Méthodes statistiques pour l'inférence causale dans les études de cohortes en présence de données longitudinales : applications au vieillissement

Kateline Le Bourdonnec

université de BORDEAUX

THÈSE PRÉSENTÉE

POUR OBTENIR LE GRADE DE

DOCTEURE
DE L'UNIVERSITÉ DE BORDEAUX

École Doctorale Sociétés, Politique, Santé Publique

Spécialité Santé Publique, option Biostatistique

Par **Kateline Le Bourdonnec**

Méthodes statistiques pour l'inférence causale dans les études de cohortes en présence de données longitudinales : applications au vieillissement

Sous la direction de : **Cécile Proust-Lima**

Soutenue le 30 Avril 2024

Membres du jury :

Mme Karen LEFFONDRÉ	Professeure des universités	Bordeaux	Présidente
Mme Agathe GUILLOUX	Professeure des universités	Paris	Rapportrice
M Vivian VIALLON	Maître de conférence	Lyon	Rapporteur
Mme Nicola COLLEY	Directrice de recherche	Toulouse	Examinatrice
Mme Cécile PROUST-LIMA	Directrice de recherche	Bordeaux	Directrice de thèse

À l'ange que tu es,

Remerciements

Parce qu'il y a toujours quelqu'un à remercier : un grand merci à chacun d'entre vous, ainsi qu'à toi lecteur.

Dire que ma thèse a été *un long fleuve tranquille* serait faux. Je l'ai adoré autant que je l'ai detesté, et pourtant, si je devais recommencer, je le referais - avec la même directrice de thèse, la même équipe et les mêmes collègues qui pour beaucoup sont maintenant des amis et voire même davantage pour une personne !

À ma directrice de thèse, Cécile Proust-Lima,

Cécile, cette thèse n'aurait jamais pu être ce qu'elle est aujourd'hui sans toi. Lorsque je me suis lancée dans cette aventure, tu m'as dit que le choix du directeur de thèse était très important (un peu comme un mariage) - je confirme, il l'était et quel plaisir d'avoir pu évoluer à tes côtés. J'ai certes parfois baissé les bras, que ce soit pour des raisons personnelles ou à cause des codes, des simulations, (ou des reviewers) qui me prenaient la tête durant des jours et des mois, mais tu as toujours su trouver les mots justes. Au-delà de tes qualités humaines qui m'ont été d'une aide précieuse durant cette thèse, j'ai eu la chance de bénéficier de ton immense expertise en statistique (et ce n'est pas pour te flatter). Je t'en suis très reconnaissante. Quel bonheur de pouvoir continuer un peu mon chemin à tes côtés - j'ai encore beaucoup à apprendre de toi. Je te dois énormément.

Aux membres du jury,

À Karen Leffondré, merci d'avoir accepté d'être présidente du jury. Il y a six ans tu corrigeais mes copies, il y a deux ans j'enseignais à tes côtés et aujourd'hui, quelle chance de t'avoir de nouveau pour évaluer mon travail. À Vivian Viallon et Agathe Guilloux, merci

d'avoir accepté d'être les rapporteurs et d'avoir pris le temps de lire mon travail. À Nicola Colley d'avoir accepté d'examiner cette thèse.

À Linda Valeri,

Linda, quelques mots en Français pour te remercier infiniment d'avoir accepté de collobarer avec moi sur les analyses de médiation. Merci également de m'avoir accueilli pendant trois mois à Columbia University à New-York. J'espère sincérement te revoir en France ou à New-York un jour !

À l'équipe Biostatistique,

Mes connaissances en causalité me permettent aujourd'hui d'affirmer que la joie et l'envie de me rendre au travail sont la conséquence d'une équipe exceptionnelle. Que serait ma thèse sans les visites de **Pierre** pour parfois simplement regarder par la fenêtre, les "coucou" toujours enthousiastes d'**Antoine**, reconnaissables à sa façon de frapper à la porte ou bien les jeux de mots et la bonne humeur contagieuse d'**Anaïs**.

À mes collègues de bureau : *Le placard à balais*

Au fond d'un couloir, dans une ancienne cage d'escalier, à côté d'une fenêtre souvent confondue avec une cabine téléphonique, sans que les gens ne se doutent qu'on puisse entendre toutes leurs conversations, se niche un bureau où se déroulent : des discussions scientifiques, une ambiance studieuse (à beaucoup d'exception près), des pauses, de la nourriture, du thé remplacé par le café en fin de thèse, des recommandations musicales mais pas que..., et surtout des amis ! **Tiphaine**, six ans que l'on partage le même bureau, ça commence en tant que collègue stagiaire avec des soirées Netflix-UberEat puis ça fini en tant que copine à partir en voyage ensemble. Merci pour tous ces moments et surtout merci d'avoir toujours été présente. **Ariane** *"les recommandations musicales mais pas que..."* te sont belles et bien destinées, merci pour ta folie, ta joie et ta bonne humeur ! Merci également de nous avoir appris à être des bébés artistes et merci pour ces moments à "cabaner" : hâte d'y retourner. **Lisa**, notre "maman" du bureau. Merci pour tes conseils avisés en médécine ou sur la vie quotidienne en jouant le cupidon : *méthode offensive.*

Maintenant c'est à moi de prendre soin de "vous". **Anthony**, même si aujourd'hui tu ne partages plus ce bureau avec nous, tu as fait la majorité de cette thèse avec moi. Merci d'avoir pu partager avec moi la vraie *galère* de la thèse. **Léa**, qu'est-ce-que je me sens vieille et sur le téco avec toi... mais merci d'insuffler ta jeunesse et tes expressions dans ce bureau.

Aux doctorant.e.s des autres bureaux : *Les trois petits cochons*

À mes copines de thèse : **Manel**, une vraie Parisienne à Bordeaux... Merci pour tes conseils (même si je n'ai pas toujours été prête à les accepter). Bien que tu sembles être plus stressée que moi pour ma propre thèse, n'oublie pas que la perfection est inatteignable. Ne te brûle pas les ailes à la chercher. **Léonie**, plus qu'une aventure de thèse, j'ai eu la chance de partager avec toi plusieurs défis sportifs. La thèse finalement c'est comme le sport, ça se joue au mental. **Sara**, merci pour les sorties extra-thèse qui se font un peu rares maintenant mais qui vont pouvoir reprendre, je l'espère, une fois cette thèse aboutie.

À **Fédé, Romain, Valentine, Blandine, Justine** (*mais où est Justine ?*), merci pour les conversations que ce soit stat, surf ou bébé. **Coco**, finalement d'élève à collègue il n'y a qu'un pas. Il y a fort longtemps, tu as été la première thèse à laquelle j'ai assisté, alors oui on peut dire que tu es un peu un modèle pour moi ! *(j'ai dit "un peu").* **Marius**, le petit stagiaire est devenu grand, tu fais partie de la relève maintenant. Merci de m'avoir fait bien rire, parfois malgré toi !

À mes copines de Bretagne (et pas que),

Amandine, Anaïs, Bérengère, Cindy, Gwendoline, Juliette B., Manon H.,, d'années en années, de nouvel an en nouvel an, de visio en visio, vous êtes là. J'ai vraiment beaucoup de chance d'avoir des amies comme vous. **Manon B.,** mon binôme en Master 1, et une amie aujourd'hui. Même si on a toutes les deux *jamais le temps*, je sais que tu es là pour moi et je le suis également pour toi ! **Juliette M.,** qu'est ce que l'on a pu en vivre des moments ensembles autour d'une bouteille de vin (ou plusieurs), que ce soit pour refaire le monde ou juste parce que l'on en avait besoin. Merci pour ton soutien depuis de nombreuses années.

Le meilleur pour la fin : Un grand merci à **toute ma famille**. Vous m'avez fait vivre une fin de thèse très mouvementée, riche d'imprévus et de rebondissements, telle un roman remplis de péripéties. Dans ce roman, quoi qu'il arrive l'amour triomphe toujours et nous rend toujours plus fort. **Maman, Papa**, même si je pense que vous n'avez pas toujours compris ce que je faisais concrétement, vous avez toujours eu la patience de m'écouter, que ce soit pour faire des présentations, relire des articles, discuter médécine avec Maman ou bien plus important être là pour moi qu'importe le moment et qu'importe les décisions que je souhaitais prendre, merci. **À mes frères et ma soeur**, qui de près ou de loin sont présents dans mon coeur. Parce que l'on est quatre et qu'on le restera 🕊! **À ma nièce Gabrielle**, mon petit ange, je souhaite de tout cœur que tu puisses t'épanouir, que tu exerces plus tard un métier qui te plaise (même si ce n'est pas *biostatisticienne*), et sache que je ferai tout pour y contribuer ! **À mes grand-parents**, peut-être est-ce un signe de soutenir juste avant le 1er Mai. **À ma cousine Marine**, merci de m'avoir fait courir, pédaler et nager pendant cette thèse pour un combat commun contre la mucoviscidose. J'ai hâte de faire un nouveau défi avec toi !

Enfin, merci à toi, **Louis**. Merci d'être la personne que tu es, celui avec qui je souhaite partager ma vie et construire mon futur.

Valorisation scientifique

Publications

Le Bourdonnec K, Samieri C, Tzourio C, Mura T, Mishra A, Trégouët DA, Proust-Lima C. Addressing unmeasured confounders in cohort studies : Instrumental variable method for a time-fixed exposure on an outcome trajectory. Biom J. 2024 Jan ;66(1) :e2200358. doi : 10.1002/bimj.202200358. Epub 2023 Dec 14. PMID : 38098309.

Le Bourdonnec K, Valeri L, Proust-Lima C. Continuous-time mediation analysis for repeatedly measured mediators and outcomes. Available on arXiv *Submitted in Biometrics*

Présentations orales en conférence

Le Bourdonnec K et al. Prise en compte de la confusion non-observée dans les études de cohorte : méthode par variables instrumentales pour exposition transversale et marqueur répété. Journée de l'école doctorale - May, 2021. Bordeaux, France

Le Bourdonnec K et al. Prise en compte de la confusion non-observée dans les études de cohorte : méthode par variables instrumentales pour exposition transversale et marqueur répété. EPICLIN 2021 - Junes, 2021. Marseille, France

Le Bourdonnec K et al. Addressing unmeasured confounders in cohort studies : Instrumental variable method for a time-fixed exposure on an outcome trajectory. Journées de statistiques et santé, organisées par le GDR « Statistiques et Santé », la Société française

de biométrie et le groupe « Biopharmacie » de la Société française de statistiques. October, 2021. Paris, France

Le Bourdonnec K et al. Continuous-time mediation analysis for repeated mediators and outcomes. Journée des jeunes chercheurs. January 19, 2023. Rennes, France

Le Bourdonnec K et al. Continuous-time mediation analysis for repeated mediators and outcomes. ISCB44. August 27-31, 2023. Milan, Italy

Présentations affichées en conférence

Le Bourdonnec K et al. Continuous-time mediation analysis for repeated mediators and outcomes. EUROCim. April 19-21, 2023. Oslo, Norway

Le Bourdonnec K et al. Continuous-time mediation analysis for repeated mediators and outcomes. Journée des jeunes chercheurs (BPH). September 25, 2023. Bordeaux, France

Médiations scientifiques

"Bureau des enquêtes" - Cap Sciences. March, 2023. Bordeaux, France

"Têtes chercheuses". Dans le cadre du label « Science Avec et Pour la Société » - Université de Bordeaux et Cap Sciences. April, 2024. Agen, France

Table des matières

Liste des tableaux

Table des figures

Chapitre 1

Introduction

D'après l'Organisation Mondiale de la Santé, le processus de vieillissement résulte de l'accumulation progressive d'une diversité de dommages moléculaires et cellulaires au fil du temps. Au cours des dernières décennies, le vieillissement de la population a connu une augmentation constante de l'espérance de vie, passant de 69,2 ans chez les femmes et 63,4 chez les hommes en 1950 à respectivement 85,5 ans et 79,5 ans en 2022 selon l'INSEE.
Le vieillissement de la population suscite un intérêt croissant dans divers domaines académiques liés à la Santé Publique. En sociologie, par exemple, l'étude du vieillissement permet de comprendre les aspects sociaux et culturels de cette période de la vie (Morgan and Kunkel (2007)), en économie et en politique l'étude du vieillissement permet la mise en place de politiques de finance des retraites ou des établissements de santé (Fisher and Ryan (2018)), tandis qu'en épidémiologie, l'étude du vieillissement de la population permet, entre autre, d'explorer les impacts de ce dernier sur la santé des personnes âgées (Murman (2015)).
Selon Harman (1981), "Le vieillissement est [...] associé ou responsable de la prédisposition croissante aux maladies et à la mort qui accompagne l'avancée en âge.". Actuellement, de multiples facteurs jouant un rôle dans le vieillissement sont connus dans la littérature, tels que les modifications de l'acide désoxyribonucléique (ADN) (e.g. avec le raccourssissement des télomères (Rey-Millet et al. (2023))) causant un vieillissement accéléré ou bien l'apparition de plusieurs maladies (e.g. maladie d'Alzheimer (Drachman (2006)), cancer (Duray et al. (2014))) comme étant des résultantes de ce vieillissement. Comprendre les

causes et les conséquences sous-jacentes à ce dernier est ainsi un véritable challenge en Santé Publique. Ce challenge se rélève d'autant plus complexe que différentes composantes contribuent à ce processus de vieillissement, comme le vieillissement cérébral (Lee et al. (2011)).

Dans ce travail nous souhaitons approfondir la compréhension des causes du vieillissement cérébral, tout en considérant la nature dynamique de ce processus, susceptible d'évoluer au cours du temps. Pour répondre à cette problématique, nous proposons des approches statistiques developpées pour identifier les facteurs et les mécanismes sous-jacents pouvant contribuer de façon causale au processus de vieillissement cérébral.

1.1 Le vieillissement cérébral et la démence

1.1.1 Définition et contexte épidémiologique du vieillissement cérébral et de la démence

Avec le viellissement général de la population, sont apparues des pathologies liées au vieillissement cérébral (VC). Le vieillissement cérébral se traduit par des altérations progressives structurelles (i.e., changement anatomique) et fonctionnelles (i.e., changement dans le fonctionnement et l'activité) du cerveau au cours des années (Angel and Isingrini (2015)). Cliniquement, le vieillissement cérébral peut se traduire par des pertes de mémoire épisodique (Salthouse (2003), Buckner (2004), Nyberg et al. (2012)), une diminution de la vitesse de traitement (i.e. capacité du cerveau à traiter l'information) (Salthouse (1996)), ainsi que des altérations des fonctions exécutives (e.g. difficultés dans la planification de tâches complexes, problèmes de concentration) (West (1996)). Ces manifestations cliniques sont la résultante des altérations cérébrales (e.g. perte de mémoire conséquence de la perte neuronale (Agid (2016)).

Deux catégories de vieillissement cérébral existent, celui qualifié de **normal** et le **pathologique**. En fonction de la nature de ce vieillissement, bien que des changements structurels puissent survenir, ces variations structurelles ne se traduisent pas nécessairement par des altérations immédiates au niveau clinique des fonctions cognitives ou du comportement. Dans le VC normal, on peut évoquer un déclin cognitif léger, caracté-

risé par d'éventuels ralentissements dans la vitesse de traitement de l'information et des changements mineurs de la mémoire (Harada et al. (2013)). Le déclin cognitif devient plus substantiel lorsqu'il est associé à un vieillissement cérébral pathologique, souvent attribuable à la présence de maladies cérébrales liées à l'âge regroupées sous le terme de syndrôme de "démence" (Giffard et al. (2001)).

La démence, syndrôme évolutif, regroupe différentes maladies cérébrales qui altérent le fonctionnement cognitif (e.g. mémoire, compréhension), dont la principale est la maladie d'Alzheimer (Dartigues and Helmer (2009)). Le diagnostic de démence repose sur les critères du manuel diagnostique et statistique des troubles mentaux (Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV)) (Guze (1995)). Il exige une évaluation approfondie réalisée par des experts de la santé, impliquant des tests cognitifs, des analyses sanguines, et des examens d'imagerie à résonnance magnétique (IRM) cérébrale. D'après l'Organisation mondiale de la santé, environ 50 millions de personnes dans le monde sont actuellement touchées par la démence, et une projection estime que ce nombre pourrait s'élever à 82 millions d'ici à 2030.

1.1.2 Facteurs de risque du vieillissement cérébral et de la démence

En raison de la nature dynamique et multifactorielle du vieillissement cérébral, plusieurs études ont exploré l'influence des facteurs biologiques, environnementaux et sociodémographiques sur le risque de démence (van der Flier and Scheltens (2005), Jorm (2001), McCullagh et al. (2001), Livingston et al. (2017)). Ces recherches ont révélé que les facteurs de risque (i.e. facteurs qui augmentent le risque) de la démence peuvent être classés en modifiables ou non modifiables, et qu'ils se manifestent à différents stades de la vie.

Parmi les facteurs non modifiables, le vieillissement est reconnu comme le principal facteur de risque, avec un risque plus de six fois supérieur chez les individus âgés de plus de 80 ans par rapport à ceux âgés de 65 à 79 ans, selon l'INSERM. Des disparités de genre ont aussi été observées, indiquant un risque plus élevé de 50% chez les femmes par rapport aux hommes, à un âge équivalent (Livingston et al. (2020)). Certains facteurs génétiques de la démence ont été identifiés, le principal étant l'apolipoprotéine E4 (APOE4) (Lambert

et al. (2013)). Plus spécifiquement, l'apolipoprotéine E est une protéine qui joue un rôle essentiel dans le transport des lipides, tels que le cholestérol, dans le sang. Ces lipides sont vitaux pour les cellules du cerveau. L'apolipoprotéine E (APOE) se présente sous diverses formes déterminées par les allèles codants (E2, E3 ou E4). Pour être identifiée comme porteuse de l'APOE4, une personne doit hériter d'au moins une copie de l'allèle APOE4. Chaque individu hérite de deux copies du gène APOE, une de chaque parent. Par conséquent, une personne est considérée comme porteuse de l'APOE4 si elle possède au moins une copie de l'allèle APOE4. L'APOE4 est reconnue comme la variante défavorable de l'APOE et est présente chez 10 à 20 % de la population.

Les facteurs modifiables ont un rôle majeur dans le développement de la démence. Le Lancet (Livingston et al. (2017)) a indiqué dans son rapport qu'il était possible de prévenir ou de retarder jusqu'à 40% des cas de démence en ciblant des facteurs potentiellement modifiables. Dans une mise à jour récente de cette revue, Livingston et al. (2020) ont representé ces facteurs modifiables au cours de la vie (c.f. Figure 1.1), tels que l'inactivité physique, le tabagisme, la consommation excessive d'alcool, la pollution atmosphérique, les traumatismes crâniens, la faible fréquence des contacts sociaux, la dépression et la déficience auditive.

Certains facteurs sont présents dès l'enfance comme le niveau d'éducation. Intervenir sur le niveau d'éducation, c'est-à-dire avoir un meilleur niveau d'éducation durant l'enfance permettrait de réduire de 7% la prévalence de la démence (figure 1.1). Valenzuela and Sachdev (2006) ont montré que les individus ayant un niveau d'étude avancé ont un risque de démence diminué de 47% par rapport à ceux ayant un niveau d'étude plus faible. Le niveau d'éducation élévé apparait donc comme étant un facteur protecteur de la démence. Cela peut s'expliquer par l'accroissement de la capacité cognitive globale tout au long des études, atteignant son apogée à la fin de l'adolescence, période où la plasticité cérébrale est maximale (Kremen et al. (2019)). Ce phénomène est communément désigné sous le terme de "réserve cognitive".

Le diabète est quant à lui un facteur de risque en âge avancé, avec une diminution possible de 1% de la prévalence de la démence si le diabète était éliminé (c.f Figure

1.1). Plusieurs études ont montré que chez les personnes âgées, le diabète est associé à un déclin cognitif plus important (Awad et al. (2004),Biessels et al. (2006)). Le diabète est souvent accompagné d'autres facteurs de risque vasculaire tels que l'hypertension et l'obésité (Kloppenborg et al. (2008)), ayant un impact plus important sur le vieillissement cérébral en milieu de vie (c.f Figure 1.1).

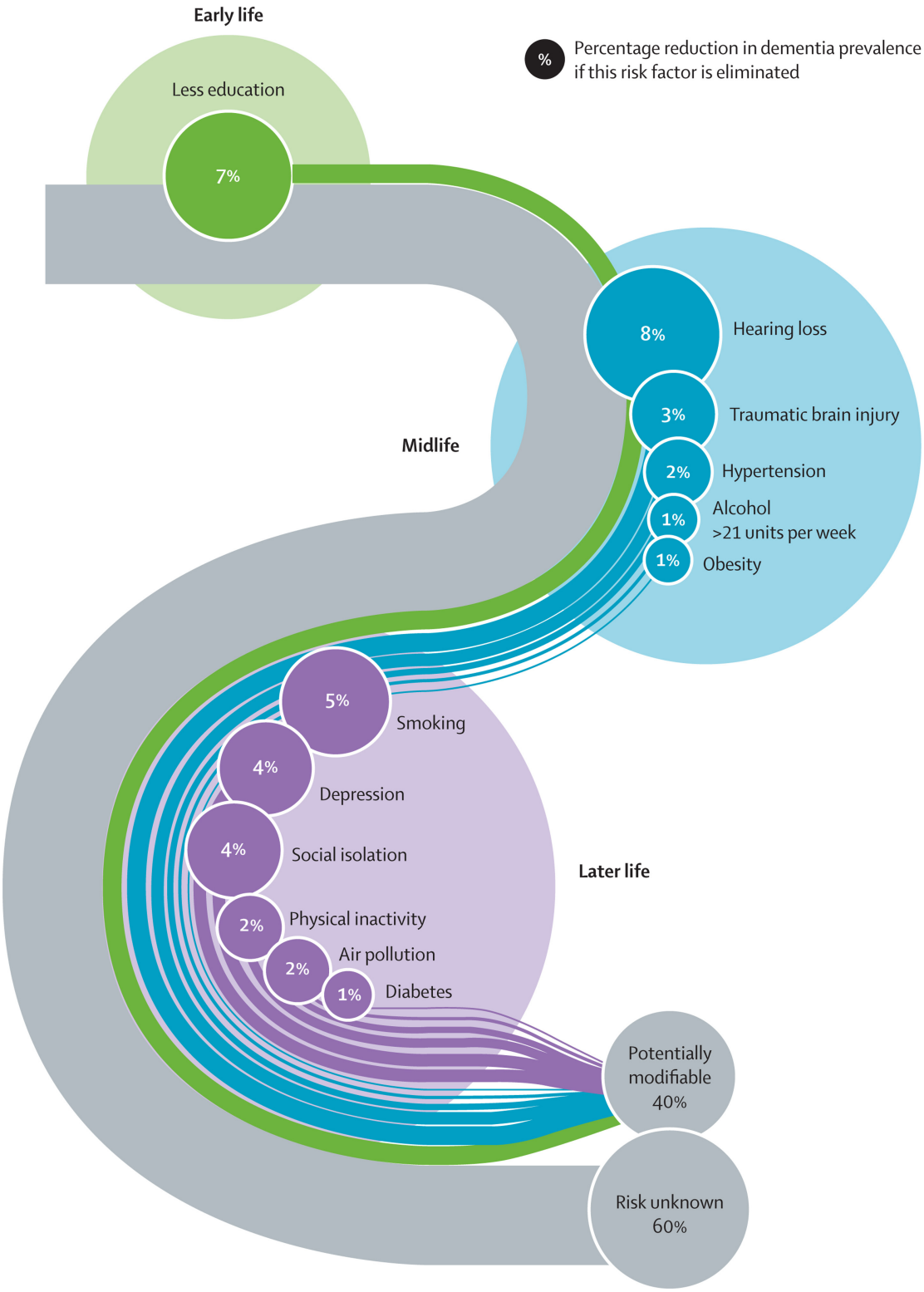


FIGURE 1.1 – Part modifiable des facteurs de risque de la démence au cours de la vie (Extrait de "Dementia prevention, intervention, and care : 2020 report of the Lancet Commission")

Bien que des facteurs modifiables majeurs émergent de la littérature, la littérature rapporte fréquemment des résultats discordants. Diverses raisons peuvent mener à des conclusions contradictoires. Par exemple, sur la Figure 1.1 l'obésité est montré comme

étant un facteur de risque de la démence en milieu de vie, cependant l'étude de Nourhashémi et al. (2003) a montré sur une population âgée qu'un indice de masse corporel (IMC) faible était associé à un risque de démence plus élévé. Wagner et al. (2021) ont démontré que selon la période de vie étudiée, l'association entre l'indice de masse corporelle et le déclin cognitif pouvait avoir des effets opposés. Alors que l'IMC évalué très en amont de l'évolution cognitive montrait une association délétère, l'IMC élevé était associé à un déclin cognitif plus prononcé. A l'approche de l'évaluation l'association s'inversait avec un IMC élevé associé à un déclin cognitif moindre après ajustement sur l'histoire de l'IMC.

De manière similaire, les individus avec un haut niveau d'étude montrent un déclin cognitif rapide post-signes de démence (Meng and D'Arcy (2012)), tandis que Livingston et al. (2017)) identifient un faible niveau d'éducation comme étant un facteur de risque. Cette disparité pourrait s'expliquer par une réserve cognitive plus élevée chez les individus éduqués, entraînant un déclin cognitif retardé mais plus prononcé une fois initié, en raison d'une charge pathologique accrue.

L'hétérogéneité des résultats souligne l'importance de l'aspect temporel dans l'étude du vieillissement et de ses causes. Cette variabilité des résultats peut également être dû au type d'étude effectué (suivi court *versus* suivi long), à l'erreur de mesure des facteurs de risque, ou la prise en compte ou non d'autres facteurs de risques associés.

1.2 Challenges méthodologiques dans l'étude du vieillissement cérébral

L'étude des facteurs de risque du vieillissement cérébral induit de nombreux challenges statistiques. La principale difficulté qui nous intéresse dans ce travail est que des facteurs augmentent la probabilité de développer une démence sans systématiquement être causalement liés à celle-ci.

1.2.1 Le vieillissement cérébral : un processus multifactoriel

Le vieillissement cérébral est un processus complexe et multifactoriel, influencé par une multitude de facteurs interdépendants (Livingston et al. (2020)). Dans le cas de la démence, de nombreux facteurs peuvent influencer la maladie, notamment les facteurs cardiométaboliques tels que le diabète, l'hypercholestérolémie et l'hypertension. Ces facteurs sont interdépendants les uns des autres, ce qui signifie qu'ils peuvent se renforcer mutuellement ou partager des mécanismes pathologiques communs. Par exemple, le diabète et l'hypercholestérolémie sont souvent présents chez les mêmes individus en raison de facteurs de risque partagés comme l'obésité, le manque d'exercice et une alimentation peu équilibrée.

Lorsque l'on étudie l'association entre un facteur de risque et une maladie, cette multifactorialité peut mener à des problèmes de confusion, dû à des variables qui impactent à la fois le facteur de risque et la maladie étudiée. C'est ce que l'on nomme facteur de confusion. Par exemple, dans la relation entre le diabète et la démence, l'IMC est un facteur de confusion souvent cité, étant associé à la fois au diabète et à la démence. Négliger l'IMC peut engendrer des biais dans l'estimation de l'association entre le diabète et la démence, en effaçant ou en surestimant l'effet réel du diabète sur le risque de démence. Il est donc important de prendre en compte les facteurs de confusion dans les études.

De la même manière, l'association entre le facteur de risque et la maladie peut parfois passer par à une variable intermédiaire, également appelée variable médiatrice. Il s'agit d'une variable qui se situe sur le chemin causal entre le facteur de risque et la maladie. L'effet global du facteur de risque sur la démence à deux composantes, celle passant par le médiateur et celle passant par la démence. Lorsque l'on s'intéresse à l'effet total il est important de garder ces deux chemins. Toutefois si l'on souhaite connaître la part expliquée par la démence alors des analyses de médiation sont nécessaires.

La multifactorialité constitue ainsi un aspect important à prendre en compte dans les études sur le vieillissement, notamment lorsque l'on souhaite interpréter les effets de manière causale.

1.2.2 Le vieillissement cérébral : un processus multidimensionnel

Un processus est dit multidimensionnel lorsque qu'il est composé de plusieurs dimensions, autrement dit lorsque plusieurs domaines peuvent être atteints. Le vieillissement de la population en est un exemple. Dans la littérature différentes études ont montré que le vieillissement de la population n'était pas un concept unidimensionnel (Lee et al. (2011), Cosco et al. (2013)). Dans l'étude menée par Zanjari et al. (2017), le vieillissement de la population est défini à travers de cinq dimensions principales : la santé physique, le bien-être social, le bien-être psychologique, la spiritualité, et l'environnement et sécurité économique.

Similairement, le vieillissement cérébral ne se limite pas simplement à la détérioration de la mémoire, par conséquent ne possède pas qu'une seule dimension. Il est composé de plusieurs dimensions interconnectées. Tigano et al. (2019) ont montré que le vieillissement cérébral implique différents aspects, notamment fonctionnels, sociaux et cognitifs. La dimension sociale comprend des changements dans les interactions sociales, tels que liés à la retraite, l'isolement ou la dépression. L'aspect fonctionnel et l'aspect cognitif sont eux-même multidimensionnels. L'aspect fonctionnel par exemple, peut être caractérisé par des altérations dans la mobilité ou à une dépendance (elle-même multidimensionnelle (Edjolo (2014)). En ce qui concerne la cognition, elle englobe des composantes telles que la vitesse de la mémoire, l'attention et la fluence verbale (Park and Uno (2012)).

Étudier le mécanisme des différentes dimensions, y compris leurs temporalités, leurs dépendances et leurs successions, est essentiel pour une meilleure compréhension d'un phénomène multidimensionnel. Le modèle sur la progression de la maladie d'Alzheimer proposé par Jack et al. (2010) illustre parfaitement nos propos. Ce modèle (figure 1.2) permet de décrire l'évolution temporelle de la maladie. Il met en évidence l'implication de divers processus qui interviennent successivement à des moments distincts. Dans le processus pré-démentiel, plusieurs marqueurs interviennent, tout d'abord avec un changement dans les biomarqueurs tels que les protéines $A\beta$ dans le cerveau, puis avec un changement de la structure cérébrale, entraînant une atrophie de certaines régions telles que le lobe

médio-temporal ou l'hippocampe. Puis enfin, par un déclin cognitif et fonctionnel.

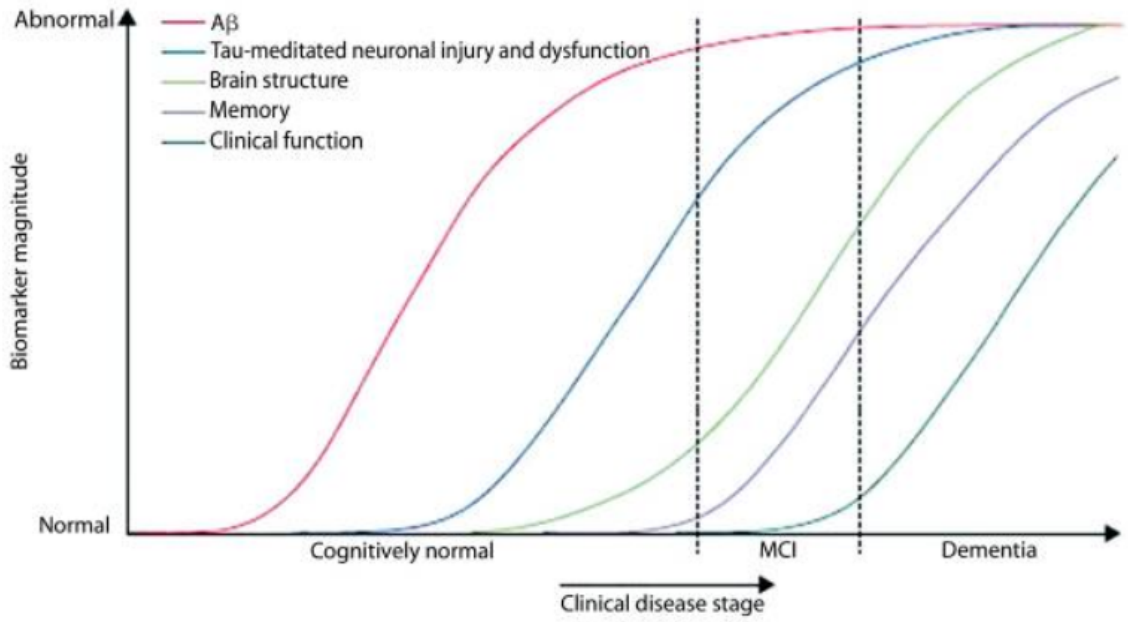


FIGURE 1.2 – Modèle hypothétique des biomarqueurs dynamiques de la cascade de la maladie d'Alzheimer. *Issue de Jack CR Jr, Knopman DS, Jagust WJ, Shaw LM, Aisen PS, Weiner MW, Petersen RC, Trojanowski JQ. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. Lancet Neurol. 2010 Jan ;9(1) :119-28.*

Néanmoins, bien que la multidimensionnalité soit un aspect présent dans de nombreux processus, se manifestant à l'intérieur même de certains processus, comme démontré dans des domaines tels que le vieillissement cérébral et la cognition, ou encore la dépendance, sa modélisation demeure un défi du point de vue statistique. Pourtant, il est nécessaire de la prendre en compte pour appréhender la problématique de médiation et des mécanismes causaux du vieillissement.

Des approches ont proposé de modéliser simultanément plusieurs processus possiblement corrélés, soit en relation avec un événement spécifique (par exemple, la cognition, la dépression et la dépendance par rapport à l'apparition de la démence, Proust-Lima et al. (2016)), soit pour comprendre la structure des dépendances entre ces processus (comme la structure cérébrale, la cognition et la fonction ; Taddé et al. (2020)). Cependant, dans ces études, l'aspect causal de ces associations n'a pas été pleinement exploré.

1.2.3 Challenges liés au design des données

Les challenges précédents se rapportent à l'interprétation causale des associations mises en évidence avec la problématqiue de confusion et de médiation. A cela s'ajoute des challenges liés au design des données.

1.2.3.1 Données issues des cohortes observationnelles

En épidémiologie, il existe différents schémas d'études, tels que les essais cliniques randomisés, les études transversales, les études cas-témoins, les études de cohortes. Montrer qu'un facteur de risque est une des causes d'une maladie peut être réalisé dans des essais cliniques randomisés. Il s'agit d'études expérimentales dans lesquelles on intervient sur un unique facteur de risque dans plusieurs groupes comparables pour observer les effets de cette manipulation sur une maladie. Par exemple Mapelli et al. (2013), ont montré à l'aide d'un essai clinique que la prise d'un traitement de stimulation cognitive conduit à une amélioration cognitive, avec une amélioration des performances aux tests cognitifs et une diminution significative des symptômes comportementaux avec la prise du traitement.

