_k}(u_k|x,\boldsymbol{u}\backslash u_k) du_k,$$

which equals $f_{Y|X,U}(y|x,\boldsymbol{u})$.

Thus if $\delta_k = 0$ in (4.1) then $Y \perp\!\!\!\perp U_k|X,\boldsymbol{U}\backslash U_k$. Also, it is obvious that if $Y \perp\!\!\!\perp U_k|X,\boldsymbol{U}\backslash U_k$, then $\delta_k = 0$. Therefore, assuming model (4.1) is correctly specified we have that $Y \perp\!\!\!\perp U_k|X,\boldsymbol{U}\backslash U_k$ if and only if $\delta_k = 0$. Recall that no assumptions were made concerning the distribution of X and $\boldsymbol{U}\backslash U_k$.

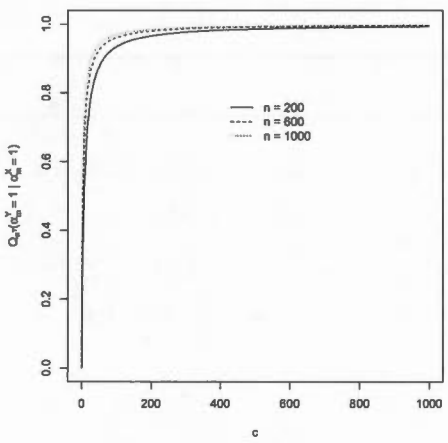
4.7.4 The behavior of Q_{α^Y}

In Figure 4.2 we examine how the term $Q_{\alpha^Y}(\alpha_m^Y = 1|\alpha_m^X = 1)$ in the definition of $P^B(\boldsymbol{\alpha}^{\boldsymbol{Y}})$ behaves as a function of the constant c, the sample size n and the standardized parameter $\tilde{\delta}_m^{\alpha^Y} s_{U_m}/s_Y$. Specifically, we take $\omega = c\sqrt{n}$, as suggested in Section 4.3.2, and plot the $Q_{\alpha^Y}(\alpha_m^Y = 1|\alpha_m^X = 1)$ values as a function of $c \in [0, 1000]$ for fixed values of n ($n = 200, 600, 1000$) and $\tilde{\delta}_m^{\alpha^Y} s_{U_m}/s_Y$ ($\tilde{\delta}_m^{\alpha^Y} s_{U_m}/s_Y = 0.1, 0.05, 0.01$).

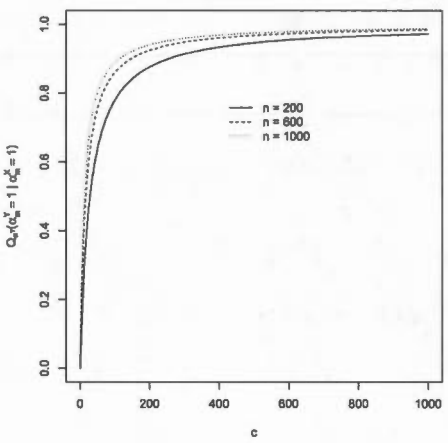
In Figure 2 (a), we see that, for all sample sizes considered, $Q_{\alpha^Y}(\alpha^Y = 1|\alpha^X = 1)$ rapidly increase from 0 to the limit 1 as c goes from 0 to 1000. This behavior is desirable since a standardized regression parameter of 0.1 is non negligible. A similar pattern is seen in Figure 2 (b), although the progression of $Q_{\alpha^Y}(\alpha^Y = 1|\alpha^X = 1)$ from 0 to 1 is slightly less rapid. In Figure 2 (c), the progression of $Q_{\alpha^Y}(\alpha^Y = 1|\alpha^X = 1)$ is much slower, especially for the smaller sample size. This behavior is desirable as well since an effect size of 0.01 would usually be considered as negligible.

Figure 4.2: $Q_{\alpha^Y}(\alpha_m^Y = 1|\alpha_m^X = 1)$ with $\omega = c\sqrt{n}$ as a function of $c \in [0, 1000]$ for $n = 200, 600, 1000$ and $\tilde{\delta}_m^{\alpha^Y} s_{U_m}/s_Y = 0.1$ (a), $\tilde{\delta}_m^{\alpha^Y} s_{U_m}/s_Y = 0.05$ (b) and $\tilde{\delta}_m^{\alpha^Y} s_{U_m}/s_Y = 0.01$ (c) .

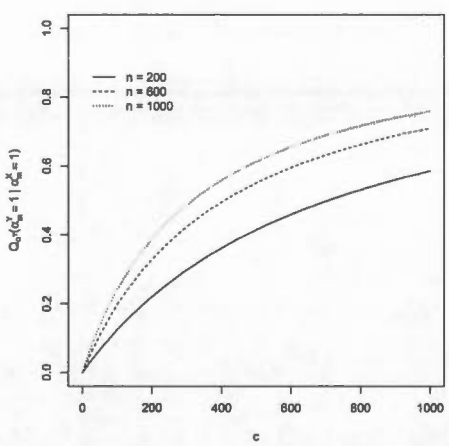
(a)



(b)



(c)



4.7.5 Marginal posterior probabilities of inclusion of potential confounding covariates

Table 4.11: Marginal posterior probability of inclusion of potential confounding covariate U_m, $m = 1, ..., 5$, for BMA, BAC, TBAC, N-BCEE, and A-BCEE for 500 Monte Carlo replicates of the first data-generating process (DGP1). The covariates included in the true outcome model are $\{U_3, U_4, U_5\}$.

n	*Method*	U_1	U_2	U_3	U_4	U_5
200	BMA	0.11	0.11	1.00	0.18	1.00
200	BAC ($C_V^m(\omega)$)	0.35	0.35	1.00	0.41	1.00
200	BAC ($\omega = \infty$)	1.00	1.00	1.00	1.00	1.00
200	TBAC ($\omega = \infty$)	1.00	1.00	1.00	1.00	1.00
200	N-BCEE (c = 100)	0.19	0.24	1.00	0.37	1.00
200	N-BCEE (c = 500)	0.36	0.41	1.00	0.54	1.00
200	N-BCEE (c = 1000)	0.44	0.49	1.00	0.61	1.00
200	A-BCEE (c = 100)	0.29	0.35	1.00	0.44	1.00
200	A-BCEE (c = 500)	0.51	0.56	1.00	0.63	1.00
200	A-BCEE (c = 1000)	0.60	0.64	1.00	0.70	1.00
600	BMA	0.06	0.06	1.00	0.24	1.00
600	BAC ($C_V^m(\omega)$)	0.32	0.33	1.00	0.47	1.00
600	BAC ($\omega = \infty$)	1.00	1.00	1.00	1.00	1.00
600	TBAC ($\omega = \infty$)	1.00	1.00	1.00	1.00	1.00
600	N-BCEE (c = 100)	0.11	0.15	1.00	0.44	1.00
600	N-BCEE (c = 500)	0.22	0.30	1.00	0.60	1.00
600	N-BCEE (c = 1000)	0.28	0.37	1.00	0.66	1.00
600	A-BCEE (c = 100)	0.15	0.21	1.00	0.45	1.00
600	A-BCEE (c = 500)	0.34	0.42	1.00	0.63	1.00
600	A-BCEE (c = 1000)	0.44	0.51	1.00	0.70	1.00
1000	BMA	0.05	0.04	1.00	0.33	1.00
1000	BAC ($C_V^m(\omega)$)	0.38	0.37	1.00	0.61	1.00
1000	BAC ($\omega = \infty$)	1.00	1.00	1.00	1.00	1.00
1000	TBAC ($\omega = \infty$)	1.00	1.00	1.00	1.00	1.00
1000	N-BCEE (c = 100)	0.09	0.11	1.00	0.55	1.00
1000	N-BCEE (c = 500)	0.19	0.22	1.00	0.69	1.00
1000	N-BCEE (c = 1000)	0.25	0.28	1.00	0.74	1.00
1000	A-BCEE (c = 100)	0.12	0.15	1.00	0.54	1.00
1000	A-BCEE (c = 500)	0.30	0.34	1.00	0.70	1.00
1000	A-BCEE (c = 1000)	0.39	0.44	1.00	0.75	1.00

Table 4.12: Marginal posterior probability of inclusion of potential confounding covariate U_m, $m = 1, ..., 5, 7, 8$, for BMA, BAC, TBAC, N-BCEE, and A-BCEE for 500 Monte Carlo replicates of the second data-generating process (DGP2). The covariates included in the true outcome model are $\{U_3, U_4, U_5\}$.

n	*Method*	U_1	U_2	U_3	U_4	U_5	U_7	U_8
200	BMA	0.14	0.12	0.20	0.18	0.31	0.10	0.10
200	BAC ($C_V^m(\omega)$)	0.44	0.41	0.48	0.18	0.32	0.12	0.11
200	BAC ($\omega = \infty$)	1.00	1.00	1.00	0.22	0.40	0.15	0.15
200	TBAC ($\omega = \infty$)	1.00	1.00	1.00	0.18	0.29	0.14	0.14
200	N-BCEE (c = 100)	0.45	0.39	0.55	0.26	0.34	0.14	0.14
200	N-BCEE (c = 500)	0.64	0.58	0.72	0.28	0.35	0.17	0.17
200	N-BCEE (c = 1000)	0.71	0.66	0.78	0.29	0.36	0.18	0.18
200	A-BCEE (c = 100)	0.64	0.56	0.66	0.19	0.30	0.13	0.13
200	A-BCEE (c = 500)	0.79	0.73	0.81	0.19	0.30	0.14	0.14
200	A-BCEE (c = 1000)	0.84	0.79	0.85	0.19	0.30	0.14	0.14
600	BMA	0.12	0.09	0.39	0.25	0.64	0.08	0.07
600	BAC ($C_V^m(\omega)$)	0.63	0.59	0.75	0.22	0.62	0.09	0.07
600	BAC ($\omega = \infty$)	1.00	1.00	1.00	0.21	0.57	0.07	0.06
600	TBAC ($\omega = \infty$)	1.00	1.00	1.00	0.19	0.53	0.09	0.08
600	N-BCEE (c = 100)	0.37	0.26	0.73	0.28	0.65	0.09	0.08
600	N-BCEE (c = 500)	0.56	0.45	0.84	0.29	0.63	0.11	0.09
600	N-BCEE (c = 1000)	0.64	0.54	0.88	0.29	0.61	0.12	0.10
600	A-BCEE (c = 100)	0.56	0.43	0.76	0.22	0.59	0.08	0.07
600	A-BCEE (c = 500)	0.74	0.63	0.86	0.21	0.56	0.09	0.08
600	A-BCEE (c = 1000)	0.80	0.70	0.90	0.20	0.56	0.09	0.08
1000	BMA	0.12	0.08	0.55	0.33	0.82	0.06	0.06
1000	BAC ($C_V^m(\omega)$)	0.69	0.66	0.86	0.28	0.75	0.06	0.06
1000	BAC ($\omega = \infty$)	1.00	1.00	1.00	0.25	0.71	0.05	0.05
1000	TBAC ($\omega = \infty$)	1.00	1.00	1.00	0.25	0.69	0.07	0.07
1000	N-BCEE (c = 100)	0.34	0.23	0.83	0.34	0.78	0.07	0.07
1000	N-BCEE (c = 500)	0.50	0.39	0.90	0.34	0.76	0.08	0.08
1000	N-BCEE (c = 1000)	0.58	0.47	0.92	0.34	0.75	0.08	0.08
1000	A-BCEE (c = 100)	0.50	0.37	0.84	0.28	0.75	0.06	0.07
1000	A-BCEE (c = 500)	0.69	0.57	0.91	0.27	0.72	0.07	0.07
1000	A-BCEE (c = 1000)	0.75	0.65	0.93	0.26	0.71	0.07	0.07

Table 4.13: Marginal posterior probability of inclusion of potential confounding covariate U_m, $m = 1, ..., 4$, for BMA, BAC, TBAC, N-BCEE, and A-BCEE for 500 Monte Carlo replicates of the third data-generating process (DGP3). The covariates included in the true outcome model are $\{U_1, U_2\}$.

n	*Method*	U_1	U_2	U_3	U_4
200	BMA	0.17	0.26	0.08	0.09
200	BAC ($C_V^m(\omega)$)	0.48	0.29	0.09	0.10
200	BAC ($\omega = \infty$)	1.00	0.28	0.08	0.10
200	TBAC ($\omega = \infty$)	1.00	0.30	0.14	0.15
200	N-BCEE (c = 100)	0.57	0.32	0.14	0.16
200	N-BCEE (c = 500)	0.75	0.34	0.17	0.19
200	N-BCEE (c = 1000)	0.80	0.35	0.18	0.20
200	A-BCEE (c = 100)	0.62	0.30	0.13	0.14
200	A-BCEE (c = 500)	0.77	0.30	0.13	0.14
200	A-BCEE (c = 1000)	0.83	0.30	0.13	0.15
600	BMA	0.29	0.50	0.07	0.05
600	BAC ($C_V^m(\omega)$)	0.75	0.53	0.07	0.06
600	BAC ($\omega = \infty$)	1.00	0.52	0.07	0.05
600	TBAC ($\omega = \infty$)	1.00	0.51	0.09	0.08
600	N-BCEE (c = 100)	0.68	0.52	0.10	0.08
600	N-BCEE (c = 500)	0.82	0.53	0.11	0.10
600	N-BCEE (c = 1000)	0.87	0.53	0.12	0.10
600	A-BCEE (c = 100)	0.68	0.51	0.09	0.07
600	A-BCEE (c = 500)	0.81	0.51	0.09	0.08
600	A-BCEE (c = 1000)	0.85	0.51	0.09	0.08
1000	BMA	0.41	0.68	0.05	0.05
1000	BAC ($C_V^m(\omega)$)	0.86	0.70	0.06	0.05
1000	BAC ($\omega = \infty$)	0.99	0.69	0.06	0.05
1000	TBAC ($\omega = \infty$)	1.00	0.68	0.07	0.07
1000	N-BCEE (c = 100)	0.76	0.69	0.07	0.07
1000	N-BCEE (c = 500)	0.88	0.69	0.08	0.08
1000	N-BCEE (c = 1000)	0.91	0.69	0.09	0.09
1000	A-BCEE (c = 100)	0.75	0.68	0.07	0.07
1000	A-BCEE (c = 500)	0.85	0.68	0.07	0.07
1000	A-BCEE (c = 1000)	0.89	0.68	0.07	0.07

Table 4.14: Marginal posterior probability of inclusion of potential confounding covariate U_m, $m = 1, ..., 6$, for BMA, BAC, TBAC, N-BCEE, and A-BCEE for 500 Monte Carlo replicates of the fourth data-generating process (DGP4). The covariates included in the true outcome model are $\{U_4, U_5, U_6\}$.

n	*Method*	U_1	U_2	U_3	U_4	U_5	U_6
200	BMA	0.15	0.14	0.13	0.22	1.00	1.00
200	BAC ($C_V^m(\omega)$)	0.32	0.15	0.31	0.22	1.00	1.00
200	BAC ($\omega = \infty$)	0.97	0.23	0.98	0.24	1.00	1.00
200	TBAC ($\omega = \infty$)	0.87	0.23	0.99	0.25	1.00	1.00
200	N-BCEE (c = 100)	0.36	0.17	0.36	0.29	1.00	1.00
200	N-BCEE (c = 500)	0.53	0.22	0.57	0.32	1.00	1.00
200	N-BCEE (c = 1000)	0.60	0.25	0.66	0.33	1.00	1.00
200	A-BCEE (c = 100)	0.50	0.19	0.48	0.23	1.00	1.00
200	A-BCEE (c = 500)	0.65	0.21	0.66	0.24	1.00	1.00
200	A-BCEE (c = 1000)	0.70	0.21	0.73	0.24	1.00	1.00
600	BMA	0.14	0.10	0.08	0.35	1.00	1.00
600	BAC ($C_V^m(\omega)$)	0.44	0.16	0.41	0.29	1.00	1.00
600	BAC ($\omega = \infty$)	0.99	0.24	1.00	0.27	1.00	1.00
600	TBAC ($\omega = \infty$)	0.96	0.23	1.00	0.26	1.00	1.00
600	N-BCEE (c = 100)	0.33	0.12	0.22	0.41	1.00	1.00
600	N-BCEE (c = 500)	0.50	0.18	0.41	0.41	1.00	1.00
600	N-BCEE (c = 1000)	0.58	0.21	0.50	0.41	1.00	1.00
600	A-BCEE (c = 100)	0.48	0.16	0.32	0.31	1.00	1.00
600	A-BCEE (c = 500)	0.67	0.19	0.53	0.29	1.00	1.00
600	A-BCEE (c = 1000)	0.73	0.20	0.61	0.28	1.00	1.00
1000	BMA	0.13	0.09	0.07	0.52	1.00	1.00
1000	BAC ($C_V^m(\omega)$)	0.48	0.18	0.42	0.42	1.00	1.00
1000	BAC ($\omega = \infty$)	0.99	0.30	1.00	0.35	1.00	1.00
1000	TBAC ($\omega = \infty$)	0.99	0.30	1.00	0.34	1.00	1.00
1000	N-BCEE (c = 100)	0.30	0.12	0.19	0.57	1.00	1.00
1000	N-BCEE (c = 500)	0.45	0.19	0.36	0.56	1.00	1.00
1000	N-BCEE (c = 1000)	0.53	0.22	0.44	0.54	1.00	1.00
1000	A-BCEE (c = 100)	0.47	0.18	0.26	0.44	1.00	1.00
1000	A-BCEE (c = 500)	0.67	0.23	0.47	0.40	1.00	1.00
1000	A-BCEE (c = 1000)	0.73	0.25	0.56	0.38	1.00	1.00

Table 4.15: Marginal posterior probability of inclusion of potential confounding covariate U_m, $m = 1, ..., 5$, for BMA, BAC, TBAC, and A-BCEE for 500 Monte Carlo replicates of the fifth data-generating process (DGP5). The true outcome model includes only U_5.

n	*Method*	U_1	U_2	U_3	U_4	U_5
200	BMA	0.11	0.12	0.12	0.11	1.00
200	BAC ($C_V^m(\omega)$)	0.42	0.42	0.42	0.42	1.00
200	BAC ($\omega = \infty$)	1.00	1.00	1.00	1.00	1.00
200	TBAC ($\omega = \infty$)	1.00	1.00	1.00	1.00	1.00
200	A-BCEE (c = 100)	0.21	0.22	0.22	0.21	1.00
200	A-BCEE (c = 500)	0.45	0.45	0.46	0.45	1.00
200	A-BCEE (c = 1000)	0.55	0.55	0.56	0.55	1.00
600	BMA	0.08	0.07	0.08	0.08	1.00
600	BAC ($C_V^m(\omega)$)	0.42	0.41	0.42	0.42	1.00
600	BAC ($\omega = \infty$)	1.00	1.00	1.00	1.00	1.00
600	TBAC ($\omega = \infty$)	1.00	1.00	1.00	1.00	1.00
600	A-BCEE (c = 100)	0.12	0.11	0.12	0.12	1.00
600	A-BCEE (c = 500)	0.30	0.29	0.31	0.31	1.00
600	A-BCEE (c = 1000)	0.40	0.40	0.41	0.41	1.00
1000	BMA	0.06	0.07	0.06	0.06	1.00
1000	BAC ($C_V^m(\omega)$)	0.41	0.41	0.41	0.40	1.00
1000	BAC ($\omega = \infty$)	1.00	1.00	1.00	1.00	1.00
1000	TBAC ($\omega = \infty$)	1.00	1.00	1.00	1.00	1.00
1000	A-BCEE (c = 100)	0.08	0.09	0.09	0.08	1.00
1000	A-BCEE (c = 500)	0.22	0.23	0.23	0.22	1.00
1000	A-BCEE (c = 1000)	0.32	0.33	0.33	0.32	1.00

4.7.6 Simulation results for scenarios with $\beta = 0$

Table 4.16: Comparison of estimates of β obtained from the true outcome model, the fully adjusted model, BMA, BAC, TBAC, N-BCEE and A-BCEE for 500 Monte Carlo replicates of the first data-generating process (DGP1).

n	*Method*	*Mean*	$\overline{SEE}$	SDE	$\sqrt{MSE}$	CP
200	True model	0.000	0.045	0.044	0.044	97
200	Fully adjusted model	0.000	0.072	0.070	0.070	96
200	BMA	0.014	0.047	0.043	0.045	96
200	BAC ($C_V^m(\omega)$)	0.006	0.055	0.058	0.058	95
200	BAC ($\omega = \infty$)	0.000	0.072	0.070	0.070	96
200	TBAC ($\omega = \infty$)	0.000	0.072	0.070	0.070	96
200	N-BCEE (c = 100)	0.008	0.051	0.050	0.051	96
200	N-BCEE (c = 500)	0.004	0.055	0.057	0.057	95
200	N-BCEE (c = 1000)	0.003	0.057	0.060	0.060	95
200	A-BCEE (c = 100)	0.008	0.055	0.049	0.050	97
200	A-BCEE (c = 500)	0.005	0.061	0.056	0.056	98
200	A-BCEE (c = 1000)	0.004	0.063	0.058	0.058	97
600	True model	0.001	0.026	0.027	0.027	96
600	Fully adjusted model	0.001	0.041	0.042	0.042	96
600	BMA	0.012	0.027	0.028	0.030	92
600	BAC ($C_V^m(\omega)$)	0.004	0.032	0.037	0.037	93
600	BAC ($\omega = \infty$)	0.001	0.041	0.042	0.042	96
600	TBAC ($\omega = \infty$)	0.001	0.041	0.042	0.042	96
600	N-BCEE (c = 100)	0.008	0.029	0.031	0.032	92
600	N-BCEE (c = 500)	0.005	0.031	0.034	0.034	94
600	N-BCEE (c = 1000)	0.004	0.032	0.035	0.035	94
600	A-BCEE (c = 100)	0.008	0.030	0.030	0.031	95
600	A-BCEE (c = 500)	0.006	0.034	0.033	0.033	96
600	A-BCEE (c = 1000)	0.005	0.035	0.034	0.035	96
1000	True model	0.000	0.020	0.019	0.019	96
1000	Fully adjusted model	0.001	0.032	0.030	0.030	97
1000	BMA	0.010	0.021	0.021	0.023	92
1000	BAC ($C_V^m(\omega)$)	0.002	0.026	0.030	0.030	93
1000	BAC ($\omega = \infty$)	-0.001	0.032	0.030	0.030	97
1000	TBAC ($\omega = \infty$)	-0.001	0.032	0.030	0.030	97
1000	N-BCEE (c = 100)	0.006	0.022	0.023	0.023	95
1000	N-BCEE (c = 500)	0.003	0.023	0.024	0.024	94
1000	N-BCEE (c = 1000)	0.003	0.024	0.025	0.025	95
1000	A-BCEE (c = 100)	0.006	0.023	0.022	0.023	96
1000	A-BCEE (c = 500)	0.003	0.025	0.024	0.024	96
1000	A-BCEE (c = 1000)	0.003	0.027	0.024	0.024	96

LEGEND: $Mean$ is the mean estimated value of β where the true value is 0, $\overline{SEE}$ is the mean standard error estimate, SDE is the standard deviation of the estimates of β, $\sqrt{MSE}$ is the squared-root of the mean squared error, CP is the coverage probability in % of 95% confidence intervals.

Table 4.17: Comparison of estimates of β obtained from the true outcome model, the fully adjusted model, BMA, BAC, TBAC, N-BCEE and A-BCEE for 500 Monte Carlo replicates of the second data-generating process (DGP2).

n	*Method*	*Mean*	$\overline{SEE}$	*SDE*	$\sqrt{MSE}$	*CP*
200	True model	-0.003	0.045	0.045	0.045	94
200	Fully adjusted model	0.000	0.073	0.073	0.073	95
200	BMA	0.022	0.039	0.038	0.044	89
200	BAC ($C_V^m(\omega)$)	0.007	0.047	0.061	0.061	89
200	BAC ($\omega = \infty$)	0.000	0.071	0.072	0.072	94
200	TBAC ($\omega = \infty$)	-0.001	0.072	0.072	0.072	94
200	N-BCEE (c = 100)	0.007	0.050	0.057	0.057	89
200	N-BCEE (c = 500)	0.002	0.056	0.065	0.065	90
200	N-BCEE (c = 1000)	0.001	0.058	0.068	0.068	91
200	A-BCEE (c = 100)	0.008	0.060	0.056	0.057	97
200	A-BCEE (c = 500)	0.004	0.065	0.063	0.063	96
200	A-BCEE (c = 1000)	0.003	0.067	0.065	0.065	96
600	True model	-0.002	0.026	0.026	0.026	95
600	Fully adjusted model	0.000	0.041	0.041	0.041	95
600	BMA	0.014	0.023	0.024	0.028	87
600	BAC ($C_V^m(\omega)$)	0.003	0.029	0.036	0.036	90
600	BAC ($\omega = \infty$)	0.000	0.041	0.041	0.041	94
600	TBAC ($\omega = \infty$)	0.000	0.041	0.041	0.041	95
600	N-BCEE (c = 100)	0.006	0.027	0.030	0.031	90
600	N-BCEE (c = 500)	0.002	0.031	0.037	0.037	92
600	N-BCEE (c = 1000)	0.002	0.031	0.037	0.037	92
600	A-BCEE (c = 100)	0.007	0.032	0.031	0.031	95
600	A-BCEE (c = 500)	0.004	0.036	0.034	0.035	95
600	A-BCEE (c = 1000)	0.003	0.037	0.036	0.036	96
1000	True model	-0.001	0.020	0.020	0.020	95
1000	Fully adjusted model	0.000	0.032	0.031	0.031	96
1000	BMA	0.012	0.018	0.019	0.022	87
1000	BAC ($C_V^m(\omega)$)	0.002	0.022	0.028	0.028	89
1000	BAC ($\omega = \infty$)	0.000	0.032	0.030	0.030	96
1000	TBAC ($\omega = \infty$)	0.000	0.032	0.030	0.030	96
1000	N-BCEE (c = 100)	0.006	0.021	0.022	0.023	89
1000	N-BCEE (c = 500)	0.004	0.023	0.025	0.025	92
1000	N-BCEE (c = 1000)	0.003	0.023	0.026	0.027	93
1000	A-BCEE (c = 100)	0.006	0.024	0.022	0.023	94
1000	A-BCEE (c = 500)	0.004	0.027	0.025	0.025	97
1000	A-BCEE (c = 1000)	0.003	0.028	0.026	0.026	97

LEGEND: See Table 4.16

Table 4.18: Comparison of estimates of β obtained from the true outcome model, the fully adjusted model, BMA, BAC, TBAC, N-BCEE and A-BCEE for 500 Monte Carlo replicates of the third data-generating process (DGP3).

n	*Method*	*Mean*	$\overline{SEE}$	*SDE*	$\sqrt{MSE}$	*CP*
200	True model	0.001	0.100	0.101	0.101	95
200	Fully adjusted model	-0.001	0.105	0.107	0.107	95
200	BMA	0.045	0.083	0.091	0.101	88
200	BAC ($C_V^m(\omega)$)	0.017	0.085	0.102	0.104	89
200	BAC ($\omega = \infty$)	0.001	0.100	0.101	0.101	95
200	TBAC ($\omega = \infty$)	0.001	0.101	0.102	0.102	94
200	N-BCEE (c = 100)	0.013	0.091	0.101	0.101	92
200	N-BCEE (c = 500)	0.006	0.095	0.102	0.102	93
200	N-BCEE (c = 1000)	0.004	0.096	0.102	0.102	94
200	A-BCEE (c = 100)	0.015	0.095	0.099	0.100	94
200	A-BCEE (c = 500)	0.009	0.098	0.100	0.101	95
200	A-BCEE (c = 1000)	0.007	0.099	0.101	0.101	95
600	True model	-0.002	0.057	0.061	0.060	96
600	Fully adjusted model	-0.003	0.058	0.062	0.062	95
600	BMA	0.036	0.053	0.063	0.072	80
600	BAC ($C_V^m(\omega)$)	0.005	0.054	0.066	0.066	87
600	BAC ($\omega = \infty$)	-0.003	0.058	0.061	0.061	96
600	TBAC ($\omega = \infty$)	-0.003	0.058	0.061	0.061	95
600	N-BCEE (c = 100)	0.009	0.055	0.066	0.066	88
600	N-BCEE (c = 500)	0.002	0.063	0.063	0.063	91
600	N-BCEE (c = 1000)	0.000	0.056	0.063	0.063	93
600	A-BCEE (c = 100)	0.011	0.057	0.065	0.065	91
600	A-BCEE (c = 500)	0.005	0.057	0.063	0.064	93
600	A-BCEE (c = 1000)	0.003	0.057	0.063	0.063	94
1000	True model	0.002	0.044	0.045	0.045	95
1000	Fully adjusted model	0.002	0.045	0.045	0.045	95
1000	BMA	0.035	0.044	0.050	0.061	77
1000	BAC ($C_V^m(\omega)$)	0.007	0.059	0.068	0.068	91
1000	BAC ($\omega = \infty$)	0.002	0.044	0.044	0.044	95
1000	TBAC ($\omega = \infty$)	0.002	0.044	0.045	0.045	95
1000	N-BCEE (c = 100)	0.010	0.044	0.049	0.050	89
1000	N-BCEE (c = 500)	0.005	0.044	0.047	0.047	92
1000	N-BCEE (c = 1000)	0.004	0.044	0.046	0.046	94
1000	A-BCEE (c = 100)	0.012	0.045	0.048	0.050	90
1000	A-BCEE (c = 500)	0.007	0.045	0.047	0.047	93
1000	A-BCEE (c = 1000)	0.006	0.045	0.046	0.047	94

LEGEND: See Table 4.16

Table 4.19: Comparison of estimates of β obtained from the true outcome model, the fully adjusted model, BMA, BAC, TBAC, N-BCEE and A-BCEE for 500 Monte Carlo replicates of the fourth data-generating process (DGP4).