Néanmoins, les essais cliniques ne sont pas toujours adaptés pour l'étude des facteurs de risque du vieillissement cérébral. Ce sont généralement des suivis très court alors que des maladies, telles que la démence, se développent sur des décennies. Ils ont souvent des critères d'inclusion restrictifs pour cibler une population à risque et ne permettent donc pas d'étudier des mécanismes très en amont de la maladie. Pourtant comme vu en Figure 1.1, certains facteurs de risque peuvent arriver très en amont du processus de vieillissement pathologique et l'étude des facteurs de vieillissement demandent des suivis sur plusieurs décennies (Livingston et al. (2020)).

C'est pourquoi, l'étude des facteurs de risque est souvent menée à travers des cohortes en population. Dans ces cohortes, les individus, issus de la population générale, sont suivis à intervalles réguliers dans le temps. Ces suivis longitudinaux permettent de collecter de façon standardisée des informations sur leurs expositions potentielles et leurs indicateurs de santé. Ces suivis longitudinaux, riches en informations, présentent de aussi des défis et difficultés méthodologiques.

Certains participants vont sortir de l'étude au fil du temps. Cela peut venir d'un refus de continuer à participer ou d'un événement comme le décès. Ce phénomène peut introduire un biais de sélection, altérant ainsi la validité des résultats. Par exemple, dans une étude de cohorte portant sur la cognition, si les individus quittent l'étude au début de l'apparition des premiers problèmes de mémoire, cela risque d'altérer l'évaluation de facteurs de risque sur la trajectoire cognitive.

Parallèlement, en raison de la nature longitudinale de ces études, les données manquantes sont fréquentes, qu'il s'agisse de participants perdus de vue, de visites ponctuellement ratées ou d'examens non prévu de façon systématique à chaque suivi. Ce phénomène est d'autant plus courant dans les études de cohorte très longues. La gestion des données manquantes est ainsi importante pour minimiser les biais potentiels sur les résultats.

De plus, étant dans une étude de cohorte, les participants ne sont pas suivis tous exactement au même temps. Cette variabilité des intervalles de suivi pour chaque participant empêche de traiter l'aspect temporel de façon discrète en ne considérant que les temps théoriques de visite.

Ces défis sont particulièrement fréquents dans les études de cohorte. D'autres challenges sont à considérer lorsqu'il s'agit d'études spécifiques aux sujets âgés.

1.2.3.2 Etude chez le sujet âgé

Deux challenges majeurs sont à prendre en compte lorsque l'on étudie le sujet âgé.

L'entrée retardée : Généralement l'âge est un des critères d'inclusion dans les études de cohorte sur le vieillissement. Par exemple, dans la cohorte "Trois-cités" (Alpérovitch et al. (2002))(présentée en détail dans le Chapitre 2), le critère d'inclusion sur l'âge est qu'il soit supérieur à 65 ans. Ainsi, au moment des inclusions, les individus peuvent rejoindre l'étude à différents moments de leur vie, et donc, à différents âges. Le phénomène d'entrée retardée doit être pris en compte lors de l'étude de ces cohortes, car il peut entraîner des biais dans les estimations des événements étudiés. Les personnes recrutées à des âges plus avancés peuvent présenter un risque de survenue de ces événements différent de celui des personnes recrutées à des âges plus jeunes.

Présence du décès : Les études de cohorte sur le vieillissement sont plus complexes en raison de la présence du décès. Premièrement cela peut entrainer un biais de survie. Les personnes inclusent dans la cohorte sont de fait vivantes et indemnes de l'événement d'intérêt au moment de leur inclusion. Les personnes qui survivent jusqu'à un âge avancé peuvent avoir des caractéristiques différentes de celles qui décèdent plus tôt. De même, ces études nécessitent souvent un suivi sur une période prolongée pour comprendre les changements liés au vieillissement. Cependant, le décès peut entraîner la censure des participants, c'est-à-dire qu'ils peuvent sortir de l'étude en raison de la fin de leur vie. Des méthodologies statistiques doivent alors être adaptées à ce type de données.

1.2.3.3 Mesures en temps discret de processus en temps continu

Dans les études de cohorte, il est fréquent d'étudier des processus qui évoluent en temps continu. Par exemple, le vieillissement est un processus au long cours, progressant en temps continu. Les changements liés à l'âge se produisent de manière progressive et continue tout au long de la vie d'un individu. Par exemple, la perte de mémoire est un processus dynamique, considérée comme un continuum. Cela implique qu'elle peut fluctuer en intensité et en sévérité tout au long de la vie d'un individu.

Étudier les causes des processus implique d'analyser les déterminants de leur changement au cours du temps. Dans les cohortes, cette analyse s'effectue à l'aide de mesures répétées de ces processus à intervalles plus ou moins réguliers (par exemple, tous les 2 ou 3 ans), avec une certaine erreur de mesure pouvant être dûe à l'utilisation des instruments de mesure.

Il est donc important de tenir compte à la fois du processus étudié et de sa mesure irrégulière afin de minimiser les biais potentiels. De plus, les données analysées sont généralement "groupées", car chaque sujet possède plusieurs mesures du processus sur la durée de l'étude. Par conséquent, il est nécessaire d'utiliser des méthodes statistiques appropriées pour prendre en considération la corrélation entre les données d'un même sujet.

Enfin, le fait de récolter des données uniquement à des temps de visites peut poser des

défis lorsque l'on s'intéresse à la survenue d'un événement. D'une part, lorsque l'on étudie l'apparition de la démence, les participants peuvent développer la maladie à n'importe quel moment entre deux visites. Ce phénomène est appelé censure par intervalle. Ne pas prendre en compte cette caractéristique peut mener à un biais dans les interprétations (Lindsey and Ryan (1998)). D'autre part, il est important de considérer que le décès peut parfois survenir avant qu'une démence ne soit diagnostiquée chez une personne. Intégrer cette possibilité dans la modélisation est essentiel pour une compréhension précise des facteurs de risque et des trajectoires de la maladie (Leffondré et al. (2013)).

1.3 Objectifs de la thèse

1.3.1 Objectifs épidémiologique de la thèse

D'un point de vue épidémiologique, cette thèse vise à approfondir la compréhension des facteurs de risque du vieillissement cérébral et d'explorer les mécanismes étiologiques qui leur sont associés.

Cette thèse a pour ambition de répondre à plusieurs questions, telles que : *Est-ce que le diabète est une cause du déclin cognitif ? ; Quel est le rôle des lésions vasculaires dans la relation entre l'APOE4 et le déclin cognitif ? Quelle est la part de l'effet de facteurs cardiométaboliques sur le décès, qui est expliquée par l'effet sur la démence ?*

1.3.2 Objectifs statistique de la thèse

Pour répondre à ces questions l'objectif de cette thèse est de proposer des méthodologies statistiques permettant la réalisation d'analyses d'inférence causale au sein d'études de cohortes prospectives, en tenant compte de l'aspect temporel des processus étudiés.

Les contributions de cette thèse sont articulées autour de deux travaux majeurs. Le premier se concentre sur les analyses de médiation avec l'exploration des mécanismes de médiation en présence de données longitudinales (répétées dans le temps) et aborde succinctement l'étude des mécanismes dans le contexte de temps d'événements censurés par intervalle. Le deuxième travail vise à estimer l'effet d'expositions dans des études longi-

tudinales en prenant en compte des facteurs de confusion non observés.

L'objectif de chacun de ces travaux est de fournir une méthodologie adaptée, accompagnée de solution informatique, et d'un article didactique, permettant d'appréhender l'inférence causale dans l'analyse épidémiologique des données de cohortes.

Les travaux de ma thèse sont motivés par l'analyse des données de la cohorte populationnelle "Trois-cités" mise en place en 1999 pour étudier le vieillissement cérébral en France

1.4 Plan de la thèse

La structure de cette thèse se décline comme suit :

Dans le **chapitre 2**, sont décrites les données de la cohorte en population "Trois-Cités" sur lesquelles s'appuient toutes les applications des méthodes proposées.

Le **chapitre 3** présente l'état de l'art statistique en scindant la modélisation statistique des données de cohorte des approches pour traiter la causalité. Nous présentons d'abord comment modéliser des données longitudinales à l'aide de modèles mixtes, des extensions des modèles mixtes ainsi qu'à partir des modèles dynamiques, puis comment modéliser des données de survie, en présence d'un temps d'évènement ou de plusieurs temps d'évènement. Une revue non exhaustive des développements méthodologiques en inférence causale est ensuite réalisée en définissant conceptuellement la causalité, puis présentant les analyses de médiation et enfin présentant la méthode par variables instrumentales pour les facteurs de confusion non observés.

Le **chapitre 4** propose un développement méthodologique d'analyse de médiation lorsque la variable d'exposition est fixe dans le temps et que le médiateur et l'outcome sont des variables répétées au cours du temps. Ce travail est soumis pour publication. Le chapitre décrit ensuite une approche d'analyse de médiation lorsque la variable d'exposition est fixe dans le temps et que le médiateur et l'outcome sont des temps d'évènements, avec le médiateur possiblement censuré par intervalles.

Le **chapitre 5** propose une extension des analyses par variables instrumentales lorsque la variable d'exposition est fixe dans le temps et que l'outcome est une variable répétée dans le temps, afin de pouvoir faire de l'inférence causale dans les études longitudinales en présence de facteurs de confusion non mesurés. Ce travail est publié dans le journal *Biometrical Journal*

Enfin, le **chapitre 6** aborde la discussion générale de travaux de la thèse du point de vue statistique, en montrant leurs avancées et leurs limites, ainsi que les perspectives de ces travaux.

Chapitre 2

La cohorte "Trois Cités"

Les méthodes et outils développés dans cette thèse sont appliqués aux données de la cohorte en population "Trois Cités".

2.1 Présentation de la cohorte

La cohorte "Trois Cités" (3C), débutée en 1999, a été initialement développée pour l'étude et la compréhension des maladies vasculaires cardiaques et cérébrales. Plus précisement, l'objectif de cette cohorte est "d'étudier le lien qui relie la maladie d'Alzheimer et les affections vasculaires et les facteurs génétiques qui peuvent influencer la fréquence des maladies cérébrales liées à l'âge" (Alpérovitch et al. (2002)).

Il s'agit d'une cohorte multicentrique fermée. Les inclusions ont eu lieu en 1999-2000, dans trois villes françaises : Bordeaux, Dijon et Montpellier. Des personnes âgées de 65 ans et plus vivant à domicile et tirées au sort sur la liste électorale de ces trois villes (et communes avoisinantes) ont été incluses. Au total, 9294 individus participent à cette étude sur le vieillissement cérébral en population générale, avec 2104 participants à Bordeaux, 4931 à Dijon, et 2259 participants à Montpellier.

2.2 Recueil des données

Les participants de la cohorte 3C ont eu des entretiens réalisés par un psychologue ou une infirmière qualifiée en face à face à l'inclusion. Y ont été récoltées des données via des questionnnaires standardisés, des examens et des mesures, communs aux trois centres. Comme présenté sur la Figure 2.1, les participants de Bordeaux et Montpellier ont été suivis pendant 17 ans, le suivi des participants de Dijon s'est quant-à-lui arrêté après 12 ans.

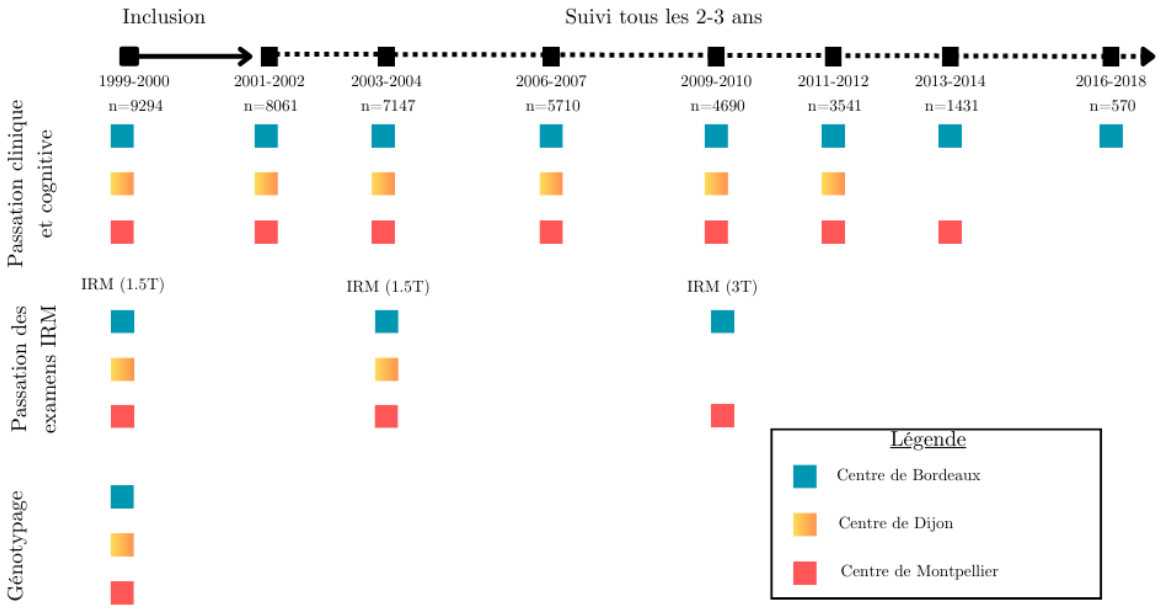


FIGURE 2.1 – Schéma représentant les différents types de visites effectuées par les participants de la cohorte 3C en fonction du temps et du centre d'appartenance

Les participants ont été suivis tous les deux-trois ans environ. Sur la figure 2.2 est illustrée la distribution des temps de visite des participants à chaque suivi depuis leur inclusion dans l'étude, montrant ainsi une diminution progressive du nombre de participants au cours du temps, et également la variabilité des délais de visites entre les individus.

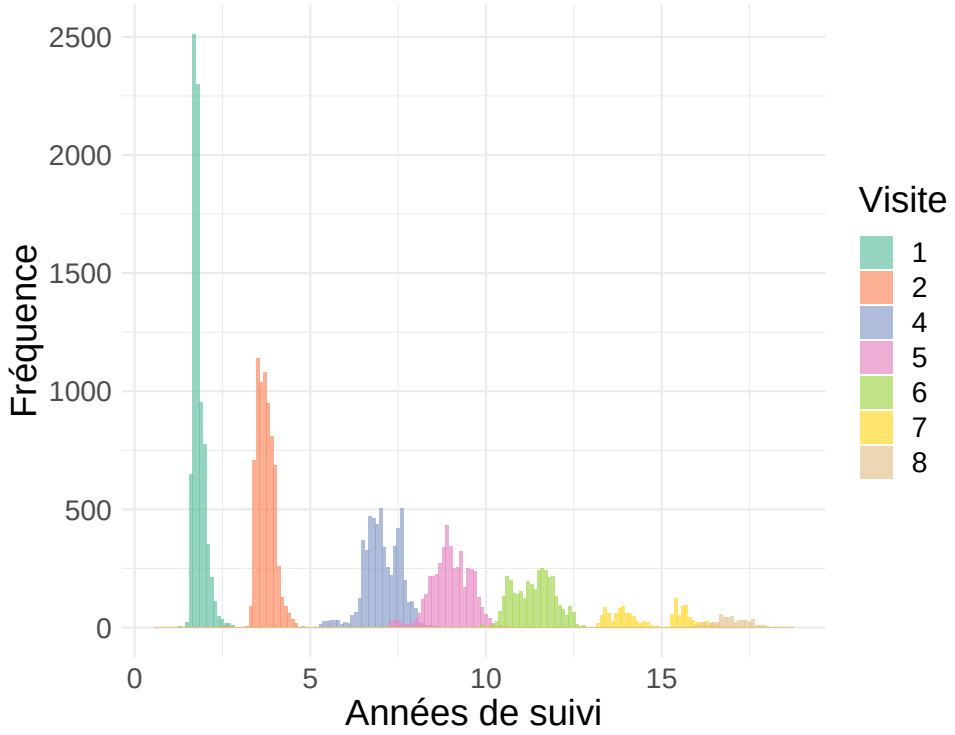


FIGURE 2.2 – Distribution des temps des visites des participants de la cohorte 3C depuis leur inclusion (n=9294)

Le protocole de l'étude a été approuvé par le comité éthique du Kremlin-Bicêtre, avec un consentement signé par chaque participant.

2.2.1 Données générales, psychologiques et cognitives

Lors des visites, de nombreux facteurs ont été recueillis à l'aide d'un questionnaire. Parmi ces facteurs, des données telles que les caractéristiques socio-démographiques (e.g., l'âge, le sexe, le niveau d'étude), les antécédants de maladies et pathologies (e.g., diabète, infarctus du myocarde, cancer, hypertension) et le mode de vie (e.g., consommation de tabac, consommation d'alcool, prise de médicaments, alimentation, activité physique) ont été collectées.

2.2.1.1 Evaluation psychologique

Chaque entretien incluait une évaluation psychologique à l'aide de l'échelle Center for Epidemiologic Studies (CES-D) pour évaluer la symptomatologie dépressive. Cette échelle comporte 20 items évaluant divers aspects des symptômes dépressifs tels que la tristesse, la fatigue, la perte d'intérêt, la perte d'appétit et les troubles du sommeil. Les participants

devaient indiquer la fréquence à laquelle ils ont ressenti chaque symptôme au cours de la semaine précédente, utilisant une échelle de notation de 4 points allant de "rarement ou jamais" à "presque toujours". Les scores totaux de la CES-D peuvent varier de 0 à 60, avec des scores plus élevés reflétant une sévérité accrue des symptômes dépressifs.

2.2.1.2 Performances cognitives

Une évaluation des performances cognitives a été réalisée à chaque suivi à l'aide d'une batterie de tests diversifiée. Cela comprenait des tests évaluant les fonctions exécutives avec le Trail Making Test parties A et B (TMT-A, TMT-B) Reitan (1955), la mémoire visuelle avec le Benton Visual Retention Test (BVRT) Benton (1945), les fonctions cognitives avec le Mini Mental State Examination (MMSE) Folstein et al. (1975) et la fluence verbale avec le Isaacs Set Test (IST) Isaacs and Akhtar (1972). Le test de fluence verbal correspond à la capacité d'une personne à générer rapidement un grand nombre de mots répondant à un certain critère donné, en un temps limité. Dans le test IST, les participants doivent nommer autant de mots que possible appartenant à une catégorie spécifique en l'espace de 15 secondes. Quatre catégories sont explorées : les animaux, les fruits, les couleurs et les villes. Le score obtenu correspond au nombre total de mots correctement générés dans les quatre catégories dans le temps imparti.

2.2.1.3 Performances fonctionnelles

L'évaluation de l'incapacité dans les activités de la vie quotidienne permet d'estimer l'impact du vieillissement et des affections qui y sont liées sur le fonctionnement quotidien (Pérès et al. (2008)). Dans la cohorte 3C, l'incapacité fonctionnelle est évaluée à l'aide de l'échelle Lawton, qui mesure les activités instrumentales de la vie quotidienne. Cette échelle comprend une série de questions portant sur la capacité des individus à effectuer diverses tâches telles que l'utilisation du téléphone, les courses, la mobilité, la gestion des médicaments et la gestion du budget. Dans le cadre de cette thèse, nous utilisons un score de dépendance fonctionnelle, calculé comme la somme des dépendances associées à ces différentes activités. Un score plus élevé indique une dépendance accrue.

D'autres questions ont été posées concernant la gestion des repas et les tâches ménagères, mais elles ne s'appliquent qu'à l'un des deux sexes, ce qui explique pourquoi nous

ne les prenons pas en considération dans notre analyse.

2.2.2 Données d'Imagerie par Résonance Magnétique

Dans la cohorte 3C, les trois centres (Bordeaux, Dijon et Montpellier) ont bénéficié de la collecte des données d'imagerie par résonance magnétique (IRM). Toutefois, dans cette thèse nous nous interessons qu'aux données IRM issues des centres de Bordeaux et Dijon. Deux IRMs ont été effectuées (une à baseline et une à 2 ans de suivi) pour Dijon. Bordeaux possède une IRM de plus réalisée entre 2010 et 2011. Les IRMs ont été réalisées sur une IRM 1,5 T Philips Intera® pour les deux premières passations, la troisième passation a eu lieu sur une IRM 3T Philips Achieva®. Ils ont permis d'obtenir de nombreux marqueurs de la structure cérébrale (e.g. le volume hypoccampique, le volume de la substance grise) et des marqueurs de lésions vasculaires (e.g. hyperintensités de la substance blanche).

2.2.3 Diagnostic de la démence

Lors de l'inclusion dans l'étude les participants ont été évalués par un neurologue. À Dijon, en raison du grand nombre de participants, seuls ceux présentant des suspicions de démence, en raison de leurs performances au MMSE et à l'IST, ont été vus par un neurologue. Puis à chaque visite le processus de diagnostic de la démence s'est déroulé en trois étapes.

Premièrement, une évaluation neuro-psychologique a été effectuée. Les performances aux tests cognitifs, la capacité à effectuer les activités de la vie quotidienne, la gravité des troubles cognitifs et les résultats de l'IRM, ont ainsi été reccueillis.
Ensuite, à partir de l'évaluation neuro-psychologique, le diagnostic de démence conforme aux critères du DSM-IV est effectué par un neurologue.
Enfin, chaque cas de démence suspecté a été soumis à une validation par un comité de neurologues indépendants, permettant de déterminer le type de démence en se basant sur des critères spécifiques NINCDS-ADRA (National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association).

2.2.4 Décès

Le décès de participants de la cohorte 3C est collecté en temps continu sur l'ensemble de la période de suivi et après une éventuelle sortie d'étude.

2.3 Description de la population

À l'inclusion dans l'étude, les 9294 participants de la cohorte 3C présentaient un âge moyen de 74 ans, avec des âges d'entrée dans l'étude variables. Cette cohorte est majoritairement composée de femmes, représentant 60% de l'échantillon. Plus de 60% des participants ont un niveau d'éducation inférieur au baccalauréat. L'analyse génétique a révélé que 1778 participants de la cohorte portent le génotype APOE4. Au moment de l'inclusion, 10% des sujets présentaient un diagnostic de diabète.

Le suivi des participants s'étend sur une période allant de 0 à 18 ans, avec une médiane de suivi de 8,5 ans (sd=5.0). Au cours du suivi 1370 personnes ont été diagnostiquées avec une démence et 3561 sont décédées. La figure 2.3 illustre la trajectoire de la probabilité de survie au cours du suivi, montrant un risque de décès qui aboutit à une probabilité de survie inférieure à 50% après 15 ans.

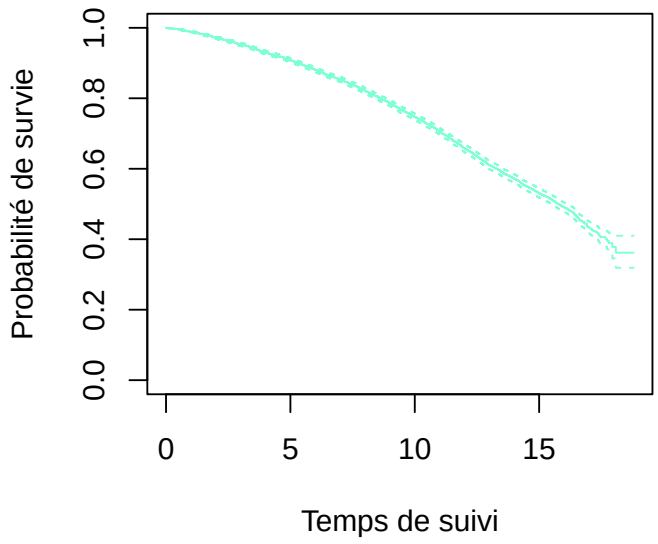


FIGURE 2.3 – Courbe de survie des 9294 participants de la cohorte 3C depuis l'entrée dans l'étude

FIGURE 2.4 – Trajectoires individuelles de six participants de la cohorte 3C en fonction de leur statut de démence pour : (a) évolution du score au test de fluence vers bale (IST), (b) évolution du score de symptomatologie dépressive (CES-D), (c) l'évolution du score de dépendance aux activités de la vie quotidienne

Le suivi au cours du temps de la cohorte permet d'explorer l'évolution de diverses mesures dans le temps. À titre d'exemple, les figures 2.4 (a),(b) et (c) représentent respectivement l'évolution des scores de fluence verbale ISAAC (IST), de symptomatologie dépressive (CES-D) et de dépendance fonctionnelle (IADL) en fonction du statut de démence, chez 6 participants de la cohorte 3C. Dans cette thèse, je vais étudier l'effet de certaines expositions comme le diabète, l'APOE4 ou le niveau d'éducation sur ces processus ainsi que leurs inter-relations.

Chapitre 3

Etat de l'art

Dans ce chapitre, je décris les méthodes statistiques sur lesquelles s'appuient mes travaux de thèse. Ce chapitre se divise en deux grandes sections. La première aborde la modélisation statistique des données longitudinales ainsi que celles basées sur des temps d'événement. La seconde partie présente les méthodes d'inférence causale en mettant un accent particulier sur deux méthodes importantes dans le cadre de cette thèse : l'analyse de médiation et l'utilisation de la méthode des variables instrumentales.

3.1 Modélisation statistique

"La modélisation statistique est le filtre qui transforme le chaos des données en un tableau clair et significatif, révélant ainsi les lois cachées de la nature." **George Box**

3.1.1 Modélisation des données longitudinales

Dans les études de cohorte, les participants sont suivis à plusieurs visites si bien que plusieurs données sont mesurées au cours du temps pour chaque participant. Cela definit le cadre des données longitudinales, également appelées données répetées. L'analyse de ces données permet de décrire leur évolution au cours du temps.

Notons Y la variable d'intérêt, que nous appellons également marqueur ou bien marqueur d'intérêt et Y_{ij} la valeur de Y pour l'individu i à la visite j, avec i allant de 1 à N, et j allant de 0 à n_i où N et n_i correspondent respectivement au nombre de sujets et au

nombre de visites pour sujet i.

Dans cette section, considérons la Figure 3.1 illustrant les trajectoires individuelles des taux de glycémie de quatre individus.

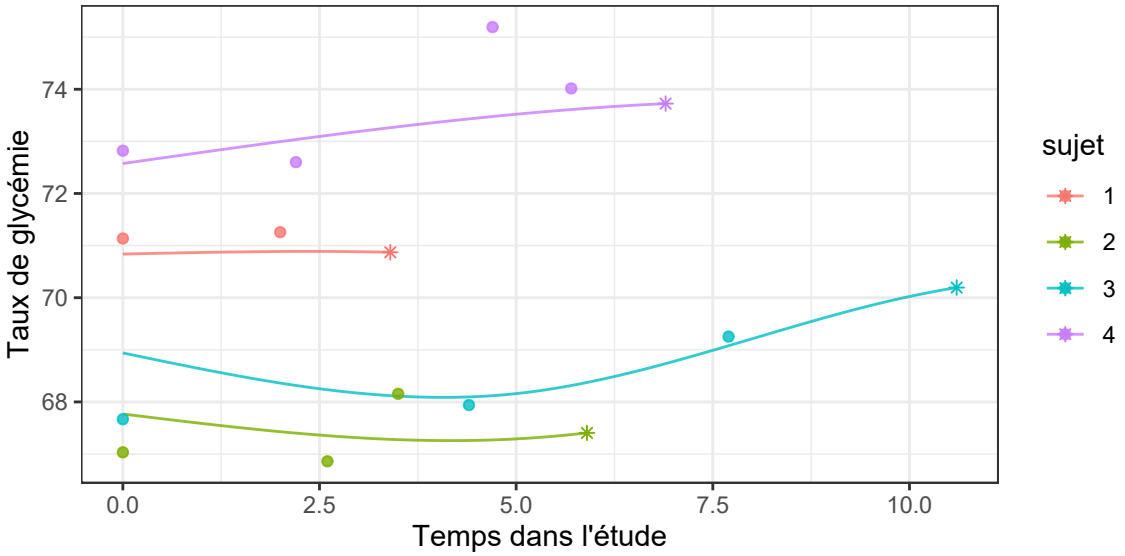


FIGURE 3.1 – Trajectoires individuelles simulées/fictives du taux de glycémie au cours du temps, chez quatre sujets

Si l'on s'intéresse qu'à un temps de visite, par exemple au taux de glycémie à baseline Y_{i0} (i.e. la visite 0), il est possible d'expliquer les variations du taux de glycémie selon des variables explicatives X_i, à l'aide d'une régression linéaire, telle que :

$$Y_{i0} = X_i^T \beta + \epsilon_i \tag{3.1}$$

où X_i représente les variables explicatives, β correspond au vecteur des coefficients de regression et ϵ est l'erreur de mesure indépendante, le plus souvent avec $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Cela implique différentes hypothèses, dont le fait que les observations Y_{i0} sont indépendantes. Cette hypothèses est plausible lorsque les unités sont des sujets. Cependant, cette hypothèse n'est plus valide si plusieurs mesures sont effectuées sur le même sujet. Par conséquent, l'utilisation de la régression linéaire limite l'analyse des données à un seul temps. Toutefois, considérer un seul temps ne permet pas d'étudier l'évolution de marqueurs d'intérêt. Ainsi, inclure des données répétées d'un sujet nécessite de tenir compte d'une possible corrélation entre les mesures d'un même sujet (i.e. corrélation intra-sujet).

Des méthodes d'analyse plus appropriées ont été considérées afin de mieux prendre en compte la structure complexe de ces données longitudinales, principalement les modèles à équations d'estimation généralisée (GEE) et les modèles mixtes. Dans ce travail nous détaillons uniquement les modèles mixtes.

3.1.1.1 Le modèle linéaire mixte

Présentation du modèle linéaire mixte

Les données observées pour les sujets (e.g. Figure 3.1) sont des données bruitées, mesurées avec un terme d'erreur ϵ_{ij}. Cette erreur peut être causée par l'instrument de mesure utilisé. La valeur observée pour chaque sujet se compose de la véritable valeur non observée à laquelle s'ajoute le terme d'erreur.

Les modèles linéaires mixtes ont été introduits par Laird and Ware (1982) pour analyser des variables d'intérêt gaussiennes longitudinales, en tenant compte d'une corrélation intra-sujet. Il s'agit d'une extension de la régression linéaire pour des données groupées, c'est-à-dire quand chaque unité ou individu a plusieurs données.

Notons Y_{ij} la valeur observée de la variable d'intérêt Y pour l'individu i (avec $i = 1, ..., N$) à la visite j. Le modèle linéaire mixte s'écrit de la manière suivante :

$$Y_{ij} = X_{ij}^T \beta + Z_{ij}^T b_i + \epsilon_{ij}$$

où X_{ij} est le vecteur des variables explicatives du sujet i à la visite j et β le vecteur de coefficients associé aux variables explicatives. Z_{ij} est un sous-vecteur des variables explicatives X_{ij} du sujet i à la visite j et b_i correspond à des paramètres individuels qui traduisent une variation aléatoire d'un individu à l'autre autour des coefficients β. Dans la suite, $b_i \sim \mathcal{N}(0, B)$ où B est la matrice de variance-covariance de dimension $q \times q$. L'erreur de mesure $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$.

Ce modèle est décomposé en deux partie : une partie populationnelle et une partie individuelle.

$$Y_{ij} = \underbrace{X_{ij}^T \beta}_{\text{Partie populationnelle}} + \underbrace{Z_{ij}^T b_i}_{\text{Partie individuelle}} + \epsilon_{ij}$$

Les coefficients β représentent les effets fixes du modèle, fonctionnant de manière similaire à une régression linéaire classique. Ils servent à décrire l'évolution moyenne de la variable Y. En revanche, les termes b_i représentent les effets aléatoires propres à chaque sujet. Leur rôle est de modéliser l'écart individuel par rapport à la moyenne de Y. Ils permettent de prendre en considération la corrélation intra-sujet.

La variance des observations reflète la variabilité individuelle au sein d'un groupe d'individus. Dans ce modèle, la variabilité est divisée en deux parties : la variabilité due aux effets aléatoires (représentée par $Z_i B Z_i^\top$, avec Z_i la matrice de ligne j, Z_{ij}) et la variabilité due aux erreurs de mesure (représentée par $\Sigma_i = \sigma^2 In_i$). Soit $Y_i = (Y_{i1}, \ldots, Y_{ij}, \ldots, Y_{ini})$ le vecteur des observations pour l'individu i. La variance de Y_i est définie telle que :

$$\text{var}(Y_i) = V_i = Z_i B Z_i^\top + \Sigma_i$$

Prenons l'exemple le plus simple d'une trajectoire linéaire. Dans le contexte de l'étude du vieillissement, les effets aléatoires dans les modèles sont souvent sur l'intercept et sur la pente mesurant l'évolution en fonction du temps. En effet, le niveau initial de la variable d'intérêt peut varier significativement d'un individu à un autre, on considère alors b_{0i} l'écart individuel à l'intercept et b_{1i} l'écart individuel à la pente. Par exemple, les valeurs initiales du taux de glycémie à l'inclusion peuvent différer entre les sujets. De plus, la pente de la variable d'intérêt peut également varier individuellement. L'évolution du taux de glycémie au cours du temps pourrait être plus ou moins marquée d'un individu à l'autre, voire positive pour un sujet tandis qu'elle serait négative pour un autre.
Considérons Y_{ij} le taux de glycémie de l'individu i à la visite j. Le temps d'observation de la variable Y_{ij} est représenté par t_{ij}. Le modèle peut être formulé tel que :

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \beta_2 t_{ij} + \beta_3 X_{ij} \times t_{ij} + b_{0i} + b_{1i} t_{ij} + \epsilon_{ij}$$

Dans ce modèle, β_0 correspond à l'intercept. X_{ij} représente l'indice de masse corporel de l'individu i à la visite j et β_1 est son coefficient de régression associé. t_{ij} correspond au temps d'observation de la variable Y_{ij} et β_2 est son coefficient associé. L'interaction entre l'indice de masse corporel et le temps est associé au coefficient β_3 . Les termes b_{0i} et b_{1i} représentent respectivement les effets aléatoires de l'intercept et de la pente pour l'individu i. Le vecteur des effets aléatoires b_i suit une distribution normale multivariée $\mathcal{N}(0, B = \begin{smallmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{smallmatrix})$, où σ_0^2 représente la variance de l'intercept aléatoire, σ_{01} la covariance entre l'intercept et la pente, et σ_1^2 la variance de la pente aléatoire.

Estimation du modèle linéaire mixte

Pour estimer les paramètres d'un modèle statistique, nous pouvons utiliser la méthode du maximum de vraisemblance (i.e. maximum likelihood (ML)). L'objectif du maximum de vraisemblance est de trouver les valeurs des paramètres qui maximisent la probabilité d'observer les données réelles étant donné le modèle statistique. Cette approche cherche à ajuster le modèle de manière à rendre les données observées les plus probables, selon les paramètres du modèle.

Considérons les paramètres (β, ϕ) d'un modèle mixte, où β est le vecteur des coefficients de regression des effets fixes et ϕ est le vecteur des paramètres de covariance intervant dans la variance des observations V_i, avec $\phi = (var(B), var(\sigma^2))$. La log-vraisemblance est définie par :

$$L(\beta, \phi) = -\frac{1}{2} \sum_{i=1}^{N} \left\{ n_i \log(2\pi) + \log |V_i(\phi)| + (Y_i - X_i\beta)^\top V_i(\phi)^{-1} (Y_i - X_i\beta) \right\}$$

où $|V_i(\phi)|$ est le déterminant de la matrice $V_i(\phi)$ et X_i est la matrice vecteur ligne X_{ij}^T.

La plupart des logiciels font une optimisation numérique de $\theta = (\beta, \phi)$ via un algorithme itératif.

Dans ce travail, la maximisation se fait en utilisant l'algorithme de Marquardt-Levenberg (Levenberg (1944) ; Marquardt (1963)) implémenté sous R dans le package **marqlevalg** (Philipps et al. (2021)). Les variances sont, elles, estimées à partir de l'inverse de la matrice hessienne observée.

Avantages du modèle linéaire mixte

Le modèle linéaire mixte permet de traiter les données longitudinales en y intégrant la dépendance entre les observations répétées d'un même sujet. En permettant la modélisation de la variabilité individuelle, il offre une représentation précise des disparités entre les individus. Sa capacité à gérer efficacement les données manquantes renforce sa pertinence dans des contextes où des observations peuvent ne pas être disponibles à tous les instants (e.g. Figure 3.1). En effet, comme toute estimation par ML, les estimations du modèle mixte sont robustes aux données manquantes tant que le processus de données manquantes peut être prédit par les observations. De plus, les modèles linéaires mixtes ne requièrent pas que les observations des sujets soient effectuées simultanément ou que l'intervalle entre deux mesures soit régulier. Un avantage supplémentaire des modèles mixtes est qu'ils permettent l'analyse de toutes les données répétées, quel que soit le moment dans le temps. Ceci s'explique par le fait que le modèle peut être formulé de la manière suivante :

$$Y_{ij} = Y^*(t_{ij}) + \epsilon_{ij} = X^T(t_{ij})\beta + Z^T(t_{ij})b_i + \epsilon_{ij}$$

où Y_{ij} est la valeur observée et définie à partir de la vraie valeur sous jacente $Y^*(t_{ij})$. $Y^*(t_{ij})$ est ainsi défini en tout temps $t \in R$

De nombreuses extensions ont été développées afin de permettre une plus grande flexibilité du modèle, notamment les modèles mixtes généralisés (Clayton et al. (1996)) qui reposent sur les mêmes principes que les modèles linéaires mixtes. Des modèles dit curvilinéaires qui incluent dans le modèle mixte une normalisation du Y estimée en même temps que les autres paramètres du modèle. Cela permet de traiter des marqueurs continus qui ne suivent pas forcément une loi normale. Ont également été proposés des modèles mixtes multivariés, pour prendre en compte simultanément plusieurs marqueurs.

3.1.2 Modélisation conjointe de plusieurs marqueurs

Dans les études de cohortes longitudinales, différentes variables sont collectées à chaque visite. Ces variables peuvent être corrélées entre elles. Leur modélisation simultanée permet d'analyser simultanément leurs évolutions, leurs déterminants et leur structure de dépendance. Par exemple, il pourrait être intéressant d'étudier l'évolution simultanée des taux de glycémie et de la pression artérielle ou bien du cholestérol. Dans les recherches sur le virus de l'immunodéficience humaine (VIH), beaucoup d'études se sont interessées aux évolutions simultanées des cellules CD4 et de la charge virale (Thiébaut et al. (2002)). Dans cette section, nous présentons différentes méthodes destinées à traiter ces données répétées et multivariées en nous appuyant sur l'exemple du taux de glycémie et en intégrant un second marqueur le taux de cholestérol.

3.1.2.1 Modèle mixte multivarié

Le modèle mixte multivarié (Sy et al. (1997)) est une extension directe du modèle mixte pour un marqueur. Il permet d'étudier plusieurs variables, possiblement corrélées, en les modélisant simultanément. L'analyse conjointe de leur trajectoire permet d'approfondir la compréhension de leurs interrelations.

Soit K le nombre de marqueurs longitudinaux. Pour des raisons de clarté, nous détaillons la modélisation de la corrélation entre deux marqueurs uniquement ($K = 2$). Cependant la procédure est la même pour un nombre plus important de marqueurs. Considérons Y_1 (e.g. taux de glycémie) et Y_2 (e.g taux de cholestérol), deux marqueurs longitudinaux. On note Y_{ki} $(k = 1, ..., K)$ le vecteur de taille n_{ki} des observations du $k^{\text{ème}}$ marqueur du sujet i aux temps de mesure $(t_{kij})_{j=1,...,n_{ik}}$, avec t_{kij} pouvant être différent de $t_{k'ij}$. Un modèle linéaire mixte est utilisé pour modéliser l'évolution de chaque marqueur.

$$Y_{1ij} = X_{1ij}^T \beta_1 + Z_{1ij}^T b_{1i} + \epsilon_{1ij}$$
$$Y_{2ij} = X_{2ij}^T \beta_2 + Z_{2ij}^T b_{2i} + \epsilon_{2ij}$$

avec X_{kij}, β_k, Z_{kij}, b_{ki} et ϵ_{kij} sont spécifiques au marqueur k et sont définis de la même façon que dans la section 3.1.1.1.

La possibilité de corrélation entre les marqueurs est introduite grâce à l'inclusion d'une corrélation entre les effets aléatoires b_{ki}, à l'aide d'une distribution jointe de tous les effets aléatoires, telle que : $b_i = \begin{bmatrix} b_{1i} \\ b_{2i} \end{bmatrix} \sim N(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} B_1 & B_{12} \\ B_{12}^T & B_2 \end{bmatrix})$.

Soit $Y_i = \begin{bmatrix} Y_{1i} \\ Y_{2i} \end{bmatrix}$, un modèle multivarié peut être réécrit simplement comme un modèle mixte classique en définissant :

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}, X_i = \begin{bmatrix} X_{1i} & \mathbf{0} \\ \mathbf{0}^T & X_{2i} \end{bmatrix}, Z_i = \begin{bmatrix} Z_{1i} & \mathbf{0} \\ \mathbf{0}^T & Z_{2i} \end{bmatrix}, b_i = \begin{bmatrix} b_{1i} \\ b_{2i} \end{bmatrix}, \text{et } \epsilon_i = \begin{bmatrix} \epsilon_{1i} \\ \epsilon_{2i} \end{bmatrix},$$

où X_{1i} est la matrice avec X_{1ij} en vecteur ligne, X_{2i} est la matrice avec X_{2ij} en vecteur ligne. De la même manière Z_{1i} est la matrice avec Z_{1ij} en vecteur ligne et Z_{2i} est la matrice avec Z_{2ij} en vecteur ligne. ϵ_{1i} et ϵ_{2i} correspondent respectivement aux vecteurs de ϵ_{1ij} et ϵ_{2ij}.

Alors le modèle mixte multivarié s'écrit tel que :

$$Y_i = X_i\beta + Z_i b_i + \epsilon_i$$

avec $\Sigma_i = \begin{bmatrix} \Sigma_{1i} & \Sigma_{12i} \\ \Sigma_{12i}^T & \Sigma_{2i} \end{bmatrix}$.

L'estimation du modèle mixte multivarié est alors similaire à celle du modèle mixte univarié, expliquée en section 3.1.1.1.

Le modèle mixte multivarié permet de tenir compte de la corrélation pour évaluer l'effet des variables explicatives et décrire les trajectoires. Néammoins, ce modèle ne permet pas d'étudier comment les marqueurs s'influencent les uns sur les autres.

3.1.2.2 Modèle mixte multivarié avec influence d'un marqueur sur l'autre

Une alternative au modèle mixte multivarié classique, élaborée par van Oudenhoven et al. (2022), est de modéliser la dépendance entre les marqueurs non pas en corrélant directement les effets aléatoires, mais en introduisant l'influence d'un marqueur sur un autre. Plus précisément, un modèle linéaire mixte est réalisé pour chaque marqueur afin d'estimer sa trajectoire, en incluant le prédicteur linéaire du premier marqueur comme une covariable dans le second modèle, tel que :

$$
\begin{aligned}
Y_{1ij} &= \underbrace{X_{1ij}\beta_1 + Z_{1ij}b_{1ij}}_{Y_1^*(t_{ij})} + \epsilon_{1ij} \\
Y_{2ij} &= \underbrace{X_{2ij}\beta_2 + Z_{2ij}b_{2i} + Y_1^*(t_{2ij})\xi}_{Y_2^*(t_{2ij})} + \epsilon_{2ij}
\end{aligned}
$$

avec X_{kij}, β_k, Z_{kij}, b_{ki} et ϵ_{kij} qui sont spécifiques au marqueur k et sont définis de la même façon que dans la section 3.1.1.1.

Contrairement au modèle mixte multivarié, les b_{ki} sont ici indépendants : la corrélation entre Y_1 et Y_2 est captée par l'influence de Y_1 sur Y_2, où le paramètre ξ quantifie la variation de la valeur prédite de $Y_2^*(t)$ dûe à une augmentation d'une unité de la valeur prédite de $Y_1^*(t)$.

3.1.2.3 Modèle mixte à équations différentielles

Les modèles mécanistes permettent d'établir des liens entre un ensemble de processus au cours du temps à l'aide d'équations différentielles. Les systèmes d'équations différentielles sont largement utilisés en physique, en ingénierie et en épidémiologie. Ces systèmes sont composés d'équations qui décrivent comment une quantité change en fonction d'autres variables, mettant en évidence les relations dynamiques entre les éléments d'un système au cours du temps (Prague et al. (2013)).

Classiquement, une équation différentielle ordinaire de premier ordre peut être écrite

de la manière suivante :

$$\frac{\delta_y}{\delta_t} = f(t, y)$$

Cette équation permet de décrire comment le processus $y(t)$ varie en fonction du temps t. $\frac{\delta_y}{\delta_t}$ représente la dérivée de y par rapport à t, et $f(t, y)$ est une fonction donnant la variation instantanée de y par rapport à t et y.

Pour modéliser l'influence entre plusieurs processus mesurés de façon répétée au cours du temps, différentes méthodes ont été proposées. Dans cette thèse, nous avons considéré le travail de Taddé et al. (2020). Ils proposent de mêler les modèles mixtes et les équations différentielles pour quantifier les influences temporelles entre un ensemble de processus latents mesurés à l'aide de données répétées de marqueurs.
En repartant de nos deux marqueurs, le taux de glycémie et le taux de cholestérol, nous notons Y_{1ij} pour j allant de 1 à n_{1i} et Y_{2ij} pour j allant de 1 à n_{2i}, les mesures répétées observées aux temps t_{1ij} et t_{2ij} des deux processus sous-jacent $(Y^*_{1i}(t))_{t\in R}$ et $(Y^*_{2i}(t))_{t\in R}$. Le modèle statistique définit le niveau sous-jacent comme dans le modèle mixte classique :

$$Y_{1ij} = Y^*_{1i}(t_{1ij}) + \epsilon_{1ij}$$
$$Y_{2ij} = Y^*_{2i}(t_{2ij}) + \epsilon_{2ij}$$

Simultanément, les trajectoires des deux processus sont définies par une équation différentielle en spécifiant un niveau initial à l'instant 0 et le changement instantané au fil du temps, tous deux modélisés dans le cadre des modèles mixtes en utilisant des effets aléatoires. Les dépendances temporelles entre les processus sont modélisées par l'effet du niveau actuel d'un processus sur le changement instantané de l'autre. Le modèle peut être formulé comme suit :

$$Y^*_1(t) : \begin{cases} Y_{1i}(0) = \mathbf{X}_i^{Y_1^0 T}\beta^{Y_1} + u_i^{Y_1} \\ \dfrac{\partial Y_{1i}(t)}{\partial t} = \mathbf{X}_i^{Y_1^t T}(t)\gamma^{Y_1} + \mathbf{Z}_i^{Y_1 T}(t)\mathbf{v}_i^{Y_1} + \alpha_{11}(t)Y_{1i}(t) + \alpha_{12}(t)Y_{2i}(t) \end{cases}$$

$$Y_2^*(t) : \begin{cases} Y_{2i}(0) = \mathbf{X}_i^{Y_2(0)T}\beta^{Y_2} + u_i^{Y_2} \\ \dfrac{\partial Y_{2i}(t)}{\partial t} = \mathbf{X}_i^{Y_2^t T}(t)\gamma^{Y_2} + \mathbf{Z}_i^{Y_2 T}(t)\mathbf{v}_i^{Y_2} + \alpha_{21}(t)Y_{1i}(t) + \alpha_{22}(t)Y_{2i}(t) \end{cases}$$

où $\mathbf{X}_i^{Y_k^0}$ pour $k = (1, 2)$ est le vecteur de covariables associées aux niveaux initiaux des processus à travers le paramètre $\boldsymbol{\beta}_k^Y$. Ce vecteur inclut l'intercept, des variables explicatives fixes dans le temps, par exemple l'exposition X et les facteurs de confusion $\mathbf{C}$ à l'inclusion. Le vecteur $\mathbf{X}_i^{Y_k(t)}$, inclut les covariables associées aux changements au cours du temps des processus à travers le vecteur de paramètres $\boldsymbol{\gamma}_k^Y$. Il comprend l'intercept, les fonctions temporelles (en cas de changement non linéaire au cours du temps), des variables explicatives comme une exposition d'intérêt, des facteurs de confusion et les éventuelles interactions avec les fonctions temporelles. Le vecteur $\mathbf{Z}_i^Y{}_k(t)$ inclut l'intercept et éventuellement des fonctions temporelles. Il est associé aux effets aléatoires individuels $\mathbf{v}_i^Y{}_k$, pour tenir compte de la corrélation intra-individuelle. L'ensemble des effets aléatoires $({u_i^Y}_{k=1,..,K})$ et $({v_i^Y}_{k=1,..,K})$, suivent une distribution multivariée normale telle que $\nu_i = \begin{bmatrix} u_i \\ v_i \end{bmatrix} \sim N(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} G_u & \\ G_{uv} & G_v \end{bmatrix})$.

L'effet des processus les uns sur les autres est capté par les paramètres α_{pq} pour l'effet du processus p sur le changement instantané du processus q. Ce modèle permet de prendre en compte comment l'histoire d'un processus joue sur le niveau d'un autre.

L'estimation des modèles mixtes à équations différentielles est complexe. Des techniques ont été développées, principalement en pharmacométrie (e.g. avec Monolix, Prague et al. (2019)). Dans Taddé et al. (2020), ils ont proposé de s'appuyer sur des approximations des équations différentielles par des équations de différence avec un pas petit (indépendant du schéma d'observation) pour obtenir une écriture analytique de la distribution des Y et obtenir l'estimation du modèle mixte de façon classique par la maximisation de la vraisemblance ainsi trouvée par un algorithme d'optimisation, comme Marquardt-Levenberg (Levenberg (1944) ; Marquardt (1963)).

3.1.3 Modélisation des données d'événement

Dans la section précédente, nous avons dressé un aperçu non-exhaustif des modèles statistiques utilisés pour analyser les données longitudinales, en illustrant cela avec un exemple de données répétées (taux de glycémie / taux de cholestérol). À présent, intérressons nous au temps d'apparition d'un événement clinique, noté T_i, par exemple le temps jusqu'au diagnostic de démence.

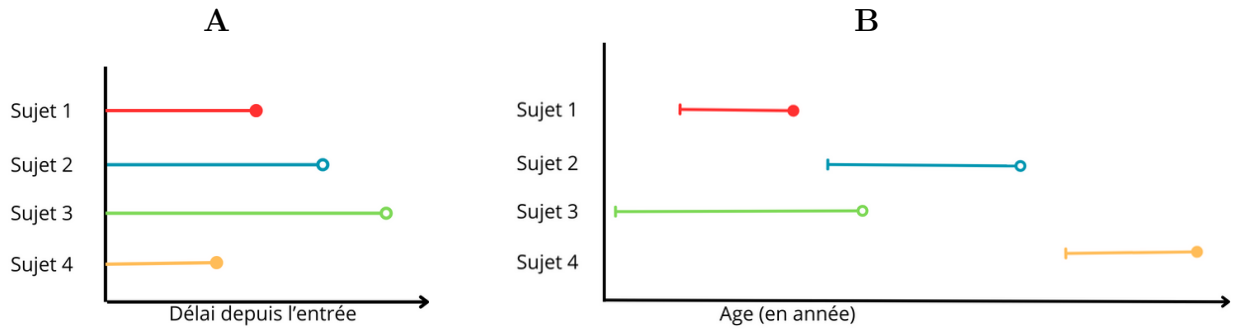


FIGURE 3.2 – Durée de suivi en années de quatre sujets (• symbolise la survenue de démence / ∘ symbolise la censure à droite) (**A**) lorsque le temps d'intérêt est le délai depuis l'inclusion (**B**) lorsque le temps d'intérêt est l'âge à l'entrée dans l'étude.

L'analyse des temps d'événement se concentre sur la modélisation du temps entre différents états. Ce temps peut se présenter sous différentes formes : délai depuis un diagnostic, la durée depuis l'inclusion dans une cohorte, l'âge au moment d'un événement, ou encore une période de temps. La Figure 3.2 illustre deux types de temps d'événement couramment utilisés dans les analyses de temps d'événement dans les études de cohorte sur le vieillissement : le délai depuis l'inclusion (Figure 3.2 (A)) et l'âge (Figure 3.2 (B)).

Lors de l'analyse des temps d'événement, différents mécanismes interviennent, empêchant d'observer le temps de passage entre deux états pour tous les individus (e.g. absence de démence *versus* survenue de démence).

Censure

La censure est une notion fréquemment rencontrée dans les analyses de survie axées sur le vieillissement. Deux types de censure sont particulièrement observés dans ces analyses : la censure à droite et la censure par intervalle.

Censure à droite : Dans les études de cohorte, il est courant que certains indi-

vidus n'aient pas encore expérimenté l'événement d'intérêt à leur dernière visite. Cette situation peut se produire soit parce que le suivi a pris fin comme prévu, soit parce que les participants ont quitté prématurément l'étude, ce que l'on appelle une sortie d'étude. Lorsqu'un individu n'a pas encore rencontré l'événement à la date de sa dernière visite, la variable T est considérée comme censurée à droite. L'information sur T est partielle, on sait uniquement que T est plus grand que le temps de censure. Graphiquement, cette situation est représentée par le symbole $\circ$ sur la Figure 3.2.

Censure par intervalles : Les cohortes sur le vieillissement se basent sur des suivis au cours desquels les participants sont évalués. Ainsi en fonction de la nature de l'événement, on peut être confronté à de la censure par intervalle. Cette situation se produit lorsque le moment précis de l'événement n'est pas connu de façon exacte ; nous avons simplement l'information qu'il s'est produit entre deux visites, sans précision quant à l'exactitude de sa survenue. Par exemple, dans la cohorte 3C, si l'événement étudié est le décès, il n'y aura pas de censure par intervalle car nous disposons de la date précise du décès grâce au registre des décès. En revanche, ce n'est pas le cas si l'on s'intéresse à la survenue de la démence, qui est diagnostiquée aux visites de suivi.

En reprenant la Figure 3.2 (A), intéressons nous à la survenue de la démence chez le sujet 1 (Figure 3.3).

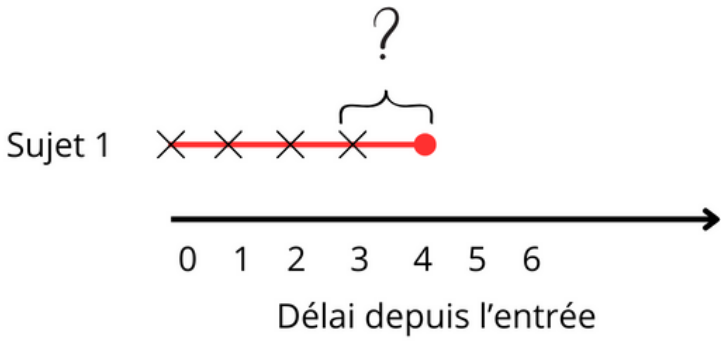


FIGURE 3.3 – Statut de la démence du sujet 1 depuis son entrée dans l'étude

Cette illustration démontre que le sujet 1 ne présentait pas de signes de démence à son inclusion dans l'étude. Aucun diagnostic de démence n'a été posé aux visites 1, 2 et 3, mais un diagnostic a été établi à la visite 4. Toutefois, la démence ne s'est pas manifestée à la visite 4, elle s'est déclarée entre deux visites, entre la visite 3 et la visite 4, sans pouvoir déterminer le moment exact. Le temps d'événement est alors dit "censuré par intervalles".

On peut alors définir un intervalle T_L, T_U observé tel que $T_L < T < T_U$, avec T_L et T_U représentant respectivement les temps de visite précédant le diagnostic de démence et le moment du diagnostic de démence.

Troncature à gauche

La troncature à gauche est un phénomène courant dans les études de cohortes. Elle survient lorsque les participants ne rentrent pas dans l'étude au même moment, ce qui est le cas des études ne débutant pas dès la naissance.
Si la population d'étude est correctement définie et si le temps étudié correspond au délai depuis l'inclusion alors cela ne pose pas de problème. Le problème intervient lorsque le délai d'entrée est différent d'un sujet à l'autre. C'est le cas lorsque le temps étudié est l'âge.

Les sujets sont selectionnés selon qu'ils ont survécu jusqu'à un temps spécifique noté T_0, ce qui signifie que $T > T_0$, avec T la variable aléatoire du temps d'intérêt. Dans une cohorte, comme dans la cohorte 3C qui inclut toute personne de 65 ans et plus en 1999, on exclut de fait toute personne de 65 ans et plus qui serait décédé ou démente avant 1999. Une personne incluse à 80 ans a donc nécessairement survécu jusqu'à 80 ans, à la différence de la censure qui induit une information partielle, la troncature entraîne une sélection de l'échantillon.

3.1.3.1 Modèle de survie pour un unique temps d'événement

En analyse de survie, on modélise généralement le temps de survenue d'un événement censuré à droite par le biais du risque instantané de l'événement, noté $\lambda(t)$ qui permet d'estimer la probabilité qu'un événement survienne à un moment donné pour les individus qui ont survécu jusqu'à ce moment.

Il existe deux façons principales de modéliser le risque instantané d'intérêt, soit par une approche multiplicative et soit par une approche additive.

Modèle de Cox à risques proportionnels

Le modèle à risques proportionnels de Cox (Cox (1972)) est un des modèles de référence dans les analyses de survie en épidémiologie. Ce modèle fait l'hypothèse que la variable

explicative augmente le risque de l'événement de façon multiplicative.

Mathématiquement, le modèle de Cox s'écrit :

$$\lambda_i(t) = \lambda_0(t) \exp(X_i^T \beta)$$

avec $\lambda_i(t)$ le risque instantané du sujet i, $\lambda_0(t)$ la fonction de risque instantané de base (i.e. lorsque les variables explicatives du sujet i sont nulles), X_i le vecteur des variables explicatives du sujet i et β le vecteur des coefficients associés.

La popularité du modèle de Cox est attribuable à deux aspects. Premièrement son interprétation est simple. L'effet d'une variable explicative est quantifié par le changement du risque instantané pour l'augmentation d'une unité pour une variable X donnée. Ce changement est appelé le risque relatif (RR). Il s'écrit tel que :

$$\frac{\lambda(t|X = x + 1)}{\lambda(t|X = x)} = \exp(\beta)$$

Deuxièmement il s'agit d'une approche semi-paramétrique, c'est-à-dire que seul l'effet des variables explicatives est modélisé, $\lambda_0(t)$ reste non spécifié. Cela est possible grâce à la technique d'estimation par vraisemblance partielle dans laquelle $\lambda_0(t)$ peut disparaître.

Le modèle de Cox repose cependant sur l'hypothèse majeure de proportionnalité des risques, c'est-à-dire que le changement de risque instantané pour une augmentation de variable explicative est constant dans le temps. Cette hypothèse forte est à vérifier car elle peut ne pas être vérifiée dans certaines études de survie et peut dans ce cas biaiser les résultats (Kragh Andersen et al. (2021)).

Modèle d'Aalen à risques additifs

Une alternative au modèle de Cox est de modéliser l'effet de la variable explicative avec un risque additif plutôt que multicatif. Nous nous intéressons ici au modèle additif proposé par Aalen (1989).

La fonction de risque s'écrit pour le sujet i comme :

$$\lambda_i(t) = \beta_0(t) + \beta_1 X_{1i} + ... + \beta_p X_{pi}$$

Avec $\beta_0(t)$ la fonction de risque de base, et les β_j $(j : 1, ..., p)$ représentent l'augmentation du risque à l'instant t pour une augmentation d'une unité de la variable explicative X_{ij}).

Ce modèle est nettement moins utilisé que le modèle de Cox, dû à sa complexité d'interprétation. En effet il permet de mesurer l'effet des covariables sur une échelle absolue et non plus de manière relative comme pour le modèle de Cox. Pourtant il a l'avantage de supposer la linéarité par rapport aux covariables, facilitant la détection des variations de coefficient à chaque instant de survie (Henderson and Milner (1991)). De plus, la forme linéaire de ce modèle représente un atout majeur en simplifiant les calculs lorsqu'on souhaite obtenir des quantités à partir de ces modèles, ce qui est particulièrement intéressant en causalité.

3.1.3.2 Modèle de survie avec plusieurs évènements en compétition

Les méthodes évoquées précédemment sont conçues pour modéliser un seul temps d'événement. Cependant, dans le contexte des études sur le vieillissement, il peut être intéressant voire nécessaire de modéliser simultanément plusieurs temps d'événements d'intérêt. Par exemple, lors de l'étude de la démence chez les personnes âgées, le risque de décès est non négligeable. Ne pas prendre en compte simultanément ces deux événements pourrait induire un biais dans l'analyse, car cela ne tient pas compte du fait qu'un individu peut décéder avant de développer une démence, ce qui empêche alors l'observation de la démence chez cet individu. Dans ces cas, la méthode d'analyse pour le risque de démence doit prendre en compte que les participants peuvent mourir avant de développer une démence.

Différentes techniques permettent de modéliser deux événements en compétition simultanément, tels que la démence et le décès. La technique du modèle de Cox peut être exploitée dans le cas d'événement compétitif. En effet, le risque instantané d'événement peut être traité de façon similaire même en présence de plusieurs événements. Seulement,

les différents risques instantanés de chaque cause doivent être combinés pour déduire la probabilité d'événement et doivent tenir compte du fait que l'autre événement n'a pas eu lieu.

Un autre modèle a été proposé pour directement modéliser la probabilité d'événement en présence d'événement compétitif. Il s'agit du modèle de Fine and Gray (1999) qui inclut les variables explicatives dans les risques instantanés de sous-distribution. Nous ne détaillons pas plus cette technique qui ne sera pas utilisée dans cette thèse.

3.1.3.3 Modèle pour une séquence d'événement

L'approche par risque en compétition ne traite que le cas de deux événements mutuellement exclusifs. Dans le cas de la démence, il s'agit de la démence et du décès avant la démence.
Pour traiter l'ensemble de l'histoire du vieillissement, il peut être pertinent de modéliser plus généralement le décès avant ou après démence. Cela peut être résalisée par les modèles multi-états.

Un modèle multi-états est composé par des états et des transitions entre chaque état. Ce modèle consiste à analyser le risque intantané de transition d'un état à l'autre (appelé intensités de transition). Sur la Figure 3.4 les états sont représentés par les rectangles, avec trois états : sain, démence et décès. Les transitions sont elles représentées par des flèches, avec les intensités de transitions α_{01}, α_{12},α_{02}.

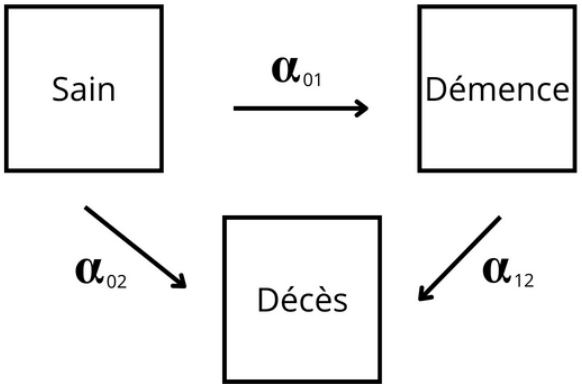


FIGURE 3.4 – Modèle multi-états Sain-Dément-Décédé

L'intensité α_{01} représente le risque instantané de passer de l'état sain à l'état dément,

α_{02} représente le risque instantané de passer de l'état sain à l'état décédé et α_{12} est le risque instantané de passer de l'état dément à l'état décédé.

Plus spécifiquement, le modèle représenté sur la Figure 3.4 est également connu sous le nom de modèle "illness-death" irréversible (i.e. sans rétablissement possible). Dans ce modèle un sujet peut transiter de l'état sain à l'état décédé directement ou via l'état malade, réprésenté ici par la démence. Le modèle n'est pas réversible car il suppose qu'on ne peut pas guérir de la maladie (i.e. passer de la maladie à sain).

Considérons l'évolution d'un système à travers différents états en définissant le processus à temps continu $X = X(t)_{t \in R^+}$, à espace d'états fini dans $\{0, 1, 2\}$. Notons k et l deux états différents et s et t deux temps où $s < t$. Le processus peut être caractérisé par des probabilités de transition entre les différents états (e.g., entre les états k et l) entre deux temps (e.g. temps s et t) :

$$p_{kl}(s,t) = P\left(X(t) = l \mid X(s) = k, H_s^-\right)$$

$p_{kl}(s,t)$ représente la probabilité conditionnelle que le processus $X(t)$ soit dans l'état l à l'instant t, sachant qu'il était dans l'état k à l'instant s et que H_s^- représente l'histoire du processus jusqu'à l'instant s.

Les modèles multi-états peuvent être estimés de la même façon que les risques des modèles vus précédement en modélisant séparement le risque de chaque transition par un modèle de Cox. On peut ensuite déduire les probabilités d'être dans chaque état à un temps donné.

En fonction du contexte, le processus X peut reposer sur une hypothèse semi-markovienne ou markovienne. L'hypothèse la plus fréquente est l'hypothèse Markovienne. Elle stipule que les probabilités futures dépendent uniquement de l'état de X au temps présent et non du passé (Hougaard (1999)). On parle de processus homogène si les probabilités sont indépendantes du temps aussi ou de processus non homogène si elles peuvent varier avec le temps. Toutefois cette hypothèse n'est pas réellement adaptée pour les applications au vieillissement. Par exemple la probabilité de décéder après une démence peut aussi

dépendre du temps dans l'état de démence. Dans ce cas, le processus X est dit semi-markovien. C'est-à-dire que le futur du processus dépend non seulement de l'état actuel du processus, mais aussi du temps écoulé depuis l'entrée dans cet état (Pérez-Ocón et al. (1999)).

Les différents modèles statistiques présentés dans cette section permettent d'analyser des données longitudinales et des temps d'événements dans les études de cohortes, pour établir des associations entre des facteurs de risque et des événements d'intérêt. Cependant, ces associations ne peuvent pas nécessairement être interprétées de manière causale. L'interprétation causale des associations identifiées nécessite une approche méthodologique adaptée.

3.2 Inférence causale

"La causalité est le fondement de la compréhension et de la prédiction dans le monde complexe qui nous entoure." - ***Judea Pearl***

L'inférence causale est une notion fondamentale dans le processus de compréhension d'un événement. Elle permet la reconstitution des liens causaux entre divers événements, répondant ainsi à la question du "Pourquoi" : *Pourquoi cet événement s'est-il produit ?*. Identifier les mécanismes sous-jacents à la survenue d'un événement est essentiel pour pouvoir intervenir sur la ou les causes.

En recherche biomédicale, la référence pour établir qu'une variable en entraîne une autre est l'essai clinique randomisé. Considérons un exemple où l'on souhaite établir la relation entre un traitement (variable binaire : traitement ou placebo) et la réduction des symptômes d'une maladie. Pour ce faire, on constitue deux groupes de manière aléatoire parmi les participants d'une étude : un groupe qui recevra le traitement et un autre qui aura le placebo. La présence de symptômes est mesurée au cours du temps, permettant ainsi de comparer l'évolution des symptômes entre les deux groupes. La randomisation, c'est-à-dire l'attribution aléatoire des participants à l'un des groupes, permet d'avoir des groupes comparables sur toutes les caractérisques, à l'exception du traitement (intervention) étudié.

Comme vu en introduction, l'épidémiologie du vieillissement est surtout basée sur des études de cohorte et non des essais cliniques randomisés. Cela induit des complexités pour étudier les causes des maladies. Cette partie vise alors à définir la causalité, et répertorier les méthodes statistiques actuelles sur lesquelles nous nous sommes appuyées pour étudier l'inférence causale dans les études de cohortes.

3.2.1 Association et causalité

Établir la relation entre plusieurs variables représente une démarche cruciale dans le domaine de la recherche en Santé Publique. Il existe deux types de liens majeurs : l'association et la causalité. Bien que ces deux relations présentent des similitudes et puissent parfois être confondues, il est essentiel de les distinguer, car elles conduisent à des conclusions différentes. Une association se manifeste lorsque deux variables présentent une corrélation mutuelle, c'est-à-dire lorsque les variations d'une variable sont liées aux variations d'une autre variable, cela pouvant être le résultat de divers mécanismes.
Il est important de noter que la simple présence d'une association ne garantit pas nécessairement une relation causale. Pour qu'une association soit qualifiée de "causale", il faut que la variation de l'une des deux variables soit **directement responsable** de la variation de l'autre, établissant ainsi un lien de cause à effet. Cette condition est respectée dans les essais cliniques grâce à la randomisation, où seule l'intervention sur le traitement diffère. Ainsi, toutes associations démontrées dans ce contexte peut être qualifiées de causales.

La distinction entre association et causalité est souvent illustrée par des exemples paradoxaux, appelés paradoxes de Simpson (Hernán et al. (2011)). L'un de ces paradoxes est représenté par la légende alsacienne selon laquelle "les cigognes apportent les bébés". Des statisticiens ont identifié une association entre la fréquence d'observation de cigognes et le taux de natalité (cf. Figure 3.5).

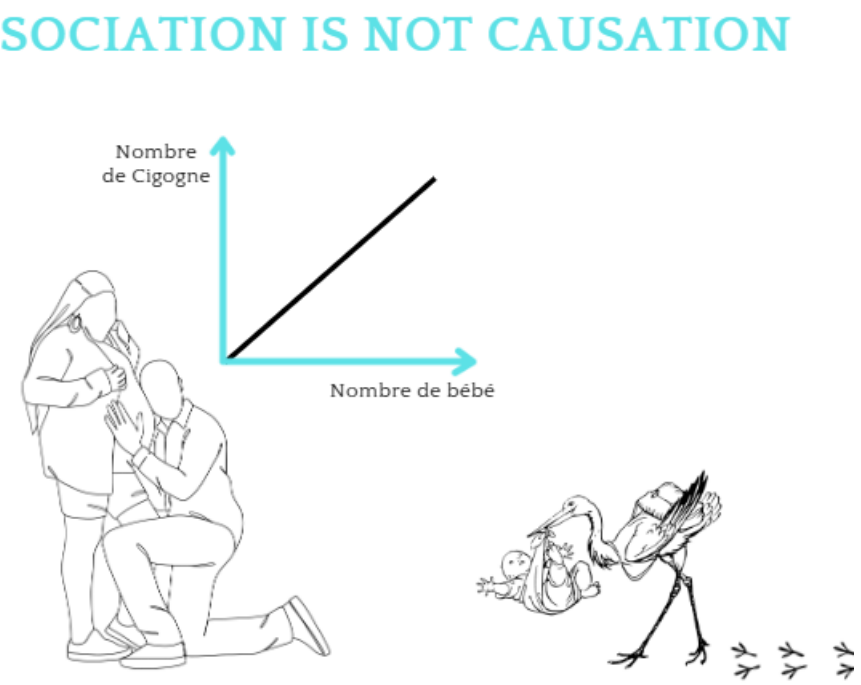


FIGURE 3.5 – Corrélation entre les taux de naissances et la présence des cigognes

Bien que cette association ait été observée, elle ne démontre évidement pas de lien de cause à effet. Cette association pourrait s'expliquer par exemple, par la présence migratoire saisonnière des cigognes, qui a lieu généralement au printemps. Cette période coïncide avec une hausse potentielle de la fréquence des naissances. La saisonnalité joue un rôle à la fois sur la présence des cigognes et sur la fréquence des naissances, mais elle n'est pas la raison directe des deux - on dit que c'est un facteur de confusion.