n	*Method*	*Mean*	$\overline{SEE}$	*SDE*	$\sqrt{MSE}$	*CP*
200	True model	-0.004	0.054	0.055	0.055	95
200	Fully adjusted model	-0.007	0.072	0.072	0.073	94
200	BMA	0.011	0.060	0.057	0.058	95
200	BAC ($C_V^m(\omega)$)	0.002	0.062	0.067	0.067	93
200	BAC ($\omega = \infty$)	-0.009	0.072	0.072	0.073	95
200	TBAC ($\omega = \infty$)	-0.007	0.071	0.071	0.071	95
200	N-BCEE (c = 100)	0.000	0.062	0.066	0.066	93
200	N-BCEE (c = 500)	-0.004	0.064	0.069	0.069	93
200	N-BCEE (c = 1000)	-0.006	0.066	0.070	0.070	93
200	A-BCEE (c = 100)	0.000	0.066	0.065	0.065	95
200	A-BCEE (c = 500)	-0.003	0.068	0.067	0.067	95
200	A-BCEE (c = 1000)	-0.004	0.069	0.068	0.068	95
600	True model	0.000	0.031	0.032	0.032	94
600	Fully adjusted model	0.000	0.041	0.041	0.041	95
600	BMA	0.012	0.035	0.036	0.037	92
600	BAC ($C_V^m(\omega)$)	0.003	0.037	0.040	0.041	92
600	BAC ($\omega = \infty$)	-0.001	0.041	0.041	0.041	95
600	TBAC ($\omega = \infty$)	0.000	0.041	0.041	0.041	96
600	N-BCEE (c = 100)	0.004	0.035	0.039	0.039	91
600	N-BCEE (c = 500)	0.002	0.037	0.040	0.040	92
600	N-BCEE (c = 1000)	0.001	0.037	0.040	0.040	93
600	A-BCEE (c = 100)	0.005	0.038	0.038	0.039	94
600	A-BCEE (c = 500)	0.003	0.039	0.039	0.039	95
600	A-BCEE (c = 1000)	0.002	0.040	0.040	0.040	95
1000	True model	0.000	0.024	0.024	0.024	95
1000	Fully adjusted model	0.002	0.032	0.032	0.032	95
1000	BMA	0.009	0.028	0.029	0.030	93
1000	BAC ($C_V^m(\omega)$)	0.002	0.030	0.032	0.032	92
1000	BAC ($\omega = \infty$)	0.000	0.032	0.032	0.032	95
1000	TBAC ($\omega = \infty$)	0.000	0.032	0.032	0.032	95
1000	N-BCEE (c = 100)	0.003	0.027	0.030	0.030	93
1000	N-BCEE (c = 500)	0.002	0.028	0.031	0.031	92
1000	N-BCEE (c = 1000)	0.002	0.029	0.031	0.031	92
1000	A-BCEE (c = 100)	0.004	0.030	0.030	0.030	93
1000	A-BCEE (c = 500)	0.003	0.030	0.031	0.031	94
1000	A-BCEE (c = 1000)	0.002	0.031	0.031	0.031	94

LEGEND: See Table 4.16

4.7.7 Simulation results for scenarios with exponential errors

Table 4.20: Comparison of estimates of β obtained from the true outcome model, the fully adjusted model, BMA, BAC, TBAC and A-BCEE for 500 Monte Carlo replicates of the sixth data-generating process (DGP6).

n	*Method*	*Mean*	$\overline{SEE}$	*SDE*	$\sqrt{MSE}$	*CP*
200	True model	0.097	0.046	0.045	0.045	95
200	Fully adjusted model	0.095	0.073	0.070	0.070	96
200	BMA	0.111	0.048	0.044	0.045	96
200	BAC ($C_V^m(\omega)$)	0.100	0.055	0.057	0.057	95
200	BAC ($\omega = \infty$)	0.095	0.073	0.070	0.070	96
200	TBAC ($\omega = \infty$)	0.095	0.073	0.070	0.070	95
200	A-BCEE (c = 100)	0.106	0.055	0.050	0.050	97
200	A-BCEE (c = 500)	0.102	0.061	0.056	0.056	98
200	A-BCEE (c = 1000)	0.100	0.064	0.058	0.058	98
600	True model	0.101	0.026	0.027	0.027	95
600	Fully adjusted model	0.101	0.041	0.041	0.041	95
600	BMA	0.112	0.027	0.027	0.029	93
600	BAC ($C_V^m(\omega)$)	0.103	0.032	0.036	0.036	93
600	BAC ($\omega = \infty$)	0.101	0.041	0.041	0.041	95
600	TBAC ($\omega = \infty$)	0.101	0.041	0.041	0.041	95
600	A-BCEE (c = 100)	0.108	0.030	0.029	0.030	94
600	A-BCEE (c = 500)	0.106	0.033	0.032	0.033	96
600	A-BCEE (c = 1000)	0.105	0.035	0.033	0.034	95
1000	True model	0.098	0.020	0.020	0.020	95
1000	Fully adjusted model	0.097	0.032	0.031	0.032	94
1000	BMA	0.108	0.021	0.022	0.023	93
1000	BAC ($C_V^m(\omega)$)	0.099	0.026	0.028	0.028	93
1000	BAC ($\omega = \infty$)	0.097	0.032	0.031	0.032	94
1000	TBAC ($\omega = \infty$)	0.096	0.031	0.032	0.032	94
1000	A-BCEE (c = 100)	0.105	0.023	0.023	0.023	95
1000	A-BCEE (c = 500)	0.102	0.026	0.024	0.024	96
1000	A-BCEE (c = 1000)	0.101	0.027	0.025	0.025	97

LEGEND: See Table 4.3.

Table 4.21: Comparison of estimates of β obtained from the true outcome model, the fully adjusted model, BMA, BAC, TBAC and A-BCEE for 500 Monte Carlo replicates of the seventh data-generating process (DGP7).

n	*Method*	*Mean*	$\overline{SEE}$	*SDE*	$\sqrt{MSE}$	*CP*
200	True model	0.096	0.046	0.046	0.046	96
200	Fully adjusted model	0.094	0.074	0.072	0.072	96
200	BMA	0.145	0.044	0.042	0.061	74
200	BAC ($C_V^m(\omega)$)	0.109	0.054	0.070	0.071	82
200	BAC ($\omega = \infty$)	0.095	0.072	0.071	0.071	96
200	TBAC ($\omega = \infty$)	0.094	0.073	0.070	0.070	97
200	A-BCEE (c = 100)	0.114	0.063	0.059	0.061	95
200	A-BCEE (c = 500)	0.105	0.067	0.064	0.064	96
200	A-BCEE (c = 1000)	0.102	0.069	0.065	0.065	96
600	True model	0.099	0.026	0.027	0.027	94
600	Fully adjusted model	0.099	0.042	0.044	0.044	94
600	BMA	0.132	0.030	0.034	0.047	71
600	BAC ($C_V^m(\omega)$)	0.102	0.036	0.045	0.045	84
600	BAC ($\omega = \infty$)	0.099	0.041	0.043	0.043	95
600	TBAC ($\omega = \infty$)	0.099	0.041	0.043	0.043	94
600	A-BCEE (c = 100)	0.112	0.036	0.039	0.041	90
600	A-BCEE (c = 500)	0.106	0.038	0.041	0.041	92
600	A-BCEE (c = 1000)	0.104	0.039	0.041	0.041	92
1000	True model	0.100	0.020	0.019	0.019	96
1000	Fully adjusted model	0.101	0.032	0.032	0.032	95
1000	BMA	0.121	0.024	0.026	0.034	80
1000	BAC ($C_V^m(\omega)$)	0.102	0.028	0.031	0.031	89
1000	BAC ($\omega = \infty$)	0.101	0.032	0.032	0.032	95
1000	TBAC ($\omega = \infty$)	0.101	0.032	0.032	0.032	95
1000	A-BCEE (c = 100)	0.109	0.027	0.028	0.029	93
1000	A-BCEE (c = 500)	0.106	0.029	0.029	0.030	95
1000	A-BCEE (c = 1000)	0.105	0.030	0.030	0.030	96

LEGEND: See Table 4.3.

Table 4.22: Comparison of estimates of β obtained from the true outcome model, the fully adjusted model, BMA, BAC, TBAC and A-BCEE for 500 Monte Carlo replicates of the eighth data-generating process (DGP8).

n	*Method*	*Mean*	$\overline{SEE}$	*SDE*	$\sqrt{MSE}$	*CP*
200	True model	0.105	0.054	0.053	0.053	95
200	Fully adjusted model	0.103	0.072	0.070	0.070	95
200	BMA	0.121	0.059	0.055	0.059	94
200	BAC ($C_V^m(\omega)$)	0.110	0.061	0.064	0.065	93
200	BAC ($\omega = \infty$)	0.101	0.072	0.070	0.070	95
200	TBAC ($\omega = \infty$)	0.104	0.071	0.069	0.069	95
200	A-BCEE (c = 100)	0.110	0.066	0.063	0.063	96
200	A-BCEE (c = 500)	0.107	0.068	0.065	0.066	97
200	A-BCEE (c = 1000)	0.106	0.069	0.066	0.067	96
600	True model	0.101	0.031	0.031	0.031	95
600	Fully adjusted model	0.099	0.041	0.041	0.041	96
600	BMA	0.112	0.035	0.035	0.037	92
600	BAC ($C_V^m(\omega)$)	0.102	0.037	0.040	0.040	92
600	BAC ($\omega = \infty$)	0.098	0.041	0.041	0.041	96
600	TBAC ($\omega = \infty$)	0.098	0.041	0.040	0.040	96
600	A-BCEE (c = 100)	0.104	0.038	0.038	0.038	94
600	A-BCEE (c = 500)	0.102	0.039	0.039	0.039	95
600	A-BCEE (c = 1000)	0.101	0.040	0.039	0.039	96
1000	True model	0.101	0.024	0.023	0.023	96
1000	Fully adjusted model	0.100	0.032	0.032	0.032	95
1000	BMA	0.109	0.028	0.028	0.029	93
1000	BAC ($C_V^m(\omega)$)	0.101	0.029	0.031	0.031	92
1000	BAC ($\omega = \infty$)	0.099	0.032	0.032	0.032	95
1000	TBAC ($\omega = \infty$)	0.099	0.032	0.032	0.032	95
1000	A-BCEE (c = 100)	0.104	0.029	0.029	0.030	95
1000	A-BCEE (c = 500)	0.102	0.031	0.030	0.030	95
1000	A-BCEE (c = 1000)	0.101	0.031	0.030	0.030	95

LEGEND: See Table 4.3.

4.7.8 Comparison of the distribution of $\hat{\beta}$ obtained from A-BCEE and BAC

Figure 4.3: Comparison of the distribution of $\hat{\beta}$ obtained from BAC ($\omega = \infty$) and A-BCEE (c = 100, 500, and 1000) for all four data-generating processes and a sample size $n = 600$. The red line corresponds to the true value $\beta = 0.1$

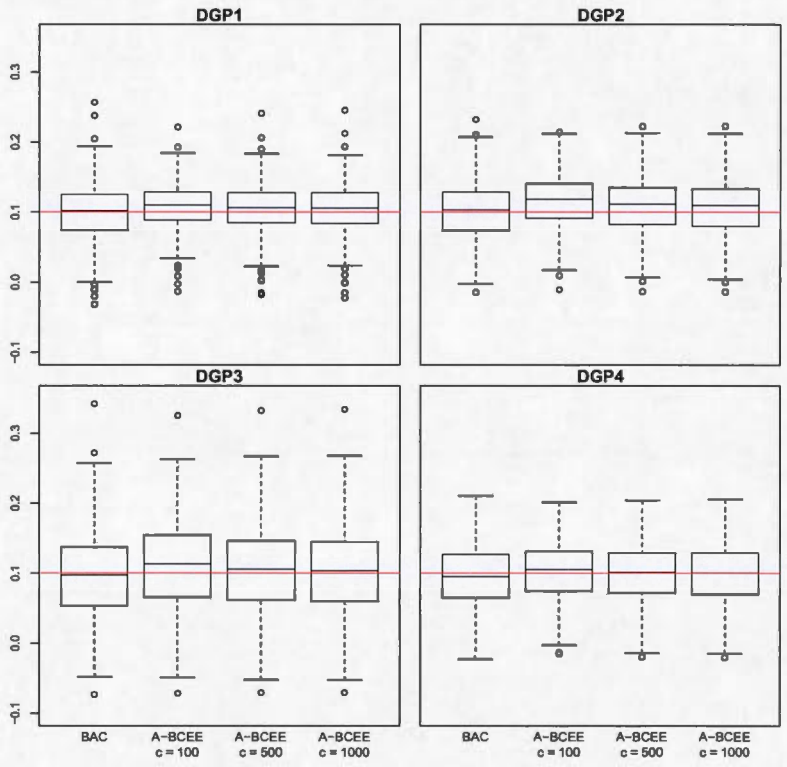
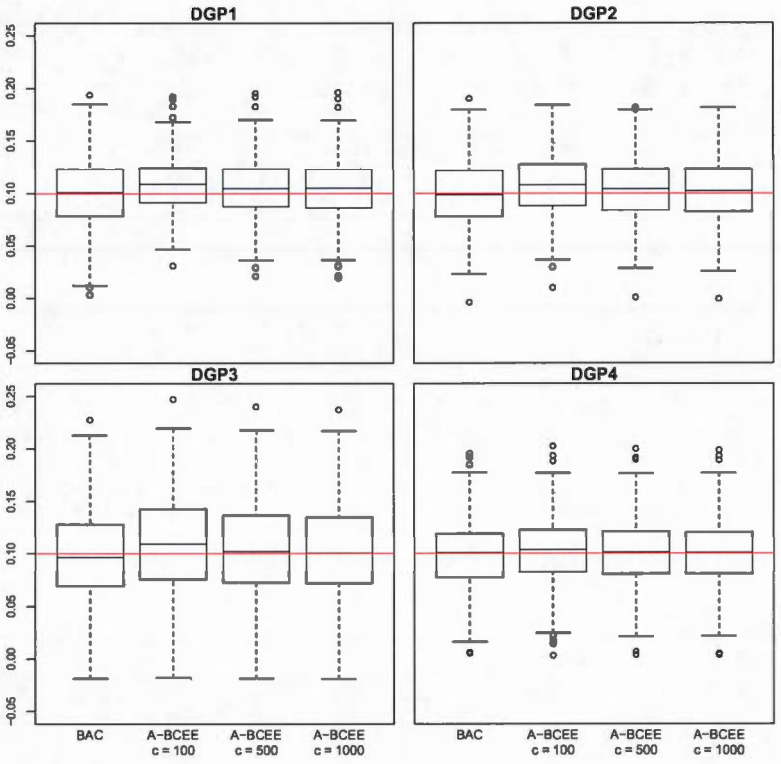


Figure 4.4: Comparison of the distribution of $\hat{\beta}$ obtained from BAC ($\omega = \infty$) and A-BCEE (c = 100, 500, and 1000) for all four data-generating processes and a sample size $n = 1000$. The red line corresponds to the true value $\beta = 0.1$



Deuxième partie

Identification de modèles pour l'inférence causale guidée par la littérature scientifique

CHAPITRE V

INTRODUCTION AU *HONOLULU HEART PROGRAM* ET AUX MODÈLES STRUCTURAUX MARGINAUX

La première partie de la thèse portait sur la sélection de modèles pour l'inférence causale dans un contexte où les connaissances du domaine d'application sont peu développées. La deuxième partie de la thèse, composée de ce chapitre et des deux suivants, porte encore sur la sélection de modèles pour l'inférence causale, mais cette fois dans un contexte où la compréhension du problème substantiel est plus avancée. Les travaux qui y sont présentés découlent d'une analyse des données du *Honolulu Heart Program* (HHP).

Les sections 5.1 et 5.2 du présent chapitre décrivent respectivement les objectifs de notre étude et les données du HHP. À la section 5.3, nous présentons les modèles utilisés pour les analyses, les modèles structuraux marginaux (MSMs), alors que les résultats obtenus sont brièvement exposés à la section 5.4. Une présentation et une discussion plus détaillée des résultats substantiels sont disponibles dans un article scientifique dont je suis le deuxième auteur. Cet article a été soumis à une revue scientifique du domaine de la santé et se trouve en appendice A. Le prochain chapitre est formé d'une version longue d'un article scientifique qui sera soumis à la revue *Epidemiology* décrivant la méthodologie statistique que j'ai élaborée pour analyser les données du HHP. Cette méthodologie comporte plusieurs éléments novateurs, notamment parce qu'une sélection de variables basée sur des DAGs validés et améliorés en utilisant les données est effectuée. La méthodologie proposée peut

ainsi faciliter l'implantation des MSMs en simplifiant le processus d'identification des variables potentiellement confondantes. Le chapitre 7 est formé d'un article dont je suis le premier auteur publié dans *Statistics in Medicine* et portant également sur l'implantation des MSMs. Plus spécifiquement, cet article s'intéresse au type de pondération utilisé pour estimer les paramètres des MSMs en relation avec le modèle structurel choisi.

5.1 Objectifs de l'analyse des données du *Honolulu Heart Program*

Les objectifs de cette analyse secondaire des données du HHP ont été élaborés par Amanda Rossi, étudiante au doctorat en sciences de l'exercice à l'Université Concordia, et par son directeur de recherche, le professeur Simon Bacon également de l'Université Concordia. Ces objectifs consistaient à étudier les effets causaux de l'activité physique sur la pression artérielle, la mortalité et le risque d'événements cardiaques indésirables majeurs (ÉCIMs) ainsi que les effets de la pression artérielle sur la mortalité et sur le risque d'ÉCIMs. On souhaitait aussi explorer la possibilité que l'effet de l'activité physique sur la mortalité soit médié par un effet de l'activité physique sur la pression artérielle.

Bien que de nombreuses études ont déjà porté sur ces sujets (par exemple Vatten *et al.* (2006), Lee & Skerrett (2001), Prospective Studies Collaboration (2002)), ces études ont généralement procédé en prenant des mesures initiales concernant le niveau d'activité physique ou la pression artérielle, ainsi que pour des facteurs potentiellement confondants, et ont suivi les participants pour une période donnée afin de déterminer le moment du décès. Ces études n'ont pas pris en compte la possibilité que le niveau d'activité physique et la pression artérielle peuvent varier au cours de la vie d'un même individu.

Notre analyse vise ainsi, entre autres, à offrir une nouvelle perspective en tenant compte de l'impact des changements des niveaux d'activité physique et de pression artérielle s'opérant en cours de vie.

5.2 Les données du *Honolulu Heart Program*

Le HHP est une étude de cohorte prospective ayant suivi des hommes Japonais-Américains entre 1965 et 1994. Les participants vivaient sur l'île d'Oahu à Hawaï et ont été recrutés à l'aide d'une liste d'inscrits à la réserve de l'armée américaine pour la Deuxième Guerre mondiale. Les participants devaient être nés entre 1900 et 1919 pour pouvoir participer à l'étude et devaient donc avoir entre 45 et 68 ans au moment du début de l'étude (Worth & Kagan, 1970). Nos analyses secondaires des données ont principalement porté sur trois examens de suivis pour lesquels des mesures similaires de l'activité physique, de la pression artérielle systolique et de la pression artérielle diastolique ont été prises : l'examen 1 (1965-1968), l'examen 2 (1968-1971) et l'examen 4 (1991-1993). Le protocole de collecte des données a été précédemment décrit par Kagan *et al.* (1974).

Les variables d'intérêt principal sont le niveau d'activité physique, la pression artérielle systolique, la pression artérielle diastolique, le temps de survie et le temps avant le premier ÉCIM. D'autres variables ont été sélectionnées en raison de leur importance clinique ou parce qu'elles constituaient des variables potentiellement confondantes selon notre connaissance de la littérature scientifique. Les variables sélectionnées devaient également avoir été mesurées de façon similaire aux trois examens retenus.

5.2.1 Mesure de l'activité physique

L'activité physique a été mesurée à l'aide d'un questionnaire auto-administré. Aux examens 1 et 4, les participants rapportaient le nombre d'heures par jour passées à effectuer cinq niveaux d'activité physique allant d'aucune activité physique (par exemple, dormir, se coucher) à des activités exigeantes (par exemple, pelleter, faire des poids et haltères). Aux examens 1 et 2, les participants devaient qualifier séparément leur niveau d'activité physique au travail et à la maison comme étant « surtout assis », « modéré » ou « élevé ».

Pour nos analyses, le niveau d'activité physique a été catégorisé comme étant « actif » ou « inactif ». À l'examen 1 et à l'examen 4, les participants étaient réputés comme actifs s'ils effectuaient au moins 1 heure d'activité physique modérée ou très exigeante et inactifs autrement. À l'examen 2, les participants ayant rapporté être modérément ou très actifs à la maison ou au travail étaient considérés comme actifs et, autrement, comme inactifs. Puisqu'il était possible d'utiliser ces deux approches de catégorisation du niveau d'activité physique à l'examen 1, nous avons pu mesurer le niveau de concordance entre l'approche basée sur le nombre d'heures passées à effectuer des activités modérées ou très exigeantes et celle basée sur le niveau d'activité à la maison et au travail. Le niveau de concordance nous semblait adéquat (taux de concordance = 84%, κ de Cohen = 0.42).

5.2.2 Mesure de la pression artérielle

Les pressions artérielles systolique et diastolique ont été mesurées à l'aide d'un manomètre au mercure par un professionnel à tous les examens. Une série de mesures de la pression artérielle était prise alors que les participants étaient en position assise. Pour nos analyses, la moyenne de la série de mesures prises à chaque examen pour chaque type de pression artérielle a été utilisée.

5.2.3 Mortalité et événements cardiaques

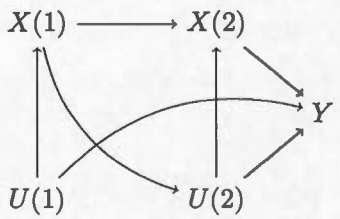
La mortalité et les événements cardiaques ont été répertoriés entre le moment d'entrée dans l'étude et la fin de la période de suivi, en décembre 1994, par le biais des registres hospitaliers, des avis de décès et des certificats de décès produits par le département de la santé d'Hawaï. Les ÉCIMs ont été définis comme n'importe quel événement fatal ou non fatal, incluant les infarctus du myocarde, les attaques, les pontages, les insuffisances coronariennes, les angioplasties coronariennes et autres chirurgies cardiaques.

5.3 Modèles structuraux marginaux

L'étude de l'effet causal d'une exposition variant dans le temps sur une issue d'intérêt est généralement impossible à l'aide des modèles statistiques habituels (par exemple, les modèles linéaires généralisés). En effet, il arrive fréquemment qu'une variable confondant la relation entre l'exposition et la réponse à un temps donné soit sur le chemin entre l'exposition passée et la réponse. Il s'agit d'une situation souvent nommée comme étant la confusion dépendante du temps (*time-dependent confounding*, Robins *et al.* (2000)).

Pour illustrer ce phénomène, considérons une exposition mesurée lors de deux visites de suivis ($X(1)$, $X(2)$) ainsi qu'une variable potentiellement confondante également mesurée à deux reprises ($U(1), U(2)$). On désire estimer l'effet de l'exposition aux deux visites sur la réponse observée à la fin de l'étude (Y). Le DAG de la figure 5.1 représente les liens causaux entre ces différentes variables.

Figure 5.1: Illustration du phénomène de confusion dépendante du temps



On pourrait imaginer estimer l'effet causal total de $X(1)$ et de $X(2)$ sur Y, par exemple, à l'aide d'un modèle de régression linéaire de la forme $E[Y|X(1), X(2)]$ avec ajustement approprié pour des facteurs confondants. Cependant, cette approche se révèle être insatisfaisante. En effet, si on considère un modèle ajustant pour $U(2)$, par exemple $E[Y|X(1), X(2)] = \beta_0 + \beta_1 X(1) + \beta_2 X(2) + \beta_3 U(1) + \beta_4 U(2)$, le chemin causal $X(1) \rightarrow U(2) \rightarrow Y$ est bloqué et ainsi β_1 ne représente pas l'effet causal total de $X(1)$ sur Y. Toutefois, en considérant un modèle n'ajustant pas pour $U(2)$, par exemple $E[Y|X(1), X(2)] = \beta_0 + \beta_1 X(1) + \beta_2 X(2) + \beta_3 U(1)$, le chemin porte-arrière $X(2) \leftarrow U(2) \rightarrow Y$ demeure ouvert et β_2 ne représente pas l'effet causal de $X(2)$ sur

Y. On se retrouve ainsi dans une situation où il faudrait à la fois ajuster pour $U(2)$ et ne pas ajuster pour $U(2)$.

Afin de pallier à ce problème des modèles statistiques classiques, deux nouvelles classes de modèles causaux ont été introduites par Robins juste avant le début du dernier siècle : les modèles structuraux emboîtés (*structural nested models*) et les MSMs (e.g. Robins (2000)). Les MSMs ont été introduits quelques années après les modèles structuraux emboîtés et sont décrits comme étant « une classe plus simple » de modèles causaux (Robins, 1997), notamment parce que leur implantation correspond davantage à celle des modèles statistiques utilisés en absence de confusion dépendante du temps (Robins, 2000).

Nous avons utilisé des MSMs afin d'analyser les données du HHP. Ces modèles sont habituellement décrits à l'aide du paradigme contrefactuel à l'inférence causale. Nous effectuons ici une courte présentation générique des MSMs à mesures répétées, des MSMs classiques et des MSMs de Cox, en considérant ce paradigme. Néanmoins, tel que mentionné, nous avons plutôt considéré le paradigme causal graphique pour l'implantation des MSMs. Le prochain chapitre porte sur l'implantation spécifique des modèles que nous avons effectuée pour l'analyse des données du HHP.

5.3.1 MSMs à mesures répétées

Pour présenter les MSMs à mesures répétées, nous suivons de près la notation de Hernán & Brumback (2002). Nous adoptons la convention consistant à représenter les variables aléatoires par des lettres majuscules et leurs réalisations par des lettres minuscules.

Nous considérons une étude portant sur n sujets échantillonnés d'une population pour lesquels différentes variables sont mesurées lors de $t = 0, ..., T$ visites ou temps de suivi, en plus de la réponse mesurée à la fin du suivi (à $T+1$). Soit $Y_i(t+1)$ la réponse du sujet i, $i = 1, ..., n$, à la $t+1$ième période de suivi et soit $X_i(t)$ l'exposition pour le sujet

i à la période t. Pour simplifier la présentation, nous supposons que X est une variable binaire (0/1). Nous utilisons une barre pour représenter l'historique d'une variable, par exemple $\bar{X}(t) = \{X(u)|u = 0, 1, ..., t\}$. Soit $Y_{i\bar{x}}(t+1)$ la réponse contrefactuelle pour le sujet i qui aurait été observée si, possiblement contrairement aux faits, le sujet i avait eu l'historique d'exposition $\bar{x}(t)$ au lieu de son historique d'exposition observé, $\bar{x}_i(t)$. Pour un individu donné, seule la réponse contrefactuelle correspondant à son historique d'exposition réel est observée, c'est-à-dire que lorsque $\bar{x}(t) = \bar{x}_i(t), Y_{i\bar{x}}(t+1) = Y_i(t+1)$, mais si $\bar{x}(t) \neq \bar{x}_i(t)$, $Y_{i\bar{x}}(t+1)$ n'est pas observée.

Un MSM à mesures répétées est un modèle pour l'espérance de $Y_{\bar{x}}(t+1)$, $t = 0, .., T$, en fonction de l'historique d'exposition $\bar{x}(t)$:

$$E[Y_{\bar{x}}(t+1)] = g(\bar{x}(t), \gamma), \tag{5.1}$$

où g est une fonction choisie par l'utilisateur, par exemple $g(\bar{x}(t), \gamma) = \gamma_0 + \gamma_1 x(t) + \gamma_2 t$ ou encore $g(\bar{x}(t), \gamma) = \gamma_0 + \gamma_1 \sum_{k=0}^{t} x(k) + \gamma_2 t$. Puisqu'une seule réponse contrefactuelle est observée par individu, le modèle (5.1) ne peut pas être ajusté directement sur les données observées. Cependant, les paramètres γ de (5.1) correspondent aux paramètres β du modèle

$$E[Y(t+1)] = g(\bar{x}(t), \beta) \tag{5.2}$$

en absence de confusion. Les paramètres β, et donc γ, peuvent, par exemple, être estimés de façon semi-paramétrique à l'aide d'équations d'estimation généralisées (*generalized estimating equations*, GEE). En présence de confusion, les paramètres peuvent tout de même être estimés en ajustant le modèle (5.2) sur les données observées, à condition qu'elles soient pondérées selon les poids

$$w_i(t) = \prod_{k=0}^{t} \frac{1}{P(X(k) = x_i(k) | \bar{X}(k-1) = \bar{x}_i(k-1), \bar{L}(k) = \bar{l}_i(k))}, \tag{5.3}$$

où $\bar{L}(k)$ est un ensemble de variables telles que l'hypothèse d'ignorabilité séquentielle suivante est respectée

$$Y_{\bar{x}}(t+1) \perp\!\!\!\perp X(k) | \bar{X}(k-1), \bar{L}(k) \ \forall \ \bar{x} \text{ et } t \geq k.$$

Intuitivement, l'effet de la pondération est de répliquer une étude où l'exposition est séquentiellement randomisée à chaque visite de suivi. En effet, sur les données pondérées, l'exposition à la visite t devient indépendante de l'exposition antérieure et indépendante des variables potentiellement confondantes $\bar{L}(t)$. Ainsi, l'effet causal de l'exposition peut être directement estimé sur ces données (voir par exemple l'annexe 1 de Robins *et al.* (2000)).

En pratique, les poids (5.3) peuvent avoir une variance très élevée, ce qui conduit également à des estimateurs des paramètres γ dont la variance est élevée. Il est ainsi recommandé d'effectuer une stabilisation des poids (e.g. Robins *et al.* (2000)). Un phénomène moins connu est que la stabilisation des poids peut non seulement avoir un impact sur la variance des estimateurs, mais également sur leur biais. L'article présenté au chapitre 7 porte sur cette problématique.