Ne pas prendre en compte les facteurs de confusion peut amener à des interprétations biaisées. Dans cet exemple, l'évidence réside dans notre connaissance que ce ne sont pas les cigognes qui apportent les bébés. Cependant, lorsqu'on aborde des phénomènes tels que le vieillissement cérébral, nos connaissances ne sont pas aussi claires. Comment pouvons-nous prétendre qu'il y a une relation causale dans de telles circonstances ? Comment garantir que les résultats que nous obtenons ne sont pas affectés par des biais ?

C'est dans ce contexte que nous présenterons plus précisement dans la suite ce qu'est l'inférence causale.

3.2.2 Le monde contrefactuel

Pour obtenir l'effet causal entre deux variables, c'est-à-dire que "$X \xrightarrow{cause} Y$", il faut qu'une variation de X impacte la variable Y. Considérons la variable d'exposition binaire X comme étant le statut diabétique, où $X = 1$ signifie que le sujet est diabétique et $X = 0$ indique qu'un sujet est non diabétique. Le taux de glycémie est représenté par la variable Y. Si l'on souhaite connaître l'effet du diabète sur la glycémie il suffit de regarder si le statut diabétique X impacte le taux de glycémie Y. En observant les données réelles du tableau 3.1 reportant le statut diabétique de quatre sujets ainsi que leur taux de glycémie, nous pouvons constater que la moyenne du taux de glycémie des personnes non diabétiques (i.e. 90 mg/dL) est plus faible que les personnes diabétiques (i.e 150 mg/dL). Une association est montrée dans cet exemple car les deux groupes de sujets (diabétique *versus* non diabétique) ne sont pas identiques. Pour savoir si elle est de nature causale, une approche proposée par Rubin (1974) est de se placer dans un monde contrefactuel, c'est-à-dire un monde hypothétique.

Dans ce contexte de monde contrefactuel, on se pose la question de ce qui aurait pu se produire **si** une variable avait adopté une valeur spécifique. Prenons l'exemple du sujet 1 : quelle aurait été la valeur de son taux de glycémie **si** il avait été diabétique ? Dans le monde réel, basé sur des données observées, la valeur du taux de glycémie du sujet 1 **si** il avait été diabétique reste inconnue. A contrario, dans le monde contrefactuel, cette information est supposée connue. Le tableau 3.2 présente les valeurs prises par les variables contrefactuelles. Ainsi nous pouvons dire que le sujet 1 aurait eu un taux de glycémie de 155mg/dL **si** il avait été diabétique. Connaître les valeurs des variables contrefactuelles permet de la même manière qu'un essai clinique, d'avoir des groupes comparables sur toutes les caractéristiques à l'exception de celle étudiée. Ici les sujets sont exactement les mêmes, à l'exception de la valeur du statut diabétique.

TABLE 3.1 – Observations réelles du statut diabétique et du taux de glycémie de 4 sujets.

	Sujet 1	**Sujet 2**	**Sujet 3**	**Sujet 4**
Statut Diabétique	0	1	1	0
Taux de Glycémie (en mg/dL)	80	180	130	100

TABLE 3.2 – Observations contrefactuelles du statut diabétique et du taux de glycémie de 4 sujets.

	Sujet 1	**Sujet 2**	**Sujet 3**	**Sujet 4**
Statut Diabétique	1	0	0	1
Taux de Glycémie (en mg/dL)	155	80	85	165

L'approche contrefactuelle offre ainsi une perspective importante pour comprendre l'impact causal des variables sur un phénomène donné.

3.2.3 Les étapes d'une démarche causale

Chercher à étudier un effet causal s'inscrit dans une démarche scientifique de raisonnement rigoureuse et claire (Goetghebeur et al. (2020)).

La première étape du raisonnement causal réside dans la clarification de la nature de la question d'intérêt, à savoir si elle est causale, descriptive ou prédictive. Pour ce faire, des recherches bibliographiques approfondies doivent être entreprises afin d'examiner la nature des liens (Snowden et al. (2018)). Par exemple, anticiper l'apparition de la démence chez des patients âgés est une question de nature prédictive. Il n'est donc pas necessaire d'envisager une analyse causale. A contrario comprendre le cheminement qui mène à la survenue d'une démence, est une question causale.
La question d'intérêt doit être formulée de manière rigoureuse et claire. D'après Rubin (1974) : "If you are not talking about intervention, you can't talk about causality", à savoir pour étudier la causalité, il est nécessaire de pouvoir intervenir. Par exemple que se serait-il passé si l'on avait eu (...) ou si l'on était intervenu sur (...). C'est d'ailleurs le nom du livre d'Hernan and Robins (2020) abordant la causalité "What if ?".

La démarche scientifique suivante a été proposée par Goetghebeur et al. (2020), elle est résumée dans la figure 3.6.

Suite à la définition de la question causale, nous pouvons identifier la population **cible de l'étude**, ainsi que les **variables potentiellement impliquées** et identifier les **variables que nous avons à notre disposition**.

Par exemple, si la question causale est d'étudier l'impact du diabète sur le déclin cognitif chez les personnes âgées, la population cible correspond aux personnes âgées,

d'autres critères peuvent également être mis comme l'âge (e.g. 65 ans et plus).

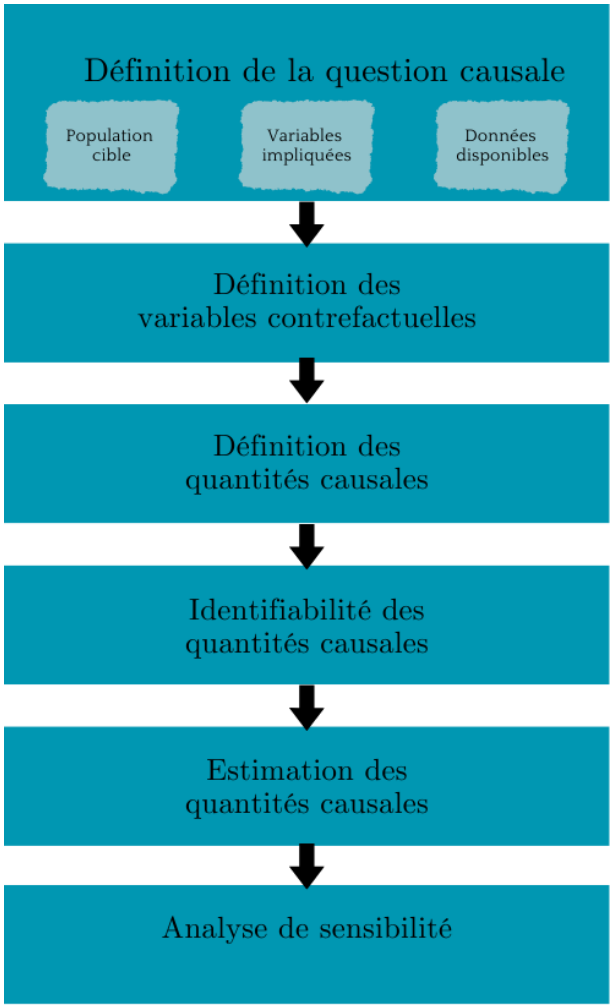


FIGURE 3.6 – Différentes étapes de la démarche scientique causale

La clarification des variables contrefactuelles est nécessaire afin de définir ultérieurement les quantités causales pertinentes, en fonction de l'interprétation que l'on souhaite obtenir et d'établir les hypothèses essentielles à leur identification. Les quantités causales définies préalablement peuvent ensuite faire l'objet d'estimations, souvent à partir de modèle statistique, dit "modèle de travail". Il est également recommandé d'effectuer des analyses de sensibilité par la suite.

3.2.4 Les graphes causaux

La définition de la question causale est une étape majeure dans la démarche causale. A partir de cette question il est possible de faire une synthèse du système dans lequel évoluent les variables impliquées et les relations entre elles. Cela peut être fait par l'élaboration de graphes causaux. Ces graphes revêtent une importance fondamentale dans

la conceptualisation de l'inférence causale. Ils simplifient la compréhension des structures causales inhérentes aux données, éclaircissent les relations entre les variables, intègrent les biais potentiels, et orientent de manière plus précise les analyses statistiques en vue d'évaluer les relations de cause à effet.

3.2.4.1 Les diagrammes dirigés acycliques

Diverses représentations graphiques existent pour illustrer les relations causales entre les variables. Parmi elles, les **diagrammes dirigés acycliques** (DAG) constituent l'une des approches les plus répandues. Ils permettent de représenter les liens causaux au sein d'un système où les nœuds représentent les variables, tandis que les flèches indiquent les relations causales (Digitale et al. (2022)). Par exemple, $X \rightarrow Y$ représente un système où la variable X cause la variable Y.

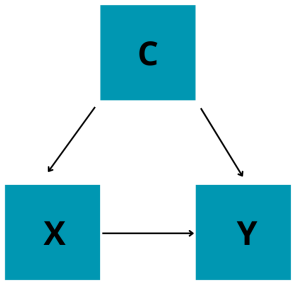


FIGURE 3.7 – Diagramme dirigé acyclique de la relation entre une variable d'exposition X, un variable d'intérêt Y et une variable de confusion C

Considérons la Figure 3.7, ce DAG représente le système où X cause Y (i.e. $X \rightarrow Y$), C cause X (i.e. $C \rightarrow X$) et C cause Y (i.e.$C \rightarrow Y$). Si l'on s'intéresse à la relation entre X et Y, nous devons tenir compte de la variable C. La variable C est à la fois liée à X et à Y sans être sur le chemin causal des deux (i.e. ~~$X \rightarrow C \rightarrow Y$~~), on dit alors qu'il s'agit d'un facteur de confusion. Plus généralement, on parle de facteur de confusion lorsqu'une variable est à la fois associée à la variable explicative (X) et à la variable à expliquer (Y) mais qu'elle n'est pas située sur le chemin causal des deux. La non prise en compte de cette variable conduit à une possible introduction de biais dans l'inférence causale. Cela reviendrait à capter dans l'effet de X une part d'effet qui est en réalité dûe à C et non à X. Cette omission peut compromettre la validité des conclusions dans une analyse causale.

Cette situation est très fréquente dans les études épidémiologiques. Par exemple, Busby et al. (1994) ont identifié une association entre la pression artérielle et le risque de décès, avec un risque presque deux fois supérieur entre les personnes âgées ayant une pression artérielle faible *versus* une pression artérielle normale. En prenant en compte des facteurs de confusion tels que la présence de maladies cardiaques, cette association n'a pas été retrouvée.

La Figure 3.8 représente un autre système de variable dans lequel X cause Y (i.e. $X \rightarrow Y$), X cause M (i.e. $X \rightarrow M$) et M cause Y (i.e. $M \rightarrow Y$). Dans cette configuration M "intervient" dans la relation entre X et Y.

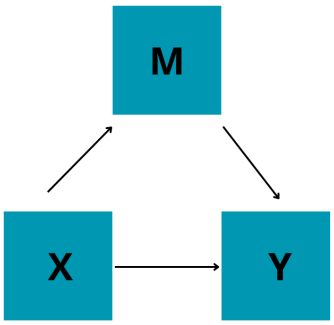


FIGURE 3.8 – Diagramme dirigé acyclique de la relation entre une variable d'exposition X, une variable d'intérêt Y et une variable intermédiaire M

Il s'agit d'une variable médiatrice, c'est-à-dire qu'elle est causée par X mais qu'elle cause à son tour Y. C'est donc une variable intermédiaire sur le chemin entre X et Y. Cette situation est courante lorsque l'on souhaite comprendre les mécanismes sous-jacents existants entre différentes variables de santé. Par exemple, Grande et al. (2020) ont montré qu'une exposition prolongée à la pollution de l'air était associée à un risque accru de démence et que l'accident vasculaire cérébral jouerait un rôle intermédiaire important dans l'association entre l'exposition à la pollution de l'air et la démence. Dans ce cas de figure, prendre en compte la variable M dans un modèle statistique en ajustant dessus revient à décomposer l'effet de X pour ne récuperer que l'effet ne passant pas par M. Cela empêche d'identifier l'effet total passant par les deux chemins de X et Y.

Dans la figure 3.9, le système est similaire à celui de la figure 3.7, c'est-à-dire X cause

Y (i.e. $X \rightarrow Y$), C cause X (i.e. $C \rightarrow X$), C cause Y (i.e. $C \rightarrow Y$). Cependant sur ce graphique, la variable Z a été ajoutée à ce système avec Z cause X (i.e. $Z \rightarrow X$).

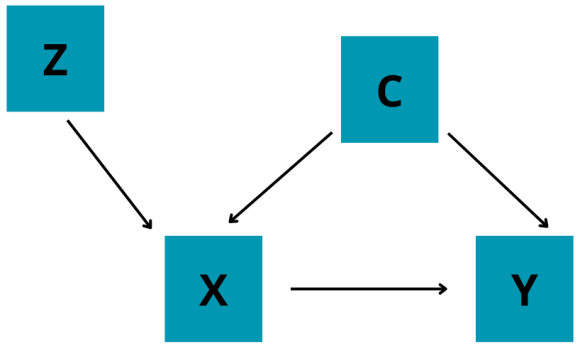


FIGURE 3.9 – Diagramme dirigé acyclique de la relation entre une variable instrumentale Z, une variable d'exposition X, une variable d'intérêt Y et une variable de confusion C

A la différence de C, la variable Z cause X mais n'est pas associée à Y autrement que par X. On appelle cela une varibale instrumentale pour la relation entre X et Y. Maintant, supposons que nous ne connaissons pas la variable C, nous ne pouvons donc pas ajuster sur les facteurs de confusion C et la relation entre X et Y reste biaisée. Une façon de contourner ce problème est de considérer la variable Z. Comme Z n'est pas influencé par les facteurs de confusion C et que Z impacte Y uniquement via la variable d'exposition X, alors on peut retrouver l'effet causal de X sur Y en s'appuyant sur la variable instrumentale Z.
La présence de facteurs de confusion non observés C, constitue un véritable challenge dans l'épidémiologie du vieillissement, car il s'agit d'un processus multifactoriel (e.g. de multiples facteurs cardiométaboliques), avec des mécanismes pas forcément connus et les cohortes ne peuvent pas être exhaustives.

Comme illustré dans ces trois cas simples, la conception d'un graphe causal permet d'identifier l'ensemble des variables impliquées et les liens qui les unissent. Les graphes permettent de définir les phénomènes en jeu et les analyses statistiques adaptées à cette

configuration, telles que l'analyse de médiation pour la figure 3.8 ou l'analyse par variables instrumentales pour la figure 3.9.

3.2.4.2 Limites des diagrammes dirigés acycliques

Les DAGs sont un outil extrêment utile pour appréhender une problématique épidémiologique et définir une stratégie d'analyse. Ils restent cependant un outil graphique synthétique et conceptuel qui repose sur des hypothèses et présentent des limites.

Premièrement l'interprétation causale dans les graphes dirigés acycliques tels que présentés sur les figures 3.7, 3.8, et 3.9 reposent toujours sur l'hypothèse qu'ils sont correctement spécifiés. Imaginons qu'un facteur de confusion ne soit pas représenté sur le DAG alors qu'il est effectivement présent. Cela conduira au choix d'un modèle statistique inapproprié, engendrant ainsi un biais dans les résultats. De même si les flèches sont mal orientées, un facteur de confusion C peut être pris pour une variable médiatrice M, impliquant une modélisation erronnée.
Deuxièmement les DAGs sont utiles pour détecter l'existence de relations causales, mais ils ne spécifient pas la forme exacte de ces relations (Digitale et al. (2022)). Cela doit être supposé lors de l'estimation.
Enfin, il est important de noter que les DAGs traditionnels ne capturent pas la dimension temporelle. Dans ces graphes, les nœuds représentent des variables et non des processus pouvant évoluer dans le temps.

Les DAGs étant comme le nom l'indique des diagrammes "acycliques" cela signifie qu'ils n'ont pas de cycle. Ils sont conçus pour représenter des relations causales dans une seule direction. Or, un système dans lequel X implique Y et Y implique X existe, et sa représentation serait celle donnée par la Figure 3.10.

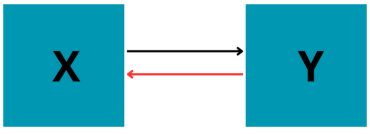


FIGURE 3.10 – Exemple de diagramme représentant la causalité inverse entre les variables X et Y

Lorsque l'on étudie des phénomènes évolutifs tels que le vieillissement, les relations entre les variables peuvent s'avérer bidirectionnelles. Illustrativement, l'indice de masse corporelle (IMC) peut être à la fois à l'origine du processus pathologique conduisant à la démence et, simultanément, le processus prédementiel peut induire des modifications alimentaires conduisant à une diminution de l'indice de masse corporelle (Wagner et al. (2021)).

Cette problématique est intrinsèquement liée à la dynamique inhérente des processus en santé, nécessitant ainsi d'explorer l'intégration de la dimension temporelle dans les DAGs.

3.2.4.3 DAG étendus aux données longitudinales

Pour aborder l'analyse des données répétées dans le temps, une extension des DAGs a été réalisée en définissant une variable distincte à chaque temps. Par exemple, à trois temps distincts $T1$, $T2$ et $T3$, la variable X devient $X1$, $X2$ et $X3$, représentant le niveau d'exposition à ces trois temps. De manière similaire, la variable Y est devient $Y1$, $Y2$ et $Y3$. La relation entre ($X1$,$X2$,$X3$) et ($Y1$,$Y2$,$Y3$) peut alors prendre plusieurs formes. D'abord, au sein de chaque variable, il y a fréquemment une dépendance sérielle, c'est-à-dire que la valeur future est dépendante du passé (e.g. $X1$ implique $X2$ implique $X3$) (Figure 3.11).

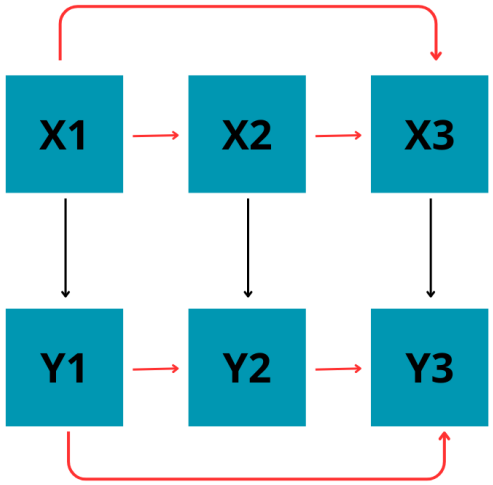


FIGURE 3.11 – Exemple de diagramme dirigé acyclique avec dépendance sérielle des variables X et Y

Entre les variables X et Y, différentes situations peuvent aussi être rencontrées telles que :

— **une influence simultanée entre les variables**

C'est le cas sur la Figure 3.12 où la valeur d'une variable à un temps, influe une autre variable au même temps (e.g. $X1$ implique $Y1$, $X2$ implique $Y2$, $X3$ implique $Y3$).

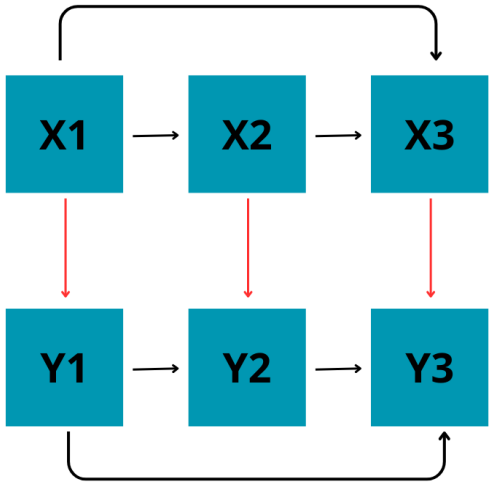


FIGURE 3.12 – Exemple de diagramme dirigé acyclique avec dépendance simultanée de la variable X sur la variable Y

— **Une influence unidirectionnelle sur le futur**

C'est le cas sur la Figure 3.13 correspondant à une situation où une variable à un temps impacte l'autre variable au temps suivant (e.g. $X1$ implique $Y2$, $X2$ implique $Y3$).

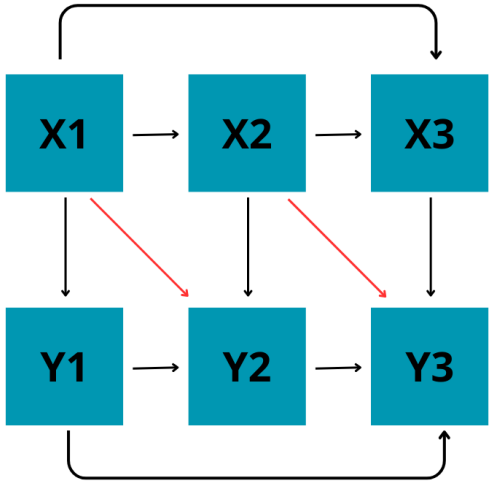


FIGURE 3.13 – Exemple de diagramme dirigé acyclique avec dépendance unidirectionnelle des valeurs du passé de la variable X sur la variable Y

— **Une influence bidirectionnelle sur le futur**

Elle est représentée par une dépendance entre les valeurs du passé et du futur pour les eux variables (e.g. $X1$ implique $Y2$, $X2$ implique $Y3$ et $Y1$ implique $X2$, $Y2$ implique $X3$), comme représenté sur la Figure 3.14.

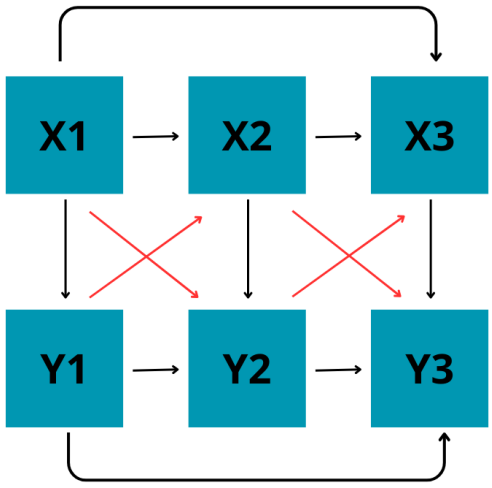


FIGURE 3.14 – Exemple de diagramme dirigé acyclique avec dépendance bidirectionnelle des valeurs du passé des variables X et Y sur les valeurs du futur.

Lorsque les variables longitudinales sont mesurées à trois temps et que seulement deux variables interviennent, le graphe reste lisible. Toutefois avec un plus grand nombre de suivis, ainsi qu'avec un plus grand nombre de variables, comme c'est souvent le cas en épidémiologie, les graphes deviennent rapidement illisibles, comme illustré en Figure 3.15, avec des variables longitudinales mesurées à six temps ($t : 1, ..., 6$). Dans cet exemple nous supposons une dépendance sérielle **totale** au sein d'une même variable, c'est-à-dire qu'une variable au temps t influera sur toutes les futures variables (e.g. $X3 \rightarrow (X4, X5, X6)$), une influence simultanée entre les variables (e.g $X_t \rightarrow Y_t$) ainsi qu'une influence bidirectionnelle sur le futur entre les variables (e.g $X_t \rightarrow Y_{t+1}$).

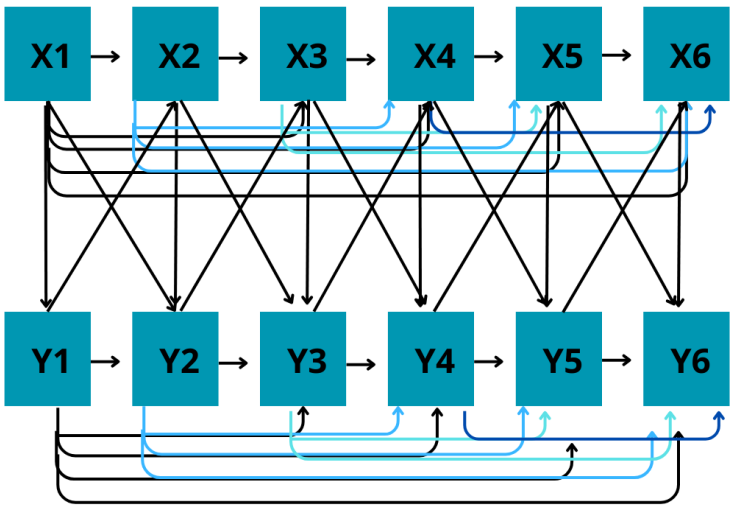


FIGURE 3.15 – Exemple de diagramme dirigé acyclique avec les variables X et Y mesurées à 6 temps

De nombreux travaux sur la causalité se basent sur des graphiques (DAGs) qui permettent de traiter l'indépendance entre les variables, à l'aide d'ajustement sur des variables observées à des temps discrets. Cependant, ils ne rendent pas compte de manière intuitive de la dépendance dynamique entre les processus. De même ces représentations, ne traduisent pas forcément le design de l'étude. Dans une étude de cohorte, les temps de mesure peuvent varier d'un sujet à l'autre, comme illustré dans la cohorte 3C avec le délai depuis l'entrée dans l'étude variables d'un sujet à l'autre et des intervalles de temps entre les mesures différents d'un suivi à l'autre. De plus, même si les observations se font en temps discret, les processus d'intérêt qu'elles mesurent restent le plus souvent définis en temps continu comme le niveau cognitif ou bien le taux de glycémie.

Les graphes d'indépendance locale permettent d'étendre les graphes acycliques dirigés à des processus en traitant l'indépendance locale entre les processus (Schweder (1970)). Ils représentent les dépendances entre le passé du processus multidimensionnel et son état actuel (Aalen and Frigessi (2007),Didelez (2008),Gégout-Petit and Commenges (2010)), afin de comprendre comment le système change avec le temps (Voelkle et al. (2018)).

Soit $\mathcal{X}^*(t)$ et $\mathcal{Y}^*(t)$ deux processus définis en temps continu et mesurés en temps discret, le schéma Figure 3.15, peut être étendu à la Figure 3.16.

Les graphes d'indépendance locale sont beaucoup moins répandus que les DAGs et pré-

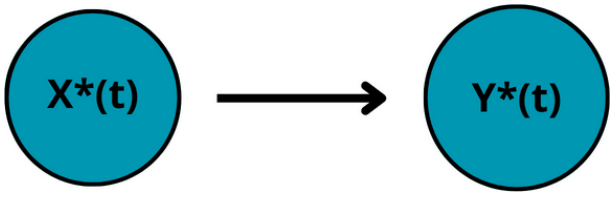


FIGURE 3.16 – Exemple de graphe de dépendance local entre les processus X^* et Y^*

sentent une complexité en termes de compréhension. Néanmoins, ils sont indispensables pour étudier des phénomènes définis en temps continu et mesurés à des temps discrets. Nous nous appuierons sur ces concepts dans la suite de nos travaux.

3.2.5 Du graphe à l'estimation

Les DAGs donnent la structure causale conceptuelle des systèmes. Cependant, ils ne permettent pas d'estimer les effets causaux. L'estimation nécessite une approche statistique, avec des modèles adaptés à la question causale d'intérêt.

3.2.5.1 Régression linéaire

Considérons à nouveau l'exemple de la Figure 3.7 dans laquelle les variables X et Y sont confondues par les facteurs de confusion C. L'objectif est d'estimer l'effet de X sur Y, avec X qui est une variable d'exposition mesurée à un temps, Y est la variable d'intérêt continue mesurée à un temps et C est un facteur de confusion dans la relation entre X et Y. Pour obtenir l'effet causal entre X et Y, il faut fermer tous les chemins impliquant X et Y, avec une flèche pointée sur X. On appelle ces chemins des chemins "back door". Une méthode statistique permettant de bloquer ces chemins, consiste à ajuster sur les variables de ces chemins, c'est-à-dire les chemins par la variable C dans la Figure 3.7. Elle peut être traduite par le modèle de regression linéaire suivant :

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 C_i + \epsilon_i \tag{3.2}$$

En supposant que le modèle est correctement spécifié, nous obtenons l'effet causal de X sur Y à l'aide du coefficient β_1.

3.2.5.2 Réseaux bayésiens

Une autre approche proposée par Pearl (2012) pour estimer l'effet causal de X sur Y dans la Figure 3.7 s'appuie sur les réseaux bayésiens. Il s'agit d'une approche probabiliste des graphes. Plus spécifiquement, les réseaux bayésiens reprennent la même représentation que les DAGs et consistent à quantifier les probabilités conditionnelles de chaque noeud selon ses noeuds parents.

Dans la Figure 3.7, X a pour noeud parent C, et Y a pour noeuds parents C et X. Ainsi, la structure causale présentée peut se traduire par la probabilité $P(X, Y, C) = P(Y|C, X)P(X|C)P(C)$ d'après la formule de Bayes. Les réseaux bayésiens n'étant pas intrinsèquement des modèles causaux, ils sont controversés en causalité. Ils se basent sur une approche probabiliste, dans laquelle la causalité est définie comme la probabilité qu'un évènement soit influencé par un autre événement (Mellor (1995)). Pearl (2000) a montré qu'il était difficile à partir de cette définition de distinguer une réelle association causale d'une association.

3.2.5.3 Modèle à équations structurelles

Le modèle à équations structurelles (SEM) sont une approche très populaire pour l'analyse causale dans les sciences sociales. Cette approche reprend la structure de répresentation des DAGs mais en distinguant les variables observées (e.g., $X1$,$X2$,$X3$,C,Y), des variables latentes (e.g., X^*, Y^* : les niveaux sous-jacents). Chaque flèche est ensuite traduite par un modèle linéaire. Les regressions qui expliquent les observations en fonction des variables latentes sont appelées équations d'observations. Les regressions qui expliquent les variables latentes en fonction d'autres variabes latentes ou observées sont appelées des équations structurelles. Il s'agit donc d'une approche d'analyse statistique permettant de tester des hypothèses sur les relations entre des variables (observées ou latentes).

Les modèles à équations structurelles, initialement développés en psychologie et en sciences de l'éducation par Wright (1934), représentent une approche statistique reposant sur un schéma de relations entre variables. Cette méthode, largement adoptée dans les

sciences sociales, offre la possibilité de travailler avec des variables observées comme dans les réseaux, tout en permettant la conceptualisation des relations entre ces variables à travers l'utilisation de graphes.

Les SEM sont très controversés en inférence causale. Deux écoles se font face. Les défenseurs des modèles à équations structurelles considèrent que les relations causales peuvent être établies à partir de modèles basés sur des corrélations entre les variables observées et qu'il n'est pas nécessaire d'intervenir ou de manipuler des variables pour faire de la causalité. Ils soutiennent que les modèles basés sur des corrélations peuvent offrir une compréhension suffisante des relations causales. En revanche, d'autres mettent fortement l'accent sur l'importance de prendre en compte des scénarios contrefactuels pour parvenir à une véritable inférence causale, s'appuyant sur le principe bien connu "No causation without manipulation" (Holland (1986) ;Rubin (1974)). Dans leur livre, Bollen and Pearl (2013) en désaccord avec cette citation, soulignent que même si ce principe était vrai, cela n'éliminerait pas la possibilité de mener une analyse par modèles d'équations structurelles.

3.2.6 Analyse de médiation

3.2.6.1 Contextualisation des analyses de médiation

L'analyse de médiation est une méthodologie statistique visant à décomposer l'effet d'une exposition X sur un événement Y (Robins and Greenland (1992)). Comme représenté sur la Figure 3.17, cette approche vise à décomposer l'effet total de X sur Y en un effet indirect passant par un ou plusieurs médiateurs (représenté par les flèches en vert allant de X à M puis de M à Y) et un effet direct de X sur Y (symbolisé par la flèche en rouge directe allant de X à Y).

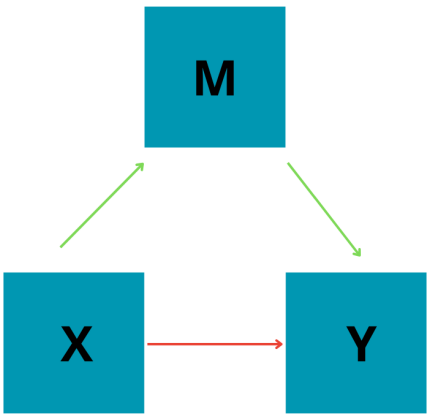


FIGURE 3.17 – Diagramme dirigé acyclique décomposant l'effet direct (flèche rouge) de X sur Y par son effet indirect (flèches vertes) passant par la variable intermédiaire M

La décomposition de l'effet total, défini comme l'association totale entre X et Y incluant à la fois le chemin direct et le chemin indirect, permet entre autres de mieux comprendre les mécanismes d'action d'un facteur de risque dans une maladie, mais également d'évaluer la pertinence d'intervention. Mathématiquement exprimée avec TE représentant l'effet total, DE l'effet direct, et IE l'effet indirect, cette relation s'exprime telle que :

$$TE = DE + IE \tag{3.3}$$

En analyse de médiation, les effets totaux, indirect et direct, sont des contrastes de quantités d'intérêt causales, appelées estimandes. Ces estimandes représentent des quantités potentielles ou contrefactuelles qui permettent de contraster la variable d'intérêt Y ou une fonction dérivée en réponse à un changement hypothétique du niveau d'exposition X. L'analyse de médiation nous situe dans un monde hypothétique où se pose la question de ce qui aurait pu se produire si une intervention avait été réalisée ou si une variable avait adopté une valeur particulière.

La modélisation des effets causaux peut s'effectuer au travers de méthodes traditionnelles faisant intervenir le résultat de régression, ou bien au travers d'approches contrefactuelles, faisant intervenir des variables hypothétiques.

3.2.6.2 Analyse de médiation traditionnelle

Initiallement dévelopées en psychologie et en sciences sociales, notamment grâce aux travaux de Baron and Kenny (1986), les méthodes d'analyse de médiation ont été déve-

loppées pour des cadres de données très simples, à savoir une exposition binaire X, une variable d'intérêt Y, et un médiateur M continu tous collectés à un temps (pas forcément le même). Deux approches sont particulièrement utilisées : l'approche par différence et l'approche par produit.

L'approche par différence de coefficients a été proposée par Judd and Kenny (1981). Elle consiste en la réalisation de deux regressions sucessives.