5.3.2 MSMs classiques

Nous n'avons pas utilisé de MSMs classiques afin d'analyser les données du HHP. Toutefois, ces modèles sont étudiés au chapitre 7. Les MSMs classiques sont très similaires aux MSMs à mesures répétées. La principale différence est que dans les

MSMs classiques, au lieu de modéliser la réponse contrefactuelle à chaque temps de suivi en fonction de l'historique d'exposition antérieur, seule la réponse contrefactuelle à la fin du suivi, $Y_{\bar{x}} = Y_{\bar{x}}(T+1)$, est modélisée en fonction de l'ensemble de l'historique d'exposition, $\bar{x} = \bar{x}(T)$:

$$E[Y_{\bar{x}}] = g(\bar{x}, \gamma), \tag{5.4}$$

où g est une fonction choisie par l'utilisateur, par exemple $g(\bar{x}, \gamma) = \gamma_0 + \gamma_1 \sum_{t=0}^{T} x(t)$ ou encore $g(\bar{x}, \gamma) = \gamma_0 + \gamma_1 x(T) + ... + \gamma_{T+1} x(0)$. Similairement aux MSMs à mesures répétées, le modèle (5.4) ne peut pas être ajusté directement sur les données observées, mais les paramètres γ correspondent aux paramètres β du modèle de régression linéaire

$$E[Y] = g(\bar{x}, \beta) \tag{5.5}$$

en absence de confusion. Lorsqu'il y a présence de confusion, les paramètres peuvent être estimés en ajustant le modèle (5.5) sur les données observées pondérées selon les poids

$$w_i = \prod_{k=0}^{T} \frac{1}{P(X(k) = x_i(k) | \bar{X}(k-1) = \bar{x}_i(k-1), \bar{L}(k) = \bar{l}_i(k))}, \tag{5.6}$$

sous l'hypothèse d'ignorabilité séquentielle

$$Y_{\bar{x}} \perp\!\!\!\perp X(k)|\bar{X}(k-1), \bar{L}(k) \ \forall \ \bar{x} \text{ et } k.$$

5.3.3 MSMs de Cox

Les MSMs de Cox partagent plusieurs similarités avec les MSMs à mesures répétées ; nous effectuons ainsi une présentation plus brève, axée sur les particularités des MSMs de Cox. Pour présenter ces modèles, nous suivons la notation de Xiao *et al.* (2010) et de Robins (1997) en y apportant quelques modifications mineures.

Une fois de plus, nous considérons une étude longitudinale portant sur n sujets échantillonnés d'une même population. Soit T_i le temps de survie observé (ou le temps s'écoulant avant un événement d'intérêt) pour le sujet i, $i = 1, ..., n$, et soit $X_i(t)$ l'exposition du sujet i au temps t, $t \geq 0$. Ainsi, dans un MSM de Cox, $X_i(t)$ est considérée comme un processus stochastique à temps continu, alors que dans un MSM à mesures répétées, $X_i(t)$ est considérée comme un processus stochastique à temps discret.

Toutefois, en pratique, $X_i(t)$ n'est pas mesurée de façon continue, mais plutôt à des moments spécifiques. Notons par $k = 1, ..., m_i$ les visites de suivi de l'individu i et par $m_i(t)$ le nombre de visites auquel le sujet i a participé dans l'intervalle $[0, t]$. Notons de plus par $t_i(k)$ le temps t correspondant à la visite k pour l'individu i, c'est-à-dire $t_i(k) = \underset{t}{\text{argmax}}\{m_i(t) \leq k\}$. Par ailleurs, nous utilisons de nouveau une barre afin de représenter l'historique d'une variable, par exemple $\bar{X}(t) = \{X(k)|0 \leq k < t\}$.

Soit $T_{i\bar{x}}$ le temps de survie contrefactuel du sujet i si, possiblement contrairement aux faits, son historique d'exposition avait été $\bar{x}$ plutôt que $\bar{x}_i$. Un MSM de Cox est un modèle pour le risque instantané contrefactuel (*hazard rate*) au temps t correspondant à l'historique d'exposition $\bar{x}$:

$$\lambda_{T_{\bar{x}}}(t) = \lambda_0(t) \exp(g(\bar{x}(t), \gamma)),$$

où $\lambda_0(t)$ est une fonction de risque de base non-spécifiée pour $\bar{x}(t) = \bar{0}$ et g est une fonction spécifiée par l'utilisateur. Par exemple, $g(\bar{x}(t), \gamma) = \gamma_1 X(t)$ ou encore $g(\bar{x}(t), \gamma) = \gamma_1 \int_0^t X(k)dk$.

Les paramètres γ du MSM de Cox peuvent être estimés par les paramètres β d'un modèle de Cox sur les données observées pondérées avec des covariables dépendantes du temps

$$\lambda(t) = \lambda_0(t) \exp(g(\bar{x}(t), \beta)).$$

Les poids utilisés pour la pondération sont très similaires à ceux pour les MSMs à mesures répétées (5.3) :

$$w_i(t) = \prod_{k=1}^{m_i(t)} \frac{1}{P(X(t_i(k)) = x_i(t_i(k)) | \bar{X}(t_i(k-1)) = \bar{x}_i(t_i(k-1)), \bar{L}(t_i(k)) = \bar{l}_i(t_i(k)))},$$

où $\bar{L}(t_i(k))$ est un ensemble de variables telles que l'hypothèse d'ignorabilité séquentielle suivante est respectée

$$Y_{\bar{x}}(t) \perp\!\!\!\perp X(t) | \bar{L}(t), \bar{X}(t), \ \forall\ t \text{ et } \bar{x},$$

où

$$Y_{\bar{x}}(t) = \begin{cases} 0 & \text{si } t < T_{\bar{x}} \\ 1 & \text{sinon.} \end{cases}$$

5.4 Résumé des analyses statistiques effectuées

Puisque le prochain chapitre décrit en détail les analyses statistiques effectuées, nous n'en faisons ici qu'une description très brève, mais suffisante pour résumer les résultats obtenus.

Pour estimer l'effet de l'activité physique sur la pression artérielle systolique et sur la pression artérielle diastolique, des MSM à mesures répétées ont été utilisés. Dans ces modèles, la pression artérielle à l'examen t, $t = 1, 2, 4$, était modélisée en fonction du niveau d'activité physique le plus récent (à l'examen t).

Pour estimer l'effet de l'activité physique sur la mortalité et sur le risque d'ÉCIMs ainsi que pour estimer les effets des pressions artérielles systolique et diastolique sur la mortalité et sur le risque d'ÉCIMs, des MSMs de Cox ont été utilisés.

Les résultats obtenus indiquent que le fait d'être actif physiquement est associé à une diminution de 2,5 mmHg de pression artérielle systolique (intervalle de confiance (IC) à 95% : -3,5 mmHg à -1,5 mmHg). Cependant, les données sont compatibles avec une absence d'effet de l'activité physique sur la pression artérielle diastolique (différence = 0.3 mmHg; IC à 95% : -0.2 mmHg à 0.8 mmHg). Par ailleurs, le fait d'être actif physiquement est associé avec une réduction du risque instantané de mortalité de 32% (rapport de risque instantané (RR) = 0,68 ; IC à 95% : 0,60 à 0,76) et de 16% du risque instantané d'ÉCIMs (RR = 0,84 ; IC à 95% : 0,75 à 0,93).

Les résultats suggèrent finalement qu'il existe une relation dose-réponse entre la pression artérielle et la mortalité ainsi que le risque d'ÉCIMs, où des valeurs plus

élevées de pression artérielle sont associées à des issues négatives. Les figures 5.2 et 5.3 illustrent les résultats obtenus.

Figure 5.2: Rapport de risque instantané de mortalité selon le niveau de pression artérielle systolique (à gauche) et selon le niveau de pression artérielle diastolique (à droite). La catégorie de référence est < 120 mmHg pour la pression artérielle systolique et < 80 mmHg pour la pression artérielle diastolique. Les barres représentent des intervalles de confiance à 95%.

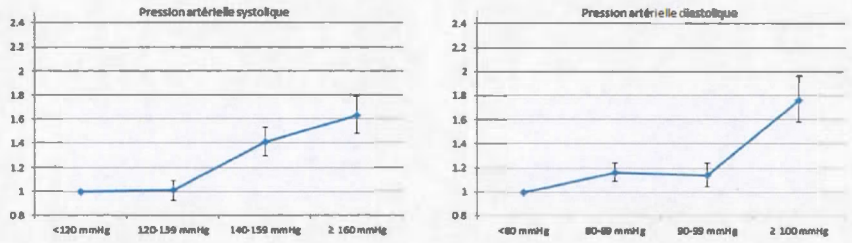
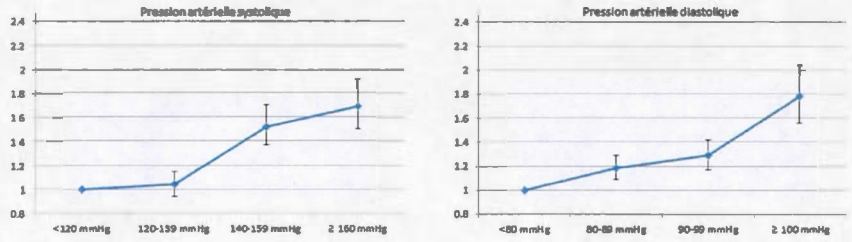


Figure 5.3: Rapport de risque instantané d'ÉCIMs selon le niveau de pression artérielle systolique (à gauche) et selon le niveau de pression artérielle diastolique (à droite). La catégorie de référence est < 120 mmHg pour la pression artérielle systolique et < 80 mmHg pour la pression artérielle diastolique. Les barres représentent des intervalles de confiance à 95%.



Les résultats des analyses statistiques sont davantage détaillés et discutés dans l'article de Rossi *et al.* (2015) disponible en appendice A. Tel que mentionné précédemment, la méthodologie statistique que j'ai élaborée est précisée au prochain chapitre.

CHAPITRE VI

DEUXIÈME ARTICLE : A GRAPHICAL PERSPECTIVE OF MARGINAL STRUCTURAL MODELS WHEN ESTIMATING THE CAUSAL RELATIONSHIPS BETWEEN PHYSICAL ACTIVITY, BLOOD PRESSURE, AND MORTALITY

Denis Talbot, Amanda M. Rossi, Simon L. Bacon, Juli Atherton, Geneviève Lefebvre

Abstract: Estimating causal effects requires important prior subject-matter knowledge and, sometimes, sophisticated statistical tools. The latter is especially true when targeting the causal effect of a time-varying exposure in a longitudinal study. Marginal structural models (MSMs) are a relatively new class of causal models which effectively deal with the estimation of the effects of time-varying exposures. MSMs have traditionally been embedded in the counterfactual framework to causal inference. In this paper, we use the causal graph framework to enhance the implementation of MSMs. We illustrate our approach using data from a prospective cohort study, the Honolulu Heart Program. These data consist of 8,006 men at baseline for which measurements of physical activity and blood pressure were taken at three time-points. Our study focused on the estimation of the causal effects of physical activity on blood pressure, mortality and major adverse cardiovascular events (MACE), and the causal effects of blood pressure on mortality and MACE. First, causal graphs were built to encompass prior knowledge. Those graphs were then validated and improved utilizing structural equation models. We estimated the aforementioned causal effects using MSMs for repeated measures and marginal structural Cox models and guided the implementation of the models with the causal graphs.

Keywords: Marginal structural models, causal diagrams, time-dependent confounding, time-varying exposure, structural equation models.

6.1 Introduction

Estimating a causal effect in an observational study is not an easy task. This is true despite the existence of well-established frameworks for causal inference in observational setups (Pearl, 2009; Rubin, 1974). In fact, to simply identify confounding covariates, causal inference techniques require a substantial knowledge of the domain of application. When the goal is to estimate the causal effect of a time-varying exposure, the task is even more complicated. The standard approach of adjusting for confounding covariates (e.g., in a linear regression or in a survival model for the outcome) can lead to biased estimates. This happens when a time-varying confounding covariate is an effect of previous exposure (Robins *et al.*, 2000) .

Marginal structural models (MSMs) are a relatively new class of causal models that specifically address this issue (Robins, 1997). Instead of directly adjusting for the confounding covariates in the model for the outcome, the causal effects estimated from MSMs are obtained using inverse probability weighting (Hernán & Brumback, 2002; Xiao *et al.*, 2010). Correctly selecting the variables used to calculate the weights in order to eliminate confounding can be challenging (Cole & Hernán, 2008).

MSMs have traditionally been embedded in Rubin's counterfactual framework to causal inference (Rubin, 1974), even though causal graphs have previously been used to illustrate the relationships between variables in MSMs analyses (VanderWeele *et al.*, 2012; Robins *et al.*, 2000). In this paper, we propose to embed MSMs in the graphical framework to enhance the implementation of these models (Pearl, 2009). We illustrate our approach using data from the Honolulu Heart Program (HHP). The HHP is a cohort study that followed 8,006 Japanese-American men from 1965 until 1994. The main objective of our analyses was to estimate the causal effects of physical activity on blood pressure (BP), mortality and major adverse cardiovascular events (MACE), and

the causal effects of BP on mortality and MACE. As a secondary objective, we wished to explore the potential mediating role of BP on the causal effects of physical activity on survival and MACE. The substantive results are presented in our companion paper (Rossi *et al.*, 2015). The primary aim of the current paper is twofold: 1) provide a thorough presentation of the statistical methodology used to obtain these results; 2) compare the results obtained using our graphical approach with those obtained using a naive approach for the selection of the variables in the weight models. A secondary aim is to show the validity of fitting conditional marginal structural models for repeated measures (MSMRMs) in the context implied by our data.

The outline of the paper is as follows. In Section 6.2, we describe the data from the Honolulu Heart Program. We begin Section 6.3 by discussing how causal graphs can be used to encompass prior knowledge and to identify variables that are sufficient to avoid confounding. We then explain how structural equation models (SEMs) were used to verify if the suggested graphs were supported by the observed data. In Section 6.4, we revisit MSMRMs from a graphical perspective (Robins, 2000; Hernán & Brumback, 2002). Additionally, we introduce a conditional version of MSMRMs which adjusts for time-varying covariates without introducing bias in the estimation of the average causal effect (Hernán & Brumback, 2002). Section 6.5 describes marginal structural Cox models (MSCMs) using a graphical perspective (Xiao *et al.*, 2010). Using the HHP data, we contrast the naive MSMRMs results with those from our proposed approach in Section 6.6. In Section 6.7, we compare conditional and unconditional MSMRMs using both simulated and the HHP data. We discuss our methodology in Section 6.8.

6.2 Data

The Honolulu Heart Program is a cohort study that followed 8,006 Japanese-American men living on the island of Oahu, Hawaii from 1965 until 1994. The participants were initially recruited between 1965 and 1968 from a listing of selective service registrants.

The data collection protocol has been described elsewhere (Kagan *et al.*, 1974). Our analyses were based on three examinations for which comparable measures of physical activity, and both systolic and diastolic blood pressures (SBP and DBP, respectively) were taken: Exam 1 (1965-1968), Exam 2 (1968-1971) and Exam 4 (1991-1993). For those subjects who did not participate at Exam 4, a fourth examination (Exam 3, 1971-1975) was used to estimate their right censorship times due to lost to follow-up. To simplify the presentation, we denote Exam 1, 2 and 4 as Visit 1, 2 and 3, respectively, throughout.

The variables of main interest were self-reported physical activity (active or inactive), SBP (in mmHg), DBP (in mmHg), survival time (in days since birth) and time before a MACE (in days since birth). The HHP variables that were identified as clinically relevant or as potential confounders, and that were measured in a similar manner at all three visits were selected for the analyses. Those variables, which are all time-varying, are: age (in years), employment status (currently employed or not), body mass index (in kg/m^2), smoking status (current smoker, previous smoker or never smoker) and anti-hypertension medication usage (yes or no). More information about how the variables were measured is available elsewhere (Rossi *et al.*, 2015).

6.2.1 Data treatment

We used age in days as the time-scale for both time-to-event variables (survival and time to MACE) and considered them to be left truncated at the time of Visit 1 (Thiébaut & Bénichou, 2004; Kom *et al.*, 1997). For individuals whose event was not recorded during the study, the time-to-event was right censored at the elapsed time between birth and either the time of their last examination, if they did not attend Visit 3, or one year after Visit 3 otherwise. Note that for each individual, we do not know the exact amount of time elapsed between Visit 3 and the end of monitoring. However, we know it to be at least one year and at most four years. Sensitivity analyses (not presented) were performed to verify if varying the estimated elapsed time between Visit 3 and the end of monitoring from one to four years changed any conclusions.

The results we obtained were very similar, and hence we took the time between Visit 3 and end of monitoring to be one year. The time to MACE for individuals who died before experiencing a MACE was considered to be right censored at death time (see Bakoyannis & Touloumi (2012) for a discussion and simulations of this approach).

When SBP and DBP were used as exposure variables in the statistical analyses, we divided each of them in four categories according to a common BP classification scheme: < 120mmHg, $120 - 139$mmHg, $140 - 159$mmHg, and $>= 160$mmHg for SBP and < 80mmHg, $80-89$mmHg, $90-99$mmHg, and $>= 100$mmHg for DBP (Chobanian *et al.*, 2003).

For every MSCM and MSMRM analysis, we built an augmented dataset where each subject-visit corresponded to one row. If a row contained missing values for at least one variable required to estimate a given effect, then it was ignored for that estimation (listwise deletion was performed). A table of the amount of data available for each MSM analysis is provided in Rossi *et al.* (2015). For the SEMs analyses, which were performed prior to the MSMs analyses, the data was kept in a wide format where every subject corresponded to one row.

In the next section, we describe how causal graphs were constructed to represent the causal links between the HHP covariates considered in our study.

6.3 Building causal graphs

The issue of confounding is particularly challenging in the context of longitudinal data, such as the HHP, where intermediate covariates in the pathway between the exposure and the outcome can also act as confounding covariates. Therefore, before performing any statistical analyses, we needed to determine how to correctly account for potential confounding. Using substantive prior knowledge, we began by drawing directed acyclic graphs (DAGs) to represent the causal relationships between the selected variables at all visits (Hernán *et al.*, 2002). In these DAGs, the observed

variables are depicted by nodes. The variables may be connected to each other by directed arrows (e.g., $A \rightarrow B$) that represent cause to effect relationships (e.g., A causes B) or by dashed two-sided arrows that are used to represent unobserved common causes between two variables ($A \leftarrow - \rightarrow B \equiv A \leftarrow C \rightarrow B$). These two-sided arrows are a notational shortcut since DAGs do not permit bi-directional causal links *per se*.

The main objective in building the DAGs was to identify sets of variables that could be used to eliminate confounding. First, we used SEMs to determine if our initial postulated DAGs were supported by the data. We then used information from the data to help us modify the initial DAGs into their final form.

6.3.1 Building the initial DAGs

We started by building DAGs to represent the relationships between the time-varying variables listed in Section 6.2. Because the secondary objective of our study was to investigate whether SBP and DBP mediate the effects of physical activity on survival and MACE, we constructed one DAG for the relationships between physical activity, SBP, DBP and time of survival (DAG for survival), and one for the relationships between physical activity, SBP, DBP and time to MACE (DAG for MACE). The inclusion or exclusion of arrows between variables and the directionality of the included arrows were carefully decided based on prior knowledge in the scientific literature.

Because the HHP study features relatively few visits, most of which are very far apart in time, we allowed for the existence of cause-to-effect relationships between variables measured at a same visit, even though this is not strictly possible because a cause always precedes its effect in time. Had we not done so, we would have had no pre-treatment variables for the data at Visit 1. Moreover, it was unlikely that variables measured many years in the past could effectively eliminate confounding bias, whereas visit-specific variables could be used as proxies to pre-treatment versions of themselves.

6.3.2 Assessing the fit of and improving the initial DAGs

We verified if our proposed DAGs fitted the data well using SEMs. SEMs are statistical models that combine qualitative cause-effect assumptions with data to test causal models and estimate causal relationships. Most current SEM packages assume linear relationships between variables and multivariate normality. We used the `lavaan` package in R to fit the SEMs (R Core Team, 2014; Rosseel, 2012). The multivariate normality assumption is untenable in our case since many of our variables are not continuous (e.g., *Smoking* and *Employment*), hence we assessed the goodness-of-fit of our proposed causal models with Bollen-Stine bootstrap (Bollen & Stine, 1992), a statistical test that is robust to non normality of data.

We note, however, that we were partially restrained in our ability to test the proposed DAGs. For instance, only subjects without any missing data at any visits can be included in the SEMs when using the Bollen-Stine bootstrap in `lavaan` v0.5-16. Moreover, censored variables are also not currently handled. Despite these limitations, we believed that any input we could get from the data to assess the correctness of our initial DAGs was valuable. Indeed, we were able to test the appropriateness of the postulated relationships between every variable at every visit, except the relationships involving time of survival or time to MACE.

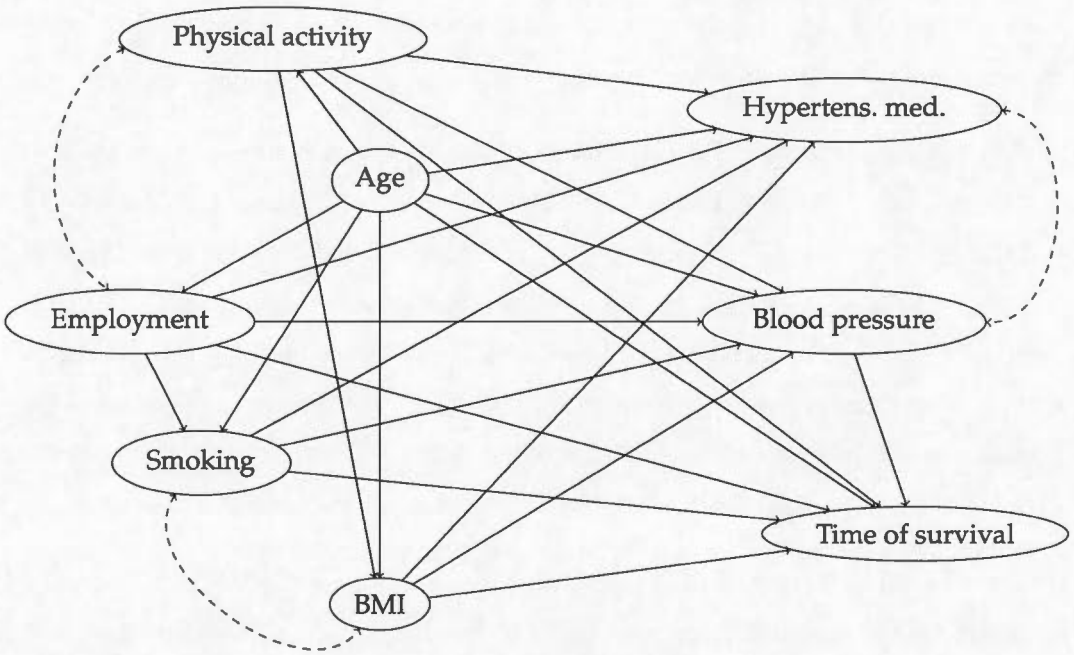
Because the number of available subjects is largest at Visit 1 and smallest at Visit 3, we took full advantage of the available information by sequentially fitting larger and larger models. This strategy wherein SEMs of increasing complexity are considered, was also guided by the longitudinal aspect of the data. We began by fitting SEMs that only involved the relationships between the variables at Visit 1, then we fit SEMs for Visits 1 and 2, and lastly, SEMs for Visits 1, 2 and 3. Moreover, since SBP and DBP are often strongly correlated, we fitted separate SEMs for these variables. Thus, we tested a total of six SEMs (1: SBP Visit 1; 2: DBP Visit 1; 3: SBP Visits 1 and 2; 4: DBP Visits 1 and 2; 5: SBP Visits 1, 2 and 3; 6: DBP Visits 1, 2 and 3).

The two initial DAGs we had proposed did not fit the data well according to the chi-square statistics from the six SEMs. The chi-square statistic tests whether the observed data could be compatible with the proposed DAG by comparing the observed covariance matrix of the variables in the SEM representing the DAG with the covariance matrix that is generated by the SEM. Because the fits of the SEMs were poor, we included additional causal links (cause-effect paths and unobserved common causes) between the variables. These links made sense from a substantive point of view and were found using modification indices. A modification index represents the expected improvement in the chi-square statistic that would occur if a causal link were added to a SEM.

The final SEMs for Visit 1 and the final SEMs for Visits 1 and 2 had non-significant chi-square statistics ($p > 0.05$), meaning we could not reject the null hypothesis that the proposed DAGs generated the observed data. Despite the modifications we made, the final SEMs for Visits 1, 2 and 3 still had significant chi-square statistics. We could find no further modifications to the SEMs that made sense from a theoretical point of view. However, we found some observations that were highly influential in the calculations of chi-square statistics. Fitting the models on the data without 69 such observations (2% of the total data) yielded non-significant chi-square statistics. Hence, our SEMs appeared to be reasonable representations of the causal process between the selected variables for most of the data. Using the six final SEMs, we updated our two initial DAGs. Figure 6.1 presents a part of the final DAG for survival, showing nodes at Visit 1 only. The nodes for SBP and DBP have been joined into a single BP node in Figure 6.1 to simplify the presentation. The complete final DAG for the aforementioned relationships is detailed in Rossi *et al.* (2015). Essentially, the core structure for Visits 2 and 3 is the same as for Visit 1 (as depicted in Figure 6.1). Naturally, the final DAG also features extra causal links from variables at Visit 1 to variables at Visits 2 and 3, and links from variables at Visit 2 to variables at Visit 3. The complete final DAG for MACE is exactly the same as the complete final DAG for survival we have just described, except for the time of survival node that is replaced

by a time to MACE node.

Figure 6.1: A close-up of the final DAG for time of survival at Visit 1. The nodes for SBP and DBP have been joined into a single BP node to simplify the presentation.



6.3.3 Identifying confounding variables

If a time-varying confounding variable is on the causal pathway between the exposure and the outcome, direct adjustment for this confounding variable in an outcome model could lead to biased estimates (Robins *et al.*, 2000). The two complete final DAGs obtained in the previous section confirmed that we were in the presence of such time-varying confounding variables. For instance, *BMI* at Visit 1 confounds the relationship between *Physical activity* at Visit 2 and *BP* at Visit 2 (*BP* at Visit 2 $\leftarrow$ *BMI*

at Visit 1 $\rightarrow$ *Physical activity* at Visit 2), and *BMI* at Visit 1 is also an effect of *Physical activity* at Visit 1 (*Physical activity* at Visit 1 $\rightarrow$ *BMI* at Visit 1).

On the basis of a causal DAG, Pearl's back-door criterion provides sufficient conditions to identify sets of variables that eliminate confounding when estimating the causal effect of an exposure variable on an outcome variable (Pearl, 2009). Specifically, it is sufficient to block all back-door paths from the exposure to the outcome and not adjust for any descendants of the exposure.

A path is already blocked if it contains a collider, that is, a variable which has two arrows pointing to it on the path. For instance, in Figure 6.1, the path *Physical activity* $\leftarrow$ *Age* $\rightarrow$ *Hypertension medication* $\leftarrow - - \rightarrow$ *BP* is a back-door path from *Physical activity* to *BP* that is blocked by the collider *Hypertension medication*. Note that a path that is blocked by a collider becomes unblocked if one adjusts for the collider or for one of its descendants. A path that is not blocked by a collider can be blocked by appropriately adjusting for a non-collider on the path. For instance, the back-door path *Physical activity* $\leftarrow$ *Age* $\rightarrow$ *BP* can be blocked by appropriately adjusting for *Age*.

To identify the causal effect of *Physical activity* on *BP* from the DAG in Figure 6.1 for example, we note that there are only two arrows lying on a back-door path and pointing toward *Physical activity*. One stems from *Age* and the other stems from unknown causes which also have arrows pointing toward *Employment*. Therefore, assuming the DAG is correct, all back-door paths from *BP* to *Physical activity* can be blocked by adjusting for *Age* and *Employment*. Moreover, we cannot adjust for *BMI* or *Hypertension medication*, because they are descendants of *Physical activity*. We further remark that adjusting for *Smoking*, in addition to *Age* and *Employment*, though not necessary to avoid confounding, can be done without harm.

In the next two sections, we present the MSMs we used to estimate the causal effects of interest. We also explain how our MSMs were embedded in the graphical framework we have just described. As subsequently detailed, Pearl's back-door criterion was

invoked to identify sets of covariates sufficient to satisfy the sequential randomization assumption underlying each MSM analysis.

6.4 Marginal structural models for repeated measures

In this section, we describe the MSMRMs used to estimate the causal effects of physical activity on current SBP and DBP. In the sequel, we generically explain the modeling process in terms of BP, since it is the same for both SBP and DBP. To simplify the presentation, we proceed for now as if all subjects were observed at every visit.

We first introduce some notation for MSMs. Our notation is very similar to that in Hernán & Brumback (2002) but eliminates the reference to counterfactual outcomes to accommodate the causal graphical framework we consider. Let $i = 1, ..., n$ denote the individuals, $Y(t)$ be the random variable representing the BP value at Visit $t = 1, 2, 3$, and $X(t)$ be the random variable representing the physical activity level at Visit t ($X(t) = 1$ denotes physically active, whereas $X(t) = 0$ denotes physically inactive). We modelled the effect of current and prior physical activity history on current BP as a function of current physical activity (recall the long delay between Visit 2 and Visit 3). We thus considered the following model:

$$E[Y(t)] = \beta_0 + \beta_1 X(t) + \beta_2 Age(t), \tag{6.1}$$

where β_0 is the unknown intercept, β_1 is the unknown parameter associated with the physical activity level and β_2 is the unknown slope parameter associated with the age of subjects at Visit t. Note that it is common in MSMRMs to introduce a parameter associated with t, the Visit number, to allow the intercept to vary with time (Hernán & Brumback, 2002). Because we have considered age as being the time-scale for both survival and time to MACE, it was natural to instead consider a parameter associated with age. An additional reason for preferring this approach was that the range in the

ages of subjects at a given visit was similar to the time elapsed between the start and the end of the study.

Ignoring the complications arising from missing data and possible informative censoring, the parameters of model (6.1) can be directly estimated by fitting a GEE regression to an augmented dataset, where each line corresponds to a given subject at a given visit. However, for β_1 to have a causal interpretation, time-varying confounding must be adequately dealt with. This is done by attributing an inverse probability of treatment weight (IPTW) to each subject-visit.

As seen next in Equation (6.3), sets of variables $\boldsymbol{L}^{XY}(t)$, $t = 1, 2, 3$, were used to calculate the subject-specific IPTWs. Let $y_i(t)$, $x_i(t)$ and $\boldsymbol{l}_i^{XY}(t)$ be the respective observed realizations of $Y(t)$, $X(t)$ and $\boldsymbol{L}^{XY}(t)$ for subject i. We use overbars to denote the history of any time-varying variable. For instance, $\bar{X}(t)$ represents the physical activity levels at Visit t and all prior visits.

The specification of $\boldsymbol{L}^{XY} = \{\boldsymbol{L}^{XY}(1), \boldsymbol{L}^{XY}(2), \boldsymbol{L}^{XY}(3)\}$ can be very challenging. In the counterfactual framework, the variables $\boldsymbol{L}^{XY}$ entering the weight models are chosen so that the sequential (conditional) randomization assumption holds (Hernán & Brumback, 2002). Because of how model (6.1) is specified, this assumption can be simplified as:

$$Y_{\bar{x}}(t) \perp\!\!\!\perp X(t) | \boldsymbol{L}^{XY}(t), \ \forall \bar{x}, t \in \{1, 2, 3\}, \tag{6.2}$$

where $Y_{\bar{x}}(t)$ is the counterfactual BP value at Visit t that would have been observed if, possibly contrary to the fact, the physical activity history $\bar{x}$ had been observed. As discussed in Pearl (2009) Section 3.6.3 and Section 11.3.2, the randomization assumption holds for a given $\boldsymbol{L}^{XY}(t)$ if the back-door criterion holds for this $\boldsymbol{L}^{XY}(t)$. Hence, on the basis of the complete final DAGs mentionned in Section 6.3, we selected the variables in $\boldsymbol{L}^{XY}(t)$ as described in Section 6.3.3. A complete list of the variables in $\boldsymbol{L}^{XY}$ is available in Appendix 6.9.1.

We considered the weighted GEE regression model (6.1) with stabilized weights

$$W_i^{XY}(t) = \prod_{k=1}^{t} \frac{P(X(k) = x_i(k))}{P(X(k) = x_i(k)|\boldsymbol{L}^{XY}(k) = \boldsymbol{l}_i^{XY}(k))} \tag{6.3}$$

and estimated $P(X(k) = x_i(k))$ and $P(X(k) = x_i(k)|\boldsymbol{L}^{XY}(k) = \boldsymbol{l}_i^{XY}(k))$ using logistic regression. Our choice of stabilized weights follows the recommendation given in Talbot *et al.* (2015) when the structural model only includes partial treatment history [1]. In Section 6.6, we show how the estimated causal effect of physical activity on BP when $\boldsymbol{L}^{XY}$ is specified using the graphical approach detailed above differs from the estimate obtained when following the naive approach where all potentially confounding covariates available are selected.

6.4.1 Estimation with incomplete data

Up until now, we have presented the MSMRMs we would have fitted to estimate the effect of physical activity on BP had there been no deaths or losses to follow-up. Recall that the HHP is a longitudinal study that spanned over a very long period of time. Inevitably, many subjects died before the end of the study or were lost to follow-up. Therefore, we did not have a complete dataset where every subject participated at every visit. Subjects who did not participate in all visits could have different characteristics than subjects that did. Biased estimates can arise if incomplete data is not adequately dealt with (Little & Rubin, 2002). Because a weighting scheme is already used to account for confounding, a convenient approach to deal with incomplete follow-up in MSMs is to use inverse probability of censoring weights (IPCWs) (Hernán & Brumback, 2002; Moodie *et al.*, 2008). This is the approach we retained and now describe.

---

1. Note : Le chapitre 7 de la thèse est formé de cet article

Let $C(t)$ be a random variable representing the censoring at Visit t, with $C(0) \equiv 0$, and let $c_i(t)$ be the observed realization for subject i ($c_i(t) = 0$ if subject i is still in the study at Visit t and $c_i(t) = 1$ otherwise). Also, let $\boldsymbol{Z}(t)$ denote the covariates available at Visit t and $\boldsymbol{z}_i(t)$ be their observed values for subject i. Our weights for censoring are

$$W_i^C(t) = \prod_{k=1}^{t} \frac{P(C(k) = 0|C(k-1) = 0)}{P(C(k) = 0|C(k-1) = 0, \boldsymbol{Z}(k) = \boldsymbol{z}_i(k))}.$$

We estimated $P(C(k) = 0|C(k-1) = 0)$ and $P(C(k) = 0|C(k-1) = 0, \boldsymbol{Z}(k) = \boldsymbol{z}_i(k))$ using logistic regression. For $i = 1, ..., n$, we computed the total weights as $W_i^{Total}(t) = W_i^C(t) \times W_i^{XY}(t)$, and then calculated the corresponding normalized weights $NW_i^{Total}(t)$ as described in Equation (4) in Xiao *et al.* (2010). Finally, the GEE regression (6.1) was fitted with weights $NW_i^{Total}(t)$. We used an independent working correlation matrix and a robust variance estimator to account for the repeated measures in the GEE regression (Tchetgen Tchetgen *et al.*, 2012a,b).

6.4.2 Conditional marginal structural models for repeated measures

It is usually recommended not to include time-varying variables in the outcome model (6.1) of a MSMRM (Hernán & Brumback, 2002). This is because some of these variables can act both as confounders and intermediate variables over time (Robins *et al.*, 2000). In this section, we argue that it is safe to include time-varying variables $\boldsymbol{U}(t)$ in the model we consider, even if $\boldsymbol{U}(t)$ includes such time-dependent confounders.

In our study, we also considered the following conditional model to estimate the causal effect of physical activity on BP:

$$E[Y(t)|\boldsymbol{U}(t)] = \beta_0 + \beta_1 X(t) + \beta_2 Age(t) + \boldsymbol{\beta_3 U}(t), \tag{6.4}$$

where β_3 is a vector of unknown parameters. With the back-door criterion (see Section 6.3.3) in mind, the variables $\boldsymbol{U}(t)$ we selected were such that they were not descendants of $X(t)$ according to our complete final DAGs, and that all back-door paths between $X(t)$ and $Y(t)$ remained blocked after conditioning on $\boldsymbol{U}(t)$. These variables are *Employment* and *Smoking* at Visit t. Note that $\boldsymbol{U}(t)$ may have included variables on the causal pathway between $X(s)$ and $Y(t)$, $s < t$, without introducing bias in the estimation of β_1. This is because model (6.4) only considers the effect of $X(t)$ on $Y(t)$. As seen in Section 6.7, which presents a simulation study that validates our methodology, one potential advantage to conditioning on $\boldsymbol{U}(t)$ in model (6.4) is to reduce the standard error of the estimated causal effect.

We estimated the corresponding causal effect of physical activity on BP as presented in Section 6.4.1. That is, we built an augmented dataset and fitted the weighted GEE regression model (6.4) using the same normalized weights as before. In the sequel, we refer to model (6.4) we have just introduced as a conditional MSMRM, as opposed to the unconditional MSMRM presented previously. Arguably, because age is a time-varying covariate, model (6.1) could also be considered as a conditional MSMRM.

This completes our presentation of the MSMRM methodology used to estimate the effect of physical activity on SBP and DBP. In the next section, we present the MSCM methodology utilized to estimate the separate causal effects of physical activity and BP on survival time and time to MACE.

6.5 Marginal structural Cox models

We used MSCMs to estimate the causal effects that involved the two time-to-event outcomes of interest, that is, survival time and time to MACE. We describe in detail the process we followed for the estimation of the causal effect of physical activity on survival time. The estimation process for each of the three other relationships investigated was similar (additional precisions are provided at the end of this section).

Our MSCM methodology has strong connections with the one proposed by Xiao *et al.* (2010). It also shares similarities with the MSMRM methodology described in the previous section.

We believe that the causal effect of physical activity history on survival time is mostly a function of current physical activity level. Hence, we considered the following model for the hazard at age τ

$$\lambda(\tau) = \lambda_0(\tau) \exp(\beta_1 X(\tau)), \tag{6.5}$$

β_1 is the unknown parameter associated to the physical activity level $X(\tau)$ and $\lambda_0(\tau)$ is the unspecified baseline hazard at age τ. While it is assumed that $X(\tau)$ is a stochastic process in continuous time, $X(\tau)$ was really only measured at the ages corresponding to examinations. Thus, we took $X(\tau)$ as a step function with steps at the ages corresponding to examinations. Once again, the time-varying confounding problem is solved by using inverse probability weighting.

We define $\boldsymbol{L}^{XT}(t)$ and $W_i^{XT}(t)$ analogously to $\boldsymbol{L}^{XY}(t)$ and $W_i^{XY}(t)$ (see Equation (6.3)), only replacing BP (Y) by survival time (T). To satisfy the conditional ignorability assumption of the MSCMs (Robins, 1997), we selected the variables $\boldsymbol{L}^{XT}(t)$ on the basis of the complete final DAG for survival and the back-door criterion described in Section 6.3.3. The list of the selected variables is again provided in Appendix 6.9.1.

We normalized the weights $W_i^{XT}(t)$ as in Equation (4) from Xiao *et al.* (2010) and fitted a weighted Cox model with hazard (6.5) utilizing those normalized weights. We used a robust estimator for the estimation of the standard errors; this estimator accounts for the dependence between the rows associated to a same subject in the augmented dataset.

We used exactly the same approach to estimate the causal effect of physical activity on time to MACE, only replacing survival time by time to MACE. Moreover, only minor

changes to the methodology were done to estimate the causal effects of SBP and DBP on survival time and on time to MACE. As mentioned in Section 6.2.1, we divided the SBP and DBP values into four categories when these BP variables were used as exposure variables. The probabilities $P(X(k) = x_i(k))$ and $P(X(k) = x_i(k)|\boldsymbol{L}^{XT}(k) = \boldsymbol{l}_i^{XT}(k))$ required in the calculation of $W_i^{XT}(t)$ were estimated using ordinal logistic regression models.

Note that it is not possible to propose a conditional version of the MSCMs, as was done for the MSMRMs. In fact, the Cox proportional hazards model is not collapsible for β_1 over predictors of survival (or time to MACE), even if those predictors are not associated with the exposure (Struthers & Kalbfleisch, 1986). In other words, the causal parameter β_1 defined in a conditional MSCM would generally not have the same true value, and thus not the same interpretation, as the causal parameter β_1 defined in an unconditional MSCM.

6.6 Contrasting our approach with a naive approach

We have presented in the previous sections a graphical approach to MSMs where the covariates selected for estimating the IPTWs are identified using DAGs and the back-door criterion. A more naive approach for estimating the IPTWs is to use every potentially confounding covariates available at a given visit. Using the HHP data, we now illustrate how the results obtained with both approaches can differ. For this illustration, we focus on the estimation of the causal effects of physical activity on SBP and DBP.

The first line of Table 6.1 presents the results obtained by estimating the aforementioned causal effects using the unconditional MSMRM described in Section 6.4. For the naive approach, the causal effects were estimated similarly, only replacing $\boldsymbol{L}^{XY}(t)$ by $\boldsymbol{L}_N^{XY}(t)$ in the IPTWs (6.3). The variables in $\boldsymbol{L}_N^{XY}(t)$, $t = 1, 2, 3$, are listed in Appendix 6.9.1. The results obtained using the naive approach are presented

in the second line of Table 6.1. All other substantive results of our analysis of the HHP data are presented in our companion paper (Rossi *et al.*, 2015).

The estimated causal effects of physical activity on SBP obtained with the naive and the graphical approaches are both compatible with a decrease in SBP when being currently physically active. However, the interpretation of the results for DBP differs. Indeed, the results obtained using the graphical approach are compatible with no effect of physical activity on DBP, whereas the results pertaining to the naive approach suggest that being physically active *increases* DBP. That physical activity would increase DBP is not supported by the current scientific knowledge (Cornelissen & Smart, 2013). The observed divergence in conclusions lends support to our proposed approach.

Table 6.1: Results from the graphical and naive approaches to estimate the causal effect of current physical activity on SBP and DBP (95% confidence intervals in parenthesis).

Approach	SBP	DBP
Graphical	-2.47 (-3.46, -1.48)	0.26 (-0.22, 0.75)
Naive	-1.64 (-2.64, -0.64)	0.96 (0.47, 1.44)

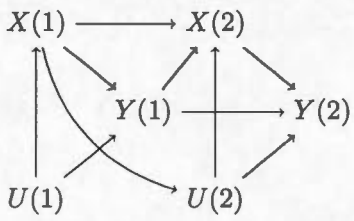
6.7 Comparing conditional and unconditional MSMRMs

Conditional MSMRM estimates of the effects of physical activity on SBP and DBP are not reported in our companion paper. Before presenting these additional HHP estimates, we describe a simulation study devised to investigate the validity of this conditional version of the MSMRMs. In Section 6.7.1, we first describe the four simulation scenarios that were considered in the study. The results from the simulations are presented in Section 6.7.2. Finally, in Section 6.7.3, we compare the results from using the conditional and unconditional MSMRMs to estimate the causal effect of physical activity on BP based on the HHP data.

6.7.1 Simulation scenarios

All the simulation scenarios are compatible with the DAG depicted in Figure 6.2; the exact data-generating equations however differ slightly between scenarios. Although this DAG is simple, it is sufficient to illustrate the main properties of our conditional MSMRM approach.

Figure 6.2: DAG for Scenarios 1-4



Scenario 1 is (essentially) the same as Scenario 3 in Talbot *et al.* (2015). The equations that generated the data are:

$$
\begin{aligned}
U(1) &= \varepsilon_{U(1)}, \\
P(X(1) = 1) &= expit(0.5U(1)), \\
Y(1) &= X(1) + U(1) + \varepsilon_{Y(1)}, \\
U(2) &= 0.5X(1) + \varepsilon_{U(2)}, \\
P(X(2) = 1) &= expit(0.5X(1) + 0.5Y(1) + 0.5U(2)), \\
Y(2) &= X(2) + 0.5Y(1) + U(2) + \varepsilon_{Y(2)},
\end{aligned}
$$

where $expit(z) = exp(z)/(1 + exp(z))$, and $\varepsilon_{U(1)}$, $\varepsilon_{Y(1)}$, $\varepsilon_{U(2)}$, $\varepsilon_{Y(2)}$ are independent $N(0, 1)$ random variables.

Scenario 2 is the same as Scenario 1, but replaces the strong links from $U(1)$ to $Y(1)$, and from $U(2)$ to $Y(2)$ with weak links:

$$Y(1) = X(1) + 0.1U(1) + \varepsilon_{Y(1)}, \text{ and}$$
$$Y(2) = X(1) + 0.5Y(1) + 0.1U(2) + \varepsilon_{Y(2)}.$$

All other data-generating equations are the same as in Scenario 1.

Scenario 3 is also the same as Scenario 1, but features an even stronger link from $U(1)$ to $Y(1)$, and from $U(2)$ to $Y(2)$:

$$Y(1) = X(1) + 2U(1) + \varepsilon_{Y(1)}, \text{ and}$$
$$Y(2) = X(1) + 0.5Y(1) + 2U(2) + \varepsilon_{Y(2)}.$$

Scenario 4 is very similar to Scenario 1, but presents an *interaction* between $U(1)$ and $X(1)$, and between $U(2)$ and $X(2)$. Before introducing Scenario 4, we first define $U^*(2)$ as a centered (to 0) version of $U(2)$. This centering is done for convenience and to ensure that the marginal total effect of $X(2)$ on $Y(2)$ equals 1 (see Appendix 6.9.2). Also, remark that the marginal total effect of $X(1)$ on $Y(1)$ equals 1. The data-generating equations for Scenario 4 that differ from Scenario 1 are:

$$Y(1) = X(1) + U(1) + X(1) \times U(1) + \varepsilon_{Y(1)}, \text{ and}$$
$$Y(2) = X(2) + 0.5Y(1) + U^*(2) + X(2) \times U^*(2) + \varepsilon_{Y(2)}.$$

6.7.2 Simulation results

For each simulation scenario, we generated 10,000 datasets of size $n = 500$. For each dataset, we estimated the causal effect of $X(t)$ on $Y(t)$, $t = 1, 2$, using the estimated parameter associated to $X(t)$, namely $\hat{\beta}_1$, in 1) an unweighted GEE regression (crude analysis), 2) an unconditional MSMRM, and 3) a conditional MSMRM. The true causal effect equals 1 for each scenario. We computed the mean and the standard deviation of $\hat{\beta}_1$ across generated datasets for each method. The outcome models fitted in the unconditional and conditional MSMRMs analyses were the same for all simulation scenarios, and were respectively:

$$E(Y(t)) = \beta_0 + \beta_1 X(t) + \beta_2 t, \text{ and}$$
$$E(Y(t)|U(t)) = \beta_0 + \beta_1 X(t) + \beta_2 t + \beta_3 U(t).$$

Note that even though the true structural equations of Scenario 4 involve interaction terms and a centered version of $U(2)$, the fitted structural model does not. The fitted model is therefore misspecified.

Each subject-visit was attributed a normalized version of the following weights:

$$W_i^{XY}(1) = \frac{P(X(1) = x_i(1))}{P(X(1) = x_i(1)|U(1) = u_i(1))}, \text{ and}$$
$$W_i^{XY}(2) = \frac{P(X(1) = x_i(1))}{P(X(1) = x_i(1)|U(1) = u_i(1))} \times \frac{P(X(2) = x_i(2))}{P(X(2) = x_i(2)|U(1) = u_i(1), Y(1) = y_i(1), U(2) = u_i(2))},$$

where the normalization was performed as in Equation (4) from Xiao *et al.* (2010). Those weights were used to fit both the conditional and unconditional MSMSRMs.

The specification of the crude GEE regression was the same as the specification of the unconditional MSMRM, but with $W_i^{XY}(1) \equiv W_i^{XY}(2) \equiv 1$.

The results of the simulation study are presented in Table 6.2. Those results confirm that the conditional MSMRM (6.4) can yield unbiased estimates of β_1 if the weights are correctly specified and if $\boldsymbol{U}(t)$ does not include descendants of $X(t)$. The results for Scenario 4 support that this conclusion holds even when there are interactions between some variables in $\boldsymbol{U}(t)$ and $X(t)$. Moreover, the conditional MSMRM is often more efficient than the unconditional MSMRM when estimating β_1. In fact, the results for Scenarios 1, 2 and 3 suggest that the more $\boldsymbol{U}(t)$ predicts $Y(t)$, the greater is the reduction in the variance of $\hat{\beta}_1$.

Table 6.2: Results from simulation Scenarios 1-4 obtained by generating 10,000 datasets of size n = 500. The mean and the standard deviation (in parenthesis) of $\hat{\beta}_1$ are provided. The true causal effect is 1.

Model	Scenario 1	Scenario 2	Scenario 3	Scenario 4
Crude	1.808 (0.103)	1.223 (0.074)	2.558 (0.161)	2.138 (0.129)
Unconditional MSMRM	1.001 (0.087)	1.000 (0.074)	1.009 (0.149)	1.006 (0.114)
Conditional MSMRM	1.000 (0.077)	1.000 (0.074)	1.005 (0.100)	1.003 (0.092)

6.7.3 HHP results

We present in Table 6.3 a comparison of the estimates obtained using the unconditional MSMRM (6.1) and the conditional MSMRM (6.4), with 95% confidence intervals. The conditional MSMRM involves the time-varying covariates $\boldsymbol{U}(t) = \{$*Employment* and *Smoking* at Visit $t\}$. The results obtained using conditional and unconditional MSMRMs are consistent, although a reduction of more than 1 mmHg in SBP is observed for the conditional effect. No clear benefit was seen with the use of a conditional MSMRM for this application, although the simulation's results we presented suggest that conditional MSMRM yields more precise estimates in some situations.

Table 6.3: Results from using unconditional and conditional MSMRMs to estimate the causal effect of physical activity on SBP and DBP (95% confidence intervals in parenthesis). Note that only the parameter associated with physical activity has a causal interpretation.

Parameter	Unconditional SBP	Conditional SBP
Physical activity	**-2.47 (-3.46, -1.48)**	**-1.33 (-2.28, -0.38)**
Age	0.70 (0.61, 0.78)	0.44 (0.36, 0.52)
Employed	NA	-12.71 (-13.65, -11.78)
Current smoker	NA	-1.14 (-2.15, -0.14)
Previous smoker	NA	1.69 (0.65, 2.74)

Parameter	Unconditional DBP	Conditional DBP
Physical activity	**0.26 (-0.22, 0.75)**	**0.16 (-0.33, 0.65)**
Age	-0.07 (-0.11, -0.03)	-0.04 (-0.09, 0.00)
Employed	NA	1.72 (1.24, 2.20)
Current smoker	NA	-1.79 (-2.34, -1.25)
Previous smoker	NA	-0.23 (-0.79, 0.33)

6.8 Discussion

Using the Honolulu Heart Program to illustrate our approach, we have devised and implemented MSMs in the graphical framework to causal inference. This graphical framework can be particularly helpful when selecting variables used to construct the IPTWs, which are central to fitting MSMs to data. Using substantive prior knowledge, our approach first consisted in drawing DAGs to represent the links between the selected clinically relevant and potential confounding variables. Structural equation models were then used to assess the correctness of the postulated DAGs. Because these DAGs did not fit the data well, structural equation models were again used to update the DAGs and improve their fit. Finally, confounding covariates were conveniently identified upon the examination of the DAGs and by invoking Pearl's back-door criterion.

Selecting variables to calculate IPTWs has previously been recognized as a challenge in the implementation of MSMs (Cole & Hernán, 2008). This was further illustrated

in Section 6.6 of our paper, where a naive approach to variable selection was shown to yield implausible results. Contrariwise, the graphical approach we have developed for the analysis of the HHP data gave results more consistent with current scientific knowledge. Therefore, we believe the additional insight brought by our approach can be very valuable.

We have also proposed a conditional version of the MSMRM to estimate the causal effects of physical activity on SBP and DBP. Although no clear advantages were seen for the HHP data, the use of conditional MSMRMs ought not be neglected in practice. Indeed, the simulations we performed resulted in unbiased conditional estimators with smaller standard errors than the unconditional ones. The extent of the reduction in standard error was seen to be especially important when the MSMRM is conditional on strong predictors of the outcome. It is important to keep in mind that this conditional version of the MSMRM was introduced in a very specific context in which the physical activity history was summarized using only the most recent level of physical activity. However, our approach could easily be generalized to other situations, for instance where physical activity history is summarized using the two most recent levels of physical activity. Following a reasoning similar to the one presented in Section 6.4.2, one would select the variables used for conditioning in the structural model utilizing the back-door criterion to ensure the conditioning does not introduce bias in the estimation of the causal parameters.

6.9 Appendix

6.9.1 Variables used in IPTWs

Here is a list of the variables used in the calculations of the IPTWs.

- To estimate the causal effects of physical activity on SBP or DBP ($\boldsymbol{L}^{XY}$), and survival or MACE ($\boldsymbol{L}^{XT}$) :
  - $\boldsymbol{L}^{XY}/\boldsymbol{L}^{XT}(1) = \{$age at Visit 1, employment at Visit 1$\}$
  - $\boldsymbol{L}^{XY}/\boldsymbol{L}^{XT}(2) = \{$age at Visit 2, BMI at Visit 1, employment at Visit 2, hypertension medication usage at Visit 1, physical activity level at Visit 1$\}$
  - $\boldsymbol{L}^{XY}/\boldsymbol{L}^{XT}(3) = \{$age at Visit 3, BMI at Visit 2, employment at Visit 3, hypertension medication usage at Visits 1 and 2, physical activity level at Visits 1 and 2$\}$
- To estimate the causal effects of SBP on time of survival and time to MACE :
  - $\boldsymbol{L}^{XT}(1) = \{$age at Visit 1, BMI at Visit 1, employment at Visit 1, physical activity level at Visit 1$\}$
  - $\boldsymbol{L}^{XT}(2) = \{$age at Visit 2, BMI at Visits 1 and 2, employment at Visit 2, hypertension medication usage at Visit 1, physical activity level at Visits 1 and 2, SBP at Visit 1$\}$
  - $\boldsymbol{L}^{XT}(3) = \{$age at Visit 3; BMI at Visits 1, 2 and 3; employment at Visit 3; hypertension medication usage at Visits 1 and 2; physical activity level at Visits 1, 2 and 3; SBP at Visits 1 and 2$\}$
- The weights used for estimating the causal effects of DBP on time of survival and time to MACE were analogous to the preceding ones, only replacing SBP by DBP.
- The variables used for computing the naive weights, $\boldsymbol{L}_N^{XY}(t)$, $t = 1, 2, 3$, in Section 6.6 for estimating the causal effects of physical activity on SBP and DBP :
  - $\boldsymbol{L}_N^{XY}(1) = \{$age at Visit 1, BMI at Visit 1, employment at Visit 1, hypertension medication usage at Visit 1 and smoking at Visit 1$\}$

- $\boldsymbol{L}_N^{XY}(2)$ = {age at visit 2, BMI at Visits 1 and 2, employment at Visits 1 and 2, hypertension medication usage at Visits 1 and 2, smoking at Visits 1 and 2, physical activity level at Visit 1, SBP at Visit 1, DBP at Visit 1}
- $\boldsymbol{L}_N^{XY}(3)$ = {age at visit 3; BMI at Visits 1, 2 and 3; employment at Visits 1, 2 and 3; hypertension medication usage at Visits 1, 2 and 3; smoking at Visits 1, 2 and 3; physical activity level at Visits 1 and 2; SBP at Visits 1 and 2; and DBP at Visist 1 and 2}

6.9.2 Calculating the marginal causal effects in Scenario 4

To calculate the causal effect of $X(1)$ on $Y(1)$ and of $X(2)$ on $Y(2)$ in simulation Scenario 4, we consider Pearl's do-calculus (Pearl, 2009). Using this do-calculus, the marginal causal effect of $X(t)$ on $Y(t)$ is denoted by $E[Y(t)|do(X(t) = 1)] - E[Y(t)|do(X(t) = 0)]$. Referring to the data-generating equations for Scenario 4, we get:

$$\begin{aligned} E[Y(1)|do(X(1) = x)] &= x + E[U(1)] + x \times E[U(1)] + E[\varepsilon_{Y(1)}] \\ &= x, \\ E[Y(2)|do(X(2) = x)] &= x + 0.5E[Y(1)] + E[U^*(2)] + x \times E[U^*(2)] + E[\varepsilon_{Y(2)}] \\ &= x + 0.5E[Y(1)]. \end{aligned}$$

It is now direct to see that $E[Y(1)|do(X(1) = 1)] - E[Y(1)|do(X(1) = 0)] = E[Y(2)|do(X(2) = 1)] - E[Y(2)|do(X(2) = 0)] = 1$ as asserted.

CHAPITRE VII

TROISIÈME ARTICLE : A CAUTIONARY NOTE CONCERNING THE USE OF STABILIZED WEIGHTS IN MARGINAL STRUCTURAL MODELS

Denis Talbot, Juli Atherton, Amanda M. Rossi, Simon L. Bacon, Geneviève Lefebvre

Abstract: Marginal structural models (MSMs) are commonly used to estimate the causal effect of a time-varying treatment in presence of time-dependent confounding. When fitting a MSM to data, the analyst must specify both the structural model for the outcome and the treatment models for the inverse-probability-of-treatment weights. The use of stabilized weights is recommended since they are generally less variable than the standard weights. In this paper, we are concerned with the use of the common stabilized weights when the structural model is specified to only consider partial treatment history, such as the current or most recent treatments. We present various examples of settings where these stabilized weights yield biased inferences while the standard weights do not. These issues are first investigated on the basis of simulated data and subsequently exemplified using data from the Honolulu Heart Program. Unlike common stabilized weights, we find that basic stabilized weights offer some protection against bias in structural models designed to estimate current or most recent treatment effects.

Keywords: Time-dependent confounding, marginal structural models, inverse-probability-weighting, repeated measures, stabilized weights

7.1 Introduction

Marginal structural models (MSMs) (Robins, 1997, 2000; Robins *et al.*, 2000; Hernán & Brumback, 2002) are nowadays a common longitudinal data analytical approach for estimating the effects of time-varying treatments in presence of time-dependent confounding (Yang & Joffe, 2012; Fairall *et al.*, 2008; Patel *et al.*, 2008; Sampson *et al.*, 2006; Schildcrout *et al.*, 2011; VanderWeele *et al.*, 2011, 2012). When fitting a MSM to data, an analyst faces two important decisions: 1) the specification of the structural model for the outcome, done in accordance with the causal contrast of interest; 2) the specification of the treatment models which are used to calculate the inverse-probability-of-treatment received at each time point, i.e. the weights (Yang & Joffe, 2012). For the structural model, a single measure is commonly used to summarize treatment history, such as the treatment received at the last time point, a cumulative measure of the treatment or an indicator of "ever started treatment" (Yang & Joffe, 2012; Platt *et al.*, 2013). The covariates included in the treatment models are typically the baseline covariates and the histories of time-varying covariates and prior treatments. Platt *et al.* (2013) outline strategies for marginal structural model specifications and introduce a quasi-likelihood information criterion to help with the selection of the structural model on the basis of data.

Stabilized weights are recommended to be used in MSMs in place of the standard weights since they are generally less variable than the latter (Robins *et al.*, 2000). The stabilized weights are similar to the standard weights but are commonly defined so that the numerator is the marginal probability of observed treatment history predicted using prior treatments only while a numerator equal to 1 is instead used for the standard weights (Robins *et al.*, 2000; Hernán & Brumback, 2002; Yang & Joffe, 2012). The denominator is the same for both types of weights. In MSMs, it has been shown that, when saturated structural models are specified, the treatment effect estimates that result from the use of stabilized or standard weights are the same (Hernán & Brumback, 2002). In correctly specified unsaturated structural models however, the

estimates differ but this difference is only due to sampling variability (Hernán & Brumback, 2002).