La première régression est une régression simple qui permet d'estimer l'effet total (TE) de X sur Y, à savoir l'association marginale entre X et Y (sans prendre en compte d'autres variables).

$$Y = \beta_{0*} + \beta_{1*}X + \epsilon \tag{3.4}$$

Dans ce cadre TE=β_{1*}.

La deuxième regression est une régression multivariée pour prédire Y à partir de X et M :

$$Y = \beta_0 + \beta_1 X + \beta_2 M + \epsilon \tag{3.5}$$

Elle permet d'estimer l'effet direct (DE) de X sur Y après un ajustement sur le médiateur M (représenté en rouge sur la figure 3.17) : $DE = \beta_1$.

Il est alors possible grâce à l'équation 3.3 d'obtenir l'effet indirect de X sur Y, en contrastant les deux premiers effets obtenus, c'est-à-dire en soustrayant l'effet direct de l'effet total.

$$IE = TE - DE$$

$$IE = \beta_{1*} - \beta_1$$

Méthode par produit

L'approche du produit a été très utilisée suite aux travaux de Baron and Kenny (1986).

Dans cette méthode, similaire à l'approche par différence, deux régressions successives

sont également effectuées. Dans la première, on prédit Y à partir de X et M, comme présenté dans l'équation 3.5. Ensuite, dans une deuxième régression, M est prédit à partir de X, formulée comme suit :

$$M = \gamma_0 + \gamma_1 X + \epsilon \tag{3.6}$$

Dans cette approche, l'effet direct (DE) demeure égal à β_1, l'effet indirect (IE) est quant-à lui :

$$IE = \beta_1 \times \gamma_1$$

L'effet total est $TE{=}\beta_1^* + \beta_1 \times \gamma_1$. Cette approche correspond à un modèle à équations structurelles (c.f. Section 3.2.5.3)

Les approches traditionnelles se révèlent toutefois restreintes, notamment en ce qui concerne les données non Gausiennes ou bien les interactions entre les variables.
Dans cette thèse, nous porterons un intérêt particulier aux approches contrefactuelles, offrant ainsi une perspective pour aborder ces limitations méthodologiques (Robins and Greenland (1992) ; Pearl (2001)).

3.2.6.3 Approches contrefactuelles

De nombreux auteurs ont proposé dans leurs travaux de définir les effets directs et indirects, présentés dans l'équation 3.3, en utilisant un cadre contrefactuel (Robins and Greenland (1992) ; Pearl (2000) ; VanderWeele (2009) ; Imai et al. (2010a) ;Imai et al. (2010b)). Ceci permet de décomposer l'effet total dans des modèles avec des interactions et des relations non-linéaires.

Notations dans un monde contrefactuel

Considérons X la variable d'exposition, M le médiateur, Y l'outcome et C le facteur de confusion. Chaque variable est supposée fixe dans le temps. Les variables sont liées entre elles comme figuré sur la figure 3.17.

Supposons que X soit binaire et puisse prendre les valeurs x et x'. Nous noterons

$M(x)$, la valeur potentielle (i.e. contrefactuelle) qu'**aurait pris** le médiateur M **si** $X = x$ et $Y(x')$ la valeur potentielle qu'**aurait pris** l'outcome Y **si** l'on avait fixé $X = x'$. De la même manière $Y(x, M(x))$ correspond à la valeur potentielle de l'outcome **dans un monde où** la valeur de l'exposition serait égale à x, et la valeur prise par le médiateur serait celle correspondant à $X = x$.

L'approche contrefactuelle s'appuie sur des quantités potentielles qui sont ensuite identifiées à partir des données grâce à des hypothèses. Suivant la nature des hypothèses faites, différents effets sont ciblés.

Natural effect

Le "Natural Effect" décompose l'effet total naturel (NE) de l'exposition X sur l'outcome Y, en un effet direct naturel (NDE) défini comme l'effet de X sur Y seul, et un effet indirect naturel (NIE) défini comme l'effet de X sur Y passant par le médiateur. Le natural effect représente l'effet causal qui serait observé si l'on intervenait sur la variable d'exposition X, laissant tous les autres mécanismes inchangés. En comparant deux niveaux x et x', ces effets s'expriment comme suit :

$$NE = E(Y(x, \mathcal{M}(x))|C) - E(Y(x', \mathcal{M}(x'))|C)$$
$$NIE = E(Y(x, \mathcal{M}(x))|C) - E(Y(x, \mathcal{M}(x'))|C)$$
$$NDE = E(Y(x, \mathcal{M}(x'))|C) - E(Y(x', \mathcal{M}(x'))|C)$$

Les variables contrefactuelles n'étant pas observables, six hypothèses sont nécéssaires pour identifier les effets causaux à partir des observations.

L'hypothèse de **positivité** explicite que chaque individu, étant donné ses valeurs de C, doit avoir une probabilité positive de recevoir n'importe quelle valeur d'exposition, ou du médiateur, soit $P(X = x|C) > 0$ et $P(M(x) = m|X = x, C) > 0$, $\forall x$. L'hypothèse de **consistence** est utilisée pour établir une connexion entre les variables contrefactuelles et les variables observées avec $Y = Y(x, m)$ si $X = x, M(x) = m$. Les quatre hypothèses d'**indépendance conditionnelle** stipulent que :

— il n'y a pas de facteurs de confusion entre X et Y sachant C : $Y_t(x, m) \perp\!\!\!\perp X|C$

— il n'y a pas de facteur de confusion entre M et Y sachant C : $Y(x,m) \perp\!\!\!\perp M|X,C$

— M est indépendant de X sachant C : $M(x) \perp\!\!\!\perp X|C$

La quatrième hypothèse, appelée la "**cross-world independence**" indique qu'il n'y a pas de facteur de confusion entre les variables potentielles Y et M affecté par X, conditionnellement à C, elle se traduit par : $Y(x,m) \perp\!\!\!\perp M(x')|C$.

Stochastic effect

Contrairement à l'approche de l'effet naturel, qui représente l'effet causal qui serait observé en intervenant sur la variable d'exposition tout en laissant les autres inchangées, l'approche de l'effet stochastique, proposée par Stock (1989), considère des interventions stochastiques sur la variable médiatrice, c'est-à-dire que la variable d'intervention est une variable aléatoire. Bien que cette approche ne soit pas interprétée de la même manière que l'effet naturel, elle a été utilisée pour rendre la méthode plus flexible, en considérant moins d'hypothèses (Didelez et al. (2006)). Dans cette approche, plutôt que formuler les valeurs contrefactuelles individuelles de M(x) ou M(x'), les valeurs sont tirées dans la distribution de M, conditionnellement aux covariables C, avec X=x et X=x' (Rudolph et al. (2021))

Similairement à l'effet naturel, l'effet de l'intervention stochastique peut être exprimé par la différence d'espérance, avec l'effet indirect stochastique (SIE) représenté par :

$$SIE = E(Y(x, G_{Mx})|C) - E(Y(x, G_{Mx'})|C)$$

et l'effet stochastique direct (SID), par :

$$SDE = E(Y(x, G_{M_x}')|C) - E(Y(x', G_{Mx'})|C)$$

Où G_{Mx} désigne une réalisation aléatoire de la distribution de M pour $X = x$.

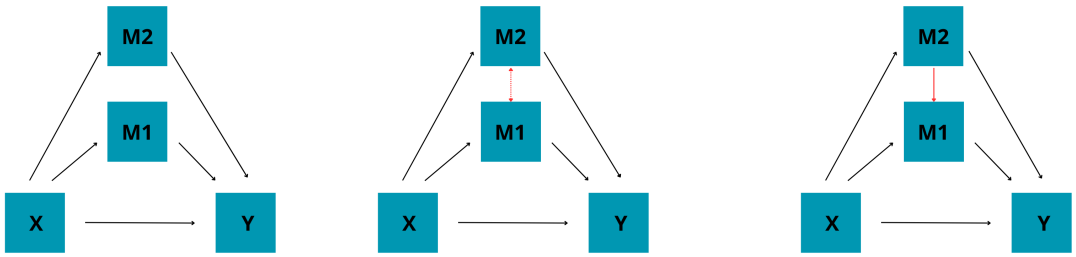
Les hypothèses nécessaires pour l'identifiablité de ces effets stochastiques sont similaires à celles exposées pour le natural effect, à l'exception de l'hypothèse de "cross-world independence" qui n'est plus nécessaire. Cette modification permet à l'approche du stochastic effect une utilisation plus large, dans des situations où cette hypothèse ne pourrait pas s'appliquer, par rapport au natural effect. Cependant, l'interprétation d'un effet sto-

chastique est aussi différente de celle des effets naturels, et répond donc à une question causale différente. On cherche ici à identifier l'effet de X sur Y si on avait pu intervenir sur M de sorte à rendre la population similaire à celle de référence. Cette approche est particulièrement utile pour évaluer les inégalités sociales ou raciales.

Les approches traditionnelles et les approches contrefactuelles ont été largement étudiées dans le cadre "simple" des données mesurées à un seul temps. Leur extension aux situations plus complexes rencontrées dans les études épidémiologiques n'est pas évidente. Deux problématiques classiques sont les multiples médiateurs et les données longitudinales.

3.2.6.4 Extension des analyses de médiation aux médiateurs multiples

En présence de médiateurs multiples, les approches présentées précédemment ne sont généralement pas utilisables. En effet, l'ajout d'une variable peut créer différents scénarios, rendant les effets causaux plus complexes à estimer, avec notamment la violation de certaines hypothèses. Par exemple, la Figure 3.18 présente le cas de deux médiateurs. Dans le cas (a), ils sont indépendants, dans le cas (b) ils sont corrélés entre eux et dans le cas (c) le médiateur $M2$ implique le médiateur $M1$.



(a) (b) (c)

FIGURE 3.18 – Diagramme dirigé acyclique de la relation entre X et Y en présence de deux médiateurs (a) indépendants (b) corrélés (c) associés

Si l'on considère l'effet des médiateurs indépendants les uns des autres, les approches proposées en section 3.2.6.2 et 3.2.6.3 restent applicables (Preacher and Hayes (2008)). Cependant, cette hypothèse est très forte et généralement peu réaliste. C'est pourquoi, différentes extensions ont été proposées en présence de médiateurs multiples. Elles varient en fonction de la nature du lien entre les médiateurs (Daniel et al. (2015)), ainsi que des

types de variables étudiées.

Par exemple, lorsque les médiateurs sont causalement liés entre eux (cas (c)), l'hypothèse du "cross world independence" n'est plus respectée, car les variables potentielles $Y(x, m)$ et $M_1(x')$ ne sont pas indépendantes, dû à M_2. C'est pourquoi plusieurs auteurs (Tchetgen and Shpitser (2012) ; VanderWeele and Vansteelandt (2014) ; VanderWeele and Chiba (2014)) ont basé leurs travaux sur des méthodes où **un médiateur** est identifié comme **"principal"** (e.g. $M1$ dans la Figure 3.18), tandis que **les autres** sont traités comme des **facteurs de confusion** de la relation entre M et l'outcome (e.g. $M2$ dans la Figure 3.18), affectés par l'exposition. Les travaux s'appuient sur la méthode dite "path-specific effect". Les effets "spécifiques aux chemins" sont une méthode d'analyse de médiation plus souple que l'effet naturel quant aux hypothèses. En effet, elle ne nécessite pas l'hypothèse du "cross world independence". Cette méthode permet ainsi de définir les quantités causales en présence de médiateurs multiples, et elle est également facilement adaptable au cas de facteurs de confusion.

Les méthodes précédentes traitent principalement des variables médiatrices et des outcomes continus. Pour permettre une plus grande **flexibilité sur la nature des variables**, Lange et al. (2014) ont proposé une méthode permettant d'étudier les effets directs et indirects naturel en présence de médiateurs multiples avec différents types de médiateurs ou d'outcome, tout comme Huang and Yang (2017) qui définissent des effets causaux en présence de médiateurs multiples et d'un outcome de type survie.

Contrairement aux extensions citées ci-dessus, Jerolon et al. (2019) ont proposé une méthode permettant d'étudier la causalité lorsque les médiateurs sont correlés entre eux de manière non causale.

3.2.6.5 Extension pour données longitudinales

Analyse de médiation pour des temps d'évènements

Récemment a été abordée la question de variables d'intérêt de type temps d'événement avec soit des médiateurs indépendants du temps ou des médiateurs dépendants du temps.

Bien que le modèle d'analyse de survie le plus connu soit le modèle à risque proportionnel de Cox (c.f. Section 3.1.3.1), les premiers travaux d'analyse de médiation pour des variables d'intérêt de type temps d'évènement se sont appuyés sur un modèle additif pour définir les effets directs et indirects (Lange and Hansen (2011)). Lange and Hansen (2011) ont montré dans leurs travaux que les changements observés dans les rapports de risques ne pouvaient pas être interprétés de manière causale. Peu de temps après VanderWeele (2011) a proposé une approche pour définir les effets directs et indirects à l'aide de modèles à risques proportionnels pour un outcome rare. Ces deux approches, qui considèrent un médiateur continu indépendant du temps, ont par la suite été comparées par Gelfand et al. (2016), montrant qu'il était plus simple d'utiliser un modèle additif dans les analyses de médiation. D'autres méthodes ont également été proposées pour traiter la survie avec des données de grande dimension dans le cadre d'une exposition fixée dans le temps, de médiateurs multiples fixés dans le temps et d'un outcome de type survie. (Zhang et al. (2021), Luo et al. (2020)).

Liu et al. (2018), Didelez (2019) et Zheng and Liu (2022) ont ensuite étendu des approches pour permettre à la fois d'étudier l'analyse de médiation pour des données répétées (exposition/médiateur) au cours du temps, ainsi que des données de survie. Dans leur étude, Zheng and Liu (2022) définissent les effets causaux en utilisant le "natural effect", en redéfinissant les hypothèses pour une exposition fixe, en considérant le médiateur comme une variable répétée et avec un outcome de type survie.

Le risque compétitif, qui se manifeste par la survenue d'un événement concurrent susceptible de modifier la probabilité de l'événement d'intérêt, constitue un phénomène fréquemment observé dans les analyses de survie. Divers auteurs ont proposé des approches permettant d'intégrer cette notion afin d'évaluer les effets causaux (Aalen et al. (2020) ;Zheng and van der Laan (2017) ; Lin et al. (2017) ; et Tai et al. (2021)).

Très récemment Valeri et al. (2023) ont proposé une méthodologie permettant d'estimer des quantités causales, en définissant des contrastes causaux en termes de fonction de

survie pour quantifier les effets directs et indirects, en présence d'une variable d'exposition fixe, et d'un médiateur et d'un outcome, tous deux de type survie. Ils traitent l'approche d'effet stochastique plutôt que d'effet naturel. Au contraire de beaucoup de méthodes d'analyse de médiation, l'estimation de ces contrastes est réalisée en une seule étape grâce à un modèle multi-états (Wreede et al. (2011)). Cependant, leur travail, motivé par l'étude de l'effet médiateur du délai de mise en place du traitement dans les disparités raciales de survie des patients atteints de cancer, nécessite que le temps exact de l'événement intermédiaire soit connu. Or dans beaucoup d'études, y compris celles sur le vieillissement, les événements intermédiaires sont quasiment tous censurés par intervalle (c.f section 3.1.3) car le diagnostic est réalisé uniquement aux visites de suivi.

Analyse de médiation pour des données longitudinales

Les analyses de médiation appliquées à des données longitudinales, ont connu des développements ces dernières années dans le but de mieux appréhender le mécanisme causal. Cependant elles sont confrontées à certaines limites.

Premièrement, comme explicité pour les DAGs, la plupart des extensions (Bind et al. (2016), Mittinty and Vansteelandt (2020)) requiert des données observées très équilibrées et ne reflète pas les processus d'intérêt qui évoluent en temps continu. Plus récemment, Albert et al. (2019) ont proposé des estimandes causales en temps continu qui permettent d'étudier le mécanisme causal des variables mesurées au cours du temps comme on le trouve dans les cohortes. Cependant leur travail n'incluait pas de technique d'estimation de ces estimandes qui soit à la fois souple et performante. Ils supposaient que seul le médiateur au temps t jouait sur l'outcome au temps t.

Malgré les nombreuses extensions qui ont été réalisées au cours de ces dernières années, des lacunes subsistent encore dans la littérature pour l'étude de phénomènes plus complexes, tels que les données liées au vieillissement. Ces lacunes incluent la nécessité de prendre en compte de multiples processus évoluant en temps continu et des données observées de façon éparses et irrégulières ainsi que la présence de censure par intervalle

des événements intermédiaires.

3.2.7 Approche par variables instrumentales

Une autre problématique récurrente, dès lors qu'on souhaite faire de l'inférence causale en Santé, est l'existence de facteurs de confusion non nécessairement observés ou identifiés.
L'approche par variables instrumentales a été proposée pour traiter cette problématique. À l'origine développée dans le domaine économique (Angrist et al. (1996)), la méthode des variables instrumentales a par la suite été adaptée à plusieurs disciplines, y compris l'épidémiologie.

Dans le contexte de la recherche en santé publique, la méthode par variables instrumentales, permet d'étudier la relation causale entre une variable explicative X et une variable d'intérêt Y lorsque cette relation est sujette à des problèmes d'endogénéité (i.e. facteurs de confusion non mesurés, causalité inverse). Elle permet, grâce à l'identification d'une variable exogène (i.e. la variable instrumentale) notée Z, d'atténuer les biais potentiels liés à des facteurs de confusion non observées ou mal mesurées U, en reconstituant le cadre de la randomisation.

Cette méthodologie repose sur différentes hypothèses pour que la variable instrumentale soit valide, illustrés sur la figure 3.19 :

- l'hypothèse (1) indique qu'une association doit exister entre la variable instrumentale Z et la variable d'exposition X, c'est-à-dire que $cor(Z, X) \neq 0$
- l'hypothèse (2) suppose que la variable instrumentale ne peut expliquer la variable d'intérêt Y que par le chemin passant par X, c'est-à-dire $cor(Z, Y|X) = 0$
- l'hypothèse (3), une des plus importantes, indique que la variable instrumentale est indépendante des facteurs de confusion non mesurés, à savoir $cor(Z, U)0$

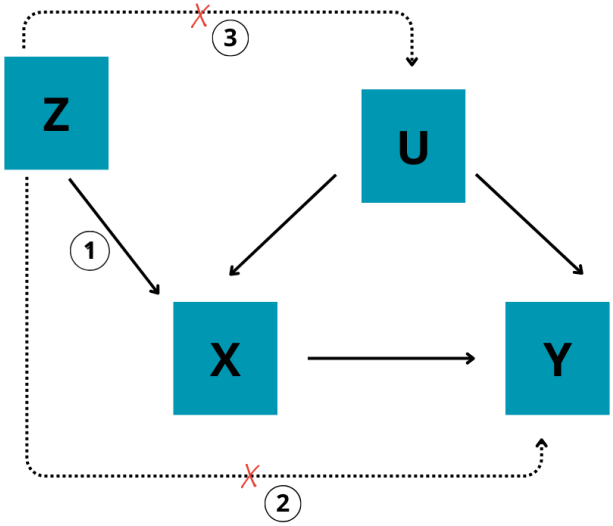


FIGURE 3.19 – Graphique dirigé acyclique de la relation causale entre X et Y ; Z est une variable instrumentale et U un facteur de confusion non observé

Lorsque les variables X et Y sont considérées à un temps fixe, plusieurs méthodes ont été développées pour estimer l'effet de X sur Y (Burgess et al. (2017) ;Baiocchi et al. (2014)). Parmi ces approches, la méthode des "Two-stage Least Squares" se distingue comme l'une des plus reconnues. Pour X et Y deux variables continues, cette méthode se décompose en deux étapes avec deux régressions linéaires successives :

1. **Première étape :** On estime la variable d'exposition X à partir des variables instrumentales Z. On peut alors prédire la valeur de X conditionnellement à Z par :

$$\hat{X} = Z(Z^T Z)^{-1} Z^T X \tag{3.7}$$

où Z est la matrice des variables instrumentales, et X est la matrice des variables explicatives.

2. **Deuxième étape :** On estime le modèle expliquant la variable d'intérêt Y à partir de X en remplaçant X par sa prédiction obtenue à l'étape précédente.

$$Y = \hat{X}\beta + \epsilon \tag{3.8}$$

où $\hat{X}$ est la variable explicative prédite, β son coefficient associé et ϵ est le terme d'erreur.

Le paramètre β estimé ainsi, fournit une estimation de l'effet de X sur Y, corrigé d'éventuels facteurs de confusion non observés. La méthode est présentée ici pour X et Y continu, mais des extensions pour X binaire existent (Burgess et al. (2017)).

Des extensions des méthodes par variables instrumentales ont été proposées en survie (Li et al. (2015), MacKenzie et al. (2014), Tchetgen Tchetgen et al. (2015),Martinussen and Vansteelandt (2020)) mais extrêmement peu ont porté sur des extensions à des données répétées. Les seuls modèles développés dans la littérature (Bond et al. (2007), Sánchez et al. (2017)) pour des données répétées s'intéressent à des cas très particuliers. Sánchez et al. (2017) ont proposé une méthode pour des expositions répétées sur un outcome transversal et Bond et al. (2007) ont développé une méthode pour des expositions répétées sur un outcome répété mais dans le cadre très spécifique d'un essai clinique dans le VIH.

Ainsi, les approches proposées actuellement ne permettent pas de traiter le cas très classique rencontré dans les cohortes populationnelles de vieillissement avec une variable mesurée à l'inclusion sur l'évolution d'un marqueur mesuré de façon répétée au cours du temps.

Chapitre 4

Analyse de médiation

Ce chapitre présente deux contributions réalisées au cours de ma thèse pour réaliser des analyses de médiation dans les études de cohorte. Ces travaux étaient motivés par l'étude des mécanismes causaux qui sous-tendent le vieillissement cérébral.

Dans un premier travail, l'objectif était d'étendre les méthodes proposées dans la littérature au cas de variables intermédiaires et finales répétées au cours du temps. Comme discuté dans le chapitre précédent, la littérature s'est limitée principalement à des méthodes pour données discrètes. Pour pouvoir traiter des données répétées mesurées de façon irrégulières dans les cohortes, et étudier directement les processus en jeu, nous avons proposé une approche en temps continu. L'objectif était d'estimer les mécanismes causaux entre une exposition fixée dans le temps (X), un processus médiateur ($\mathcal{M}_t$) et un outcome (Y_t) mesurés tous deux de manière répétée dans le temps en présence d'un processus de confusion dépendant du temps ($\mathcal{L}_t$).

Deux applications ont été envisagées dans la cohorte 3C. La première consistait à comprendre le lien entre l'apolipoprotéine APOE4 et le déclin cognitif, à travers l'atrophie cérébrale et/ou les lésions vasculaires cérébrales. La deuxième application s'est intéressée au rôle de la cognition et la dépression dans la relation entre le niveau d'éducation et le niveau de dépendance fonctionnelle.

Deux types d'effet naturels ont été proposés suivant l'absence ou présence de facteurs de confusion dépendant du temps. Dans chacune de ces approches nous avons défini

des quantités causales ainsi que leurs hypothèses d'identifiabilité associées. L'estimation de contrastes causaux nécessite l'utilisation d'un modèle de travail. Nous nous sommes focalisé ici sur le modèle mixte multivarié à équations différentielles décrit dans la section 3.1.2.3.

Ce travail a été réalisé en collaboration avec Linda Valeri (CUMC, New York). J'ai eu l'opportunité de faire une mobilité de trois dans son laboratoire d'Octobre à Décembre 2022. Ce travail est publié sur Arxiv et est soumis dans le journal *Biometrics*.

Dans un deuxième travail, l'objectif était d'étudier non plus des variables continues mesurées de façon répétée dans les cohortes mais des temps jusqu'à des événements intermédiaires et terminaux. Ce travail était motivé par l'étude du rôle de la démence dans l'effet de facteurs de risque cardiovasculaire sur le décès.

L'objectif de ce travail est d'étendre les analyses de médiation pour une variable d'exposition fixe, un médiateur de type temps d'événement (avec ou sans présence de censure par intervalle) et un outcome de type temps d'événement.

Nous nous sommes basés sur le travail de Valeri et al. (2023) qui ont proposé une approche par modèle multi-états pour étudier des interventions stochastiques sur un médiateur défini comme un événement intermédiaire censuré par l'événement final.

Nous l'avons adapté au cas d'effet naturel plutôt que d'intervention stochastique. Nous avons défini les quantités causales ainsi que les conditions pour leur identification à partir de données observées. Les contrastes causaux sont estimés dans un cadre de modélisation multi-états prenant possiblement en compte la censure par intervalle de l'événement intermédiaire et des formules analytiques pour les estimateurs des contrastes causaux sont développées.

Ce travail a été initié avec Alexandre Catherineau, stagiaire de Master 1 d'Avril à fin Juillet 2023.

Ce chapitre présente les extensions que nous avons réalisé en analyse de médiation.

4.1 Médiateurs et variables d'intérêt mesurés de façon répétées

Continuous-time mediation analysis for repeatedly measured mediators and outcomes

Le Bourdonnec Kateline[1], Valeri Linda[2,3], Proust-Lima Cécile[1]

[1]Univ. Bordeaux, Inserm, BPH, U1219, F-33000 Bordeaux, France

[2]Department of Biostatistics, Columbia University Mailman School of Public Health, 722 W 168th St, New York, NY, USA

[3]Department of Epidemiology, Harvard T.H. Chan School of Public Health, 677 Huntington Ave, Boston, MA, USA

March 16, 2024

Abstract

Mediation analysis aims to decipher the underlying causal mechanisms between an exposure, an outcome, and intermediate variables called mediators. Initially developed for fixed-time mediator and outcome, it has been extended to the framework of longitudinal data by discretizing the assessment times of mediator and outcome. Yet, processes in play in longitudinal studies are usually defined in continuous time and measured at irregular and subject-specific visits. This is the case in dementia research when cerebral and cognitive changes measured at planned visits in cohorts are of interest. We thus propose a methodology to estimate the causal mechanisms between a time-fixed exposure (X), a mediator process ($\mathcal{M}_t$) and an outcome process ($\mathcal{Y}_t$) both measured repeatedly over time in the presence of a time-dependent

confounding process ($\mathcal{L}_t$). We consider three types of causal estimands, the natural effects, path-specific effects and randomized interventional analogues to natural effects, and provide identifiability assumptions. We employ a dynamic multivariate model based on differential equations for their estimation. The performance of the methods are explored in simulations, and we illustrate the method in two real-world examples motivated by the 3C cerebral aging study to assess: (1) the effect of educational level on functional dependency through depressive symptomatology and cognitive functioning, and (2) the effect of a genetic factor on cognitive functioning potentially mediated by vascular brain lesions and confounded by neurodegeneration.

1 Introduction

Mediation analysis is commonly used in public health to assess the causal effect of an exposure on a system of variables. Mediation analysis aims to decipher the underlying mechanism by which an independent variable (X) affects a dependent variable (Y) via one or more intermediate variables (M), also called mediators. The total effect between X and Y is split into a direct effect and indirect effects through the intermediate variables M. Decomposing causal effects enhances our comprehension of the biological processes in play, and helps identify potential targets for therapeutic or prevention research.

The main framework for mediation analysis involves the definition of counterfactual or potential outcomes in a hypothetical world. Counterfactual outcomes are unobserved variables corresponding to the value the outcome Y would have taken if the exposure variable X had been modified in a certain way. Causal effects can then be defined as contrasts of counterfactual outcomes according to scenarios of intervention on X. Depending on the research question, different causal effects have been studied. For instance, the natural effects, introduced by [15], contrast counterfactual outcome values had an individual been exposed at two distinct levels of exposure (i.e. values are set at the individual level). In contrast, the stochastic effects, introduced by [3], contrast counterfactual outcome values had the distribution of the mediator been changed (i.e., intervention at the population level).

Concepts and methods for mediation analysis have been primarily developed for exposure, mediator and outcome measured at a single time point. With the inherent dynamic nature of health processes, extensions to longitudinal data have been recently proposed. For instance, [20] and [8] introduced methods for time-to-event outcomes with both the exposure variable and the mediator measured at a single time-point. Extensions to accommodate multiple mediators have followed ([9]; [6]). More recently, [23] proposed an alternative approach based on the natural effect proportional hazards model for a single-time exposure on a survival-type outcome with mediator and confounders both repeatedly measured over time. Within the stochastic intervention approach, [24] proposed a more flexible approach that handles

exposure repeatedly measured over time and [19] considered time-to-events for both the mediator and outcome.

In cerebral aging studies, the quantities of interest are usually dynamic processes such as cerebral volumes or cognitive functioning that are measured at planned visits. Considering such repeated data structure is crucial to obtain accurate results in mediation analyses but it remains a challenge. [21] defined randomized interventional analogues to natural effects (i.e., a stochastic intervention on the mediator) for a repeated exposure variable, a repeated mediator and a fixed-time outcome. [18] extended the approach to multiple longitudinal mediators. Mediation analysis techniques to repeated outcomes and mediator data are nevertheless still limited. Some authors extended the mediation methods assuming processes evolve in discrete time with regular measures ([10], [4]). Yet, processes in play lie in continuous time and may be observed only sparsely at very irregular timings across individuals. [2] considered a mediator and an outcome both defined in continuous time. They extended the identification assumptions of natural effects to continuous-time processes, and used a working model based on differential equations to estimate them. However, this was achieved under the strong assumption that only the current mediator level affects the outcome when the entire history of the mediator is likely involved. Moreover, the method did not consider time-varying confounders, and the estimation procedure was step-wise rather than simultaneous for all the variables ([16]).

Motivated by applications in cerebral aging to decipher causal mechanisms among the multidimensional spheres involved, we propose in this work a continuous-time mediation analysis framework to estimate causal effects of time-fixed exposure (X) in a system of mediator ($\mathcal{M}_t$), time-varying counfounders ($\mathcal{L}_t$) and outcome ($\mathcal{Y}_t$), all defined in continuous-time and measured at irregular and sparse visits. We consider natural effects in the absence of time-varying confounders, and path-specific effects to accommodate the presence of time-varying confounding factors, and we define for each case the identifiability assumptions required for their estimation. Under weaker identifiability assumptions, randomized interventional ana-

logues to natural effects can also be estimated under the proposed strategy. Our approach relies on a recently proposed multivariate mixed model ([17]) that quantifies the influences between dynamic processes of a network to estimate the conditional distributions of the mediator, confounder, and outcome processes from which the mediation g-formulas can be applied and the causal contrasts estimated. Two applications in cerebral aging research are considered from the population-based 3C cohort ([1]). We first estimate the pathways through cognitive functioning and depressive symptomatology involved in the association between educational level and functional dependency among olders. We secondly assess the relationship between ϵ_4 allele of the apolipoprotein E (APOE4), the main genetic factor for dementia, and cognitive functioning exploring the pathways through vascular cerebral lesions and neuro-degeneration.

2 Methods

2.1 Notation

We consider the setting described in Figure 1 with a time-fixed exposure X, time-fixed confounders C, as well as a time-dependent mediator and a time-dependent outcome, eventually (right panel) in the presence of time-dependent confounders L. The time-dependent mediator, confounder and outcome are processes defined in continuous time with their values at any time t ($t \in \mathbb{R}^+$) denoted $M(t)$, $Y(t)$, $L(t)$, and their history up to t denoted $\mathcal{M}_t$, $\mathcal{Y}_t$ and $\mathcal{L}_t$, respectively. In a cohort study, this setting translates into the collection for each subject i ($i = 1, ..., N$) of the exposure X_i and the confounder C_i at baseline, and of prone-to-error measures of the mediator $\tilde{M}_{ij}$, the outcome $\tilde{Y}_{ij}$ and the time-dependent confounders $\tilde{L}_{ij}$ at discrete visits j ($j = 1, ..., n_i$). These visits usually occur at subject-specific times t_{ij}, and eventually with different time schedules across the time-dependent variables.

To define causal effects, we introduce counterfactual variables. The counterfactual outcome $Y_t(x, \mathcal{L}_t(x''), \mathcal{M}_t(x', l_t))$ is defined as the value of the outcome that would have been

observed if X were set to x, $\mathcal{L}_t$ were set to the value it would have taken if X were set to x'', and $\mathcal{M}_t$ were set to the value it would have taken had X been set to x' and had $\mathcal{L}_t$ been set to l_t.

2.2 Causal effects definition

Depending on the framework and the objective, different causal effects can be defined using potential outcomes. We primarily consider natural effects, both in the absence of time-varying confounders, and in the presence of time-varying confounders with path-specific effects.

2.2.1 Natural effect

The natural effect decomposes the total effect (TE) of the exposure X on the outcome at time t, $Y(t)$, into a natural direct effect (NDE) defined as the effect of X on $Y(t)$ only, and a natural indirect effect (NIE) defined as the effect of X on $Y(t)$ passing through the mediator process up to t, $\mathcal{M}_t$.

Contrasting two levels x and x', these effects are expressed as follows:

$$TE = \mathbb{E}(Y_t(x, \mathcal{M}_t(x))|C) - \mathbb{E}(Y_t(x', \mathcal{M}_t(x'))|C) \tag{1}$$

$$NIE = \mathbb{E}(Y_t(x, \mathcal{M}_t(x))|C) - \mathbb{E}(Y_t(x, \mathcal{M}_t(x'))|C) \tag{2}$$

$$NDE = \mathbb{E}(Y_t(x, \mathcal{M}_t(x'))|C) - \mathbb{E}(Y_t(x', \mathcal{M}_t(x'))|C) \tag{3}$$

where TE is the difference in outcome had an individual been exposed at level x' instead of x, NIE is the difference in outcome had an individual been exposed to level x while the mediator process changed from what observed under level x' compared to level x. The NDE is the difference in outcome had an individual been exposed two different levels x and x', while keeping the mediator process fixed at what would be under exposure level x'.

Replacing $\mathcal{M}_t$ with stochastic intervention $G_{\mathcal{M}_t}$ in equations (1), (2), (3), randomized interventional analogues to natural effects can be identified under weaker assumptions de-

scribed in [21].

2.2.2 Path-specific effect

In the presence of time-dependent confounders, the path-specific effect approach decomposes the effect of the exposure X on the outcome at time t, $Y(t)$, into the direct effect of X on $Y(t)$ only (noted PSE_{XY}), the effect passing only through the mediator process up to t, $\mathcal{M}_t$ (noted PSE_{XMY}), and the effect passing first through the time-dependent confounder process up to t, $\mathcal{L}_t$ (noted $\text{PSE}_{XL(M)Y}$).

Contrasting two levels x and x', these effects are expressed as follows:

$$PSE_{XY} = \mathbb{E}(Y_t(x, \mathcal{L}_t(x'), \mathcal{M}_t(x', \mathcal{L}_t(x'))|C) - \mathbb{E}(Y_t(x', \mathcal{L}_t(x'), \mathcal{M}_t(x', \mathcal{L}_t(x'))|C) \quad (4)$$

$$PSE_{XMY} = \mathbb{E}(Y_t(x, \mathcal{L}_t(x'), \mathcal{M}_t(x, \mathcal{L}_t(x'))|C) - \mathbb{E}(Y_t(x, \mathcal{L}_t(x'), \mathcal{M}_t(x', \mathcal{L}_t(x'))|C) \quad (5)$$

$$PSE_{XL(M)Y} = \mathbb{E}(Y_t(x, \mathcal{L}_t(x), \mathcal{M}_t(x, \mathcal{L}_t(x))|C) - \mathbb{E}(Y_t(x, \mathcal{L}_t(x'), \mathcal{M}_t(x, \mathcal{L}_t(x'))|C) \quad (6)$$

For example, PSE_{XMY} (equation (5)) is the expected difference in outcome had the mediator process only changed due to an exposure change from x' to x, while keeping fixed the exposure and the time-dependent confounders.

2.3 Assumptions

The causal contrasts and expectations defined above rely on potential outcomes that are not observable. To identify these effects from the observations, it is mandatory to comply to four sets of fundamental assumptions: **consistency, positivity, sequential ignorability** and **cross-world independence**. These assumptions vary depending on the presence or absence of time-varying confounders. In the following, we make the distinction between the two cases with $\overline{TVC}$ and TVC specifying the absence and the presence of time-varying confounding factors, respectively.

(i) **consistency**: The consistency assumption establishes the connection between coun-

terfactual variables and observed variables ([12]). The value of the observed outcome $Y(t)$ is equal to the value of the corresponding counterfactual outcome, i.e. for $\overline{TVC}$: $Y(t)$=$Y_t(x, m_t)$ if X=x, $\mathcal{M}_t = m_t$; and for TVC: $Y(t) = Y_t(x, m_t, l_t)$ if X=x, $\mathcal{M}_t = m_t$ and $\mathcal{L}_t = l_t$.

(ii) **positivity** : The positivity assumption stipulates that each individual, given their values of C, has a positive probability of receiving any exposure value, and a positive probability to have any mediator and time-varying confounder history ([12]). It means that for $\overline{TVC}$: $P(X = x|C) > 0$ and $P(\mathcal{M}_t(x) = m_t|X = x, C) > 0$, and for TVC: $P(X = x|C) > 0$, $P(\mathcal{L}_t(x) = l_t|X = x, C) > 0$ and $P(\mathcal{M}_t(x, l_t) = m_t|X = x, C, \mathcal{L}_t = l_t) > 0$.

(iii) **sequential ignorability**: this assumption defines the absence of unobserved confounding in the system through 3 to 5 sub-assumptions:

(a) There is no unobserved confounding of the effect of X on Y_t given other variables, i.e.: $Y_t(x, m_t) \perp\!\!\!\perp X|C$, for $\overline{TVC}$, and $Y_t(x, l_t, m_t) \perp\!\!\!\perp X|C$ for TVC.

(b) There is no unobserved confounding of the effect of intermediate processes on Y_t given C and X, i.e., $Y_t(x, m_t) \perp\!\!\!\perp \mathcal{M}_t|C, X$ for $\overline{TVC}$, and $Y_t(x, l_t, m_t) \perp\!\!\!\perp \mathcal{L}_t, \mathcal{M}_t|C, X$ for TVC.

(c) There is no unobserved confounding of the effect of X on intermediate processes given C, i.e., $\mathcal{M}_t(x) \perp\!\!\!\perp X|C$ for $\overline{TVC}$, and $\mathcal{L}_t$,$\mathcal{M}_t(x) \perp\!\!\!\perp X|C$ for TVC.

(d) For TVC only, there is no unobserved confounding between the mediator process $\mathcal{M}_t$ and (X,$\mathcal{L}_t$) given C, i.e., $\mathcal{M}_t(x, l_t) \perp\!\!\!\perp X, \mathcal{L}_t|C$.

(e) For TVC only, there is no unobserved confounding between the counfounder process $\mathcal{L}_t$ and the mediator process $\mathcal{M}_t$ affected by X given C, i.e., $\mathcal{M}_t(x, l_t) \perp\!\!\!\perp \mathcal{L}_t(x')|X, C$.

(iv) **Cross-world independence assumption:** this assumption stipulates the absence of

unobserved confounding in the counterfactual world:

(a) There is no unobserved confounding between the potential outcome and intermediates variables, i.e., $Y_t(x, m_t) \perp\!\!\!\perp \mathcal{M}_t(x')|X, C$ for $\overline{TVC}$, and $Y_t(x, l_t, m_t) \perp\!\!\!\perp \mathcal{L}_t(x'), \mathcal{M}_t(x')|X, C$ for TVC.

(b) For TVC only, there is no unobserved confounding between the potential outcome and different potential intermediate processes, i.e., $Y_t(x, l_t, m_t) \perp\!\!\!\perp \mathcal{L}_t(x'), \mathcal{M}_t(x, l_t)|X, C$.

2.4 Identification

2.4.1 Natural Effects

The natural effects are systematically defined as a comparison of two expectations $\upsilon = \mathbb{E}(Y_t(x, \mathcal{M}_t(x'))|C)$ where x and x' can take various values depending on the effect (NE, NIE, NDE). Using the assumptions $((i), (ii), (iii.a), (iii.b), (iii.c), (iv.a))$, this expectation can be written as a function of the observations, and thus becomes identifiable. First, the estimand can be developed according to the mediator history $\mathcal{M}_t$:

$$\upsilon = \int_{m_t} \mathbb{E}(Y_t(x, m_t)|C = c, \mathcal{M}_t(x') = m_t) \times f_{\mathcal{M}_t(x')|C=c}(m_t)\ d_{\mathcal{M}_t(m_t)}$$

Second, thanks to assumption $(iv.a)$, we can remove the conditioning $\mathcal{M}_t(x') = m_t$.

$$\upsilon = \int_{m_t} \mathbb{E}(Y_t(x, m_t)|C = c) \times f_{\mathcal{M}_t(x')|C=c}(m_t)\ d_{\mathcal{M}_t(m_t)}$$

Third, in order to use the consistency assumption, we add the conditioning on $X = x$ in the expectation of Y_t and $X = x'$ in the density of $\mathcal{M}_t$ thanks to assumptions (i) and $(iii.c)$:

$$\upsilon = \int_{m_t} \mathbb{E}(Y_t(x, m_t)|C = c, X = x) \times f_{\mathcal{M}_t(x')|C=c,X=x'}(m_t)\ d_{\mathcal{M}_t(m_t)}$$

We also add the conditioning on $\mathcal{M}_t = m_t$ in the expectation of Y_t thanks to assumption

$(iii.b)$:

$$\upsilon = \int_{m_t} \mathbb{E}(Y_t(x, m_t)|C = c, X = x, \mathcal{M}_t = m_t) \times f_{\mathcal{M}_t|C=c,X=x'}(m_t)\ d_{\mathcal{M}_t(m_t)}$$

Finally, the consistency assumption i can be applied to obtain an expression that only involves the observations so that:

$$\mathbb{E}(Y_t(x, \mathcal{M}_t(x'))|C) = \int_{m_t} \mathbb{E}(Y_t|C = c, X = x, \mathcal{M}_t = m_t) \times f_{\mathcal{M}_t|C=c,X=x'}(m_t)\ d_{\mathcal{M}_t(m_t)} \quad (7)$$

2.4.2 Path-specific effect

In the presence of time-varying confounders $\mathcal{L}_t$, assumption $iii.c$ is violated and the third step of the identification of the natural effect cannot be applied anymore. Alternatively, path-specific effects are considered. They are systematically defined as a comparison of two expectations $\xi = \mathbb{E}(Y_t(x, \mathcal{L}_t(x), \mathcal{M}_t(x', l_t))|C)$, where x and x' can take various values.

Using similar developments as for the natural effects (see Web Supplementary Material, section 1 for details), we obtain:

$$\begin{aligned}\mathbb{E}(Y_t(x, \mathcal{L}_t(x), \mathcal{M}_t(x', l_t))|C) = \int_{l_t}\int_{m_t} \mathbb{E}(Y_t|C = c, X = x, \mathcal{L}_t = l_t, \mathcal{M}_t = m_t)\times \\ f_{\mathcal{L}_t|C=c,X=x'}(l_t), f_{\mathcal{M}_t|(C=c,X=x,\mathcal{L}_t=l_t)}(m_t)\ d_{m_t} d_{l_t}\end{aligned} \quad (8)$$

where a different set of assumptions is used depending on the path: assumptions (i)-$(iii.c)$ and $(iv.a)$ for the PSE_{XY} (direct path), assumptions (i)-$(iii.c)$,$(iii.e)$,$(iv.b)$ for the PSE_{XMY} (path through $\mathcal{M}_t$ only), and assumptions (i)-$(iii.a)$, $(iii.c)$, $(iv.a)$-$(iv.b)$ for the $PSE_{X(L)MY}$ (indirect paths through $\mathcal{L}_t$).

2.5 Estimation

2.5.1 Monte Carlo approximation

The causal contrasts are differences of expectation expressions developed in Equations 7 and 8 for the natural effects and for the path-specific effects, respectively. These expectations are general expressions to be estimated from the data. In some specific modeling cases, an analytical solution can be computed. Otherwise, the integrals can be approximated by the Monte Carlo approach ([5]), with B indicating the number of Mont-Carlo replicates. For the path-specific effects, this gives:

$$\mathbb{E}(Y_t(x, \mathcal{L}_t(x'), \mathcal{M}_t(x', \mathcal{L}_t(x')))) \approx \sum_{k=1}^{B} \mathbb{E}(Y_t | X = x, \mathcal{L}_t = l_t^{(k)}, \mathcal{M}_t = m_t^{(k)})$$

with random draws from conditional distributions: $f_{\mathcal{L}_t|X=x',C=c}$ and $m_t^{(k)} \sim f_{\mathcal{M}_t|C=c,X=x',\mathcal{L}_t=l_{tj}}$. The same approximation is obtained for the natural effects by removing mention to the time-varying confounder.

The Monte-Carlo approximation highlights that the causal estimands calculation requires:

- the conditional expectation of Y_t given the intermediate processes and time-fixed factors;
- the conditional distribution of $\mathcal{M}_t$ given the exposure, and eventually the time-varying confounders;
- in the presence of time-varying confounders $\mathcal{L}_t$, their distribution conditional on the exposure and the time-fixed confounders.

These quantities can be obtained from a statistical model, called working model, using the posterior distributions of the different processes in play.

2.5.2 Example of working model

Any statistical model can be used to estimate these causal effects as long as (i) the processes are jointly modelled taking into account the interrelations between processes and the effect of the time-fixed exposure, and (ii) the posterior conditional distributions of the processes can be derived. In this work, we opted for a multivariate mixed model based on differential equations ([17]). This model was developed to quantify the temporal influences between a system of latent processes measured by repeated marker data. In our case, the processes are $\mathcal{L}_t, \mathcal{M}_t$ and $\mathcal{Y}_t$ with values $L_i(t)$, $M_i(t)$ and $Y_i(t)$ at time t for subject i ($i = 1, ..., N$). Their trajectories are defined by the initial level at time 0 and the instantaneous change over time, both modelled in the mixed modeling framework using random effects ([7]). The temporal dependencies between the processes are modelled by the effect of the current level of one process on the instantaneous change of another. The model can be written as follows:

$$\text{For process } \mathcal{L}_t : \begin{cases} L_i(0) = \mathbf{X}_i^{L(0)}\boldsymbol{\beta}^L + u_i^L \\ \dfrac{\partial L_i(t)}{\partial t} = \mathbf{X}_i^L(t)\boldsymbol{\gamma}^L + \mathbf{Z}_i^L(t)\mathbf{v}_i^L \end{cases} \tag{9}$$

$$\text{For process } \mathcal{M}_t : \begin{cases} M_i(0) = \mathbf{X}_i^{M(0)}\boldsymbol{\beta}^M + u_i^M \\ \dfrac{\partial M_i(t)}{\partial t} = \mathbf{X}_i^M(t)\boldsymbol{\gamma}^M + \mathbf{Z}_i^M(t)\mathbf{v}_i^M + \alpha_i^{ML} L_i(t) \end{cases} \tag{10}$$

$$\text{For process } \mathcal{Y}_t : \begin{cases} Y_i(0) = \mathbf{X}_i^{Y(0)}\boldsymbol{\beta}^Y + u_i^Y \\ \dfrac{\partial Y_i(t)}{\partial t} = \mathbf{X}_i^Y(t)\boldsymbol{\gamma}^Y + \mathbf{Z}_i^Y(t)\mathbf{v}_i^Y + \alpha_i^{YL} L_i(t) + \alpha_i^{YM} M_i(t) \end{cases} \tag{11}$$

$$\tag{12}$$

where $\mathbf{X}_i^{L(0)}$, $\mathbf{X}_i^{M(0)}$, $\mathbf{X}_i^{Y(0)}$ are the vectors of covariates associated with the initial levels of the three processes through parameters $\boldsymbol{\beta}^L$, $\boldsymbol{\beta}^M$, $\boldsymbol{\beta}^Y$. These vectors include the intercept, the time-fixed exposure X and confounders $\boldsymbol{C}$. Vectors of covariates $\mathbf{X}_i^L(t)$, $\mathbf{X}_i^M(t)$, $\mathbf{X}_i^Y(t)$ are associated with the change over time of the processes through parameters $\boldsymbol{\gamma}^L$, $\boldsymbol{\gamma}^M$, $\boldsymbol{\gamma}^Y$. They include the intercept, time functions (allowing for nonlinear change over time), the exposure and confounders as well as their eventual interactions with the time functions.

The vectors $\mathbf{Z}_i^L(t)$, $\mathbf{Z}_i^M(t)$, $\mathbf{Z}_i^Y(t)$ include the intercept and, eventually time functions, to be associated with the changes over time through the individual random effects $\mathbf{v}_i^L$, $\mathbf{v}_i^M$, $\mathbf{v}_i^Y$. The random intercepts on the initial levels u_i^L, u_i^M, u_i^Y, and the random effects on the changes over time $\mathbf{v}_i^L$, $\mathbf{v}_i^M$, $\mathbf{v}_i^Y$ account for the intra-individual correlation and follow a zero-mean multivariate normal distribution with variance-covariance G. The influences between processes are captured by the $\alpha_i^{aa'}$ that quantify the effect of process a' on the instantaneous change over time of process a. The influences $\alpha_i^{aa'}$ can be modelled as a linear combination of covariates to account for instance for exposure interaction with intermediate processes: $\alpha_i^{aa'} = \alpha_0^{aa'} + \boldsymbol{X}_i^\top \boldsymbol{\alpha}_1^{aa'}$ with $\boldsymbol{X_i}$ a vector of covariates.

The structural models at the process level are linked to the error-prone observations in equations of observations, assuming in this work additive errors:

$$\begin{cases} \tilde{L}_{ij} = L_i(t_{ij}) + \epsilon_{ij}^L & \text{for } j = 1, ..., n_i^L \\ \tilde{M}_{ij} = M_i(t_{ij}) + \epsilon_{ij}^M & \text{for } j = 1, ..., n_i^M \\ \tilde{Y}_{ij} = Y_i(t_{ij}) + \epsilon_{ij}^Y & \text{for } j = 1, ..., n_i^Y \end{cases} \tag{13}$$

with ϵ_{ij}^L, ϵ_{ij}^L, and ϵ_{ij}^L independent zero-mean Gaussian variables with variances σ_L^2, σ_M^2, and σ_Y^2, respectively.

The model is estimated in the maximum likelihood framework in the R package CInLPN (https://github.com/bachirtadde/CInLPN) using the iterative Marquardt-Levenberg algorithm of R package marqLevAlg ([13]). To achieve an analytical likelihood calculation, the program approximates the differential equations by difference equations with a fine time grid defined by a step δ to be specified by the user.

Since this model is Gaussian and linear, analytical solutions can be found for the expectations ν and ξ without requiring a Monte Carlo approximation.

2.5.3 Confidence intervals

Since the model is parametric, the confidence intervals of the causal effects can be computed by a parametric bootstrap procedure. R random vectors of parameters $\boldsymbol{\theta}^r$ (for $r = 1, ..., R$) are repeatedly drawn from the asymptotic distribution $\mathcal{N}(\hat{\boldsymbol{\theta}}, \widehat{V(\hat{\boldsymbol{\theta}})})$ R where $\hat{\boldsymbol{\theta}}$ and $\widehat{V(\hat{\boldsymbol{\theta}})}$ are the maximum likelihood estimates of the model and their Hessian-based variance estimate, respectively. The causal effects are computed for each draw r and the 95% confidence interval of the causal effect is given by the 2.5% and 97.5% percentiles of the causal effects over the R replicates.

3 Numerical evaluation by simulations

We conducted a simulation study to assess the properties and behavior of our methodology without and with time-varying confounders, under two potential timescales: time in study (Scenarios 1), and age (Scenarios 2).

3.1 Data generating mechanism

For each individual i in a sample of size N, we generated an exposure variable X_i according to a Bernoulli distribution (with probability 0.5). Depending on the presence or absence of a time-varying confounder $\mathcal{L}_t$, we generated a system of two or three Gaussian processes using the working model defined in section 2.5.2. We assumed a constant rate of change for all the processes with random intercepts and simple effects of the covariate on both the initial level and instantaneous change of each process. We also assumed, as in Figure 1B, that only process $\mathcal{L}_t$ impacted the change in $\mathcal{M}_t$, and that processes $\mathcal{L}_t$ and $\mathcal{M}_t$ impacted the change in $\mathcal{Y}_t$. Dropout was generated according to a uniform distribution in the range of measures of each scenario (see below). Data generation process including parameters values considered are fully summarized in Web Supplementary Material Section 2. For all the scenarios, we generated 250 samples.

3.1.1 Simulation design

Scenario 1 - Main simulation design We assumed the repeated data were collected annually over a period of 5 years and used time in the study as the timescale. In scenario 1A, we considered samples of 500 subjects, a 10% dropout and a discretization step in the model of 0.1 year. In scenarios 1B-1D, we sequentially changed the discretization step, the size of the sample and the dropout rate (see web supplementary Table 2). The generating model parameters (see Web Supplementary Section 2) were randomly chosen.

Scenario 2 - Secondary simulation design We mimicked one of the application setting by considering the subjects entered the cohort at different ages (simulated according to Normal distribution $\mathcal{N}(72, 4)$), using age at the time-scale, and having some systematic missing values in markers by design. Specifically, following the 3C design, the outcome was collected at 0, 2, 4, 7, 10, 13, and 15 years after entry while mediator and confounder were collected only at 0, 4, and 10 years after entry. The sample included 500 subjects, with a discretization interval of 1 year and a fixed dropout rate of 10%. The generating model parameters, reported in Supplementary Table 4 were chosen to mimic the application.

3.2 Estimands

Our estimands were the direct effect of the exposure on the outcome not via the mediator nor via the time-varying confounder eventually (NDE and PSE_{XY} in equations (3) and (4)), the indirect effect of the exposure on the outcome via the mediator only (NIE and PSE_{XMY} in equations (2), and (5)) and, in presence of time-varying confounder, the indirect effect of the exposure on the outcome via all paths through the time-varying confounder ($PSE_{XL(M)Y}$ in equation (6)).

3.3 True generated contrast values

The true effect values were directly computed from equations (2) and (3), or equations (4), (5) and (6) under different exposure levels. In a general context, this is achieved by averaging the outcome over a large population under each scenario. However, given the linear structure of the specific working model we used, this was not necessary; the calculations were done on a single generated individual under different scenarios:

- Natural effect: value of $Y(t)$ when $X = x$ and $\mathcal{M}_t = \mathcal{M}_t(X = x')$ with x and $x' \in \{0, 1\}$
- Path-specific effect: value of $Y(t)$ when $X = x$, $\mathcal{L}_t = \mathcal{L}_t(X = x')$, and $\mathcal{M}_t = \mathcal{M}_t(X = x'', \mathcal{L}_t)$, with x, x' and $x'' \in \{0, 1\}$

3.4 Working model

In all the scenarios, the working model was the multivariate mixed model based on differential equations detailed in Section 2.5.2.

3.5 Performance measures

We computed the causal effects at 1, 2, 3, 4 and 5 years for Scenarios 1A-1D and at age 65,70,75,80,85 years for Scenario 2. The estimation quality of each effect was assessed by the distribution of the relative bias (i.e., the bias standardized by the true value, expressed in percentage) reported in violin plots, and the coverage rate of the 95% confidence interval is given in Web Supplementary Table 3.

3.6 Simulation results

The effects depicted in Figures 2 and 3 illustrate the correct estimation of the effects at all times under scenario 1A. With a sample size of 500 individuals, a discretization step of 0.1 year, and a censoring rate of 10%, the estimates show minimal bias and coverage rates close to

the 95% nominal value in the absence and in the presence of a time-dependent confounder. The results are the same when changing the simulation characteristics in scenario 1B-1D (Web supplementary Figures 1,2,3).

When considering scenario 2 in which the mediator and the confounder were measured only two to three times while the outcome was measured at six time points, the favorable results achieved previously were not replicated (Figure 4). Interestingly, this bias seemed to be driven by biased estimates in the working model.

4 Application to cerebral aging

We applied the methodology to investigate the underlying mechanisms of cognitive aging through two examples:

Study 1: We assessed the impact of the educational level on functional dependency investigating the pathways through verbal fluency and depressive symptomatology.

Study 2: We investigated the influence of cerebral vascular lesions in the relationship between the main genetic factor of dementia, Apolipoprotein E4 gene (ApoE4), and cognitive functioning, accounting for the potential confounding due to neurodegeneration.

4.1 The Three-City study sample

4.1.1 The cohort

We leveraged the data from the Three-City (3C) study, a prospective population-based cohort designed to investigate the association between vascular diseases and dementia in the elderly. Individuals aged 65 years and older were randomly enrolled in 1999 from the electoral lists of three French cities (Bordeaux, Dijon and Montpellier). A total of 9294 participants underwent a comprehensive health examination and risk factor assessment at baseline, and

at follow-up visits every 2-3 years for a duration of up to 17 years. A Magnetic Resonance Imaging (MRI) assessment was also performed on a subsample at baseline, 4 years and only for Bordeaux center at 10 years of follow-up.

The analytical sample comprised the 2,213 participants from Bordeaux and Dijon who were genotyped using genome-wide genotyping arrays, underwent at least one MRI scan, had at least one measure for each marker considered in both studies, no missing covariate for exposures or potential confounders and were free of dementia at baseline. This sample had on average a follow-up of 9.5 years.

4.1.2 Variables of interest

In Study 1 (Figure 5, top panel), the exposure was the binary educational level (high school and higher *versus* lower in reference). The final outcome was the functional dependency measured by the sum-score (range 0-5) of impairment at 5 Instrumental Activities of Daily Living (IADL) (Using the phone, transportation, medication management, finances management and shopping)(the higher, the more dependent). The mediator was the verbal fluency measured by the Isaacs Set Test (IST). The IST score equals the count of words provided across four semantic categories within a 15-second interval each. The potential confounder was the depressive symptomatology measured by the score at the Center for Epidemiologic Studies Depression Scale (CESD). The three processes were evaluated at each follow-up evaluation, although some missing data arised.

In Study 2 (Figure 5, bottom panel), the exposure was the carriership of the ApoE4 and the final outcome was the verbal fluency measured by the IST score. The mediator under investigation was the vascular cerebral lesions measured by the global volume of White Matter Hyperintensities (WMH), and the potential confounder was the global neurodegeneration as measured by the total volume of grey matter (GM). WMH and GM were only collected twice (at 1 and 4 years) in Dijon center, and three times (at 1, 4 and 10 years) in Bordeaux Center.

In both studies, potential confounding factors at baseline were sex, age at baseline and

the center (Bordeaux/Dijon).

4.1.3 Description of the analytical sample

Among the 2,213 participants of the analytical sample, 1375 (63.6%) were women, 886 (40.1%) had an educational level higher than secondary school and 417 (20.9%) were APOE4 carriers (Table 1). Participants were 72.3 years old at baseline on average, and they were followed up for 9.5 years on average with a mean of 1.7 (sd=0.7) repeated measures of WMH and GM, a mean of 6 (sd=2.2) repeated measures of IST, and a mean of 5 (sd=1.6) repeated measures of CESD and IADL.

4.2 Path-specific effects

4.2.1 Working model specification and estimands

Path-specific contrasts were estimated using the working model defined in Equations (12) and (13). In both studies, the timescale was the age and the discretization step was of 1 year. The working model systematically assumed a constant change over age. For each component, both the initial level and the change over time were adjusted for ApoE4, educational level, age at entry in the cohort, sex and center, and included an individual random-effect. In study 2, the Grey matter volume model was further adjusted for the total intra-cranial volume. In both studies, the effect of $\mathcal{L}$ on $\mathcal{M}$, and the effect of $\mathcal{L}$ and $\mathcal{M}$ on $\mathcal{Y}$ were in interaction with the exposure variable. The specifications of the two working models are detailed in Web Supplementary section 3.

To satisfy the model's assumptions, WMH volume was log transformed, and CESD and IADL scores were normalized in a preliminary step using integrated splines. In the following, IADL normalized score measuring functional dependency is expressed in Standard Deviation of the population at 65 years old.

The estimates of the working models for the two studies are reported in Web Supplementary Tables 5 and 6.

4.2.2 Identifiability assumptions

The path-specific effects can be identified under the assumptions listed in Subsection 2.3. In particular, we assume that there is no remaining confounding between the exposure and the outcome variables, between the intermediate variables and the outcomes, and between the exposure and the intermediate variables after adjustment for considered confounders.

4.2.3 Study 1

The path-specific effects of educational level on functional dependency are plotted in Figure 6. Overall, a higher educational level induced a lower functional level at all ages, with a difference increasing with age. This effect was largely mediated by the path through cognitive functioning. It was also slightly due to the path through depressive symptomatology, both in the direction of higher educational level implying lower functional level. The natural direct effect of educational level on functional level was in the opposite direction but not significant.

4.2.4 Study 2

The path-specific effects relating the presence of allele ϵ 4 of the Apolipoprotein E to cognitive functioning are displayed in Figure 6 (bottom panel). In this example, the total effect of ApoE4 on cognitive functioning was driven by its direct effect and was more pronounced as age increased. The two path-specific effects through the vascular lesions as measured by the White Matter Hyper-intensities, and through the neuro-degeneration confounding factor as measured by the total volume of Grey Matter were negligible. However these results should be interpreted with caution as they are very likely biased. As shown in the simulation scenario 2, in such a context with only a few measures for the intermediate variables (average of 1.6 (sd=0.7) measures per subject), the path-specific effects cannot be correctly retrieved.

5 Discussion

In this paper, we have expanded the methodology of mediation analysis by proposing an approach for mediator, confounder and outcome that are processes defined in continuous time and measured at sparse and possibly irregular visits in prospective studies. In addition to the natural indirect and direct effects that only hold in the absence of time-dependent confounders, we developed the path-specific effects to address mediation analyses in the presence of time-dependent confounders. A simulation study underlined that the causal contrasts, derived from a single multivariate longitudinal working model, were correctly estimated provided the repeated information collected on the processes in play was rich enough. This prevented for instance the application of mediation analysis to assess the mediating effect of a genetic factor on cognition through MRI-derived features that were measured only two or three times in a population-based cohort.

We primarily focused in this work on natural effects, defining the estimands and the hypotheses to make these quantities identifiable. However, it is important to note that the same methodology also applies to other types of effects. By replacing the process $\mathcal{M}_t$ in the contrasts definitions with a stochastic intervention $G_{\mathcal{M}_t}$, also called randomized intervention, the methodology also extends the stochastic intervention approach to mediator, confounder and outcomes defined in continuous time. In contrast with natural effects, stochastic effects don't require the cross-world independence assumption, and thus apply more broadly notably in the presence of an exposure-induced confounder ([22]). Their interpretation also differs from natural effects as in general they measure the impact of interventions at the population level, rather than mediating mechanisms. ([11]).

Causal contrasts estimation requires the use of a working model from which conditional distributions of mediator and outcome can be derived. We used in this study a multivariate mixed-effects model based on differential equations ([17]) to estimate the joint distribution of all the processes in play while taking into account the effect of the history of the mediator and confounder processes on the outcome. However any other working model adapted to mul-

tivariate interrelated processes could be considered instead to estimate the causal contrasts we proposed.

In a first application, we observed a mediating effect of cognition on the relationship between education level and functional dependency in the elderly. Although we found a non-significant trend associating higher education levels with increased functional dependency, the overall effect suggests that education may promote greater autonomy. This effect, primarily influenced by the mediation through cognition and to a lesser extent through depression, underscores the importance of considering these factors in understanding the dynamics of functional aging.

The second application exploring the impact of the main genetic factor of Alzheimer's disease ([14]), ApoE4, on cognitive level in the elderly suggested that the total effect of ApoE4 on cognitive functioning was not mediated by its effect through vascular lesions or overall grey matter atrophy. However, these results should not be interpreted further. We chose to report them in order to emphasize the limits of mediation analyses on longitudinal data. Indeed, the simulation study showed that causal contrasts could not be retrieved correctly when the repeated information was poor as in this case with a couple of repeated measures for the intermediate processes.

Mediation analysis had already been extended to longitudinal data. However, methods were mainly restricted to discrete time when processes in play usually lie in continuous time and are measured in prospective studies at irregular timings across individuals, and possibly across variables. By defining the causal contrasts at the process level and using a working model adapted to irregular longitudinal data, our methodology goes one step further to address mediation questions related to time-fixed exposures in prospective cohorts. We leave to future work extensions to time-dependent exposure variables.

5.1 Funding

This work was funded by the French government in the framework of the PIA3 ("Investment for the future") (project reference 17-EURE-0019). It was also carried out in the framework of the University of Bordeaux's France 2030 program / RRI PHDS, and of the DyMES project funded by the French National Research Agency (ANR-18-CE36-0004). This work was also supported by an EHESP Doctoral Network Fellowship.

References

[1] 3C Study Group (2003). Vascular factors and risk of dementia: design of the Three-City Study and baseline characteristics of the study population. Neuroepidemiology **22,** 316–325.

[2] Albert, J. M., Li, Y., Sun, J., Woyczynski, W. A., and Nelson, S. (2019). Continuous-time causal mediation analysis. Statistics in Medicine **38,** 4334–4347.

[3] Avin, C., Shpitser, I., and Pearl, J. (2005). Identifiability of Path-Specific Effects.

[4] Bind, M.-A. C., Vanderweele, T. J., Coull, B. A., and Schwartz, J. D. (2016). Causal mediation analysis for longitudinal data with exogenous exposure. Biostatistics **17,** 122–134.

[5] Booth, J. G. and Sarkar, S. (1998). Monte Carlo Approximation of Bootstrap Variances. The American Statistician **52,** 354–357.

[6] Huang, Y.-T. and Yang, H.-I. (2017). Causal mediation analysis of survival outcome with multiple mediators. Epidemiology **28,** 370–378.

[7] Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. Biometrics **38,** 963–974.

[8] Lange, T. and Hansen, J. V. (2011). Direct and Indirect Effects in a Survival Context. Epidemiology **22,** 575–581.

[9] Lange, T., Rasmussen, M., and Thygesen, L. C. (2014). Assessing natural direct and indirect effects through multiple pathways. Am J Epidemiol **179,** 513–518.

[10] Mittinty, M. N. and Vansteelandt, S. (2020). Longitudinal Mediation Analysis Using Natural Effect Models. Am J Epidemiol **189,** 1427–1435.

[11] Moreno-Betancur, M. and Carlin, J. B. (2018). Understanding Interventional Effects: A More Natural Approach to Mediation Analysis? Epidemiology **29,** 614.

[12] Nguyen, T. Q., Schmid, I., Ogburn, E. L., and Stuart, E. A. (2022). Clarifying causal mediation analysis: Effect identification via three assumptions and five potential outcomes. Journal of Causal Inference **10,** 246–279.

[13] Philipps, V., Hejblum, P., B., Prague, M., Commenges, D., and Proust-Lima, C. (2021). Robust and Efficient Optimization Using a Marquardt-Levenberg Algorithm with R Package marqLevAlg. The R Journal **13,** 273.

[14] Reitz, C., Brayne, C., and Mayeux, R. (2011). Epidemiology of Alzheimer disease. Nat Rev Neurol **7,** 137–152.

[15] Robins, J. M. and Greenland, S. (1992). Identifiability and Exchangeability for Direct and Indirect Effects. Epidemiology **3,** 143–155.

[16] Saunders, C. T. and Blume, J. D. (2018). A classical regression framework for mediation analysis: fitting one model to estimate mediation effects. Biostatistics **19,** 514–528.

[17] Taddé, B. O., Jacqmin-Gadda, H., Dartigues, J.-F., Commenges, D., and Proust-Lima, C. (2020). Dynamic modeling of multivariate dimensions and their temporal relationships using latent processes: Application to Alzheimer's disease. Biometrics **76,** 886–899.

[18] Tai, A.-S., Lin, S.-H., Chu, Y.-C., Yu, T., Puhan, M. A., and VanderWeele, T. (2023). Causal Mediation Analysis with Multiple Time-varying Mediators. Epidemiology **34,** 8.

[19] Valeri, L., Proust-Lima, C., Fan, W., Chen, J. T., and Jacqmin-Gadda, H. (2023). A multistate approach for the study of interventions on an intermediate time-to-event in health disparities research. Stat Methods Med Res **32,** 1445–1460.

[20] VanderWeele, T. J. (2011). Causal mediation analysis with survival data. Epidemiology **22,** 582–585.

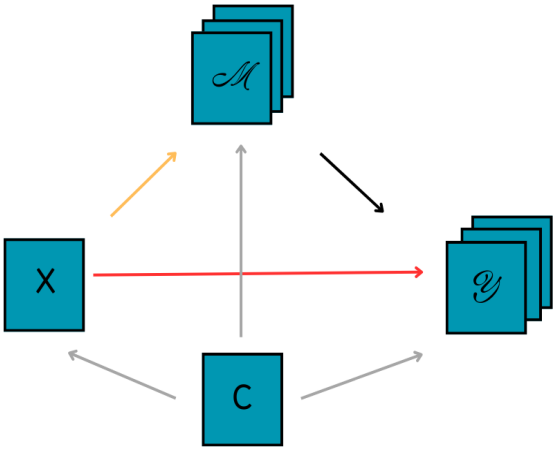
[21] VanderWeele, T. J. and Tchetgen Tchetgen, E. J. (2017). Mediation analysis with time varying exposures and mediators. J R Stat Soc Series B Stat Methodol **79,** 917–938.

[22] Vanderweele, T. J., Vansteelandt, S., and Robins, J. M. (2014). Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. Epidemiology **25,** 300–306.

[23] Vo, T.-T., Superchi, C., Boutron, I., and Vansteelandt, S. (2020). The conduct and reporting of mediation analysis in recently published randomized controlled trials: results from a methodological systematic review. J Clin Epidemiol **117,** 78–88.

[24] Zheng, W. and Laan, M. J. (2012). Causal Mediation in a Survival Setting with Time-Dependent Mediators.