This note is concerned with the impact of using the common stabilized weights under different and frequently used specifications of the structural model in ordinary MSMs. As such, we focus on the estimation of the causal effect of a static treatment regime, that is, the estimation of the causal effect that a pre-specified treatment regime would have. In contrast, inferences about a dynamic treatment would consist in estimating the causal effect of a treatment regime where the treatment a subject receives at a given time point is decided according to a pre-specified rule, which might involve time-varying covariates and prior treatments. It has already been recommended not to use stabilized weights for estimating the causal effect of dynamic treatments (Robins & Hernán, 2009). In the sequel, we present various settings where the common stabilized weights lead to biased structural model parameter estimates while the standard weights do not. This curious (and perhaps unexpected) phenomenon is observed when the structural model targets the effect of the current treatment or the most recent treatments. This result concerns both classical MSMs and MSMs with repeated measures, although MSMs with repeated measures are arguably more susceptible to this type of structural model specification.

The paper is organized as follows. In Section 7.2, we introduce the notation and review the MSMs. Section 7.3 focuses on a very simple example that captures the problem presented in this work. In Section 7.4, we present the description of a simulation study devised to illustrate the potential problems of using the common stabilized weights in MSMs. The results of the simulation study are presented in Section 7.5. In Section 7.6, we investigate these issues using data from the Honolulu Heart Program. In particular, we find that the estimated effect of the current level of physical activity on blood pressure differs depending on whether standard or stabilized weights are used. We conclude with a short discussion in Section 7.7.

7.2 Notation and MSM implementations

In the following, we distinguish between two types of implementations of MSMs: classical and repeated measures.

7.2.1 Classical marginal structural model

Based on Robins *et al.* (2000), we briefly review the classical MSM. In the sequel, we use capital letters to represent random variables and lower-case letters to represent possible realizations (values) of random variables.

Consider a follow-up study consisting of n sampled subjects from a population, along with covariates measured at $K+1$ time points (visits). Let $A_{k,i}$ be subject i's ($i = 1, \ldots, n$) treatment level at the kth visit from the start of the follow-up ($k = 0, \ldots, K$) and let Y_i be his outcome measured at end of follow-up, i.e. $Y_i = Y_{K+1,i}$. For the sake of simplicity, we consider continuous outcome and binary treatment variables (with $A_{k,i} = 1$ if subject i receives treatment at time k and $A_{k,i} = 0$ otherwise). For subject i, $L_{k,i}$ consists of the outcome at time k, $Y_{k,i}$, and the vector of all other measured risk factors for Y_i at time k, $V_{k,i}$, i.e. $L_{k,i} = (V_{k,i}, Y_{k,i})$. We suppose that $L_{k,i}$ temporally precedes $A_{k,i}$ for all i and k. Let $\bar{A}_{k,i} = (A_{0,i}, A_{1,i}, ..., A_{k,i})$ be subject i's treatment history through time k and let $\bar{A}_i = \bar{A}_{K,i}$. We define $\bar{L}_{k,i}$ and $\bar{L}_i$ similarly. Finally, $Y_{\bar{a}k,i}$ is subject i's counterfactual outcome at visit k, that is the outcome that would have been observed if, possibly contrary to the fact, subject i had received treatment regime $\bar{a}$ instead of his own treatment regime $\bar{a}_i$. Note that $Y_{\bar{a}k,i} = Y_{k,i}\ \forall\ k$ if $\bar{a} = \bar{a}_i$. As in Hernán & Brumback (2002), we assume that every subject's data are independently drawn from a common distribution; therefore we drop subscript i unless it is required for clarity.

The classical marginal structural model is defined as a model for the population's mean of the counterfactual outcome at visit $K+1$ under treatment history $\bar{a}$:

$$E[Y_{\bar{a}}] = g(\bar{a};\gamma), \tag{7.1}$$

where g is a user defined function. Possible g functions are $g(\bar{a};\gamma) = \gamma_0 + \gamma_1 a_K + \ldots + \gamma_{K+1} a_0$, $g(\bar{a};\gamma) = \gamma_0 + \gamma_1 a_K$, $g(\bar{a};\gamma) = \gamma_0 + \gamma_1 \text{cum}(\bar{a})$ where $\text{cum}(\bar{a}) = \sum_{k=0}^{K} a_k$, or $g(\bar{a};\gamma) = \gamma_0 + \gamma_1 I_{\{\text{cum}(\bar{a})\geq 1\}}$. The parameters γ of model (7.1) encode the causal effect of the treatment history on the last outcome. For example, when selecting $g(\bar{a};\gamma) = \gamma_0 + \gamma_1 \text{cum}(\bar{a})$, it is hypothesized that the effect of treatment history on the mean outcome increases linearly as a function of the cumulative treatment. Thus for two treatment regimes $\bar{a}$ and $\bar{a}'$ being compared, $\gamma_1(\text{cum}(\bar{a}) - \text{cum}(\bar{a}'))$ can be interpreted as the mean difference in outcome Y, i.e. $E[Y_{\bar{a}} - Y_{\bar{a}'}]$. In particular, if $\bar{a} = \{1,1,\ldots,1\}$ and $\bar{a}' = \{0,0,\ldots,0\}$ - corresponding to the always and never treated regimes, respectively - then the expected difference in outcome is $\gamma_1(K+1)$. Similarly, if $g(\bar{a};\gamma) = \gamma_0 + \gamma_1 a_K$ is selected, then it is hypothesized that the effect of treatment history on the mean outcome only depends on the last treatment. In this case, γ_1 corresponds to the expected difference in outcome when $\bar{a} = \{\cdot,\ldots,\cdot,1\}$ and $\bar{a}' = \{\cdot,\ldots,\cdot,0\}$, where symbol $\cdot$ is used to represent either of the two possible levels for treatment. The issues we are concerned with in this paper stem from using structural model specifications such as this one.

The parameters γ of structural model (7.1) can be consistently estimated using a weighted linear regression model for $E[Y|\bar{A}]$, where each subject is weighted by the inverse probability of his observed treatment history conditional on covariates and prior treatments. Specifically, the standard weight for subject i is

$$w_i = \left\{ \prod_{k=0}^{K} \frac{1}{P(A_k = a_{k,i} | \bar{A}_{k-1} = \bar{a}_{k-1,i}, \bar{L}_k = \bar{l}_{k,i})} \right\}, \qquad i = 1,\ldots,n, \tag{7.2}$$

where $\bar{A}_{k-1}$ is ignored in the conditioning when $k = 0$. The standard weights w are often highly variable; therefore it is usually advised to instead use stabilized weights sw, where

$$sw_i = \left\{ \prod_{k=0}^{K} \frac{P(A_k = a_{k,i} | \bar{A}_{k-1} = \bar{a}_{k-1,i})}{P(A_k = a_{k,i} | \bar{A}_{k-1} = \bar{a}_{k-1,i}, \bar{L}_k = \bar{l}_{k,i})} \right\}. \tag{7.3}$$

In both (7.2) and (7.3) the $\bar{L}$ covariates are selected to ensure that the sequential (conditional) randomized assumption holds (Robins, 2000), that is

$$Y_{\bar{a}} \perp\!\!\!\perp A_k | \bar{A}_{k-1}, \bar{L}_k \quad \forall\, \bar{a} \text{ and } k, \tag{7.4}$$

where $\perp\!\!\!\perp$ symbolizes statistical independence. Perhaps underrealized is that conditioning on $\bar{A}_{k-1}$ in (7.4) implies that, in addition to $\bar{L}_k$, the previous treatment variables should also be regarded as potential confounding variables. This last remark is crucial for understanding the possible introduction of bias when using stabilized weights sw in MSMs.

7.2.2 Marginal structural model with repeated measures

Instead of modelling the mean counterfactual outcome at the end of follow-up, a MSM with repeated measures (Hernán & Brumback, 2002) aims to model the mean counterfactual outcome at each time $k+1$ ($k = 0, \ldots, K$) as a function of treatment history up to time k, that is

$$E\left[Y_{\bar{a}(k+1)}\right] = g(\bar{a}_k; \gamma). \tag{7.5}$$

Popular choices of g function for this type of MSM implementation are $g(\bar{a}_k; \gamma) = \gamma_0 + \gamma_1 a_k + \gamma_2 k$, $g(\bar{a}_k; \gamma) = \gamma_0 + \gamma_1 a_k + \gamma_2 a_{k-1} + \gamma_3 k$, $g(\bar{a}_k; \gamma) = \gamma_0 + \gamma_1 \text{cum}(\bar{a}_k) + \gamma_2 k$, where $\text{cum}(\bar{a}_k) = \sum_{t=0}^{k} a_t$ or $g(\bar{a}_k; \gamma) = \gamma_0 + \gamma_1 I_{\{\text{cum}(\bar{a}_k) \geq 1\}} + \gamma_2 k$. Model (7.5) is then

fitted using a weighted linear generalized estimating equation (GEE) regression for $E[Y_{k+1}|\bar{A}_k]$, where person-visit $(i, k+1)$ is weighted by its standard or stabilized weight

$$w_{k,i} = \left\{ \prod_{t=0}^{k} \frac{1}{P(A_t = a_{t,i}|\bar{A}_{t-1} = \bar{a}_{t-1,i}, \bar{L}_t = \bar{l}_{t,i})} \right\} \quad \text{or}$$

$$sw_{k,i} = \left\{ \prod_{t=0}^{k} \frac{P(A_t = a_{t,i}|\bar{A}_{t-1} = \bar{a}_{t-1,i})}{P(A_t = a_{t,i}|\bar{A}_{t-1} = \bar{a}_{t-1,i}, \bar{L}_t = \bar{l}_{t,i})} \right\}, \tag{7.6}$$

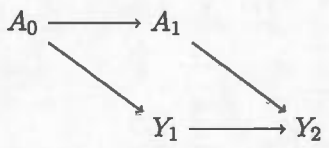
respectively. The choice of covariates $\bar{L}$ to include in these weights must also be dictated by the sequential randomized assumption (Hernán & Brumback, 2002; Brumback *et al.*, 2004):

$$Y_{\bar{a}(k+1)} \perp\!\!\!\perp A_t|\bar{A}_{t-1}, \bar{L}_t \quad \forall\, \bar{a} \quad \text{and} \quad k \geq t. \tag{7.7}$$

7.3 A striking example

The issues raised in this paper are best first illustrated with the simple directed acyclic graph depicted in Figure 7.1 (DAG1). In DAG1, Y_1 depends on A_0, A_1 depends on A_0, and Y_2 depends on both A_1 and Y_1. Here $L_0 = \varnothing$ and $L_1 = \{Y_1\}$: no L covariates other than the outcome at time 1 are considered since they are irrelevant to illustrate our problem. Covariates denoted by V are however later incorporated in our simulation scenarios presented in Section 7.4.1.

Figure 7.1: Directed acyclic graph 1 (DAG1)



Consider the implementation of a classical MSM based on data compatible with DAG1. While a first logical step would be the specification of the structural model, we momentarily delay this step and examine the definition of weights w and sw with

regard to the sequential randomization assumption (7.4). Because of the presence of the open back-door path $A_1 \leftarrow A_0 \rightarrow Y_1 \rightarrow Y_2$ (★) from A_1 to Y_2 in DAG1, it follows that $Y_{\bar{a}2} \not\perp\!\!\!\perp A_1$ and therefore the (unconditional) randomization assumption (7.4) does not hold (Pearl, 2009). This path can be closed by A_0, which leads to $Y_{\bar{a}2} \perp\!\!\!\perp A_1|A_0$. The sequential randomization assumption is achieved conditional on treatment history since for all $\bar{a}$ and $k = 0, 1$, $Y_{\bar{a}} \equiv Y_{\bar{a}2} \perp\!\!\!\perp A_k|\bar{A}_{k-1}$ (we already have $Y_{\bar{a}2} \perp\!\!\!\perp A_0$). In principle, a MSM can thus be validly implemented with the following standard and stabilized weight definitions for subject i:

$$w_i = \frac{1}{P(A_0 = a_{0,i})} \times \frac{1}{P(A_1 = a_{1,i}|A_0 = a_{0,i})}, \tag{7.8}$$

and

$$sw_i = \frac{P(A_0 = a_{0,i})}{P(A_0 = a_{0,i})} \times \frac{P(A_1 = a_{1,i}|A_0 = a_{0,i})}{P(A_1 = a_{1,i}|A_0 = a_{0,i})} = 1. \tag{7.9}$$

Note that the second denominators in (7.8) and (7.9) could have been set to $P(A_1 = a_{1,i}|A_0 = a_{0,i}, Y_1 = y_{1,i})$ to follow the generic notation (7.2) and (7.3) for the specification of the weights. However DAG1 implies that $P(A_1 = a_{1,i}|A_0 = a_{0,i}, Y_1 = y_{1,i}) = P(A_1 = a_{1,i}|A_0 = a_{0,i})$, and thus it suffices to condition on A_0 only.

The simplification of the stabilized weight sw_i to the value 1 in (7.9) indicates that, in the setting represented by DAG1, *the implementation of a classical MSM with weights sw is equivalent to the implementation of an unweighted (crude) MSM.* This leads to biased or unbiased parameter estimators depending on the form of the structural model selected.

Suppose the structural model $E[Y_{\bar{a}}] = \gamma_0 + \gamma_1 a_1 + \gamma_2 a_0$ is chosen, where parameters γ_1 and γ_2 encode the causal effect of A_1 on $Y_2 \equiv Y$ and of A_0 on $Y_2 \equiv Y$, respectively. Using stabilized weights sw with this structural model yields an unbiased estimator for both γ_1 and γ_2. The parameter γ_1 is of particular interest in this case since, recall, $Y_{\bar{a}2} \not\perp\!\!\!\perp A_1$ due to the open back-door path (★). Although the confounding introduced by this

back-door path is not handled by the weights (because $sw_i = 1 \; \forall \; i$), it is nonetheless accounted for by the inclusion of the treatment covariate A_0 in the regression model $E[Y|\bar{A}] = \beta_0 + \beta_1 a_1 + \beta_2 a_0$. This implies that the associational parameter β_1 coincides with the structural parameter γ_1, that is $\beta_1 = \gamma_1$, as desired.

Suppose we now consider the structural model $E[Y_{\bar{a}}] = \gamma_0 + \gamma_1 a_1$ and its associated regression model $E[Y|\bar{A}] = \beta_0 + \beta_1 a_1$. Although this reduced structural model is misspecified since A_0 has an effect on $Y_{\bar{a}2}$, it is much relevant to be able to obtain unbiased estimation for the effect this structural model is capable of identifying, namely, the effect of the most recent exposure effect (A_1) on $Y_{\bar{a}2}$. If the stabilized weights sw are used, then β_1 and γ_1 do not coincide anymore as the confounding is neither accounted for in the weights nor the regression model. With this structural model, unbiased γ_1 estimation can however be obtained by using the standard weights w since these weights do account for the confounding caused by A_0.

This example is simple and admittedly a bit artificial since a traditional regression-based approach could have correctly identified the causal effect targeted by the structural model $E[Y_{\bar{a}}] = \gamma_0 + \gamma_1 a_1$ (Yang & Joffe, 2012). However, it unravels a potential problem with the use of stabilized weights sw along with structural models that only include partial treatment history (e.g., current treatment or current treatment with lag 1 treatment). Indeed, a consequence of such a stabilization of the weights may be that the unconfounding achieved by the denominator is cancelled out (at least partially) by the numerator. This phenomenon is empirically demonstrated in Section 7.5. Also seen in Section 7.5 is that similar problems occur when using stabilized weights sw in MSMs with repeated measures.

7.4 Description of the simulation study

In this section, we present the four simulation scenarios investigated as well as the definitions of the standard and stabilized weights used in the classical implementation of the MSMs (the weights for the repeated mesures implementation are defined in

a similar manner). We conclude the section with a description of the analyses we performed.

7.4.1 Simulation scenarios

Scenario 1. Our first simulation scenario is compatible with DAG1 (recall Figure 7.1). The causal relationships between the variables are as follows:

$$\begin{aligned} P(A_0 = 1) &= 0.5 \\ Y_1 &= A_0 + \varepsilon_{Y_1} \\ P(A_1 = 1) &= expit(A_0) \\ Y_2 &= A_1 + Y_1 + \varepsilon_{Y_2}, \end{aligned}$$

where $expit(z) = e^z/(e^z+1)$ and ε_{Y_1} and ε_{Y_2} are independent $N(0,1)$ random variables. The standard and stabilized weights used in the classical MSM implementation are defined in (7.8) and (7.9).

Scenario 2. The second simulation scenario is only slightly more complex than the first scenario (see Figure 7.2):

$$\begin{aligned} P(A_0 = 1) &= 0.5 \\ Y_1 &= A_0 + \varepsilon_{Y_1} \\ P(A_1 = 1) &= expit(0.5A_0 + 0.5Y_1) \\ Y_2 &= A_1 + Y_1 + \varepsilon_{Y_2}, \end{aligned}$$

where ε_{Y_1} and ε_{Y_2} are independent $N(0,1)$ random variables. In this scenario, the presence of the causal link between Y_1 and A_1 makes the adjustment for Y_1 in the denominator of the weights necessary to achieve (7.4); the standard and stabilized

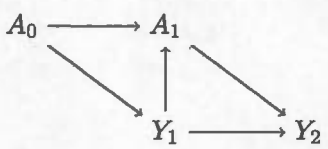
weights are thus defined as

$$w_i = \frac{1}{P(A_0 = a_{0,i})} \times \frac{1}{P(A_1 = a_{1,i} | A_0 = a_{0,i}, Y_1 = y_{1,i})},$$

and

$$sw_i = \frac{P(A_0 = a_{0,i})}{P(A_0 = a_{0,i})} \times \frac{P(A_1 = a_{1,i} | A_0 = a_{0,i})}{P(A_1 = a_{1,i} | A_0 = a_{0,i}, Y_1 = y_{1,i})}.$$

Figure 7.2: DAG for Scenario 2



Scenario 3. The third scenario is a typical MSM representation and includes a time-dependent confounder V that is affected by previous treatment (see Figure 7.3):

$$\begin{aligned}
V_0 &= \varepsilon_{V_0} \\
P(A_0 = 1) &= expit(0.5V_0) \\
Y_1 &= A_0 + V_0 + \varepsilon_{Y_1} \\
V_1 &= 0.5A_0 + \varepsilon_{V_1} \\
P(A_1 = 1) &= expit(0.5A_0 + 0.5Y_1 + 0.5V_1) \\
Y_2 &= A_1 + 0.5Y_1 + V_1 + \varepsilon_{Y_2},
\end{aligned}$$

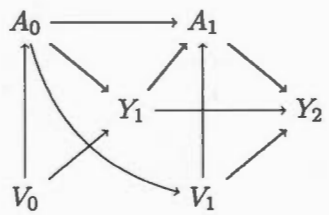
where $\varepsilon_{V_0}, \varepsilon_{Y_1}, \varepsilon_{V_1}$, and ε_{Y_2} are independent $N(0,1)$ random variables. For this scenario, we adopt the naïve strategy of including all possible covariates for the specification of the weights, that is

$$w_i = \frac{1}{P(A_0 = a_{0,i} | V_0 = v_{0,i})} \times \frac{1}{P(A_1 = a_{1,i} | A_0 = a_{0,i}, Y_1 = y_{1,i}, V_1 = v_{1,i}, V_0 = v_{0,i})},$$

and

$$sw_i = \frac{P(A_0 = a_{0,i})}{P(A_0 = a_{0,i}|V_0 = v_{0,i})} \times \frac{P(A_1 = a_{1,i}|A_0 = a_{0,i})}{P(A_1 = a_{1,i}|A_0 = a_{0,i}, Y_1 = y_{1,i}, V_1 = v_{1,i}, V_0 = v_{0,i})}.$$

Figure 7.3: DAG for Scenario 3



Scenario 4. The fourth scenario is similar to the previous scenario but generates data for an additional follow-up visit (see Figure 7.4):

$$\begin{aligned}
V_0 &= \varepsilon_{V_0} \\
P(A_0 = 1) &= expit(0.5V_0) \\
Y_1 &= A_0 + V_0 + \varepsilon_{Y_1} \\
V_1 &= 0.25A_0 + \varepsilon_{V_1} \\
P(A_1 = 1) &= expit(0.5A_0 + 0.5Y_1 + 0.5V_1) \\
Y_2 &= A_1 + 0.25A_0 + 0.5Y_1 + V_1 + \varepsilon_{Y_2} \\
V_2 &= 0.25A_1 + \varepsilon_{V_2} \\
P(A_2 = 1) &= expit(0.5A_1 + 0.3A_0 + 0.5Y_2 + 0.5V_2) \\
Y_3 &= A_2 + 0.25A_1 + 0.5Y_2 + 0.5Y_1 + V_2 + \varepsilon_{Y_3},
\end{aligned}$$

where $\varepsilon_{V_0}, \varepsilon_{Y_1}, \varepsilon_{V_1}, \varepsilon_{Y_2}, \varepsilon_{V_2}$ and ε_{Y_3} are independent $N(0,1)$ random variables. For this scenario, we also include all possible covariates for the specification of the weights,

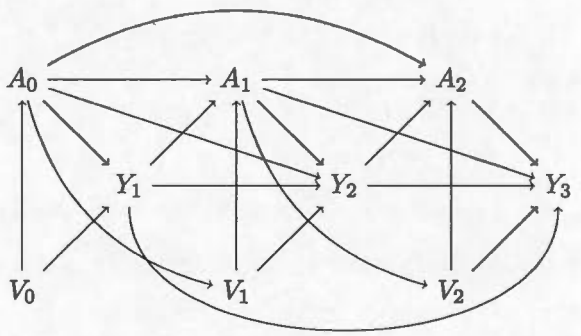
that is

$$w_i = \frac{1}{P(A_0 = a_{0,i}|V_0 = v_{0,i})} \times \frac{1}{P(A_1 = a_{1,i}|A_0 = a_{0,i}, Y_1 = y_{1,i}, V_1 = v_{1,i}, V_0 = v_{0,i})} \times \frac{1}{P(A_2 = a_{2,i}|A_1 = a_{1,i}, A_0 = a_{0,i}, Y_2 = y_{2,i}, Y_1 = y_{1,i}, V_2 = v_{2,i}, V_1 = v_{1,i}, V_0 = v_{0,i})},$$

and

$$sw_i = \frac{P(A_0 = a_{0,i})}{P(A_0 = a_{0,i}|V_0 = v_{0,i})} \times \frac{P(A_1 = a_{1,i}|A_0 = a_{0,i})}{P(A_1 = a_{1,i}|A_0 = a_{0,i}, Y_1 = y_{1,i}, V_1 = v_{1,i}, V_0 = v_{0,i})} \times \frac{P(A_2 = a_{2,i}|A_1 = a_{1,i}, A_0 = a_{0,i})}{P(A_2 = a_{2,i}|A_1 = a_{1,i}, A_0 = a_{0,i}, Y_2 = y_{2,i}, Y_1 = y_{1,i}, V_2 = v_{2,i}, V_1 = v_{1,i}, V_0 = v_{0,i})}.$$

Figure 7.4: DAG for Scenario 4



7.4.2 Description of analyses

We generated 10 000 datasets of size $n = 1000$ for each of the four scenarios described in Section 7.4.1. A series of MSM analyses was performed on each dataset. The set of structural models we considered include a variety of models that have been seen in recent classical and repeated measures MSM implementations (Fairall *et al.*, 2008; Patel *et al.*, 2008; Sampson *et al.*, 2006; Schildcrout *et al.*, 2011; VanderWeele *et al.*, 2011, 2012). For the classical version of the MSMs (*cMSM*), we considered the following three structural models:

- *Full*: $E[Y_{\bar{a}}] = \gamma_0 + \gamma_1 a_K + \gamma_2 a_{K-1} + \ldots + \gamma_{K+1} a_0$;
- *Current*: $E[Y_{\bar{a}}] = \gamma_0 + \gamma_1 a_K$;
- *Cumulative*: $E[Y_{\bar{a}}] = \gamma_0 + \gamma_1 cum(\bar{a})$.

We also considered three structural models for the repeated measures implementation of the MSMs (*rmMSM*):

- *Current*: $E[Y_{\bar{a}(k+1)}] = \gamma_0 + \gamma_1 a_k + \gamma_2 k$;
- *Current+Lag1*: $E[Y_{\bar{a}(k+1)}] = \gamma_0 + \gamma_1 a_k + \gamma_2 a_{k-1} + \gamma_3 k$;
- *Cumulative*: $E[Y_{\bar{a}(k+1)}] = \gamma_0 + \gamma_1 cum(\bar{a}_k) + \gamma_2 k$.

For Scenarios 1-3, the *Full*, *Cumulative* (*cMSM* and *rmMSM*) and *Current+Lag1* structural models are correctly specified. For Scenario 4, only the *Full* and *Cumulative* (*cMSM* and *rmMSM*) structural models are correctly specified. For every scenario and structural model (both *cMSM* and *rmMSM* implementations), the data generating equations presented in Section 7.4.1 imply that $\gamma_1 = 1$. Recall however that γ_1 has different interpretations across structural models (see Section 7.2.1).

We obtained the unweighted results (which is equivalent to setting weights equal to 1) as well as the results using the standard and stabilized weights w and sw for each scenario, implementation and structural model. Specifically, for every combination of implementation/structural model/weight, we estimated the mean and standard deviation of $\hat{\gamma}_1$ based on the 10 000 datasets generated from each scenario. As recommended, we used an independence working correlation structure for the estimation of the GEEs (Tchetgen Tchetgen *et al.*, 2012a,b). The analyses were performed using the function `geeglm` from the R (R Core Team, 2014) package `geepack` (Højsgaard *et al.*, 2006; Yan & Fine, 2004; Yan, 2002).

To comply with `geeglm`'s requirements, for every scenario we fitted the *Current+Lag1* structural model by deleting all the data pertaining to the first visit since the Lag1 treatment (i.e., a_{k-1}) is structurally missing when $k = 0$ (VanderWeele *et al.*, 2011, 2012). As a by-product of this deletion, the *rmMSM* implementation with the *Current+Lag1* structural model ends up being equivalent to the *cMSM* implementation with the *Full* structural model in the simpler scenarios (Scenarios 1-3).

7.5 Simulation results

The results of the simulation study are presented in Table 7.1.

Table 7.1: Results for Scenarios 1-4 by structural model and MSM implementation. The mean and the standard deviation (in parenthesis) of the estimates of γ_1 for each weight definition are provided (calculated from 10 000 datasets of size 1000).

	Classical MSM (*cMSM*)			Repeated measures MSM (*rmMSM*)		
Weight (by scenario):	*Full* ($\gamma_1 = 1$)	*Current* ($\gamma_1 = 1$)	*Cumulative* ($\gamma_1 = 1$)	*Current* ($\gamma_1 = 1$)	*Current+Lag1* ($\gamma_1 = 1$)	*Cumulative* ($\gamma_1 = 1$)
$S_1 : 1$	0.999 (0.094)	1.243 (0.096)	1.000 (0.057)	1.118 (0.061)	0.999 (0.094)	1.000 (0.053)
$S_1 : w$	0.999 (0.095)	0.999 (0.095)	1.000 (0.058)	0.999 (0.066)	0.999 (0.095)	1.000 (0.054)
$S_1 : sw$	0.999 (0.094)	1.243 (0.096)	1.000 (0.057)	1.118 (0.061)	0.999 (0.094)	1.000 (0.053)
$S_2 : 1$	1.474 (0.092)	1.681 (0.094)	1.179 (0.057)	1.332 (0.061)	1.474 (0.092)	1.126 (0.053)
$S_2 : w$	1.001 (0.073)	1.000 (0.074)	1.000 (0.054)	1.000 (0.053)	1.001 (0.073)	1.000 (0.051)
$S_2 : sw$	1.001 (0.071)	1.232 (0.078)	1.000 (0.054)	1.113 (0.056)	1.001 (0.071)	1.000 (0.051)
$S_3 : 1$	1.861 (0.103)	2.168 (0.103)	1.413 (0.061)	1.810 (0.074)	1.861 (0.103)	1.430 (0.056)
$S_3 : w$	1.003 (0.098)	1.002 (0.101)	1.001 (0.072)	1.002 (0.071)	1.003 (0.098)	1.001 (0.062)
$S_3 : sw$	1.003 (0.095)	1.314 (0.102)	1.001 (0.071)	1.153 (0.064)	1.003 (0.095)	1.001 (0.058)
$S_4 : 1$	2.112 (0.130)	2.900 (0.136)	1.527 (0.054)	2.133 (0.075)	2.061 (0.082)	1.486 (0.048)
$S_4 : w$	1.019 (0.185)	1.011 (0.195)	1.006 (0.110)	1.006 (0.116)	1.011 (0.138)	1.004 (0.085)
$S_4 : sw$	1.013 (0.175)	1.612 (0.193)	1.004 (0.091)	1.282 (0.079)	1.084 (0.097)	1.002 (0.065)

LEGEND. "1": unweighted; w: standard weights; sw: stabilized weights.

We first discuss the results for the classical MSM implementation. As expected, the use of either weights w or sw with the full structural model (*cMSM Full*) yields unbiased estimates for the true current effect of the treatment on the outcome ($\gamma_1 = 1$) in every scenario. Note that the slight bias of about 1% seen under the more complex Scenario 4 disappears when samples of size 5000 are considered (results not shown). The results for the cumulative structural model (*cMSM Cumulative*) are also unbiased under both types of weights. In Scenarios 1-4, when only the current treatment covariate is included in the structural model (*cMSM Current*), the standard weights w yield unbiased γ_1 estimates whereas the stabilized weights sw do not.

Now examining the results for the repeated measures MSM implementation, we observe that, as with the classical MSM implementation, the cumulative structural model (*rmMSM Cumulative*) yields unbiased γ_1 estimates under both weights w and sw. Moreover, the repeated measures MSM with only the current treatment covariate

in the model (*rmMSM* Current) similarly yields biased estimates of γ_1 when using stabilized weights sw. The repeated measures structural model with current and previous treatments (*rmMSM Current + Lag1*) produces unbiased results for weights w and sw in Scenarios 1-3 but biased results for weights sw in Scenario 4. Unlike results for *cMSM Full*, this bias does not vanish as sample size is increased (the bias remains at 8% when $n = 5000$). This last set of results does not come as a surprise given that Scenario 4 involves three post-baseline visits ($K + 1 = 3$) whereas only two visits ($K + 1 = 2$) are considered in Scenarios 1-3. More precisely, recall that the *Current + Lag1* structural model is not misspecified in Scenarios 1-3, as opposed to Scenario 4.

The biased results for weights sw under implementation/structural model *cMSM Current*, *rmMSM Current* and *rmMSM Current + Lag1* can be explained using arguments similar to those in Section 7.3. First, conditioning on the past treatment(s) in the numerators of the stabilized weights sw neutralizes some deconfounding acting through the denominators of the weights, and, second, the remaining confounding is not handled by the structural model.

It is also worthwhile to mention that, while our analyses focus on parameter γ_1 for simplicity, other parameters of the structural models considered are prone to be estimated with bias when using stabilized weights sw. For instance, in Scenario 4, $\hat{\gamma}_2$ is also biased in the implementation/structural model *rmMSM Current + Lag1* when using weights sw. Indeed, for this scenario, the mean and standard deviation (in parenthesis) of the 10 000 estimates of $\gamma_2 = 1$ under the three different weighting strategies are 1) unweighted: 1.380 (0.078); 2) standard weights w: 1.007 (0.155); 3) stabilized weights sw: 1.118 (0.107). The same reasoning as the one put forward for γ_1 explains the bias found when using weights sw to estimate γ_2.

To conclude, we observed, from our simulations, that when the structural models were correctly specified, unbiased estimators were obtained when using either stabilized weights sw or standard weights w. In this case, and as expected, a reduction in variance was also seen for the structural parameter estimators resulting from the use

of weights sw, as opposed to weights w. However, when the structural models were misspecified, only standard weights w led to unbiased estimation of the structural parameters. Given that selecting an appropriate structural model is a challenging issue, robustness of the weights to misspecification of this model is believed to be desirable. We feel this is particularly relevant for repeated measures implementations of MSMs, for which simplified structural model specifications could also be preferred to better take advantage of available data (e.g., see VanderWeele *et al.* (2011)). For instance, in our results, remark there is a decrease in variability for the current treatment effect estimator ($\hat{\gamma}_1$) in the *rmMSM Current* implementation/structural model as opposed to the same estimator in the *rmMSM Current + Lag1* implementation/structural model (as a result, in all scenarios, from the use of many more data points for the estimation of this effect in the former structural model).

In the next section, we investigate if other types of stabilized weights would consistently provide unbiased parameter estimates under differentially specified structural models.

7.5.1 Additional analyses

Although weights sw follow the typical definition for stabilized weights found in the MSM literature, other stabilization strategies could be employed. For a classical MSM for instance, basic stabilized weights which avoid conditioning on the past treatments in the numerators are

$$swb_i = \left\{ \prod_{k=0}^{K} \frac{P(A_k = a_{k,i})}{P(A_k = a_{k,i} | \bar{A}_{k-1} = \bar{a}_{k-1,i}, \bar{L}_k = \bar{l}_{k,i})} \right\}. \tag{7.10}$$

For both the classical and repeated measures implementations, we therefore also fitted the MSMs with weights swb to verify the impact of such a stabilization strategy on the distribution of $\hat{\gamma}_1$ (see Table 7.2). From these results, we observe that all estimates are unbiased and that notable variance reduction can be obtained by using the basic

stabilized weights *swb* as opposed to the standard weights w (see the results for the repeated measures MSM implementation in particular).

Table 7.2: Results from Scenarios 1-4 by structural model and MSM implementation using basic stabilized weights *swb*. The mean and the standard deviation (in parenthesis) of the estimates of γ_1 are provided (calculated from 10 000 datasets of size 1000).

	Classical MSM (*cMSM*)			Repeated measures MSM (*rmMSM*)		
Scenario	*Full* ($\gamma_1 = 1$)	*Current* ($\gamma_1 = 1$)	*Cumulative* ($\gamma_1 = 1$)	*Current* ($\gamma_1 = 1$)	*Current+Lag1* ($\gamma_1 = 1$)	*Cumulative* ($\gamma_1 = 1$)
S_1	0.999 (0.094)	0.999 (0.094)	1.000 (0.058)	1.000 (0.056)	0.999 (0.094)	1.000 (0.053)
S_2	1.001 (0.073)	1.000 (0.074)	1.000 (0.054)	1.000 (0.048)	1.001 (0.073)	1.000 (0.051)
S_3	1.003 (0.098)	1.002 (0.101)	1.001 (0.071)	1.001 (0.060)	1.003 (0.098)	1.001 (0.057)
S_4	1.012 (0.179)	1.008 (0.189)	1.003 (0.092)	1.002 (0.071)	1.006 (0.103)	1.001 (0.062)

7.6 The Honolulu Heart Program results

In this section, data from the Honolulu Heart Program (HHP) are used to illustrate how the choice of weights can influence the exposure effect estimates in non-simulated MSM analyses.

The HHP is a study of Japanese-American men living in Oahu, Hawaii, which examined 8006 participants. Participants were born between 1900 and 1919 (aged 45-68 years old at study entry) and were recruited from the selective service registry. They were evaluated at multiple time points beginning in 1965 and followed until 1994 for deaths and morbid events. Information regarding physical activity participation was collected by questionnaire at Exam 1 (1965-68), Exam 2 (1968-1971) and Exam 4 (1991-1993). Blood pressure (BP) was measured manually (in mmHg) by a trained professional during each exam. More details about HHP can be found elsewhere (Kagan *et al.*, 1974).

Repeated measures MSMs were used to estimate the causal effect of physical activity on systolic blood pressure (SBP) and diastolic blood pressure (DBP). Since physical activity was not measured at Exam 3, and since there was a long delay between Exam 2

and Exam 4, we chose to only use data from the first two exams. Our belief is that the effect of current and prior physical activity history on current BP is primarily a function of current physical activity. Our structural model for each type of BP thus has the following form:

$$E[Y_{\bar{a}k}] = \gamma_0 + \gamma_1 a_k + \gamma_2 k, \tag{7.11}$$

where unlike equation (5), which has a delayed treatment effect, the treatment effect in (7.11) is immediate. In our structural models, $Y_{\bar{a}k}$ is the counterfactual outcome (either SBP or DBP) at Exam k ($k = 1, 2$) and a_k is the physical activity level (active or inactive) reported at Exam k.

For both MSM analyses, the covariates used to calculate the visit specific weights at the first time point (Exam 1) were: age (in years) at Exam 1 and employment at Exam 1 (employed or unemployed). For the second time point (Exam 2), the weights were calculated using: employment at Exam 1, physical activity level at Exam 1, hypertension medication usage at Exam 1 (yes or no), BMI at Exam 1 (in kg/m^2), age at Exam 2 and employment at Exam 2. Note that hypertension medication usage at Exam 1 and BMI at Exam 1 were not considered in the calculation of the weights at the first time point because these variables are believed to be effects of the physical activity level at Exam 1. Subjects with missing data at a given time point were removed from the analyses (about 1% for Exam 1 and about 3% for Exam 2).

We estimated the effect of current level of physical activity on current SBP and DBP using repeated measures MSMs and the same four weights that were investigated in the simulation studies ("1", w, sw and swb). For the estimation of the GEEs, a robust variance estimator was used along with an independence working correlation structure. The results are summarized in Table 7.3.

Upon the examination of Table 7.3, we remark that the estimates of the effect of current physical activity on current SBP are relatively robust to the choice of weights. However,

Table 7.3: Estimated effect of current physical activity level on current systolic (SBP) and diastolic (DBP) blood pressure.

Weights	Estimate for SBP (95% CI)	Estimate for DBP (95% CI)
1	-2.29 (-3.35, -1.22)	-0.82 (-1.40, -0.24)
w	-1.85 (-2.94, -0.75)	-0.43 (-1.04, 0.17)
sw	-1.94 (-3.59, -0.29)	-1.29 (-2.18, -0.39)
swb	-1.56 (-2.56, -0.55)	-0.29 (-0.84, 0.26)

LEGEND. "1": unweighted; w: standard weights; sw: (common) stabilized weights; swb: basic stabilized weights.

the choice of weights has a notable impact on the estimates of the effect of physical activity on DBP. In this case, the estimates obtained using an unweighted MSM or a MSM with common stabilized weights sw exhibit a significant decrease of DBP with physical activity at level $\alpha = 0.05$, whereas a non signifiant decrease is obtained from the MSMs with standard weights w and basic stabilized weights swb. These last results are in accordance with the *rmMSM Current* results from the simulation study where the unweighted and common stabilized weights sw estimates departed from those obtained with standard weights w and basic stabilized weights swb. Because there is believed to be time-dependent confounding, the unweighted repeated measures MSM is considered to be inappropriate for estimating the causal effect of current physical activity on current DBP. We also note that the confidence intervals obtained with the basic stabilized weights swb are slightly narrower than those obtained with the standard weights w.

7.7 Discussion

Although it is widely known that the weighting scheme affects the variance of MSM estimators, it is less well known that it can also affect their bias. Using a series of simulated examples, we showed that the utilization of the most common stabilized weights (weights sw) may lead to biased parameter estimates when structural models feature only partial information on treatment history, such as the current or most recent treatments. The diffusion of this result is critical since such structural model

specifications are often seen in repeated measures MSMs, a type of MSMs which is increasingly used to perform causal inferences (Fairall *et al.*, 2008; Patel *et al.*, 2008; Sampson *et al.*, 2006; Schildcrout *et al.*, 2011; VanderWeele *et al.*, 2011, 2012).

The phenomenon documented in this paper adds to the number of subtle issues arising in the implementation of MSMs (Yang & Joffe, 2012). Indeed, our results suggest that the choice of weights needs to be done according to the structural model that is specified. Particularly, we advise analysts to avoid using the common stabilized weights when the analyses target the estimation of the current or most recent treatment causal effects. In this context, the analysts could adopt the basic stabilized weights swb put forward herein, simple weights which have been found to yield unbiased results under all scenarios and structural models investigated.

CONCLUSION

Dans cette thèse, nous nous sommes intéressés à la problématique de l'identification de modèles appropriés pour effectuer de l'inférence causale à l'aide de données d'observation. Ce dernier chapitre effectue une revue des travaux présentés dans les chapitres précédents en situant leur contribution. Des voies de recherches futures sont également proposées.

Sélection guidée par les données

Nous avons d'abord abordé la problématique de la sélection de modèles pour l'inférence causale dans un contexte où les connaissances du domaine d'application sont encore peu avancées. Dans ce contexte, nous avons plus spécifiquement étudié les approches de sélection de modèles, ou de variables, guidées par les données. Les travaux présentés dans cette thèse ont d'abord permis une compréhension plus approfondie d'une approche existante, l'algorithme BAC. Nous avons également élaboré une justification théorique plus formelle à BAC et en avons simplifié l'utilisation. Cependant, l'approche proposée pour sélectionner l'hyperparamètre ω a obtenu un succès mitigé dans les études de simulations réalisées.

Plutôt que d'étudier davantage l'algorithme BAC, nous avons développé une nouvelle approche, *Bayesian Causal Effect Estimation* (BCEE), qui partage certaines similarités avec TBAC, ainsi qu'avec le BMA. À la différence de TBAC et du BMA, BCEE est motivé par le paradigme causal graphique. Les résultats des simulations effectuées suggèrent que BCEE est prometteur puisque sa performance, en termes d'erreur quadratique moyenne, est généralement au moins légèrement supérieure à celle d'un modèle de réponse complet ou à celle de BAC et de TBAC. La performance de BCEE

peut même approcher celle du vrai modèle de réponse dans des situations idéales. Ce nouvel algorithme pourrait être particulièrement utile aux chercheurs travaillant dans des domaines de recherche émergents, d'autant plus qu'un *package* R est disponible gratuitement sur le *Comprehensive R archive network* (CRAN). L'article présentant ce nouvel algorithme a d'ailleurs été blue accepté pour publication dans le *Journal of Causal Inference*.

Malgré ses qualités, l'algorithme BCEE que nous avons développé a un certain nombre de limites. Notamment, dans sa version actuelle, l'algorithme n'est applicable qu'au cas d'une exposition continue et d'une réponse continue. Tel qu'exposé au chapitre 4, une généralisation pour d'autres types d'exposition est assez directe, mais l'extension pour d'autres types de réponse semble plus compliquée.

Par ailleurs, l'approche proposée est basée sur des modèles de régression linéaire et est donc limitée par différentes hypothèses inhérentes à ces modèles, dont l'hypothèse de linéarité des effets. Une solution simple à ce problème pourrait être d'effectuer des transformations de variables avant d'exécuter l'algorithme. L'efficacité de cette approche mériterait d'être étudiée. De plus, tel qu'exposé au chapitre 2, des méthodes non paramétriques existent et pourraient être une alternative à BCEE lorsque les hypothèses du modèle linéaire ne sont pas raisonnables. Pour l'instant, la performance de BCEE n'a cependant pas été comparée à celle d'approches non paramétriques. Des études de simulations pourraient être menées afin d'effectuer une telle comparaison dans divers scénarios où les modèles paramétriques sont plus ou moins appropriés. Nous croyons que lorsque les hypothèses de BCEE sont approximativement correctes, BCEE est plus performant que des approches non paramétriques. Lorsque les hypothèses du modèle linéaire ne sont pas respectées, la performance relative de BCEE avec transformations par rapport à des approches non paramétriques nous apparaît moins certaine.

Sélection guidée par la littérature

La thèse a également porté sur la sélection de modèles pour l'inférence causale dans un contexte où les connaissances substantielles sont mieux développées. Ces travaux ont été motivés par une analyse secondaire des données du *Honolulu Heart Program* qui visait l'estimation des relations causales entre l'activité physique, la tension artérielle et la mortalité. Pour effectuer ces analyses, nous avons utilisé des MSMs. Les travaux liés à la deuxième partie de la thèse ont mené à la rédaction de trois articles scientifiques.

Un premier article, dont je suis le deuxième auteur, porte sur les résultats substantiels de l'analyse. Cet article a été soumis à une revue scientifique du domaine de la santé et se trouve en appendice A. Les résultats obtenus abondent dans la même direction que la littérature existante ; ils confirment que l'activité peut permettre de réduire la tension artérielle ainsi que le risque de mortalité.

Le second article, présenté au Chapitre 6 et dont je suis cette fois le premier auteur, présente la méthodologie statistique développée pour analyser les données et sera soumis à *Epidemiology* prochainement. Cette méthodologie comporte plusieurs éléments novateurs. Notamment, on y effectue une utilisation élargie du paradigme graphique à l'inférence causale ainsi qu'une sélection de variables pour les MSMs basée sur des DAGs d'abord élaborés sur la base de la littérature scientifique, puis validés et améliorés grâce à l'information apportée par les données. L'utilisation du paradigme graphique nous a par ailleurs permis d'introduire des MSMs conditionnels ajustant pour des variables variant dans le temps. Ces modèles peuvent produire des intervalles de confiance appropriés et plus étroits que les MSMs non conditionnels.

Nous croyons que la méthodologie que nous avons élaborée pourra être très utile en pratique, en particulier en épidémiologie où l'utilisation des DAGs est de plus en plus répandue. Ainsi, la sélection des variables utilisées pour la pondération, qui est reconnue comme un défi important dans l'implantation des MSMs, peut être facilitée,

d'autant plus que les DAGs élaborés sur la base des connaissances substantielles peuvent être validés avec les données grâce aux modèles d'équations structurelles.

Cependant, en effectuant des modifications aux DAGs inspirées par les données, une incertitude associée à la sélection du modèle est indirectement introduite et négligée, puisque les variables sont par la suite sélectionnées sur la base du DAG modifié. Dans la méthodologie proposée, nous avons mis l'accent sur le fait que les modifications apportées aux DAGs, bien que suggérées par les données, étaient d'abord et avant tout raisonnables sur la base des connaissances du domaine d'application. Nous croyons que l'application de cette règle devrait limiter les inconvénients associés à une sélection de variables basée sur les données. Des études de simulations pourraient potentiellement être réalisées afin de vérifier cette hypothèse, bien qu'il semble difficile à première vue de répliquer à l'aide de simulations une méthodologie fortement guidée par les connaissances du domaine d'application. Une autre limitation à l'approche proposée est qu'elle utilise des modèles d'équations structurelles supposant des relations linéaires entre les variables. Lorsque cette hypothèse n'est pas raisonnable, les tests d'ajustement des modèles servant à valider les DAGs pourraient produire des résultats erronés.

Le troisième article a été publié dans *Statistics in Medicine* et porte sur le choix d'une stabilisation des poids pour les MSMs en relation avec le modèle structurel choisi. Il est déjà bien connu dans la littérature scientifique que la stabilisation des poids peut avoir un impact sur la variance des estimateurs obtenus, mais le fait que la stabilisation peut également avoir une influence sur le biais était moins connu. Nous avons illustré à l'aide d'un exemple et d'études de simulations que les poids stabilisés usuels peuvent introduire un biais lorsque les MSMs utilisés n'incluent pas l'ensemble de l'historique d'exposition alors que le modèle structurel réel dépend de l'ensemble de l'historique d'exposition. Par opposition, des poids stabilisés simples ont été trouvés robustes à une telle erreur de spécification du modèle structurel. Les MSMs à mesures répétées semblent particulièrement susceptibles au problème mis en évidence dans cet article.

Puisque ces modèles gagnent en popularité, nous croyons que les résultats que nous avons obtenus sont extrêmement importants pour les applications futures des MSMs.

Bien que nous n'avons étudié que les MSMs classiques et les MSMs à mesures répétées, il est fort probable que les conclusions auxquelles nous sommes arrivés s'appliqueraient également pour les MSMs de Cox. Il serait également intéressant de déterminer si nos conclusions sont également valides pour la stabilisation des poids par probabilité inverse de censure utilisés pour tenir compte de l'attrition. Intuitivement, la problématique nous apparaît similaire, nous croyons ainsi que les poids stabilisés usuels pourraient également introduire un biais dans cette situation.

Finalement, des études futures pourraient vérifier si la stabilisation des poids que nous avons suggérée offre une robustesse contre le biais dans d'autres situations, notamment si le modèle structurel est mal spécifié en raison de sa forme fonctionnelle et non pas parce qu'il n'inclut pas l'ensemble de l'historique d'exposition. Notre conjecture est qu'un modèle structurel ainsi mal spécifié éliminera partiellement, mais pas totalement le biais de confusion dû à l'exposition passée. Ainsi, des poids stabilisés simples pourraient offrir un léger avantage en terme de biais par rapport aux poids stabilisés usuels. Cependant, nous croyons que la différence dans le biais sera généralement mineure en pratique et potentiellement négligeable par rapport à la réduction de la variance qui pourrait être obtenue en utilisant des poids stabilisés usuels.

APPENDICE A

MARGINAL STRUCTURAL MODELS FOR ESTIMATING THE RELATIONSHIPS BETWEEN PHYSICAL ACTIVITY, BLOOD PRESSURE, AND MORTALITY IN A LONGITUDINAL COHORT STUDY: THE HONOLULU HEART PROGRAM

Amanda M. Rossi, Denis Talbot, Geneviève Lefebvre, Juli Atherton, Simon L Bacon

Abstract:

Background: The purpose of this study was to evaluate the relationships between physical activity, blood pressure (BP), mortality and major adverse cardiovascular events (MACE) during an extended period.

Methods and Results: This study comprised secondary analyses of a longitudinal, observational study, the Honolulu Heart Program (n = 8006 men). Physical activity (measured by self-report questionnaire) and BP were both assessed at three time points; Exam 1 (1965-1968), Exam 2 (1968-1971), and Exam 4 (1991-1993). Marginal structural Cox models and Marginal structural models for repeated measures were used to estimate: 1) the separate effect of physical activity and BP on mortality and MACE; and 2) the effect of physical activity on BP. Being physically active was associated with a reduced rate of mortality (Hazard Ratio (HR) = 0.68, 95% confidence interval (CI) = 0.60 to 0.76) and MACE (HR = 0.84, 95% CI: 0.75 to 0.93) by 32% and 16%, respectively. Blood pressure was shown to have a dose-dependent relationship with both mortality and MACE whereby increasing BP was related to more events. Active participants showed a significant decrease of 2.47 mmHg (95%CI, -3.46 to -1.48)

in systolic BP compared to the inactive group. No change in diastolic BP was observed. *Conclusions:* Using a large sample with an extended follow-up period, we studied the relationships between physical activity, blood pressure, mortality, and MACE, applying novel statistical models which account for covariate variation over time. The results support that being physically active is associated with better outcomes and that BP may be a mediator of the relationship between physical activity and mortality. *Clinical Trial Registration:* ClinicalTrials.gov (identifier: NCT00005123)

Keywords: Blood pressure, Hypertension, Exercise, Cardiovascular Diseases, Mortality, Epidemiology

A.1 Introduction

Findings have indicated that higher blood pressure (BP) increases risk of cardiovascular mortality, whereas physical activity decreases the risk (Vatten *et al.*, 2006). Furthermore, a dose-response association between BP and mortality, where increasing BP increases risk of death, and an inverse dose-response relationship between physical activity and mortality have been demonstrated (Glynn *et al.*, 1995; Lee & Skerrett, 2001). A variety of previous studies have shown that physical activity is associated with reduced risk of mortality in people with high BP (Rossi *et al.*, 2012). These studies have all taken a traditional approach using data collected at a single time point, usually at the point of entry into the study, then followed participants for a determined length of time, censorship point, or death (Rossi *et al.*, 2012). These studies have generally employed a standard Cox proportional hazards model to quantify the effect of BP and physical activity on mortality. Therefore, one issue with these studies is that the data and/or models did not account for changes in the exposure (e.g., BP or physical activity) occurring over time, and thus did not allow for understanding how these changes may impact survival (Rossi *et al.*, 2012).

The purpose of this study was to evaluate the following relationships: 1) the separate effects of physical activity and BP on mortality and major adverse cardiovascular

events (MACE); and 2) the effect of physical activity on BP, while allowing for the exposure and any covariates to change over time. In doing so, a secondary objective was to examine the role of blood pressure as a mediator of the physical activity-survival/MACE relationships. We explored this in the Honolulu Heart Program (HHP) dataset, which followed the same cohort of Japanese-American men for an extended period of time, from 1965 until 1994, with multiple follow-up periods between baseline and censorship. We used Marginal structural Cox models (MSCMs) and Marginal structural models (MSMs) for repeated measures to estimate the aforementioned relationships. Unlike Cox models with time-varying exposure and covariates, these recent models are recognized to be appropriate to estimate causal relationships in longitudinal settings when there exists time-dependent confounders that are affected by previous exposure (Robins *et al.*, 2000; Xiao *et al.*, 2010). Intuitively, under the assumption of no unmeasured confounder, marginal structural models allow one to replicate the results that would have been observed under a sequentially randomized experiment when utilizing observational data. Remark, however, that it would be practically impossible to carry out a true randomized experiment on physical activity with such a long follow-up period. To our knowledge, no previous studies have used MSCMs and MSMs to investigate the effect of physical activity on BP, and subsequently, on mortality and MACE.

A.2 Methods

The Honolulu Heart Program

The HHP is a longitudinal, epidemiological study of 8006 Japanese-American men living on the island of Oahu, Hawaii who were born between 1900 and 1919. Participants were initially recruited between 1965 and 1968 from a listing of selective service registrants and were between the ages of 45-68 years old at the beginning of the study (Worth & Kagan, 1970). The data collection protocol has been previously described (Kagan *et al.*, 1974). These secondary analyses of the original dataset

are based on four examinations; Exam 1 (1965-68), Exam 2 (1968-1971), Exam 3 (1971-1975) and Exam 4 (1991-1993). All variables (age, employment status, body mass index, smoking status, and anti-hypertension medication usage) included in the analyses were time-varying. These variables were selected because they are deemed clinically relevant and because they were consistently measured across three of the four examinations. Approval for these analyses was obtained from the Concordia University Human Research Ethics Committee (UH2012-025, 10000588).

Physical Activity Measurement

Physical activity was measured at three time points (Exams 1, 2, and 4) by self-report questionnaire. At Exams 1 and 4 participants reported the number of hours per day spent in each of 5 physical activity levels: no physical activity (sleeping, lying down, or reclining); sedentary activity (sitting or standing); slight activity (casual walking); moderate activity (gardening or light carpentry); and heavy activity (lifting, shoveling, or digging). At Exam 2 participants were asked two questions; first about their physical activity on the job and second about their level of physical activity at home. Participants qualified their physical activity by selecting one of the following responses: "mostly sitting," "moderate," or "much." In order to standardize physical activity across the three time points, we created a binary physical activity variable where participants were defined as active if they reported any moderate or heavy physical activity at Exams 1 and 4 and "moderate" or "much" activity at home or on the job for Exam 2. See Supplemental Materials (A.6) for validation of this method.

Blood Pressure Measurement

BP was measured using a mercury manometer by a trained individual (nurse, technician, and/or physician) at all examination points. Measurements were taken in a resting, seated position. Diastolic BP (DBP) was considered as the fifth Korotkoff sound. For the purposes of these analyses, serial BP measurements were averaged.

Additional information regarding BP measurement at each examination is available in the Supplemental Materials (A.6).

Surveillance and Outcomes

Mortality and cardiovascular morbidity were continually monitored from the inception of data collection through to the censorship point (December 1994) via hospital admission and discharge records, obituaries, and death certificates recorded with the State Department of Health. MACE was defined as any fatal or non-fatal event including myocardial infarction, stroke, coronary artery bypass graft, acute coronary insufficiency, coronary angioplasty, and other cardiac surgeries.

Data Treatment

As per recommendation we used age as the time-scale for survival time; that is, survival time was defined as the number of days between birth and death (Thiébaut & Bénichou, 2004; Kom *et al.*, 1997). For individuals who did not die during the study, survival time was right censored at the time of their last examination if they did not attend Exam 4, or at the end of follow-up (December 1994), otherwise. Time to MACE was defined analogously to survival time.

When systolic BP (SBP) and DBP were used as exposure variables in the statistical analyses, we divided them into four categories according to a standard BP classification scheme (Chobanian *et al.*, 2003). Specifically SBP was categorized as: <120mmHg, 120-139mmHg, 140-159mmHg and ≥160mmHg and DBP as: <80mmHg, 80-89mmHg, 90-99mmHg and ≥100mmHg. For both the MSCMs and the MSMs, we built an augmented dataset where each subject-Exam corresponds to one row. If a row contained missing values for at least one variable required to estimate a given effect, then it was not considered for that estimation (listwise deletion was performed). Data available for each effect estimate is detailed in the Supplemental Materials (A.6).

Building a Directed Acyclic Graph

As suggested by Hernán *et al.* (2002) we first drew directed acyclic graphs (DAGs) to represent the relationships between all clinically relevant variables using substantive prior knowledge. Two DAGs were created in total; one for the triplet of physical activity, BP, and survival, and another one for the triplet of physical activity, BP, and MACE. The DAGs were created for triplets of variables, instead of producing DAGs for each relationship of interest, in order to examine mediation of the effect of physical activity on survival and MACE through BP. Both DAGs have the same core structure with only outcome (survival or MACE) differing between DAGs. We assessed the goodness of fit of the proposed DAGs using structural equation models. Some modifications were made to the initial DAGs in order to improve the fit. For each relationship investigated, we used Pearl's back-door criterion on the final DAGs to identify the set of confounding covariates at each time point (Pearl, 2009). For a brief overview of Pearl's causal graphical framework, we refer the reader to the appendix of VanderWeele & Shpitser (2011). The final DAG equations for survival are detailed in the Supplemental Materials (A.6) as an example.

The Effects of Physical Activity, SBP and DBP on Survival Time and Time to MACE

We used a MSCM to estimate the effect of current physical activity, that is the physical activity level reported at the most recent exam, on survival time and time to MACE (Hernán *et al.*, 2001, 2000). We used normalized basic stabilized inverse probability of treatment weights in the MSCMs to account for time-dependent confounding as per the final DAGs (Xiao *et al.*, 2010; Talbot *et al.*, 2015). We used appropriate logistic regression models to calculate the weights. At each time point we truncated the weights at 100 to limit the impact of outlying individuals to notably influence the results. We also used MSCMs to estimate the effect of different levels of BP (SBP, DBP)

on time to survival and time to MACE. Due to the ordinal nature of the BP exposure covariates, the weights were calculated with ordinal logistic regression models.

The Effects of Physical Activity on SBP and DBP

We used MSMs for repeated measures to estimate the effect of current physical activity on current SBP and DBP, separately (Hernán & Brumback, 2002). Our repeated MSMs allowed for the estimation of the effect of the physical activity level reported at a given exam on the BP measured at that same exam, simultaneously for all three exams (> 18,000 person-exams). Following Hernán & Brumback (2002), we used both inverse-probability-of-treatment weights and inverse-probability-of-censoring weights to account for time-dependent confounding and censoring. Again we used logistic regression models to calculate the weights, and a similar truncation strategy was adopted.

Statistical Analysis

We used R package `lavaan` to build the DAGs (R Core Team, 2014; Rosseel, 2012). SAS version 9.2 was used for all other analyses. The `PROC PHREG` command was used to fit the MSCMs and `PROC GENMOD` to fit the MSMs for repeated measures (SAS Institute Inc., 2011).

A.3 Results

Study Participants

The current analysis examined 8006 male participants (54 $\pm$ 6 years old at baseline). Approximately 4% of participants (n = 304) had a history of cardiovascular disease (myocardial infarction, stroke, or congestive heart failure) upon entrance to the study. The second examination collected data on 7498 men and 3845 participants at the fourth examination. The average length of follow-up was 21.5 years (range: 0.1 years

to 33.1 years). A total of 4879 deaths were reported from any cause. There were 1318 cardiovascular deaths (including stroke) and 3279 individuals who experienced at least one MACE during follow-up. See Table A.1 for additional participant characteristics.