(A) Without time-varying confounders



(B) With time-varying confounders

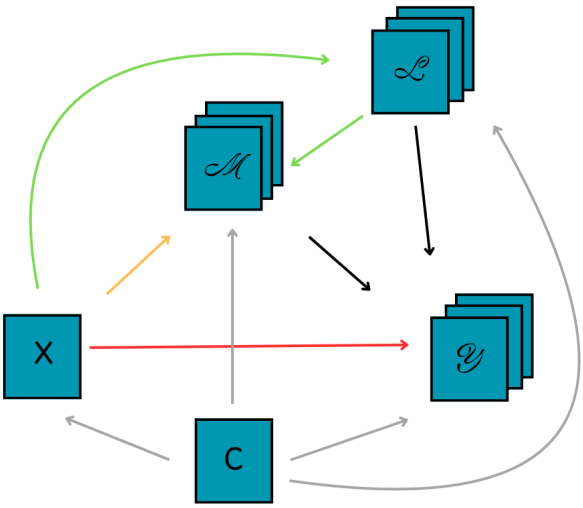
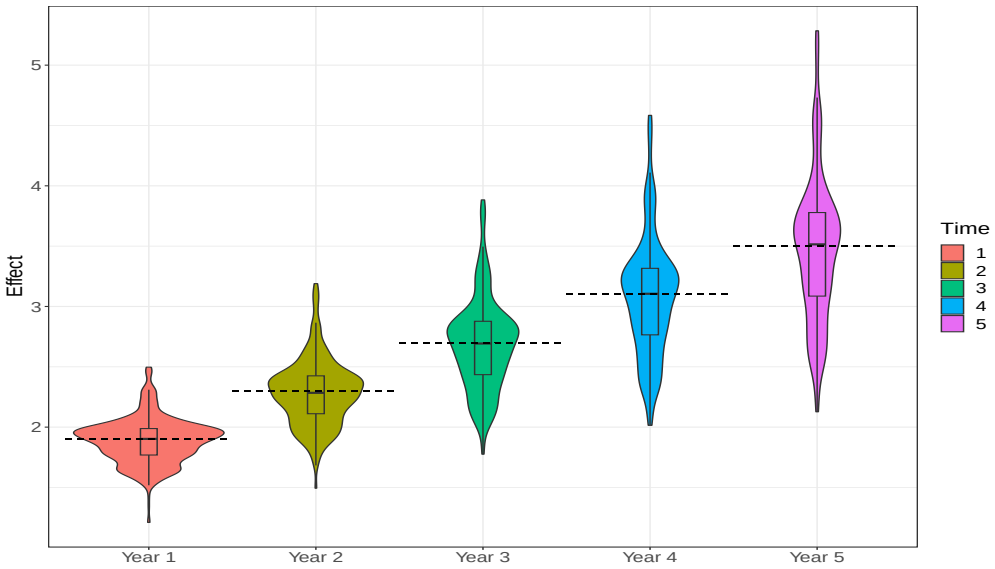


Figure 1: Causal Mediation Path Diagram: Exploring the mechanism between a time-fixed exposure X and an outcome process $\mathcal{Y}_t$ through: (A) mediator process $\mathcal{M}_t$ (B) mediator $\mathcal{M}_t$ and time-varying confounders $\mathcal{L}_t$ processes, given baseline confounders C

(A) Natural Direct Effect



(B) Natural Indirect Effect

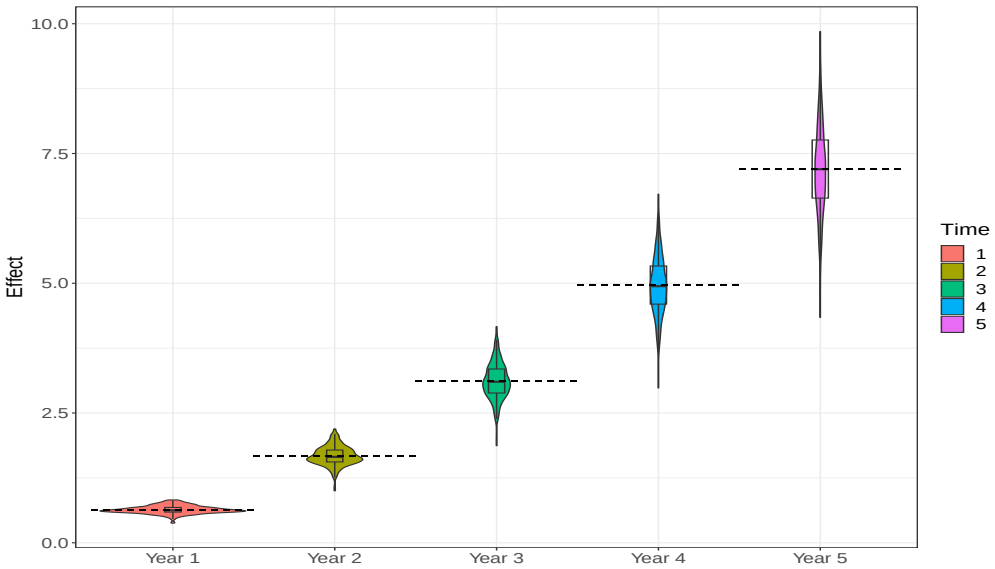
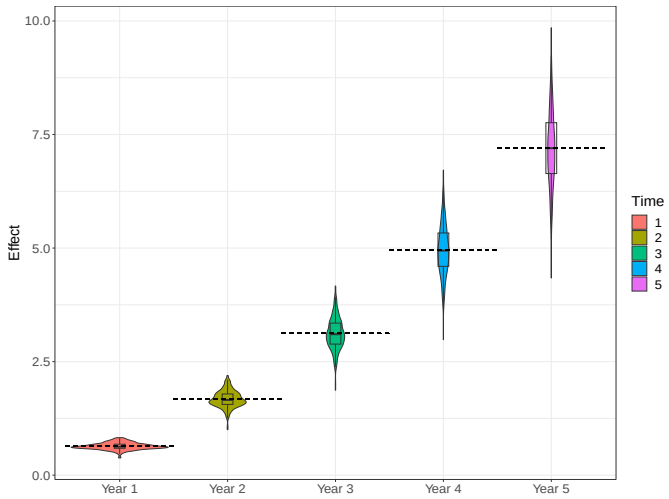
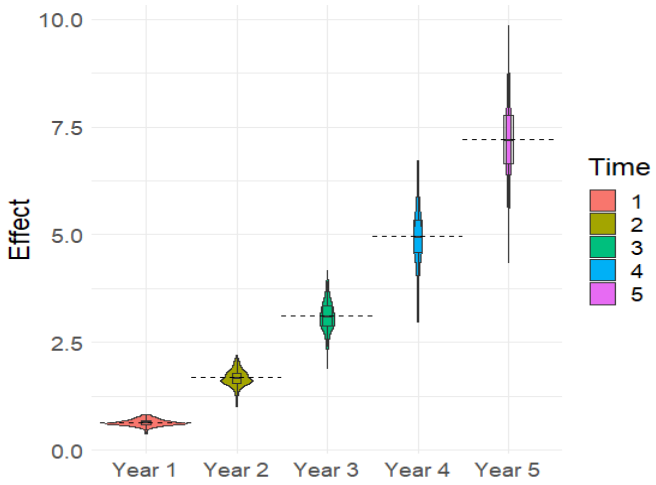


Figure 2: Violin plot across 250 Replicates of the relative bias for Scenario 1A without time-varying confounders for (A) the natural direct effect and (B) the natural indirect effect.

(A) Direct Effect



(B) Indirect Effect through $\mathcal{M}$



(C) Natural Indirect Effect through $\mathcal{L}$ and $\mathcal{M}$

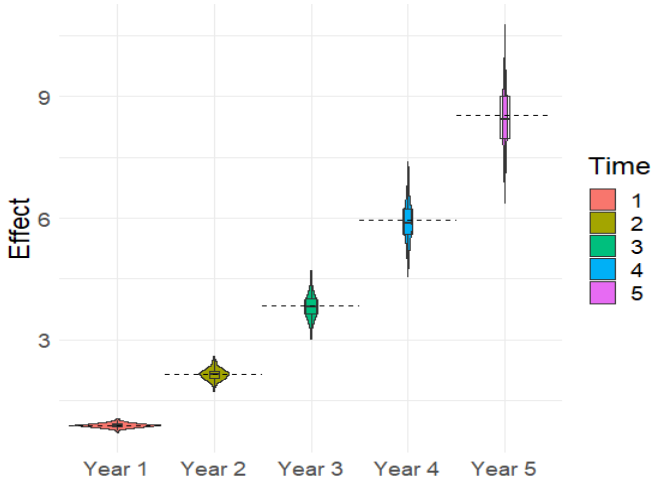
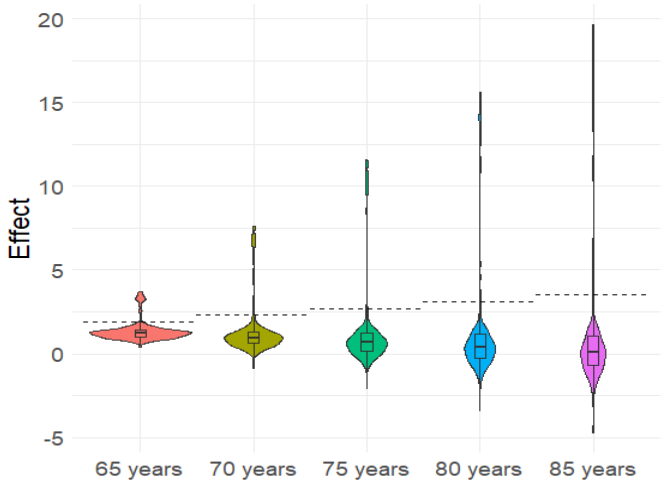
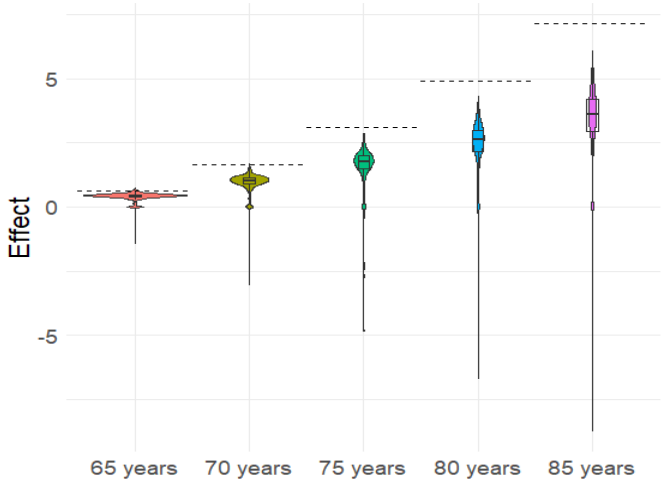


Figure 3: Violin plot across 250 Replicates of the relative bias for Scenario 1A with time-varying confounders: (A) the direct effect, (B) indirect effect through $\mathcal{M}$, and (C) indirect effect through $\mathcal{L}$ and $\mathcal{M}$

(A) Direct Effect



(B) Indirect Effect through $\mathcal{M}$



(C) Indirect Effect through $\mathcal{L}$ and $\mathcal{M}$

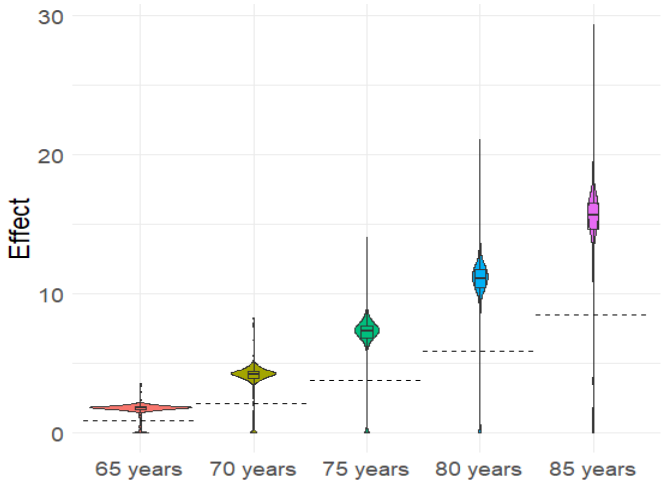
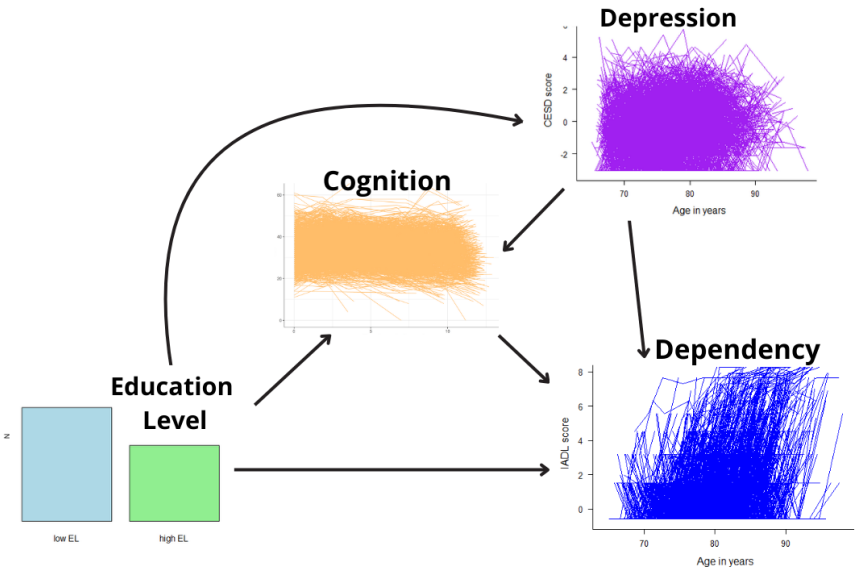


Figure 4: Violin plot across 250 Replicates of the relative bias for Scenario 2 with time-varying confounders: (A) the direct effect, (B) indirect effect through $\mathcal{M}$, and (C) indirect effect through $\mathcal{L}$ and $\mathcal{M}$

A. Effect of education level on dependency, mediated by cognition and depression



B. Effect of ApoE4 on cognition, mediated by vascular lesion and neurodegeneration

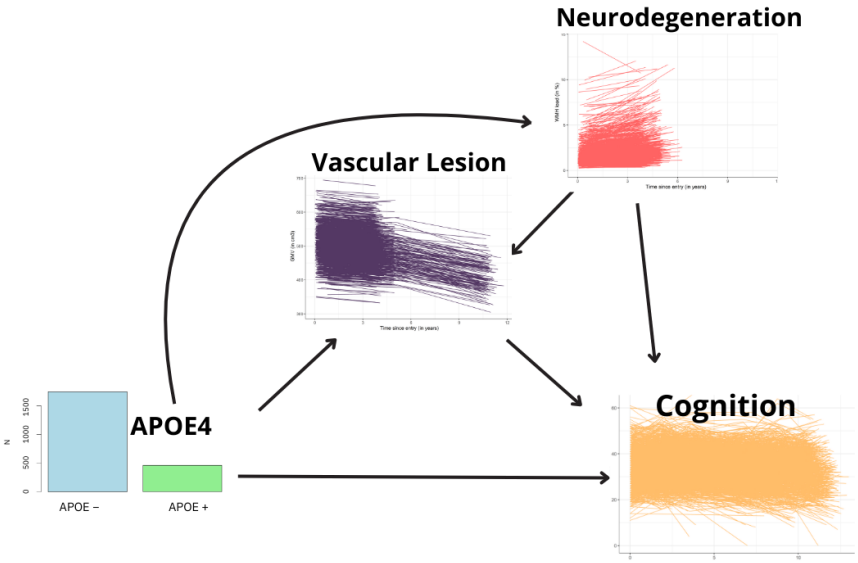


Figure 5: Directed Acyclic Graph of Studies 1 and 2; (A) effect of education level on dependency, mediated by cognition and depression (B) effect of ApoE4 on cognition, mediated by vascular lesion and global neurodegeneration

Table 1: Characteristics of the 2,213 participants of 3C sample according to their APOE4 status, educational level and overall at baseline

Characteristics	APOE4 (N=417)		No APOE4 (N=1583)		High education (N=886)		Low education (N=1327)		Overall (N = 2213)	
	N (%)	Mean (SD)	N (%)	Mean (SD)	N (%)	Mean (SD)	N (%)	Mean (SD)	N (%)	Mean (SD)
Sex										
female	60.1		62.7		54.3		67.4		62.1	
male	39.9		37.3		45.7		32.6		36.4	
Age at entry		71.9 (3.8)		72.1 (4.0)		72.3 (4.0)		72.0 (4.0)		72.0 (4.0)
WMH (cm^3)		2.1 (1.6)		2.3 (2.2)		2.3 (2.2)		2.3 (2.0)		2.3 (2.1)
Number of WMH measures/subject		1.5 (0.6)		1.5 (0.6)		1.6 (0.6)		1.5 (0.6)		1.5 (0.6)
GM (cm^3)		502.6 (51.2)		496.3 (49.1)		507.5 (49)		491.4 (49)		497.6 (49.6)
Number of GM measures/subject		1.6 (0.7)		1.6 (0.7)		1.7 (0.8)		1.6 (0.7)		1.6 (0.7)
IST score [0-66]		32.9 (7.1)		33.5 (6.7)		35.0 (6.8)		32.2 (6.5)		33.4 (6.8)
Number of IST measures/subject		6.4 (2.3)		6.5 (2.1)		6.7 (2.1)		6.4 (2.2)		6.5 (2.2)
CESD score [0-60]		9.2 (8.3)		9.1 (8.3)		8.3 (7.7)		9.6 (8.6)		9.1 (8.2)
Number of CESD measures/subject		4.8 (1.6)		5.0 (1.5)		5.1 (1.4)		4.8 (1.6)		5.0 (1.6)
IADL score [0-5]		0.5 (1.4)		0.4 (1.2)		0.4 (1.2)		0.5 (1.3)		0.4 (1.3)
Number of IADL measures/subject		5.1 (1.6)		5.2 (1.6)		5.3 (1.4)		5.0 (1.6)		5.1 (1.6)
Years of follow-up		9.1 (4.2)		9.5 (3.9)		9.8 (3.9)		9.2 (4.0)		9.5 (4.0)

Abbreviations: N=sample size, WMH=White matter hyper-intensities, GM=grey matter, IST=Isaacs Set Test,
CESD = Center for Epidemiologic Studies Depression Scale (normalized score),
IADL = Instrumental Activities of Daily Living (normalized score), SD=standard deviation

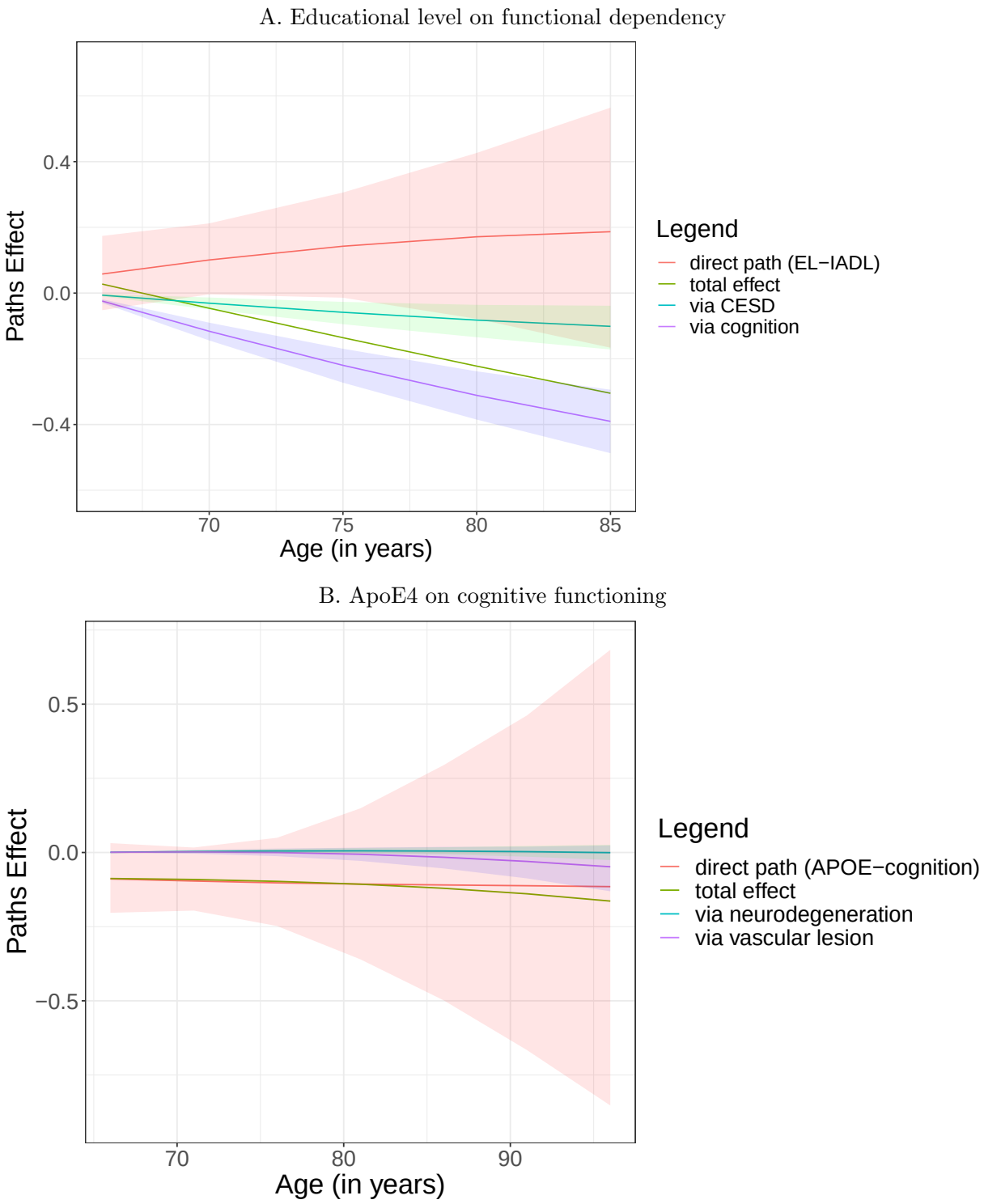


Figure 6: Estimated path-specific effects in the 3C cohort study exploring: A. educational level on functional dependency (IADL score) through cognitive functioning (IST score) and depressive symptomatology (CESD score), and B. APOE4 on cognitive functioning (IST score) through neurodegeneration (GM) and vascular lesions(WMH). Confidence bands are obtained by parametric Bootstrap with 1000 replicates.

1. Identification of path-specific effects

The path-specific effects are systematically defined as a comparison of two expectations $\mu = \mathbb{E}(Y_t(x, \mathcal{L}_t(x'), \mathcal{M}_t(x'', \mathcal{L}_t(x'))))|C)$ where x, x', and x'' can take various values. This expectation can be expressed as a function of the observations, rendering it identifiable, thanks to certain assumptions. As the assumptions vary depending on the paths, we will elaborate on the procedure for the three specific path effects: PSE_{XY}, PSE_{XMY}, $PSE_{XL(M)Y}$.

For PSE_{XY}, we have :

$$\mu = \mathbb{E}(Y_t(x, \mathcal{L}_t(x'), \mathcal{M}_t(x', \mathcal{L}_t(x'))))$$

where x can takes value x or x'

First, thank to assumption (iii.a) we can conditioned on X, and the estimand can be developed according to the time-confounder history $\mathcal{L}_t$ and the mediator history $\mathcal{M}_t$ jointly.

$$\mu = \int_{l_t, m_t} \mathbb{E}(Y_t(x, l_t, m_t)|C = c, X = x, \mathcal{L}_t(x') = l_t, \mathcal{M}_t(x', l_t) = m_t) \times f_{\mathcal{L}_t(x'), \mathcal{M}_t(x')|(C=c), X=x'} \; d_{l_t, m_t}$$

Second, thanks to assumptions (iii.c) and (iv.a), we can remove the conditioning $\mathcal{L}_t = l_t$ and $\mathcal{M}_t(x, l_t) = m_t$:

$$\mu = \int_{l_t, m_t} \mathbb{E}(Y_t(x, l_t, m_t)|C = c, X = x) \times f_{\mathcal{L}_t(x'), \mathcal{M}_t(x')|(C=c, X=x')} \; d_{l_t, m_t}$$

Third, using assumption (iii.b) we can condition on (L_t, M_t).

$$\mu = \int_{l_t, m_t} \mathbb{E}(Y_t(x, l_t, m_t)|C = c, X = x, \mathcal{L}_t = l_t, \mathcal{M}_t = m_t) \times f_{\mathcal{L}_t(x'), \mathcal{M}_t(x')|(C=c, X=x')} \; d_{l_t, m_t}$$

Finally, with the consistency assumption, we obtain :

$$\mu = \int_{l_t, m_t} \mathbb{E}(Y_t|C = c, X = x, \mathcal{L}_t = l_t, \mathcal{M}_t = m_t) \times f_{\mathcal{L}_t, \mathcal{M}_t|(C=c, X=x')} \; d_{l_t, m_t}$$

For PSE_{XMY}, we have :

$$\mu = \mathbb{E}(Y_t(x, \mathcal{L}_t(x'), \mathcal{M}_t(x, \mathcal{L}_t(x'))))$$

where x can takes value x or x'

First, the estimand can be developed according to the time-confounder history $\mathcal{L}_t$ and the mediator history $\mathcal{M}_t$.

$$\mu = \int_{l_t} \int_{m_t} \mathbb{E}(Y_t(x, l_t, m_t)|C = c, \mathcal{L}_t(x') = l_t, \mathcal{M}_t(x, l_t) = m_t)$$
$$\times f_{\mathcal{L}_t(x')|C=c}(l_t), f_{\mathcal{M}_t(x,l_t)|(C=c,\mathcal{L}_t(x')=l_t)}(m_t)\ d_{m_t} d_{l_t}$$

Second, thanks to assumptions (iii.e) and (iv.b), we can remove the conditioning $\mathcal{L}_t = l_t$ and $\mathcal{M}_t(x, l_t) = mt$:

$$\mu = \int_{l_t} \int_{m_t} \mathbb{E}(Y_t(x, l_t, m_t)|C = c) f_{\mathcal{L}_t(x')|C=c}(l_t), f_{\mathcal{M}_t(x,l_t)|(C=c}(m_t)\ d_{m_t} d_{l_t}$$

Third we add the conditioning on $X = x$ in the expectation of Y_t, $X = x'$ in the density of $\mathcal{L}_t$ and $X = x$ and $\mathcal{L}_\sqcup$ in the density of $\mathcal{M}_t$ thanks to assumptions (iii.a), (iii.c) and (iii.e):

$$\mu = \int_{l_t} \int_{m_t} \mathbb{E}(Y_t(x, l_t, m_t)|C = c, X = x)$$
$$\times f_{\mathcal{L}_t(x')|C=c,X=x'}(l_t), f_{\mathcal{M}_t(x,l_t)|(C=c,X=x,\mathcal{L}_t=l_t)}(m_t)\ d_{m_t} d_{l_t}$$

Then, we add the conditioning on $\mathcal{L}_t = l_t$ and $\mathcal{M}_t = m_t$ in the expectation of Y_t thanks to assumption (iii.b)

$$\mu = \int_{l_t} \int_{m_t} \mathbb{E}(Y_t(x, l_t, m_t)|C = c, X = x, \mathcal{L}_t = l_t, \mathcal{M}_t = m_t)$$
$$\times f_{\mathcal{L}_t(x')|C=c,X=x'}(l_t), f_{\mathcal{M}_t(x,l_t)|(C=c,X=x,\mathcal{L}_t=l_t)}(m_t)\ d_{m_t} d_{l_t}$$

By applying consistency assumption i, we obtain:

$$\mu = \int_{l_t} \int_{m_t} \mathbb{E}(Y_t | C = c, X = x, \mathcal{L}_t = l_t, \mathcal{M}_t = m_t) \times f_{\mathcal{L}_t|C=c,X=x'}(l_t), f_{\mathcal{M}_t|(C=c,X=x,\mathcal{L}_t=l_t)}(m_t)\ d_{m_t} d_{l_t}$$

For $PSE_{XL(M)Y}$ the identification process is the same.

2. Simulation study

2.1 *Simulation design*

We systematically considered the following generation structural models and observation models with the corresponding parameters given in Table 1 and Table 4 for scenario 1 and 2, respectively.

$$\text{For process } \mathcal{L}: \begin{cases} L_i(0) = \beta_0^L + \beta_1^L X_i^{L(0)} + u_i^L \\ \dfrac{\partial L_i(t)}{\partial t} = \gamma_0^L + \gamma_1^L X_i^{L(t)} + v_i^L \\ \tilde{L}_{ij} = L_i(t_{ij}) + \epsilon_{ij}^L \text{ for } j = 1, ..., n_i^L \end{cases} \tag{1}$$

$$\text{For process } \mathcal{M}: \begin{cases} M_i(0) = \beta_0^M + \beta_1^M X_i^{M(0)} + u_i^M \\ \dfrac{\partial M_i(t)}{\partial t} = \gamma_0^M + \gamma_1^M X_i^{M(t)} + v_i^M + \alpha^{ML} L_i(t) \\ \tilde{M}_{ij} = M_i(t_{ij}) + \epsilon_{ij}^M \text{ for } j = 1, ..., n_i^M \end{cases} \tag{2}$$

$$\text{For process } \mathcal{Y}: \begin{cases} Y_i(0) = \beta_0^Y + \beta_1^Y X_i^{Y(0)} + u_i^Y \\ \dfrac{\partial Y_i(t)}{\partial t} = \gamma_0^Y + \gamma_1^Y X_i^{Y(t)} + v_i^Y + \alpha^{YM} M_i(t) + \alpha^{YL} L_i(t) \\ \tilde{Y}_{ij} = Y_i(t_{ij}) + \epsilon_{ij}^Y \text{ for } j = 1, ..., n_i^Y \end{cases} \tag{3}$$

The process $\mathcal{L}$ sub-equations and corresponding parameters indicated in red were only considered in scenario 2 and in the sub-scenario 1 with time-dependent confounding factors .

[Table 1 about here.]

In the main scenario 1, labelled 1A, we considered samples of 500 subjects, a discretization step of 0.1 year, and a rate of dropout of 10%. In additional scenarios 1B to 1D in presence of a time-varying confounder, we changed each parameter one by one as listed in Table 2. In scenario 2, we considered samples of 500 subjects, a discretization step of 1 year, and a rate of dropout of 10%.

[Table 2 about here.]

2.2 *Additional results for Scenario 1*

The results of scenario 1A are detailed in the main body. Additional results for scenarios B, C, and D in presence of time-varying confounders are given in Figures 1,2, and 3, respectively. The coverage rate are given in Table 3.

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

[Table 3 about here.]

2.3 *Additional results for Scenario 2*

The estimated parameters of the working model are reported in Table 4.

[Table 4 about here.]

In this scenario with poor repeated information on the mediator and confounder, the working model fails to correctly estimate the parameters which induce bias in the resulting causal contrasts reported in the main document.

3. Additional information about the application

3.1 *Specification of the working models*

Study 1: effect of educational level on the level of functional dependency

$$\text{For process } \mathcal{L}: \begin{cases} L_i(0) = \beta_0^L + \beta_1^L EL_i^{L(0)} + \beta_2^L APOE4_i^{L(0)} + \beta_3^L SEX_i^{L(0)} + \\ \qquad \beta_4^L + CENTRE_i^{L(0)} \beta_5^L AGE0_i^{L(0)} + u_i^L \\ \dfrac{\partial L_i(t)}{\partial t} = \gamma_0^L + \gamma_1^L EL_i^L + \gamma_2^L APOE4_i^L + \gamma_3^L SEX_i^L + \\ \qquad \gamma_4^L CENTRE_i^L + \gamma_5^L AGE0_i^L + v_i^L \end{cases} \tag{4}$$

$$\text{For process } \mathcal{M}: \begin{cases} M_i(0) = \beta_0^M + \beta_1^M EL_i^{M(0)} + \beta_2^M APOE4_i^{M(0)} + \beta_3^M SEX_i^{M(0)} + \\ \qquad \beta_4^M CENTRE_i^{M(0)} + \beta_5^M AGE0_i^{M(0)} + u_i^M \\ \dfrac{\partial M_i(t)}{\partial t} = \gamma_0^M + \gamma_1^M EL_i^M + \gamma_2^M APOE4_i^M + \gamma_3^M SEX_i^M + \\ \qquad \gamma_4^M CENTRE_i^M + \gamma_5^M AGE0_i^M + v_i^M + \alpha^{ML} L_i(t) \end{cases} \tag{5}$$

$$\text{For process } \mathcal{Y}: \begin{cases} Y_i(0) = \beta_0^Y + \beta_1^Y EL_i^{Y(0)} + \beta_2^Y APOE4_i^{Y(0)} + \beta_3^Y SEX_i^{Y(0)} + \\ \qquad \beta_4^Y CENTRE_i^{Y(0)} + \beta_5^Y AGE0_i^{Y(0)} + u_i^Y \\ \dfrac{\partial Y_i(t)}{\partial t} = \gamma_0^Y + \gamma_1^Y EL_i^Y + \gamma_2^Y APOE4_i^Y + \gamma_3^Y SEX_i^Y + \\ \qquad \gamma_4^Y CENTRE_i^Y + \gamma_5^Y AGE0_i^Y + v_i^Y + \alpha^{YM} M_i(t) + \alpha^{YL} L_i(t) \end{cases} \tag{6}$$

Study 2: effect of APOE4 on the level of verbal fluency

$$\text{For process } \mathcal{L}: \begin{cases} L_i(0) = \beta_0^L + \beta_1^L APOE4_i^{L(0)} + \beta_2^L AGE0_i^{L(0)} + \beta_3^L EL_i^{L(0)} + \\ \qquad \beta_4^L SEX_i^{L(0)} + \beta_5^L VTI_i^{L(0)} + \beta_6^L CENTRE_i^{L(0)} + u_i^L \\ \dfrac{\partial L_i(t)}{\partial t} = \gamma_0^L + (\gamma_1^L APOE4_i^L + \gamma_2^L AGE0_i^L + \gamma_3^L EL_i^L + \\ \qquad \gamma_4^L SEX_i^L + \gamma_5^L VTI_i^L + \gamma_6^L CENTRE_i^L) \times \text{time}_i + \text{v}_i^L \end{cases} \tag{7}$$

$$\text{For process } \mathcal{M}: \begin{cases} M_i(0) = \beta_0^M + \beta_1^M APOE_i^{M(0)} + \beta_2^M AGE0_i^{M(0)} + \beta_3^M EL_i^{M(0)} + \\ \qquad \beta_4^M SEX_i^{M(0)} + \beta_5^M CENTRE_i^{M(0)} + u_i^M \\ \dfrac{\partial M_i(t)}{\partial t} = \gamma_0^M + (\gamma_1^M APOE_i^M + \gamma_2^M AGE0_i^M + \gamma_3^M EL_i^M + \\ \qquad \gamma_4^M SEX_i^M + \gamma_5^M CENTRE_i^M) \times \text{time}_i + \text{v}_i^M + \alpha^{ML} L_i(t) \end{cases} \tag{8}$$

$$\text{For process } \mathcal{Y}: \begin{cases} Y_i(0) = \beta_0^Y + \beta_1^Y APOE_i^{Y(0)} + \beta_2^Y AGE0_i^{Y(0)} + \beta_3^Y EL_i^{Y(0)} + \\ \qquad \beta_4^Y SEX_i^{Y(0)} + \beta_5^Y CENTRE_i^{Y(0)} + u_i^Y \\ \dfrac{\partial Y_i(t)}{\partial t} = \gamma_0^Y + (\gamma_1^Y APOE_i^Y + \gamma_2^Y AGE0_i^Y + \gamma_3^Y EL_i^Y + \\ \qquad \gamma_4^Y SEX_i^Y + \gamma_5^Y CENTRE_i^Y) \times \text{time}_i + \text{v}_i^Y + \alpha^{YM} M_i(t) + \alpha^{YL} L_i(t) \end{cases} \tag{9}$$

3.2 *Results of the working models*

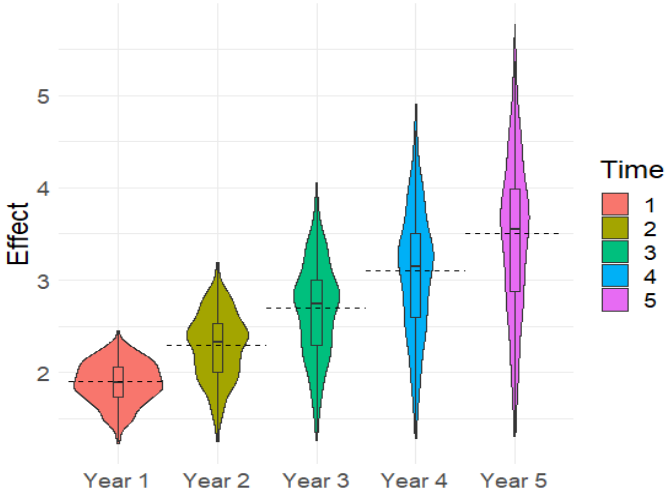
The estimates of the two working models are reported in Tables 5 and 6 for study 1 and 2, respectively.