Table A.1: Baseline participant characteristics

Characteristic	Mean ± SD
N	8006
Age (years)	54 ± 6
BMI (kg/m^2)	23.8 ± 3.1
Systolic blood pressure (mmHg)	134 ± 21
Diastolic blood pressure (mmHg)	82 ± 12
Physical activity (n)	6494 (81%)
Smoking status (n)	
Never smoker	2409 (30%)
Previous smoker	2094 (26%)
Current smoker	3502 (44%)
History of Cardiovascular Disease (n)	304 (2.5%)

Physical Activity, Survival, and MACE

Over 80% of participants were classified as active at baseline (Exam 1), whereas 88% were active at Exam 2, and 75% at Exam 4. The results of the MSCM indicated active individuals had a 32% reduced risk of mortality (Hazard Ratio (HR) = 0.68, 95% confidence interval (CI): 0.60 to 0.76) compared to the inactive participants. Risk of MACE was also significantly decreased in the physically active participants compared to the inactive group (HR = 0.84, 95% CI: 0.75 to 0.93).

Blood Pressure, Survival, and MACE

Figure 5.2 and Table A.2 display the results of BP on survival. These demonstrate a dose-response relationship between SBP, DBP and risk of all-cause mortality. The results indicated that higher BP is associated with increased risk of death. Pairwise

comparisons showed a significant difference in risk between all BP groups except for the two lowest SBP categories and between the 80-89 mmHg and 90-99 mmHg categories of DBP (see Supplemental Materials (A.6)).

Table A.2: Risk of all-cause mortality according to systolic and diastolic blood pressure categories

Systolic Blood Pressure	Hazard Ratio
< 120 mmHg	1.00 (Ref)
120-139 mmHg	1.01 (0.93, 1.09)
140-159 mmHg	1.41 (1.30, 1.53)
≥ 160 mmHg	1.63 (1.48, 1.79)
Diastolic Blood Pressure	**Hazard Ratio**
< 80 mmHg	1.00 (Ref)
80-89 mmHg	1.16 (1.09, 1.24)
90-99 mmHg	1.14 (1.05, 1.24)
≥ 100 mmHg	1.76 (1.58, 1.96)

Ref: Reference group.

Similar results were observed for MACE (see Figure 5.3 and Table A.3). Risk of MACE increased in a dose-dependent manner with higher SBP, there was a greater than 60% increased risk for participants with SBP ≥ 160 mmHg compared to the normal BP group. Risk of MACE also significantly increased with increased DBP. Pairwise comparisons also showed significant differences between SBP groups and DBP groups (see Supplemental Materials (A.6)).

Physical Activity and Blood Pressure

We observed a significant decrease of 2.47 mmHg (95% CI, -3.46 to -1.48) in SBP between physically active and inactive participants. No change in DBP was observed

Table A.3: Risk of MACE according to systolic and diastolic blood pressure categories

Systolic Blood Pressure	Hazard Ratio
< 120 mmHg	1.00 (Ref)
120-139 mmHg	1.04 (0.94, 1.15)
140-159 mmHg	1.52 (1.37, 1.70)
≥ 160 mmHg	1.69 (1.50, 1.92)

Diastolic Blood Pressure	Hazard Ratio
< 80 mmHg	1.00 (Ref)
80-89 mmHg	1.18 (1.09, 1.29)
90-99 mmHg	1.29 (1.17, 1.42)
≥ 100 mmHg	1.78 (1.56, 2.04)

Ref: Reference group.

between the physically active and the inactive groups (0.26 mmHg; 95% CI, -0.22 to 0.75).

Sensitivity Analyses

A first sensitivity analysis consisted of repeating the analyses described above, but excluding participants with a history of cardiovascular disease at baseline (n = 304). Also, MSCMs and MSMs for estimating the effect of the cumulative number of exams where participants were physically active were performed. For comparison with the main results and previous findings, we conducted crude analyses that did not account for confounding (i.e., unweighted versions of the MSCMs and MSMs). The results of the sensitivity analyses parallel the main findings above (see Supplemental Materials (A.6)).

A.4 Discussion

To our knowledge, this is the first study to use MSCMs and MSMs to simultaneously report on the effects of BP and physical activity on both survival and MACE. The results demonstrate a dose-dependent relationship between BP and the outcomes of

interest; all-cause mortality and MACE. Additionally, physical activity was found to have a relationship with survival and MACE whereby active individuals have a lower rate of both death and MACE compared to their inactive counterparts. Finally, we examined the relationship between physical activity and BP which showed a decrease in SBP (2.47 mmHg), but no change in DBP. Taken together, these analyses demonstrate that BP might mediate the physical activity-mortality relationship and the beneficial effects of physical activity on our outcome measures may in part be due to improvements in SBP.

Our analyses build from previous studies which are consistent with our findings. Several studies have previously described a relationship between BP and mortality whereby people with lower BP have lower rate of all-cause and cardiovascular mortality (Rossi *et al.*, 2012). A large meta-analysis of individual participant data has demonstrated that a strong, direct relationship exists between BP and all-cause and vascular (e.g., stroke, ischemic heart disease, etc.) mortality at all age groups above 40 years of age (Prospective Studies Collaboration, 2002). As with BP, our findings regarding physical activity are similar to those noted in association studies (Kujala *et al.*, 1998). However, to our knowledge, this is the first report to take a formal causal inference perspective to further develop this theory and investigate these relationships whereby participation in physical activity decreases risk of all-cause mortality and MACE. Whilst several studies have also shown dose-response associations between physical activity and mortality (i.e., increased volume of physical activity is associated with increased life expectancy) we were not able to confirm these because of a lack of consistent information across the follow-up periods (Moore *et al.*, 2012; Wen *et al.*, 2011). However, our coarse measure of physical activity was able to detect a significant difference in mortality and MACE rates between active and inactive participants.

Individual intervention studies of physical activity for BP lowering have shown varying results. Meta-analyses, however, have demonstrated that participation in aerobic exercise can effectively lower BP (approximately 3-4 mmHg decrease in SBP and 2-3 mmHg decrease in DBP), though there is debate as to whether resistance

training lowers BP (Cornelissen & Smart, 2013; Rossi *et al.*, 2013). Epidemiological data has been consistent with this, suggesting that being physically active (measured by self-report) is associated with lower BP levels, even in specific populations, for example people with hypertension, older women, and children (Montoye *et al.*, 1972; Paffenbarger Jr & Lee, 1997; Reaven *et al.*, 1991; Gidding *et al.*, 2006).

Our analyses found a small effect of physical activity on SBP. One explanation for this marginal effect may be that the population studied was extremely active, especially when compared to modern samples. For example, greater than 80% of the Honolulu Heart Program cohort were defined as active at baseline, whereas a recent study showed that only 15% of Canadian adults and just over 30% of American adults were active enough to meet current physical activity guidelines (Colley *et al.*, 2011; Zhao *et al.*, 2013). However, data from a recent sample of 700 Hawaiian adults showed that on average participants achieved 67.5 metabolic equivalent hours/week, exceeding the physical activity recommendations of the American College of Sports Medicine (Chai *et al.*, 2010; Garber *et al.*, 2011). Therefore, the nature of this specific population, i.e., Japanese-American men recruited from a roster of military servicemen from the Hawaiian Islands during the Second World War and who were living in a very unique setting, may potentially explain their high level of activity. Also, the measure of physical activity used in this study may not have been discreet enough to detect a greater effect of physical activity on BP. That is to say a more precise, objective measure, e.g., accelerometry, pedometry, or a more discreet self-report scale, e.g., minutes per day of activity vs. hours per day, and consistency between examinations, might have allowed for better assessment of the relationship between physical activity and BP. This point is especially important because previous association studies have shown the physical activity-BP relationship may be intensity-dependent (Paffenbarger Jr & Lee, 1997; Hu *et al.*, 2004), though there is still some debate about this (Nybo *et al.*, 2010). Thus it is important to further elucidate the role of physical activity intensity and focus on distinguishing between sedentary, light, moderate, and vigorous intensity as this may be key to understanding the relationship between physical activity and BP.

As mentioned above, previous longitudinal, observational studies examining the effect of BP and physical activity on mortality have shown strong associations between these variables, even in people with high BP (Rossi *et al.*, 2012). Meta-analyses of randomized controlled trials have shown that aerobic exercise training consistently decreased BP (Cornelissen & Smart, 2013), and is recommended for maintaining healthy BP and lowering BP in hypertension (Hackam *et al.*, 2013). Additionally, BP is related to mortality in a dose-dependent manner (Glynn *et al.*, 1995). A meta-analysis of individual data from nearly one million participants has demonstrated that BP was positively related to all-cause and cardiovascular mortality (Prospective Studies Collaboration, 2002). On this basis we hypothesized that the effect of physical activity on mortality would be mediated by BP. Though our results suggest BP may mediate the physical activity-mortality and physical activity- MACE relationships, the contribution through BP may be small. Therefore, it could be that physical activity functions to improve and maintain healthy levels of different risk factors, other than BP, which prolong longevity. For example, physical activity has been shown to help maintain weight, improve mental health, improve vascular function, and overall improved health related quality of life, amongst other health benefits (Tremblay *et al.*, 1999; Dunn *et al.*, 2001; Hambrecht *et al.*, 2003; Heesch *et al.*, 2012). In addition, evidence suggests that physical activity directly impacts vascular wall function, and therefore improving cardiovascular risk beyond traditional risk factor modification (Green *et al.*, 2008).

Although sedentary behaviours (e.g., watching TV, screen time; defined as any waking activity expending ≤ 1.5 metabolic equivalents and sitting or reclining posture) were not addressed in this study, as this would require technology not available at the time these data were collected, it could be that sedentary behavior, may be another important predictor of outcomes (Barnes *et al.*, 2012). Sedentary behaviours have been shown to be associated with various negative health outcomes (Jakes *et al.*, 2003; Beunza *et al.*, 2007; Tremblay *et al.*, 2010), and can predict both cardiovascular and all-cause mortality in adults, independent from physical activity (Wijndaele *et al.*, 2010). Therefore it is possible that the pathological mechanisms attributed to sedentary

behaviours may be responsible of the detrimental effects of physical inactivity on BP and future research should additionally focus on sedentary behaviours, e.g., television viewing and driving, and objectively measured sedentariness (measured with accelerometry).

Limitations

The results of the present series of analyses need to be interpreted within the context of some limitations of the study. Firstly, the causal interpretation of the analyses rest upon the assumption that the DAGs we have built are correct. Even though the final DAGs obtained were supported by the data, they may still not be correct. For instance, some clinically important covariates might not have been included in the DAGs, e.g., sodium consumption, because they were not consistently available in this dataset. Second, and as detailed previously, self-reported physical activity data has been shown to be less reliable than objective measurement of activity because it is subject to recall bias and can be influenced by a participants health and mood, especially in older adults (Rikli, 2000). Also, the measurement of physical activity (i.e., the questionnaire items) lacked consistency between examination points within the study and not all questions were discreet enough to develop a more comprehensive measure. However, we qualified being physically active as participating in a minimum of one hour per day of moderate activity which exceeds current guidelines and our method of standardizing the physical activity measure was shown to be valid (see Supplemental Materials (A.6)). Though BP was measured according to standards at the time of assessment, more recent data suggests that automated BP measurement is a more reliable predictor of risk and reduce white coat effect (Myers *et al.*, 2011; Myers & Godwin, 2007). Therefore, more advanced methods of measuring BP may alter the findings presented herein. Another limitation is the inclusion of only men in this cohort. Although no women were included in this study, previous findings suggest there may be a difference between sexes with respect to BP, physical activity, and mortality. Association studies have shown similar patterns in risk of mortality when

joining physical activity and BP categories; however, the magnitude of risk differed between sexes (higher in women) (Vatten *et al.*, 2006).

Despite these few limitations, there are a number of strengths; for instance, the large sample size (8006 men at baseline), the long follow-up period (> 21 years), and a good retention of participants. Most significantly, the application of MSCMs and MSMs for repeated measures, allowed us to better examine the relationships between these variables. As mentioned in the introduction, under ideal circumstances, marginal structural models can replicate the results from a sequentially randomized experiment utilizing observational data. Based on these analyses, we have garnered a more refined understanding of the relationships between physical activity, BP, and our outcomes of interest, mortality and MACE, over an extended follow-up period.

A.5 Conclusions

In summary, our analyses show strong and positive dose-dependent associations between BP and mortality/MACE. Moreover, physical activity was shown to be negatively associated with mortality/MACE. Physical activity and SBP were also found to be negatively associated. Since special attention was given to appropriately dealing with confounding, these associations could be causally interpreted under the assumption of no unmeasured confounders. Taken together, this suggests that physical activity is a determinant of mortality/MACE, with BP mediating the relationship between physical activity and mortality/MACE. Our results thus provide support for recommending physical activity as a way to reduce risk of mortality/MACE.

A.6 Supplemental Material

Methods

Consistency of the Physical Activity Measurement

In order to test the validity of our approach to classifying physical activity, we built a 2x2 table where individuals were categorized as being active/inactive according to two subjective questionnaires. Participants were classified as active in the first questionnaire if they indicated either much or moderate physical activity at home or on the job. Participants were classified as active in the second questionnaire if they reported spending any time doing moderate physical activity. We used Exam 1 as a reference because both self-report questionnaires were assessed at this time point. Formal analysis indicates 83.7% concordance between these two physical activity questionnaires ($\kappa = 0.42$). Therefore, we deemed it appropriate to use this method to create a binary physical activity variable (active/inactive).

Blood Pressure Measurement

Table A.4: Blood pressure measurement details at each examination

Examination	Number of measurements	Performed by	Arm
Exam 1	3	Nurse (2), Physician (1)	Left
Exam 2	4	Nurse (2), Physician (2)	3 Left, 1 Right
Exam 4	2	N/A	Left

Data Treatment

Table A.5: Available data for every effect estimation

	Exam 1	Exam 2	Exam 4	Total patient-exam
Alive	n = 8006	n = 7603	n = 4330	19 936
SBP ← PA	n = 7943 (99%)	n = 7410 (97%)	n = 3317 (77%)	18 670 (94%)
DBP ← PA	n = 7943 (99%)	n = 7410 (97%)	n = 3313 (77%)	18 666
Surv. ← PA	n = 7911 (99%)	n = 7410 (97%)	n = 3406 (79%)	18727 (94%)
Surv. ← SBP	n = 7906 (99%)	n = 7389 (97%)	n = 3300 (76%)	18 595 (93%)
Surv. ← DBP	n = 7906 (99%)	n = 7389 (97%)	n = 3300 (76%)	18 595 (93%)
Alive and without MACE	n = 8006	n = 7463	n = 3343	18 812
MACE ← PA	n = 7911 (99%)	n = 7295 (98%)	n = 2691 (80%)	17 897 (95%)
MACE ← SBP	n = 7906 (99%)	n = 7275 (97%)	n = 2615 (78%)	17 796 (95%)
MACE ← DBP	n = 7906 (99%)	n = 7275 (97%)	n = 2615 (78%)	17 796 (95%)

DAG Equations

Table A.6: DAG equations for survival for each examination point. Directed arrows represent cause-effect relationships. Bi-directed arrows represent an unobserved common cause between the variables. T1 = Exam 1; T2 = Exam 2; T4 = Exam 4; PhysicalActivity = physical activity status; Age = age; Employment = employment status; BMI = Body Mass Index; CurrentSmoker, PreviousSmoker = smoking status (current, previous, never); SystolicBP = systolic blood pressure; DiastolicBP = diastolic blood pressure; Survival = survival; HyperTensTrt = Hypertension medication usage.

Exam 1	
PhysicalActivityT1 ←	AgeT1
EmployementT1 ←	AgeT1
HyperTensTrtT1 ←	PhysicalActivityT1 + AgeT1 + EmployementT1 + BMIT1 + CurrentSmokerT1
CurrentSmokerT1 ←	AgeT1 + EmployementT1
PreviousSmokerT1 ←	AgeT1 + EmployementT1
BMIT1 ←	PhysicalActivityT1 + AgeT1
DiastolicBPT1 ←	PhysicalActivityT1 + AgeT1 + EmployementT1 + BMIT1 + CurrentSmokerT1 + PreviousSmokerT1
SystolicBPT1 ←	PhysicalActivityT1 + AgeT1 + EmployementT1 + BMIT1 + CurrentSmokerT1 + PreviousSmokerT1
SurvivalT1 ←	SystolicBPT1 + DiastolicBPT1 + PhysicalActivityT1 + Age1 + EmployementT1 + BMIT1 + CurrentSmokerT1 + PreviousSmokerT1

CurrentSmokerT1 ↔	PreviousSmokerT1
PhysicalActivityT1 ↔	EmployementT1
HyperTensTrtT1 ↔	DiastolicBPT1
HyperTensTrtT1 ↔	SystolicBPT1
BMIT1 ↔	CurrentSmokerT1
BMIT1 ↔	PreviousSmokerT1
Exam 2	
PhysicalActivityT2 ←	HyperTensTrtT1+ PhysicalActivityT1 + BMIT1 + AgeT2 + EmployementT2
EmployementT2 ←	EmployementT1+ AgeT2
HyperTensTrtT2 ←	SystolicBPT1 + DiastolicBPT1 + HyperTensTrtT1 + PhysicalActivityT2 + AgeT2 + EmployementT2 + BMIT2 + CurrentSmokerT2
CurrentSmokerT2 ←	CurrentSmokerT1 + PreviousSmokerT1 + AgeT2 + EmployementT2
PreviousSmokerT2 ←	CurrentSmokerT1 + PreviousSmokerT1 + AgeT2 + EmployementT2
BMIT2 ←	PhysicalActivityT1 + BMIT1 + PhysicalActivityT2 + AgeT2
DiastolicBPT2 ←	DiastolicBPT1 + HyperTensTrtT1 + PhysicalActivityT1 + BMIT1 + CurrentSmokerT1 + PreviousSmokerT1 + PhysicalActivityT2 + AgeT2 + EmployementT2 + BMIT2 + CurrentSmokerT2 +PreviousSmokerT2
SystolicBPT2 ←	SystolicBPT1 + HyperTensTrtT1 + PhysicalActivityT1 + BMIT1 + CurrentSmokerT1 + PreviousSmokerT1 + PhysicalActivityT2 + AgeT2 + EmployementT2 + BMIT2 +CurrentSmokerT2 + PreviousSmokerT2
SurvivalT2 ←	SystolicBPT1 + DiastolicBPT1 + HyperTensTrtT1 + PhysicalActivityT1 + BMIT1 + CurrentSmokerT1 + PreviousSmokerT1 + SystolicBPT2 + DiastolicBPT2 + PhysicalActivityT2 + AgeT2 + EmployementT2 + BMIT2 + CurrentSmokerT2 + PreviousSmokerT2
CurrentSmokerT2 ↔	PreviousSmokerT2
HyperTensTrtT2 ↔	DiastolicBPT2
HyperTensTrtT2 ↔	SystolicBPT2
BMIT2 ↔	CurrentSmokerT2
BMIT2 ↔	PreviousSmokerT2
Exam 4	
PhysicalActivityT4 ←	HyperTensTrtT1 + PhysicalActivityT1 + HyperTensTrtT2 + PhysicalActivityT2 + BMIT2 + AgeT4
EmployementT4 ←	EmployementT1 + EmployementT2 + AgeT4
HyperTensTrtT4 ←	SystolicBPT1+ DiastolicBPT1 + HyperTensTrtT1 + SystolicBPT2 + DiastolicBPT2 + HyperTensTrtT2 + BMIT2 + PhysicalActivityT4 + AgeT4

+ EmployementT4 + BMIT4 + CurrentSmokerT4

CurrentSmokerT4 ← CurrentSmokerT1 + PreviousSmokerT1 + CurrentSmokerT2
+ PreviousSmokerT2 + AgeT4 + EmployementT4

PreviousSmokerT4 ← CurrentSmokerT1 + PreviousSmokerT1+ CurrentSmokerT2
+ PreviousSmokerT2 + EmployementT4

BMIT4 ← PhysicalActivityT1 + BMIT1 + PhysicalActivityT2 + BMIT2 + PhysicalActivityT4
+ AgeT4

DiastolicBPT4 ← DiastolicBPT1 + HyperTensTrtT1 + PhysicalActivityT1
+ CurrentSmokerT1 + PreviousSmokerT1 + DiastolicBPT2
+ HyperTensTrtT2 + PhysicalActivityT2 + BMIT2 + CurrentSmokerT2
+ PreviousSmokerT2 + PhysicalActivityT4 + AgeT4 + EmployementT4
+ BMIT4 + CurrentSmokerT4 + PreviousSmokerT4

SystolicBPT4 ← SystolicBPT1 + HyperTensTrtT1 + PhysicalActivityT1 + CurrentSmokerT1
+ PreviousSmokerT1 + SystolicBPT2 + HyperTensTrtT2
+ PhysicalActivityT2 + BMIT2 + CurrentSmokerT2 + PreviousSmokerT2
+ PhysicalActivityT4 + AgeT4 + EmployementT4 + BMIT4
+CurrentSmokerT4 + PreviousSmokerT4

Survival T4 ← SystolicBPT1 + DiastolicBPT1 + HyperTensTrtT1 + PhysicalActivityT1
+ CurrentSmokerT1 + PreviousSmokerT1 + SystolicBPT2
+ DiastolicBPT2 + HyperTensTrtT2 + PhysicalActivityT2 + BMIT2
+ CurrentSmokerT2 + PreviousSmokerT2 + SystolicBPT4
+ DiastolicBPT4+ PhysicalActivityT4 + AgeT4 + EmployementT4
+ BMIT4 + CurrentSmokerT4 + PreviousSmokerT4

CurrentSmokerT4 ↔ PreviousSmokerT4

PhysicalActivityT4 ↔ EmployementT4

HyperTensTrtT4 ↔ DiastolicBPT4

HyperTensTrtT4 ↔ SystolicBPT4

BMIT4 ↔ CurrentSmokerT4

BMIT4 ↔ PreviousSmokerT4

Results

Table A.7: Results of pairwise comparisons for systolic and diastolic blood pressure on survival

Comparison Systolic Blood Pressure Groups	Hazard Ratio	p
$\geq$ 160 mmHg vs 140-159 mmHg	1.15	0.0017
$\geq$ 160 mmHg vs 120-139 mmHg	1.61	<0.0001
$\geq$ 160 mmHg vs <120 mmHg	1.63	<0.0001
140-159 mmHg vs 120-139 mmHg	1.40	<0.0001
140-159 mmHg vs <120 mmHg	1.41	<0.0001
120-139 mmHg vs <120 mmHg	1.01	0.8293

Comparison Diastolic Blood Pressure Groups	Hazard Ratio	p
$\geq$ 100 mmHg vs 90-99 mmHg	1.55	<0.0001
$\geq$ 100 mmHg vs 80-89 mmHg	1.51	<0.0001
$\geq$ 100 mmHg vs <80 mmHg	1.76	<0.0001
90-99 mmHg vs 80-89 mmHg	0.98	0.6146
90-99 mmHg vs <80 mmHg	1.14	0.0023
80-89 mmHg vs <80 mmHg	1.16	<0.0001

Table A.8: Results of pairwise comparisons for systolic and diastolic blood pressure on MACE

Comparison Systolic Blood Pressure Groups	Hazard Ratio	p
$\geq$ 160 mmHg vs 140-159 mmHg	1.11	0.0713
$\geq$ 160 mmHg vs 120-139 mmHg	1.63	<0.0001
$\geq$ 160 mmHg vs <120 mmHg	1.69	<0.0001
140-159 mmHg vs 120-139 mmHg	1.46	<0.0001
140-159 mmHg vs <120 mmHg	1.52	<0.0001
120-139 mmHg vs <120 mmHg	1.04	0.4193

Comparison Diastolic Blood Pressure Groups	Hazard Ratio	p
$\geq$ 100 mmHg vs 90-99 mmHg	1.39	<0.0001
$\geq$ 100 mmHg vs 80-89 mmHg	1.51	<0.0001
$\geq$ 100 mmHg vs <80 mmHg	1.78	<0.0001
90-99 mmHg vs 80-89 mmHg	1.09	0.0918
90-99 mmHg vs <80 mmHg	1.29	<0.0001
80-89 mmHg vs <80 mmHg	1.18	<0.0001

Sensitivity Analysis

The results below are those for the analyses excluding participants with a history of CVD at baseline (n = 304). Note: these findings are similar to those reported in the whole sample.

Physical Activity, Survival and MACE

The results of the analysis indicate active individuals had a reduced rate of mortality (HR = 0.78, 95% CI: 0.71 to 0.85) compared to the inactive participants. Risk of MACE was also significantly decreased in the physically active participants compared to the inactive group (HR = 0.86, 95% CI: 0.76 to 0.97).

Physical Activity and Blood Pressure

There was a significant decrease of 2.34 mmHg (95% CI, -3.35 to -1.33) in SBP between physically active and inactive participants. No change in DBP was observed between the physically active group and the inactive (0.39 mmHg; 95% CI, -0.10 to 0.88).

Blood Pressure

Table A.9: Risk of mortality according to systolic and diastolic blood pressure categories excluding participants with CVD at baseline

Systolic Blood Pressure	Hazard Ratio
< 120 mmHg	1.00 (Ref)
120-139 mmHg	1.01 (0.93, 1.09)
140-159 mmHg	1.40 (1.28, 1.52)
≥ 160 mmHg	1.63 (1.48, 1.80)

Diastolic Blood Pressure	Hazard Ratio
< 80 mmHg	1.00 (Ref)
80-89 mmHg	1.16 (1.09, 1.25)
90-99 mmHg	1.14 (1.05, 1.25)
≥ 100 mmHg	1.75 (1.56, 1.96)

Ref: Reference group.

Table A.10: Risk of MACE according to systolic and diastolic blood pressure categories excluding participants with CVD at baseline

Systolic Blood Pressure	Hazard Ratio
< 120 mmHg	1.00 (Ref)
120-139 mmHg	1.04 (0.94, 1.15)
140-159 mmHg	1.52 (1.36, 1.69)
≥ 160 mmHg	1.68 (1.48, 1.91)

Diastolic Blood Pressure	Hazard Ratio
< 80 mmHg	1.00 (Ref)
80-89 mmHg	1.23 (1.13, 1.34)
90-99 mmHg	1.36 (1.22, 1.51)
≥ 100 mmHg	1.77 (1.54, 2.05)

Ref: Reference group.

Table A.11: Results of pairwise comparisons for systolic and diastolic blood pressure on survival excluding participants with CVD at baseline

Comparison Systolic Blood Pressure Groups	Hazard Ratio	p
$\geq$ 160 mmHg vs 140-159 mmHg	1.17	0.0010
$\geq$ 160 mmHg vs 120-139 mmHg	1.62	<0.0001
$\geq$ 160 mmHg vs <120 mmHg	1.63	<0.0001
140-159 mmHg vs 120-139 mmHg	1.39	<0.0001
140-159 mmHg vs <120 mmHg	1.40	<0.0001
120-139 mmHg vs <120 mmHg	1.01	0.8605

Comparison Diastolic Blood Pressure Groups	Hazard Ratio	p
$\geq$ 100 mmHg vs 90-99 mmHg	1.53	<0.0001
$\geq$ 100 mmHg vs 80-89 mmHg	1.50	<0.0001
$\geq$ 100 mmHg vs <80 mmHg	1.75	<0.0001
90-99 mmHg vs 80-89 mmHg	0.98	0.6436
90-99 mmHg vs <80 mmHg	1.14	0.0025
80-89 mmHg vs <80 mmHg	1.17	<0.0001

Table A.12: Results of pairwise comparisons for systolic and diastolic blood pressure on MACE excluding participants with CVD at baseline

Comparison Systolic Blood Pressure Groups	Hazard Ratio	p
$\geq$ 160 mmHg vs 140-159 mmHg	1.11	0.1004
$\geq$ 160 mmHg vs 120-139 mmHg	1.61	<0.0001
$\geq$ 160 mmHg vs <120 mmHg	1.68	<0.0001
140-159 mmHg vs 120-139 mmHg	1.46	<0.0001
140-159 mmHg vs <120 mmHg	1.52	<0.0001
120-139 mmHg vs <120 mmHg	1.04	0.4340

Comparison Diastolic Blood Pressure Groups	Hazard Ratio	p
$\geq$ 100 mmHg vs 90-99 mmHg	1.31	0.0006
$\geq$ 100 mmHg vs 80-89 mmHg	1.44	<0.0001
$\geq$ 100 mmHg vs <80 mmHg	1.77	<0.0001
90-99 mmHg vs 80-89 mmHg	1.10	0.0567
90-99 mmHg vs <80 mmHg	1.36	<0.0001
80-89 mmHg vs <80 mmHg	1.23	<0.0001

Cumulative number of visits where physically active

These results pertain to MSCMs and MSMs for estimating the effect of the cumulative number of visits where physically active on SBP, DBP, mortality and MACE.

Physical Activity, Survival and MACE

The results indicate that an increase in the number of exams where the subject is active is associated with a reduced rate of mortality (HR for an increase of one exam where physically active = 0.76, 95% CI: 0.69 to 0.85). Risk of MACE was also significantly decreased for participants physically active at more exams (HR for an increase of one exam where physically active = 0.76, 95% CI: 0.67 to 0.86).

Physical Activity and Blood Pressure

There was a marginally significant decrease of 1.28 mmHg (95% CI, -2.55 to -0.01) in SBP associated to an increase of the number of exams where participants were physically active. No change in DBP was observed (0.07 mmHg; 95% CI, -0.60 to 0.75).

Crude Results

The results below are the crude results obtained using an unweighted version of the MSCMs and MSMs.

Physical Activity, Survival, and MACE

The crude results show that active individuals had a reduced rate of mortality (HR = 0.72, 95% CI: 0.70 to 0.78) compared to the inactive participants. Risk of MACE was also significantly decreased in the physically active participants compared to the inactive group (HR = 0.76, 95% CI: 0.69 to 0.84).

Physical Activity and Blood Pressure

There was a significant decrease of 2.76 mmHg (95% CI, -3.71 to -1.81) in SBP between physically active and inactive participants. No change in DBP was observed between the physically active group and the inactive group (-0.06 mmHg; 95% CI, -0.55 to 0.43).

Blood Pressure, Survival, and MACE

Table A.13: Crude results for risk of all-cause mortality according to systolic and diastolic blood pressure categories

Systolic Blood Pressure	Hazard Ratio
< 120 mmHg	1.00 (Ref)
120-139 mmHg	1.10 (1.01, 1.20)
140-159 mmHg	1.39 (1.27, 1.51)
≥ 160 mmHg	1.59 (1.44, 1.74)

Diastolic Blood Pressure	Hazard Ratio
< 80 mmHg	1.00 (Ref)
80-89 mmHg	1.06 (0.99, 1.14)
90-99 mmHg	1.14 (1.05, 1.24)
≥ 100 mmHg	1.54 (1.38, 1.71)

Ref: Reference group.

Table A.14: Crude results for risk of MACE according to systolic and diastolic blood pressure categories

Systolic Blood Pressure	Hazard Ratio
< 120 mmHg	1.00 (Ref)
120-139 mmHg	1.47 (1.32, 1.63)
140-159 mmHg	2.24 (2.01, 2.50)
≥ 160 mmHg	3.17 (2.82, 3.56)

Diastolic Blood Pressure	Hazard Ratio
< 80 mmHg	1.00 (Ref)
80-89 mmHg	1.30 (1.20, 1.42)
90-99 mmHg	1.61 (1.46, 1.77)
≥ 100 mmHg	2.49 (2.22, 2.80)

Ref: Reference group.

Table A.15: Crude results of pairwise comparisons for systolic and diastolic blood pressure on survival

Comparison Systolic Blood Pressure Groups	Hazard Ratio	p
≥ 160 mmHg vs 140-159 mmHg	1.14	0.0019
≥ 160 mmHg vs 120-139 mmHg	1.44	<0.0001
≥ 160 mmHg vs <120 mmHg	1.59	<0.0001
140-159 mmHg vs 120-139 mmHg	1.26	<0.0001
140-159 mmHg vs <120 mmHg	1.39	<0.0001
120-139 mmHg vs <120 mmHg	1.10	0.0258

Comparison Diastolic Blood Pressure Groups	Hazard Ratio	p
≥ 100 mmHg vs 90-99 mmHg	1.35	<0.0001
≥ 100 mmHg vs 80-89 mmHg	1.45	<0.0001
≥ 100 mmHg vs <80 mmHg	1.54	<0.0001
90-99 mmHg vs 80-89 mmHg	1.07	0.1052
90-99 mmHg vs <80 mmHg	1.14	0.0018
80-89 mmHg vs <80 mmHg	1.06	0.0818

Table A.16: Crude results of pairwise comparisons for systolic and diastolic blood pressure on MACE

Comparison Systolic Blood Pressure Groups	Hazard Ratio	p
$\geq$ 160 mmHg vs 140-159 mmHg	1.41	<0.0001
$\geq$ 160 mmHg vs 120-139 mmHg	2.16	<0.0001
$\geq$ 160 mmHg vs <120 mmHg	3.17	<0.0001
140-159 mmHg vs 120-139 mmHg	1.53	<0.0001
140-159 mmHg vs <120 mmHg	2.24	<0.0001
120-139 mmHg vs <120 mmHg	1.47	<0.0001

Comparison Diastolic Blood Pressure Groups	Hazard Ratio	p
$\geq$ 100 mmHg vs 90-99 mmHg	1.55	<0.0001
$\geq$ 100 mmHg vs 80-89 mmHg	1.92	<0.0001
$\geq$ 100 mmHg vs <80 mmHg	2.49	<0.0001
90-99 mmHg vs 80-89 mmHg	1.24	<0.0001
90-99 mmHg vs <80 mmHg	1.61	<0.0001
80-89 mmHg vs <80 mmHg	1.30	<0.0001

BIBLIOGRAPHIE

Agresti, A. (2013). *Categorical Data Analysis, 3rd edition.* New Jersey : John Wiley & Sons.

Apostol, T. M. (1974). Mathematical analysis, 2nd edition.

Baba, K., Shibata, R. & Sibuya, M. (2004). Partial correlation and conditional correlation as measures of conditional independence. *Australian & New Zealand Journal of Statistics, 46*(4), 657–664.

Bakoyannis, G. & Touloumi, G. (2012). Practical methods for competing risks data : A review. *Statistical Methods in Medical Research, 21*(3), 257–272.

Barnes, J., Behrens, T. K., Benden, M. E., Biddle, S., Bond, D., Brassard, P., Brown, H., Carr, L., Chaput, J.-P., Christian, H., Colley, R., Duggan, M., Dunstan, D., Ekelund, U., Esliger, D., Ferraro, Z., Freedhoff, Y., Galaviz, K., Gardiner, P., Goldfield, G., Haskell, W. L., Liguori, G., Healy, G., Herman, K. M., Hinckson, E., Larouche, R., Leblanc, A., Levine, J., Maeda, H., McCall, M., McCubbin, W., McGuire, A., Onywera, V., Owen, N., Peterson, M., Prince, S., Ramirez, E., Ridgers, N., Routen, A., Rowlands, A., Saunders, T., Schuna, J. M., Sherar, L., Spruijt-Metz, D., Taylor, B., Tremblay, M., Tuckler, J., Wijndaele, K., Wilson, J., Wilson, J. & Woodruff, S. (2012). Letter to the editor : Standardized use of the terms "sedentary" and "sedentary behaviours". *Applied Physiology Nutrition and Metabolism-Physiologie Appliquée Nutrition et Métabolisme, 37*(3), 540–542.

Beunza, J. J., Martínez-González, M. Á., Ebrahim, S., Bes-Rastrollo, M., Núnez, J., Martínez, J. A. & Alonso, Á. (2007). Sedentary behaviors and the risk of incident hypertension : The SUN cohort. *American Journal of Hypertension, 20*(11), 1156–1162.

Bollen, K. A. & Stine, R. A. (1992). Bootstrapping goodness-of-fit measures in structural equation models. *Sociological Methods & Research, 21*(2), 205–229.

Brookhart, M. A. & van der Laan, M. J. (2006). A semiparametric model selection criterion with applications to the marginal structural model. *Computational Statistics & Data Analysis, 50*(2), 475–498.

Brumback, B. A., Hernán, M. A., Haneuse, S. J. H. & Robins, J. M. (2004). Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. *Statistics in Medicine, 23*(5), 749–767.

Carlin, B. P. & Gelfand, A. E. (1990). Approaches for empirical Bayes confidence intervals. *Journal of the American Statistical Association, 85*(409), 105–114.

Carlin, B. P. & Louis, T. A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis, 2nd edition*. Boca Raton : Chapman & Hall / CRC.

Cefalu, M., Dominici, F. & Parmigiani, G. (2013). Model averaged double robust estimation. *Harvard University Biostatistics Working Paper Series*. Working Paper 149.

Chai, W., Nigg, C. R., Pagano, I. S., Motl, R. W., Horwath, C. & Dishman, R. K. (2010). Associations of quality of life with physical activity, fruit and vegetable consumption, and physical inactivity in a free living, multiethnic population in Hawaii : a longitudinal study. *International Journal of Behavioral Nutrition and Physical Activity, 7*, 83.

Chobanian, A. V., Bakris, G. L., Black, H. R., Cushman, W. C., Green, L. A., Izzo, J. L., Jones, D. W., Materson, B. J., Oparil, S., Wright, J. T. & Roccela, E. J. (2003). Seventh report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure. *Hypertension, 42*(6), 1206–1252.

Clyde, M. (2000). Model uncertainty and health effect studies for particulate matter. *Environmetrics, 11*(6), 745–763.

Clyde, M. (2003). *Subjective and Objective Bayesian Statistics, 2nd edition*. New Jersey : S. James Press : Wiley-Interscience.

Cole, S. R. & Hernán, M. A. (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology, 168*(6), 656–664.

Cole, S. R., Hernán, M. A., Robins, J. M., Anastos, K., Chmiel, J., Detels, R., Ervin, C., Feldman, J., Greenblatt, R., Kingsley, L., Lai, S., Young, M., Cohen, M. & Muñoz, A. (2003). Effect of highly active antiretroviral therapy on time to acquired immunodeficiency syndrome or death using marginal structural models. *American Journal of Epidemiology, 158*(7), 687–694.

Colley, R. C., Garriguet, D., Janssen, I., Craig, C. L., Clarke, J. & Tremblay, M. S. (2011). Physical activity of canadian adults : accelerometer results from the 2007 to 2009 canadian health measures survey. *Statistics Canada Health Reports, 22*(1).

Cornelissen, V. A. & Smart, N. A. (2013). Exercise training for blood pressure : a systematic review and meta-analysis. *Journal of the American Heart Association*.

Crainiceanu, C. M., Dominici, F. & Parmigiani, G. (2008). Adjustment uncertainty in effect estimation. *Biometrika, 95*(3), 635–651.

Dawid, A. P. (1970). On the limiting normality of posterior distributions. *Proceedings of the Cambridge Philosophical Society, 67*, 625–633.

de Luna, X., Waernbaum, I. & Richardson, T. S. (2011). Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika*, 1–15.

Dunn, A. L., Trivedi, M. H. & O'Neal, H. A. (2001). Physical activity dose-response effects on outcomes of depression and anxiety. *Medicine & Science in Sports & Exercise, 33*(6 Suppl), S587–597.

Fairall, L. R., Bachmann, M. O., Louwagie, G. M., van Vuuren, C., Chikobvu, P., Steyn, D., Staniland, G. H., Timmerman, V., Msimanga, M., Seebregts, C. J., Boulle, A., Nhiwatiwa, R., Bateman, E. D., Zwarenstein, M. F. & Chapman, R. D. (2008). Effectiveness of antiretroviral treatment in a south african program : a cohort study. *Archives of Internal Medicine, 168*(1), 86–93.

Garber, C. E., Blissmer, B., Deschenes, M. R., Franklin, B., Lamonte, M. J., Lee, I.-M., Nieman, D. C. & Swain, D. P. (2011). American college of sports medicine position stand. Quantity and quality of exercise for developing and maintaining cardiorespiratory, musculoskeletal, and neuromotor fitness in apparently healthy adults : guidance for prescribing exercise. *Medicine & Science in Sports & Exercise, 43*(7), 1334–1359.

Gidding, S. S., Barton, B. A., Dorgan, J. A., Kimm, S. Y., Kwiterovich, P. O., Lasser, N. L., Robson, A. M., Stevens, V. J., Van Horn, L. & Simons-Morton, D. G. (2006). Higher self-reported physical activity is associated with lower systolic blood pressure : the dietary intervention study in childhood (disc). *Pediatrics, 118*(6), 2388–2393.

Glynn, R. J., Field, T. S., Hebert, P., Taylor, J., Hennekens, C., Rosner, B., Rosner, B. & Hennekens, C. (1995). Evidence for a positive linear relation between blood pressure and mortality in elderly people. *Lancet, 345*(8953), 825–829.

Graham, J. W., Olchowski, A. E. & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science, 8*(3), 206–213.

Green, D. J., O'Driscoll, G., Joyner, M. J. & Cable, N. T. (2008). Exercise and cardiovascular risk reduction : time to update the rationale for exercise? *Journal of Applied Physiology, 105*(2), 766–768.

Greenland, S., Robins, J. M. & Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical Science, 14*(1), 29–46.

Gruber, S. & van der Laan, M. J. (2009). Targeted maximum likelihood estimation : A gentle introduction. Working Paper 252.

Guay, F., Larose, S., Ratelle, C., Sénécal, C., Vallerand, R. J. & Vitaro, F. (2011). *Mes amis, mes parents et mes professeurs : Une analyses comparée de leurs effets respectifs sur la motivation, la réussite, l'orientation et la persévérance scolaires.* Rapport technique 2007-PE-118485, Fonds de recherche du Québec - Société et culture, Québec. Available online : `http://www.frqsc.gouv.qc.ca/upload/editeur/Rapport_sans_annexes.pdf`.

Hackam, D. G., Quinn, R. R., Ravani, P., Rabi, D. M., Dasgupta, K., Daskalopoulou, S. S., Khan, N. A., Herman, R. J., Bacon, S. L., Cloutier, L., Dawes, M., Rabkin, S. W., Gilbert, R. E., Ruzicka, M., McKay, D. W., Campbell, T. S., Grover, S., Honos, G., Schiffrin, E. L., Bolli, P. & Wilson, T. W. (2013). The 2013 Canadian hypertension education program recommendations for blood pressure measurement, diagnosis, assessment of risk, prevention, and treatment of hypertension. *Canadian Journal of Cardiology, 29*(5), 528–542.

Hambrecht, R., Adams, V., Erbs, S., Linke, A., Kränkel, N., Shu, Y., Baither, Y., Gielen, S., Thiele, H., Gummert, J., Mohr, F. & Schuler, G. (2003). Regular physical activity improves endothelial function in patients with coronary artery disease by increasing phosphorylation of endothelial nitric oxide synthase. *Circulation, 107*, 3152–3158.

Haughton, D. M. A. (1988). On the choice of a model to fit data from an exponential family. *The Annals of Statistics, 16*(1), 342–355.

Heesch, K. C., van Uffelen, J. G., van Gellecum, Y. R. & Brown, W. J. (2012). Dose-response relationships between physical activity, walking and health-related quality of life in mid-age and older women. *Journal of Epidemiology and Community Health, 66*(8), 670–677.

Hernán, M. A., Brumback, B. & Robins, J. M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men. *Epidemiology, 11*(5), 561–570.

Hernán, M. A., Brumback, B. & Robins, J. M. (2001). Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *Journal of the American Statistical Association, 96*(454), 440–448.

Hernán, M. A. & Brumback, Babette A and, J. M. (2002). Estimating the causal effect of zidovudine on CD4 count with a marginal structural model for repeated measures. *Statistics in Medicine, 21*(12), 1689–1709.

Hernán, M. A., Hernández-Díaz, S., Werler, M. M. & Mitchell, A. A. (2002). Causal knowledge as a prerequisite for confounding evaluation : an application to birth defects epidemiology. *American Journal of Epidemiology, 155*(2), 176–184.

Hernán, M. A. & Robins, J. M. (2015). *Causal Inference*. Chapman & Hall/CRC. Version préliminaire du 24 novembre 2014.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics, 20*(1), 217–240.

Hoeting, J. A., Madigan, D., Raftery, A. E. & Volinsky, C. T. (1999). Bayesian model averaging : A tutorial. *Statistical Science, 14*(4), 382–417.

Højsgaard, S., Halekoh, U. & Yan, J. (2006). The R package geepack for generalized estimating equations. *Journal of Statistical Software, 2*(15), 1–11.

Holmes, M. D., Chen, W. Y., Li, L., Hertzmark, E., Spiegelman, D. & Hankinson, S. E. (2010). Aspirin intake and survival after breast cancer. *Journal of Clinical Oncology, 28*(9), 1467–1472.

Hu, G., Barengo, N. C., Tuomilehto, J., Lakka, T. A., Nissinen, A. & Jousilahti, P. (2004). Relationship of physical activity and body mass index to the risk of hypertension : a prospective study in Finland. *Hypertension, 43*, 25–30.

Jakes, R., Day, N., Khaw, K.-T., Luben, R., Oakes, S., Welch, A., Bingham, S. & Wareham, N. (2003). Television viewing and low participation in vigorous recreation are independently associated with obesity and markers of cardiovascular disease risk : EPIC-Norfolk population-based study. *European Journal of Clinical Nutrition, 57*, 1089–1096.

Kagan, A., Harris, B. R., Winkelstein Jr, W., Johnson, K. G., Kato, H., Syme, S. L., Rhoads, G. G., Gay, M. L., Nichaman, M. Z., Hamilton, H. B. & Tillotson, J. (1974). Epidemiologic studies of coronary heart disease and stroke in japanese men living in Japan, Hawaii and California : demographic, physical, dietary and biochemical characteristics. *Journal of Chronic Diseases, 27*(7), 345–364.

Kom, E. L., Graubard, B. I. & Midthune, D. (1997). Time-to-event analysis of longitudinal follow-up of a survey : choice of the time-scale. *American Journal of Epidemiology, 145*(1), 72–80.

Koop, G. & Tole, L. (2004). Measuring the health effects of air pollution : to what extent can we really say that people are dying from bad air ? *Journal of Environmental Economics and Management, 47*(1), 30–54.

Kujala, U. M., Kaprio, J., Sarna, S. & Koskenvuo, M. (1998). Relationship of leisure-time physical activity and mortality : the finnish twin cohort. *Journal of the American Medical Association, 279*(6), 440–444.

Laird, N. M. & Louis, T. A. (1987). Empirical Bayes confidence intervals based on bootstrap samples. *Journal of the American Statistical Association, 82*(399), 739–750.

Lee, I.-M. & Skerrett, P. J. (2001). Physical activity and all-cause mortality : what is the dose-response relation ? *Medicine & Science in Sports & Exercise, 33*(6 Suppl), S459–S471.

Lefebvre, G., Atherton, J. & Talbot, D. (2014a). The effect of the prior distribution in the bayesian adjustment for confounding algorithm. *Computational Statistics & Data Analysis, 70*, 227–240.

Lefebvre, G., Delaney, J. A. & McClelland, R. L. (2014b). Extending the bayesian adjustment for confounding algorithm to binary treatment covariates to estimate the effect of smoking on carotid intima-media thickness : the Multi-Ethnic Study of Atherosclerosis. *Statistics in Medicine, 33*(16), 2797–2813.

Little, R. J. & Rubin, D. B. (2002). *Statistical analysis with missing data*. Hoboken, NJ : Hoboken, NJ : J Wiley & Sons.

Lumley, T. (2009). leaps : regression subset selection (package R). Using Fortran code by Alan Miller. Récupéré de `http://CRAN.R-project.org/package=leaps`

Madigan, D. & Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, *89*(428), 1535–1546.

Madigan, D., York, J. & Allard, D. (1995). Bayesian graphical models for discrete data. *International Statistical Review*, *63*(2), 215–232.

McCandless, L. C., Gustafson, P. & Austin, P. C. (2009). Bayesian propensity score analysis for observational data. *Statistics in Medicine*, *28*(1), 94–112.

Montoye, H. J., Metzner, H. L., Keller, J. B., Johnson, B. C. & Epstein, F. H. (1972). Habitual physical activity and blood pressure. *Medicine & Science in Sports & Exercise*, *4*(4), 175–181.

Moodie, E. E., Delaney, J. A., Lefebvre, G. & Platt, R. W. (2008). Missing confounding data in marginal structural models : a comparison of inverse probability weighting and multiple imputation. *The International Journal of Biostatistics*, *4*(1).

Moore, S. C., Patel, A. V., Matthews, C. E., de Gonzalez, A. B., Park, Y., Katki, H. A., Linet, M. S., Weiderpass, E., Visvanathan, K., Helzlsouer, K. J., Thun, M., Gapstur, S. M., Hartge, P. & Lee, I.-M. (2012). Leisure time physical activity of moderate to vigorous intensity and mortality : a large pooled cohort analysis. *PLoS medicine*, *9*(11).

Morgan, S. L. & Winship, C. (2007). *Counterfactuals and causal inference : Methods and principles for social research*. New York : Cambridge University Press.

Myers, M. G. & Godwin, M. (2007). Automated measurement of blood pressure in routine clinical practice. *The Journal of Clinical Hypertension*, *9*(4), 267–270.

Myers, M. G., Godwin, M., Dawes, M., Kiss, A., Tobe, S. W., Grant, F. C. & Kaczorowski, J. (2011). Conventional versus automated measurement of blood pressure in primary care patients with systolic hypertension : randomised parallel design controlled trial. *BMJ*, *342*.

Neyman, J. (1923). Sur les applications de la théorie des probabilités aux expériences agricoles : Essai des principes, mémoire de maîtrise. Réédité en anglais dans la revue Statistical Science, Vol. 5, pp. 463-472 (traduction de D. M. Dabrowska, et T. P. Speed).

Nybo, L., Sundstrup, E., Jakobsen, M. D., Mohr, M., Hornstrup, T., Simonsen, L., Bülow, J., Randers, M. B., Nielsen, J. J., Aagaard, P. & Krustrup, P. (2010). High-intensity training versus traditional exercise interventions for promoting health. *Medecine & Science in Sports & Exercise*, *42*(10), 1951–8.

O'brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity, 41*(5), 673–690.

Paffenbarger Jr, R. S. & Lee, I.-M. (1997). Intensity of physical activity related to incidence of hypertension and all-cause mortality : an epidemiological view. *Blood Pressure Monitoring, 2*(3), 115–123.

Patel, K., Hernán, M. A., Williams, P. L., Seeger, J. D., McIntosh, K., Van Dyke, R. B. & Seage III, G. R. (2008). Long-term effects of highly active antiretroviral therapy on CD4+ cell evolution among children and adolescents infected with HIV : 5 years and counting. *Clinical Infectious Diseases, 46*(11), 1751–1760.

Pearl, J. (2009). *Causality : Models, Reasoning, and Inference. 2nd Edition.* New York : Cambridge University Press.

Persson, E., Häggström, J., Waernbaum, I. & de Luna, X. (2013). Data-driven algorithms for dimension reduction in causal inference : analyzing the effect of school achievements on acute complications of type 1 diabetes mellitus. *arXiv preprint arXiv :1309.4054.*

Platt, R. W., Brookhart, M. A., Cole, S. R., Westreich, D. & Schisterman, E. F. (2013). An information criterion for marginal structural models. *Statistics in Medicine, 32*(8), 1383–1393.

Prospective Studies Collaboration (2002). Age-specific relevance of usual blood pressure to vascular mortality : a meta-analysis of individual data for one million adults in 61 prospective studies. *Lancet, 360*(9349), 1903–1913.

R Core Team (2014). *R : A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Raftery, A. E., Madigan, D. & Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association, 92*(437), 179–191.

Raftery, A. E. & Zheng, Y. (2003). Discussion : Performance of bayesian model averaging. *Journal of the American Statistical Association, 98*(464), 931–938.

Reaven, P. D., Barrett-Connor, E. & Edelstein, S. (1991). Relation between leisure-time physical activity and blood pressure in older women. *Circulation, 83*(2), 559–565.

Rikli, R. (2000). Reliability, validity, and methodological issues in assessing physical activity in older adults. *Research Quarterly for Exercise and Sport, 71*(2 Suppl), S89–96.

Robins, J. M. (1997). Marginal structural models. *Proceedings of the Section on Bayesian Statistical Science.* American Statistical Association : Alexandria, VA, 1998 ; 1-10.

Robins, J. M. (2000). Marginal structural models versus structural nested models as tools for causal inference. In *Statistical models in epidemiology, the environment, and clinical trials* 95–133. Springer.

Robins, J. M. & Hernán, M. A. (2009). *Estimation of the causal effects of time-varying exposure. In : Fitzmaurice Gm, Daividian M, Verbeke G, Molenbergs G. ed. Longitudinal data analysis.* Boca Raton : CRC Press. 553-599.

Robins, J. M., Hernán, M. A. & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology, 11*(5), 550–560.

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

Rosseel, Y. (2012). lavaan : An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. Récupéré de `http://www.jstatsoft.org/v48/i02/`

Rossi, A. M., Dikareva, A., Bacon, S. L. & Daskalopoulou, S. S. (2012). The impact of physical activity on mortality in patients with high blood pressure : a systematic review. *Journal of Hypertension, 30*(7), 1277–1288.

Rossi, A. M., Moullec, G., Lavoie, K. L., Gour-Provençal, G. & Bacon, S. L. (2013). The evolution of a Canadian hypertension education program recommendation : the impact of resistance training on resting blood pressure in adults as an example. *Canadian Journal of Cardiology, 29*(5), 622–627.

Rossi, A. M., Talbot, D., Lefebvre, G., Atherton, J. & Bacon, S. L. (2015). Marginal structural models for estimating the causal relationships between physical activity, blood pressure, and mortality in a longitudinal cohort study : the honolulu heart program. *British Journal of Sports Medicine.* Submitted.

Rothman, K. J. & Greenland, S. (2005). Causation and causal inference in epidemiology. *American Journal of Public Health, 95*(S1).

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688–701.

Sampson, R. J., Laub, J. H. & Wimer, C. (2006). Does marriage reduce crime? a counterfactual approach to within-individual causal effects. *Criminology, 44*(3), 465–508.

SAS Institute Inc. (2011). *SAS/STAT Software, Version 9.2.* Cary, NC

Schildcrout, J. S., Haneuse, S., Peterson, J. F., Denny, J. C., Matheny, M. E., Waitman, L. R. & Miller, R. A. (2011). Analyses of longitudinal, hospital clinical laboratory data with application to blood glucose concentrations. *Statistics in Medicine, 30*(27), 3208–3220.

Struthers, C. A. & Kalbfleisch, J. D. (1986). Misspecified proportional hazard models. *Biometrika, 73*(2), 363–369.

Stuart, E. A. (2010). Matching methods for causal inference : A review and a look forward. *Statistical Science, 25*(1), 1–21.

Talbot, D., Lefebvre, G. & Atherton (2014). Sensitivity of the Bayesian adjustment for confounding (BAC) algorithm to the omega value (package R).

Talbot, D., Rossi, A. M., Bacon, S. L., Atherton, J. & Lefebvre, G. (2015). A cautionary note concerning the use of stabilized weights in marginal structural models. *Statisticins in Medicine*, *34*(5), 812–823.

Tchetgen Tchetgen, E. J., Glymour, M. M., Weuve, J. & Robins, J. (2012a). A cautionary note on specification of the correlation structure in inverse-probability-weighted estimation for repeated measures. *Harvard University Biostatistics Working Paper Series*. Working Paper 140.

Tchetgen Tchetgen, E. J., Glymour, M. M., Weuve, J. & Robins, J. M. (2012b). Specifying the correlation structure in inverse-probability-weighting estimation for repeated measures. *Epidemiology*, *23*(4), 644–646.

Thiébaut, A. C. M. & Bénichou, J. (2004). Choice of time-scale in cox's model analysis of epidemiologic cohort data : a simulation study. *Statistics in Medicine*, *23*(24), 3803–3820.

Tremblay, A., Doucet, E. & Imbeault, P. (1999). Physical activity and weight maintenance. *International Journal of Obesity and Related Metabolic Disorders : Journal of the International Association for the Study of Obesity*, *23*(Suppl 3), S50–4.

Tremblay, M. S., Colley, R. C., Saunders, T. J., Healy, G. N. & Owen, N. (2010). Physiological and health implications of a sedentary lifestyle. *Applied Physiology, Nutrition, and Metabolism*, *35*(6), 725–740.

van der Laan, M. J. & Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics*, *2*(1).

VanderWeele, T. J., Hawkley, L. C. & Cacioppo, J. T. (2012). On the reciprocal association between loneliness and subjective well-being. *American Journal of Epidemiology*, *176*(9), 777–784.

VanderWeele, T. J., Hawkley, L. C., Thisted, R. A. & Cacioppo, J. T. (2011). A marginal structural model analysis for loneliness : implications for intervention trials and clinical practice. *Journal of Consulting and Clinical Psychology*, *79*(2), 225–235.

VanderWeele, T. J. & Shpitser, I. (2011). A new criterion for confounder selection. *Biometrics*, *67*(4), 1406–1413.

Vansteelandt, S. (2012). Discussions. *Biometrics*. [doi : 10.1111/j.1541-0420.2011.01734.x].

Vatten, L. J., Nilsen, T. I. & Holmen, J. (2006). Combined effect of blood pressure and physical activity on cardiovascular mortality. *Journal of Hypertension*, *24*(10), 1939–1946.

Walker, A. M. (1969). On the asymptotic behavior of posterior distributions. *Journal of the Royal Statistical Society : Series B*, *31*(1), 80–88.

Wang, C., Parmigiani, G. & Dominici, F. (2012a). Bayesian effect estimation accounting for adjustment uncertainty. *Biometrics*, *68*(3), 661–671.

Wang, C., Parmigiani, G. & Dominici, F. (2012b). Rejoinder : Bayesian effect estimation accounting for adjustment uncertainty. *Biometrics*, *68*(3), 680–686.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, *44*(1), 92–107.

Wen, C. P., Wai, J. P. M., Tsai, M. K., Yang, Y. C., Cheng, T. Y. D., Lee, M.-C., Chan, H. T., Tsao, C. K., Tsai, S. P. & Wu, X. (2011). Minimum amount of physical activity for reduced mortality and extended life expectancy : a prospective cohort study. *Lancet*, *378*(9798), 1244–1253.

Wijndaele, K., Brage, S., Besson, H., Khaw, K.-T., Sharp, S. J., Luben, R., Wareham, N. J. & Ekelund, U. (2010). Television viewing time independently predicts all-cause and cardiovascular mortality : the EPIC Norfolk study. *International Journal of Epidemiology*.

Worth, R. M. & Kagan, A. (1970). Ascertainment of men of Japanese ancestry in Hawaii through World War II selective service registration. *Journal of Chronic Diseases*, *23*(5-6), 389–397.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, *20*(7), 557–585.

Xiao, Y., Abrahamowicz, M. & Moodie, E. E. (2010). Accuracy of conventional and marginal structural cox model estimators : a simulation study. *The International Journal of Biostatistics*, *6*(2).

Yan, J. (2002). geepack : yet another package for generalized estimating equations. *R News*, *2*(3), 12–14.

Yan, J. & Fine, J. (2004). Estimating equations for association structures. *Statistics in Medicine*, *23*(6), 859–874.

Yang, W. & Joffe, M. M. (2012). Subtle issues in model specification and estimation of marginal structural models. *Pharmacoepidemiology and Drug Safety*, *21*(3), 241–245.

Yuan, K.-H. & Chan, W. (2011). Biases and standard errors of standardized regression coefficients. *Psychometrika*, *76*(4), 670–690.

Zhao, G., Li, C., Ford, E. S., Fulton, J. E., Carlson, S. A., Okoro, C. A., Wen, X. J. & Balluz, L. S. (2013). Leisure-time aerobic physical activity, muscle-strengthening activity and mortality risks among US adults : the NHANES linked mortality study. *British Journal of Sports Medicine*.

Zigler, C. M. & Dominici, F. (2014). Uncertainty in propensity score estimation : Bayesian methods for variable selection and model-averaged causal effects. *Journal of the American Statistical Association, 109*(505), 95–107.