[Table 5 about here.]

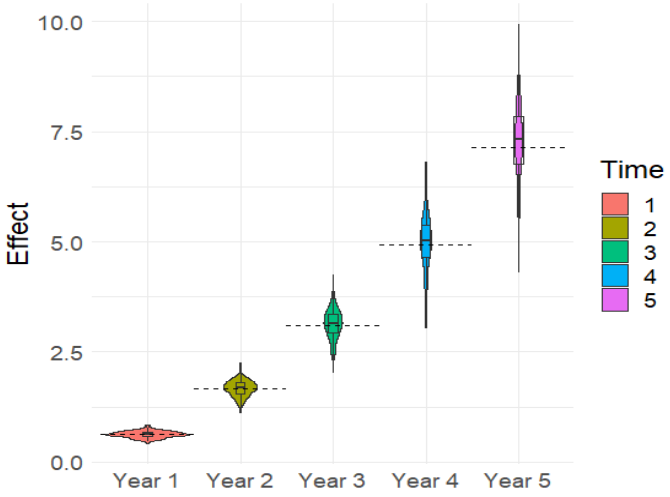
[Table 6 about here.]

[Table 7 about here.]

(A) Direct Effect



(B) Indirect Effect through $\mathcal{M}$



(C) Natural Indirect Effect through $\mathcal{L}$ and $\mathcal{M}$

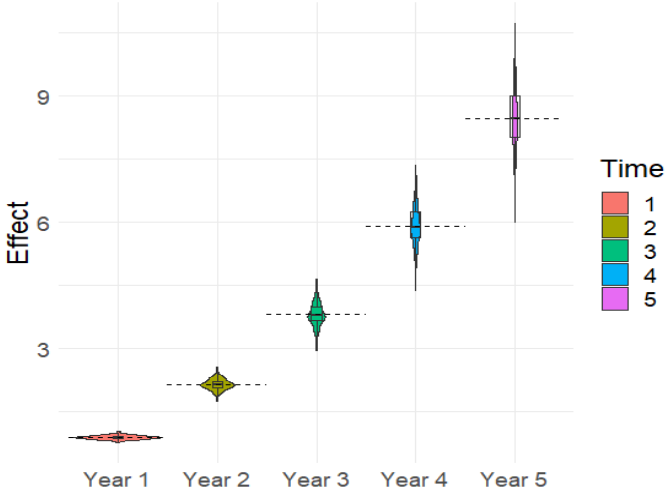
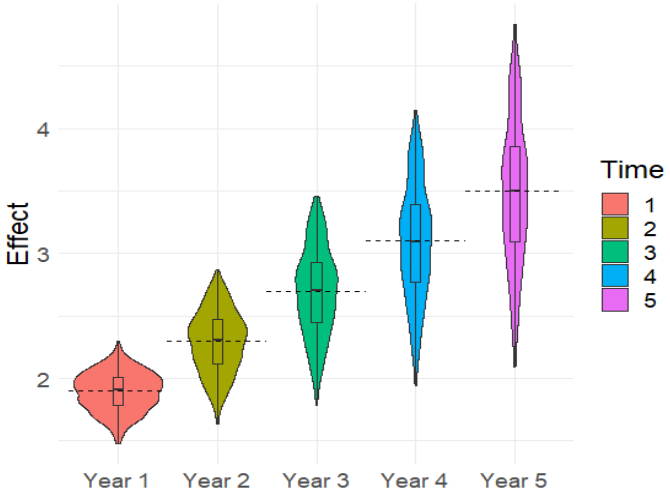
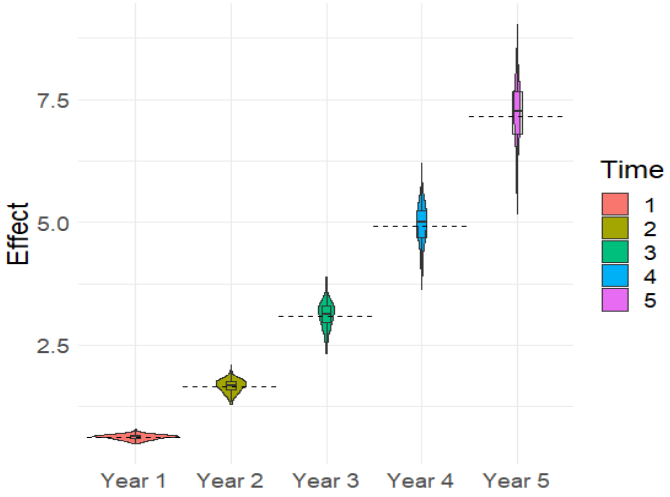


Figure 1: Median Bias of Simulations: Violin Plot Across 250 Replicates for Scenario 1B, with time-varying confounders, for (A) the direct effect, (B) indirect effect through $\mathcal{M}$, and (C) indirect effect through $\mathcal{L}$ and $\mathcal{M}$

(A) Direct Effect



(B) Indirect Effect through $\mathcal{M}$



(C) Natural Indirect Effect through $\mathcal{L}$ and $\mathcal{M}$

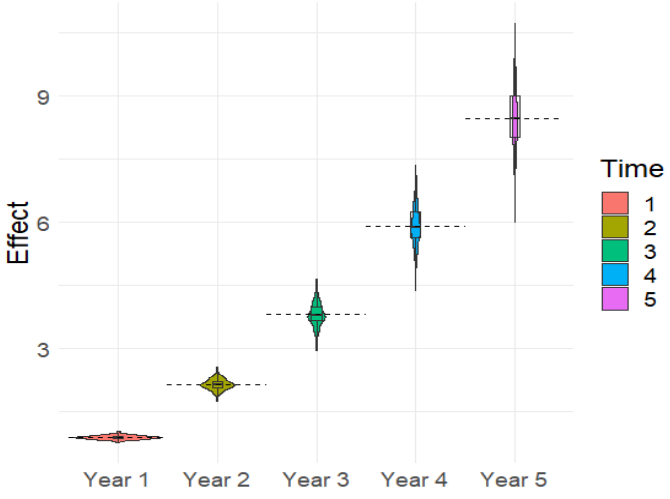
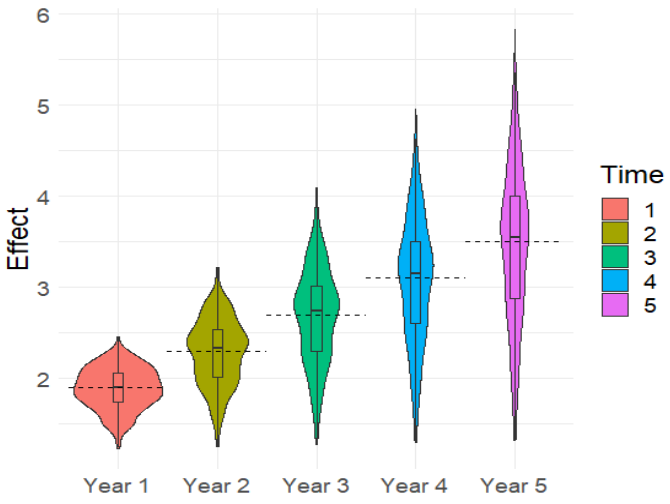
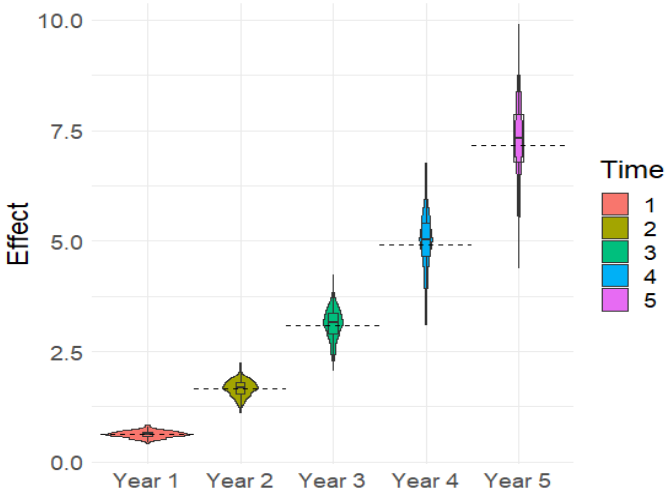


Figure 2: Median Bias of Simulations: Violin Plot Across 250 Replicates for Scenario 1C, with time-varying confounders, for (A) the direct effect, (B) indirect effect through $\mathcal{M}$, and (C) indirect effect through $\mathcal{L}$ and $\mathcal{M}$

(A) Direct Effect



(B) Indirect Effect through $\mathcal{M}$



(C) Natural Indirect Effect through $\mathcal{L}$ and $\mathcal{M}$

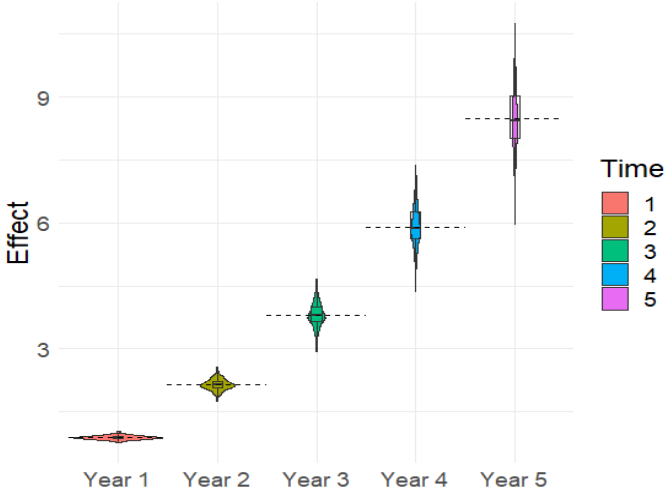


Figure 3: Median Bias of Simulations: Violin Plot Across 250 Replicates for Scenario 1D, with time-varying confounders, for (A) the direct effect, (B) indirect effect through $\mathcal{M}$, and (C) indirect effect through $\mathcal{L}$ and $\mathcal{M}$

Table 1: Parameters of the generation models in scenario 1 in absence and presence of time-varying confounder

Parameters	With L	Without L
β_0^L	0.5	-
β_1^L	1.8	-
β_0^M	0.2	0.5
β_1^M	0.9	1.8
β_0^Y	0.6	0.6
β_1^Y	1.5	1.5
γ_0^L	0.1	-
γ_1^L	0.2	-
γ_0^M	0.2	0.1
γ_1^M	0.8	0.2
γ_0^Y	0.8	0.8
γ_1^Y	0.4	0.4
chol1	1	1
chol2	0.2	0.1
chol3	0.1	-
chol6	-	3
chol7	2	-
chol8	0.1	1
chol10	-	2
chol12	3	-
chol16	1	-
chol19	2	-
chol21	3	-
α_{ML}	0.3	-
α_{YL}	0.4	-
α_{YM}	0.5	0.4
σ_L	0.3	-
σ_M	0.6	0.3
σ_Y	0.2	0.2

Table 2: Characteristics of sub-scenarios 1A-1D in presence of time-varying confounder

Scenario	N	δ **(in years)**	**Dropout rate (%)**
1A	500	0.1	10
1B	500	0.05	10
1C	250	0.1	10
1D	500	0.1	20

	Coverage rate (in %)														
	Direct effect					**Indirect effect through M**					**Indirect effect through M and L**				
Scenario	Time 1	Time 2	Time 3	Time 4	Time 5	Time 1	Time 2	Time 3	Time 4	Time 5	Time 1	Time 2	Time 3	Time 4	Time 5
A	96.2	94.8	95.0	94.2	93.8	95.6	94.6	93.8	93.4	93.6	96.8	96.8	96.2	94.6	94.4
B	96.0	96.4	96.2	96.2	96.0	94.8	93.2	93.2	94.4	95.6	95.0	93.8	92.4	92.4	92.6
C	96.4	97.2	97.2	96.8	96.8	94.2	95.2	93.2	93.4	92.2	96.0	96.2	95.6	95.2	95.4
D	96.6	95.2	95.0	94.6	94.6	95.8	95.2	95.2	94.4	94.4	96.8	95.8	95.2	94.6	94.2
Scenario 2	30.8	14.6	11.0	10.2	10.0	76.4	68.6	63.4	55.6	49.2	1.2	1.4	3.0	3.6	5.0

Table 3: Coverage rate in (%) for the different scenario

Table 4: Estimated parameters of the working model in Scenario 2

Parameter	**Value**
β_0^L	-0.20
β_1^L	0.09
β_2^L	0.01
β_0^M	6.12
β_1^M	-0.10
β_2^M	-0.07
β_0^Y	0.53
β_1^Y	-0.06
β_2^Y	-0.00
γ_0^L	-0.16
γ_1^L	-0.01
γ_2^L	0.00
γ_0^M	-0.15
γ_1^M	0.02
γ_2^M	0.00
γ_0^Y	0.10
γ_1^Y	-0.01
γ_2^Y	-0.00
Chol.1	0.86
Chol.2	0.04
Chol.3	0.10
Chol.4	-0.00
Chol.7	1.04
Chol.8	0.00
Chol.10	0.06
Chol.12	0.80
Chol.15	-0.02
Chol.16	0.03
Chol.19	-0.08
Chol.21	-0.04
α_{ML}.(Intercept)	-0.03
α_{YL}.(Intercept)	0.01
α_{YM}.(Intercept)	-0.01
σ_L	0.22
σ_M	0.48
σ_Y	-0.51

Table 5: Illustration 1 : Model estimation parameters

Variable	Coefficient	p-value
β_0^L	-2.7597	2e-16 ***
β_1^L	-0.2649	0.0004 ***
β_2^L	0.0719	0.4039
β_3^L	0.8075	2e-16 ***
β_4^L	0.0719	0.4039
β_5^L	1.0284	2e-16 ***
β_0^M	-0.0257	0.7100
β_1^M	0.3723	<2e-16 ***
β_2^M	-0.0249	0.5090
β_3^M	0.0853	0.0076 **
β_4^M	0.1316	0.0001 ***
β_5^M	0.1522	0.0005 ***
β_0^Y	-1.0434	<2e-16 ***
β_1^Y	0.0495	0.4149
β_2^Y	-0.0276	0.7015
β_3^Y	-0.1525	0.0105 *
β_4^Y	0.5056	<2e-16 ***
β_5^Y	-1.8127	<2e-16 ***
γ_0^L	0.9488	<2e-16 ***
γ_1^L	0.1107	0.0264 *
γ_2^L	0.0117	0.8403
γ_3^L	-0.0850	0.0968
γ_4^L	-0.4813	<2e-16 ***
γ_5^L	0.1848	0.0024 **
γ_0^M	-0.8743	<2e-16 ***
γ_1^M	-0.0761	0.0005 ***
γ_2^M	-0.0616	0.0188 *
γ_3^M	0.0996	<2e-16 ***
γ_4^M	0.3830	<2e-16 ***
γ_5^M	-0.1865	<2e-16 ***
γ_0^Y	1.6950	<2e-16 ***
γ_1^Y	0.0776	0.1865
γ_2^Y	0.0183	0.8030
γ_3^Y	0.0839	0.1538
γ_4^Y	-0.5950	<2e-16 ***
γ_5^Y	0.9941	<2e-16 ***
Chol.1	0.9733	<2e-16 ***
Chol.7	0.4418	<2e-16 ***
Chol.16	0.3229	<2e-16 ***
Chol.19	-0.2075	<2e-16 ***
Chol.21	0.6315	<2e-16 ***
α_{YL}.(Intercept)	0.2158	<2e-16 ***
α_{YL}.EL	0.0242	0.4699
α_{YM}.(Intercept)	-0.7694	<2e-16 ***
α_{YM}.EL	0.1185	0.0911 .
σ_L	0.9936	<2e-16 ***
σ_M	0.3921	<2e-16 ***
σ_Y	1.0402	<2e-16 ***

Table 6: Illustration 2 : Model estimation parameters

Variable	Coefficient	p-valeur
β_0^L	9.7632	<2e-16 ***
β_1^L	0.0619	0.1827
β_2^L	-0.1373	<2e-16 ***
β_3^L	0.0832	0.0350 *
β_4^L	0.1994	0.0001 ***
β_5^L	0.8815	<2e-16 ***
β_6^L	-0.2285	<2e-16 ***
β_0^M	0.6917	0.7967
β_1^M	-0.0916	0.4419
β_2^M	0.0086	0.8295
β_3^M	0.0727	0.4672
β_4^M	-0.1691	0.0972 .
β_5^M	0.3357	0.0041 **
β_0^Y	5.8808	<2e-16 ***
β_1^Y	-0.0923	0.2552
β_2^Y	-0.0947	<2e-16 ***
β_3^Y	0.5535	<2e-16 ***
β_4^Y	0.1088	0.1083
β_5^Y	0.2361	0.0017 **
γ_0^L	-0.8750	<2e-16 ***
γ_1^L	0.0013	0.8944
γ_2^L	0.0155	<2e-16 ***
γ_3^L	-0.0130	0.0997 .
γ_4^L	-0.0301	0.0024 **
γ_5^L	-0.0070	0.2166
γ_6^L	-0.0488	<2e-16 ***
γ_7^L x time	-0.0569	<2e-16 ***
γ_8^L x time	-0.0007	0.4670
γ_9^Lx time	0.0003	0.0009 ***
γ_{10}^Lx time	0.0013	0.0860 .
γ_{11}^Lx time	0.0016	0.0970 .
γ_{12}^Lx time	-0.0002	0.7721
γ_{13}^Lx time	0.0048	<2e-16 ***
γ_0^M	0.2550	0.4145
γ_1^M	0.0145	0.6279
γ_2^M	-0.0013	0.7815
γ_3^M	-0.0130	0.5909
γ_4^M	0.0207	0.4175
γ_5^M	-0.1037	0.0002 ***
γ_6^M	0.0689	0.0227 *
γ_7^Mx time	0.0005	0.8839
γ_8^Mx time	-0.0007	0.0668 .
γ_9^Mx time	-0.0004	0.8722
γ_{10}^Mx time	-0.0021	0.4511
γ_{11}^Mx time	-0.0022	0.4670

Variable	Coefficient	p-valeur
γ_0^Y	-0.5890	<2e-16 ***
γ_1^Y	0.0021	0.8731
γ_2^Y	0.0079	<2e-16 ***
γ_3^Y	-0.0163	0.0963 .
γ_4^Y	0.0049	0.6327
γ_5^Y	0.0526	<2e-16 ***
γ_6^Y	0.0125	0.0927 .
γ_7^Y x time	-0.0005	0.6594
γ_8^Y x time	-0.0002	0.0139 *
γ_9^Y x time	0.0005	0.5104
γ_{10}^Y x time	0.0003	0.6524
γ_{11}^Y x time	-0.0022	0.0060 **
Chol.1	0.4333	<2e-16 ***
Chol.2	-0.1893	<2e-16 ***
Chol.3	0.0800	0.0002 ***
Chol.4	-0.0036	0.1304
Chol.7	1.0059	<2e-16 ***
Chol.8	-0.0268	0.2392
Chol.10	0.0763	0.0021 **
Chol.12	0.7667	<2e-16 ***
Chol.15	-0.0220	<2e-16 ***
Chol.16	-0.0270	<2e-16 ***
Chol.19	-0.0499	0.4019
Chol.21	0.0373	<2e-16 ***
$\alpha_{ML}.(Intercept)$	-0.0002	0.9758
$\alpha_{ML}.APOE$	0.0019	0.8678
$\alpha_{YM}.(Intercept)$	-0.0031	0.0029 **
$\alpha_{YM}.APOE$	-0.0011	0.5614
σ_L	-0.1774	<2e-16 ***
σ_M	0.4719	<2e-16 ***
σ_Y	0.5059	<2e-16 ***

4.2 Médiateurs et événement final définis comme des temps d'événement

Dans les études de cohorte, de nombreux facteurs de risque sont communs à la démence et au décès. Parmi eux, les facteurs cardiométaboliques, tels que la pression artérielle ou l'IMC, sont associés à un risque de démence accru (Kennelly et al. (2009),Guo et al. (1996)) et, également de décès (Canoy et al. (2021)). Pour analyser cette problématique, des analyses de médiation sur des temps d'événement, avec un événement intermédiaire (non terminal) tel que la démence, en présence d'un événement terminal (e.g. le décès), peuvent être effectuées. Le décès pouvant intervenir avant ou après la démence, et la démence ne pouvant pas survenir après la décès, les événements sont dit "semi-compétitifs".

Ces risques semi-compétitifs posent des défis dans les méthodes actuelles d'analyse de la médiation. Valeri et al. (2023) ont proposé une méthode d'analyse de médiation basée sur une approche stochastique pour étudier l'effet d'une intervention sur un événement terminal en présence d'une variable médiatrice de type d'évenement. Cette approche permet d'étudier les inégalités raciales d'accès d'un traitement chirurgical après le diagnostic d'un cancer du colon aux Etats-Unis. Dans l'étude de l'effet de facteurs cardio-métaboliques sur la démence et le décès, des challenges supplémentaires interviennent. D'une part, Valeri et al. (2023) s'intéresse à une intervention stochastique sur le médiateur, ce qui est tout à fait pertinent dans le contexte d'inégalités raciales et d'accès au traitement. Cependant lorsque l'on s'intéresse au rôle de la démence dans l'association entre des facteurs de risque et le décès, le cadre est différent. Il s'agit d'étudier l'effet au niveau individuel et non populationnel. Nous nous centrons dans ce travail sur une décomposition de l'effet au niveau individuel.

D'autre part, dans les études de cohorte, les informations sur certains événements ne sont pas connues en tout temps mais uniquement rapportées au moment des visites planifiées. C'est le cas pour la démence. La modélisation du risque de démence et décès doit donc se faire par un modèle tenant compte de la censure par intervalle, afin de prendre en compte la possibilité qu'un participant ait développé une démence entre son dernier diagnostic négatif et son décès.

Dans ce travail, nous proposons une approche d'analyse de médiation basée sur des effets naturels, adaptée à une variable d'exposition fixe ainsi qu'à un médiateur et un outcome, tous deux définis en termes de temps d'événement. Nous définissons les contrastes des quantités causales et les hypothèses nécessaires pour leur identifiabilité. Les contrastes causaux sont estimés à partir d'un modèle multi-états.

Nous illustrons l'approche proposée avec les données de la cohorte $3C$ pour étudier le rôle de la démence dans la relation causale entre la santé cardiovasculaire et le décès.

4.3 Méthode

4.3.1 Notations

Considérons, la configuration de la figure 4.1 où X est une variable d'exposition mesurée à un temps fixe. D est un temps d'événement intermédiaire, correspondant au médiateur, dans notre cas la démence. Z représente un temps d'événement terminal. Dans la suite Z représentera le décès. Le vecteur des variables C inclut les facteurs de confusion de la relation entre X et Z. Dans toute la suite du travail, les temps D et Z observés sont censurés à droite, la censure étant considérée non-informative.

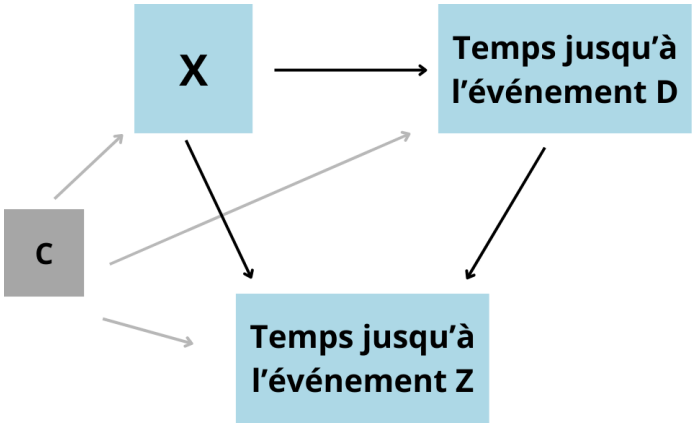


FIGURE 4.1 – DAG représentant la relation entre les variables C, X, D et Z.

Notons $Z(x)$ le temps d'événement contrefactuel de l'outcome si X avait été fixé à x. $Z(x,d)$ correspond au temps d'événement contrefactuel si X avait été fixé à x et D avait été fixé à d. De la même manière, nous définissons $D(x)$, le temps d'événement du

médiateur contrefactuel si X avait été fixé à x.

4.3.2 Estimandes causaux

4.3.2.1 Définition des quantités causales

Pour quantifier l'effet de la variable X sur l'événement, nous travaillons sur la probabilité de ne pas faire l'événement au temps t, soit la probabilité de survie $S_T(t) = P_T(T > t)$ pour un temps d'événement noté T. Soient x et x' deux valeurs de X. En considérant des effets sur la probabilité d'être indemne de l'événement au temps t, les effets totaux (TE_S), directs (NE_S) et indirects (IE_S) de X sur le temps d'événement terminal Z peuvent s'écrire tels que :

$$TE_S = P(Z(x) > t|C) - P(Z(x') > t|C) \tag{4.1}$$

$$NE_S = P(Z(x, D(x')) > t|C) - P(Z(x', D(x')) > t|C) \tag{4.2}$$

$$IE_S = P(Z(x, D(x)) > t|C) - P(Z(x, D(x')) > t|C) \tag{4.3}$$

La proportion d'effet passant par l'événement intermédiaire D se note $\frac{IE_S}{TE_S} \times 100$.

4.3.2.2 Hypothèses d'identifiabilité

Le même type d'hypothèses que pour le cas des données répétées vu dans la partie 1 de ce chapitre est nécessaire ici pour rendre identifiables ces quantités causales à partir de données observées.

Premièrement, l'hypothèse de **positivité** stipule que la probabilité de réalisation de la variable l'exposition X est strictement positive compte tenu des covariables C pour tout x, de même pour le temps d'événement de la variable D, tel que : $P(X = x|C) > 0$ et $P(D = d|C, X = x) > 0$.

L'hypothèse de **consistance** permet de créer une concordance entre les données réelles et contrefactuelles, à savoir : $Z(x, d) = Z$ si $X = x$ et $D = d$ et $D(x) = D$ si $X = x$.

L'hypothèse **d'ignorabilité séquentielle** se définit au travers de 4 sous-hypothèses :

— Il n'y a pas de facteurs de confusion dans la relation entre X et Z :

$$Z(x,d) \perp X|C,D$$

— Le temps de survie potentiel, conditionnellement à l'exposition et aux covariables C, est indépendant du temps d'événement intermédiaire, c'est-à-dire qu'il n'y a pas de facteurs de confusion :

$$Z(x,d) \perp D|C,X$$

— Le temps d'événement intermédiaire D est indépendant de l'exposition X, conditionnellement aux covariables C.

$$D(x) \perp X|C$$

Enfin, l'hypothèse **cross-world independence** indique que le temps de survie est indépendant de l'événement intermédiaire pour une autre valeur d'exposition, conditionnellement à C et à X, c'est-à-dire qu'il n'y a pas de facteurs de confusion entre l'événement intermédiaire et l'événement terminal impacté par l'exposition.

$$Z(x,d) \perp D(x')|C,X$$

Cette dernière hypothèse n'est pas requise pour une intervention stochastique mais le devient pour des effets naturels.

4.3.2.3 Ecriture des contrastes à partir des observations

Grâce à ces hypothèses, les contrastes causaux 4.1, 4.2 et 4.3 définis en section 4.3.2.1 deviennent identifiables.

En utilisant le même raisonnement que dans la section 4.1, chaque probabilité de survie potentielle peut se réécrire en fonction des observations X, D et Z. Dans le cas générique nous pouvons réécrire la probabilité $P(Z(x, D(x')) > t|C)$ telle que :

$$P(Z(x,D(x')) > t|C) = \int_0^t P(Z > t|X = x, D = u, C) \times f_D(u|X = x', C)d_u \qquad (4.4)$$

4.3.3 Estimation à partir d'un modèle multi-états

D'après l'équation 4.4, un modèle de travail est necéssaire pour obtenir la probabilité de survie conditionnellement au médiateur $P(Z > t|X = x, D = u, C)$ et la densité du médiateur $f_D(u|X = x', C)$. Le modèle multi-états offre un cadre complet pour obtenir ces quantités et permet d'estimer les contrastes définis en Section 4.3.2.1. Ce modèle analyse également l'évolution d'un individu au cours du temps dans un système d'états ainsi que les transitions entre ces états.

Nous nous plaçons dans cette section dans un contexte où l'événement intermédiaire est collecté de façon continue, ainsi le temps d'événement exact est connu. Nous discutons le cas d'un événement intermédiaire censuré par intervalle dans la section suivante.

4.3.3.1 Spécification du modèle multi-états

Lorsque l'événement intermédaire est la démence et l'événement final, le décès, le modèle multi-états est un modèle *illness-death* avec trois états envisagés : sain, malade (dans ce cas, dément) et décédé. Ces états sont représentés sur la Figure 4.2.

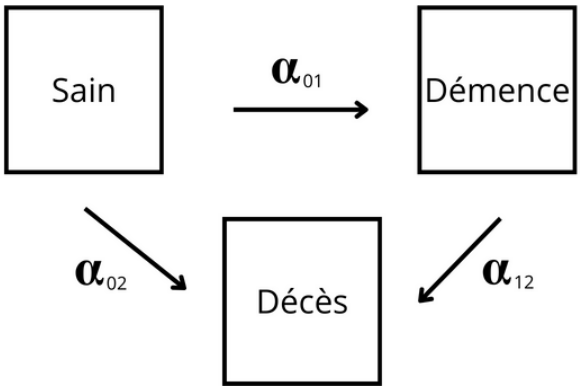


FIGURE 4.2 – Modèle multi-états Sain-Dément-Décédé

Le modèle décrit le risque de transition de sain à dément (α_{01}), de sain à décédé (α_{02}), de dément à décédé (α_{12}).

Pour modéliser le risque instantané de chaque transition d'un état à un autre (α_{01},α_{02},α_{12}), dans ce travail, nous nous sommes concentrés sur des risques instantanés de transitions proportionnels et nous avons considéré un modèle Semi-Markov pour modéliser le risque

de décès post-démence en fonction du temps depuis l'entrée en démence $\tilde{t}$. Le modèle s'écrit donc :

$$\alpha_{01i}(t) = \lambda_i^{0\to1}(t|W) = \lambda_0^{0\to1}(t)\exp(\beta_{0\to1}W_i) \tag{4.5}$$

$$\alpha_{02i}(t) = \lambda_i^{0\to2}(t|W) = \lambda_0^{0\to2}(t)\exp(\beta_{0\to2}W_i) \tag{4.6}$$

$$\alpha_{12i}(\tilde{t}) = \lambda_i^{1\to2}(t|W) = \lambda_0^{1\to2}(t)\exp(\beta_{1\to2}W_i, D_i) \tag{4.7}$$

où $\lambda_0^{a\to b}(t)$ correspond au risque instantané de base de l'état a vers l'état b. W est le vecteur des variables explicatives et $\beta_{a\to b}$ les coefficients associés. W inclut l'exposition X, les covariables C et l'événement intermédiaire D pour la transition α_{12}. Λ_{01}, Λ_{02} et Λ_{12} sont les fonctions d'intensité de transition cumulées correspondantes, telles que $\Lambda_{ab}(t|W) = \int_0^t \lambda^{a\to b}(u|W)du$.

4.3.3.2 Probabilité de survie

La probabilité de survie en t, $S_Z(t) = P(Z > t|X, C)$, s'écrit comme la somme de la probabilité de survie sans démence P_{00} et de survie après démence P_{01}. On a alors :

$$S_Z(t|X,C) = P(Z > t|C,X) = \underbrace{P(Z > t, D > t|X,C)}_{P_{00}(t)} + \underbrace{P(Z > t, D < t|X,C)}_{P_{01}(t)} \tag{4.8}$$

La probabilité d'être vivant sans démence s'écrit :

$$P_{00}(t|X,C) = e^{-\Lambda_{01}(t|X,C)-\Lambda_{02}(t|X,C)}$$

La probabilité d'être vivant avec une démence s'écrit

$$P_{01}(t|X,C) = \int_0^t \underbrace{e^{-\Lambda_{01}(u|X,C)-\Lambda_{02}(u|X,C)}}_{\text{vivre sans démence jusqu'en u}} \times \underbrace{\lambda^{0\to1}(u|X,C)}_{\text{faire une démence en u}} \times \underbrace{e^{-\Lambda_{12}(t|X,C,D)+\Lambda_{12}(u|X,C,D)}}_{\text{rester vivant entre } u \text{ et } t} du$$

Ainsi, la probabilité contrefactuelle générique s'écrit :

$$P(Z(x, D(x')) > t|C) = e^{-\Lambda_{01}(t|X=x',C)-\Lambda_{02}(t|X=x,C)} + \int_0^t e^{-\Lambda_{01}(u|X=x',C)-\Lambda_{02}(u|X=x,C)} \times \lambda^{0\to1}(u|X=x',C) \times e^{-\Lambda_{12}(t|X=x,C)+\Lambda_{12}(u|X=x,C)} du \quad (4.9)$$

4.3.3.3 Estimation à partir du modèle multi-états

Considérons, C_D et C_Z deux temps de censure supposés indépendants. L'événement intermédiaire (e.g. la démence) est observé jusqu'à C_D et l'événement final est observé jusqu'à C_Z. Nous observons donc D^* et Z^* où $D^* = \min(D, C_D)$ et $Z^* = \min(Z, C_Z)$ et δ_D et δ_Z les indicateurs, tel que $D^* = D$ et $Z^* = Z$ respectivement.
Les paramètres d'un tel modèle multi-états peuvent etre obtenus par l'estimation de trois modèles de Cox séparés, l'un pour chaque transition en considérant pour 01 et 02 l'autre événement comme une censure.

4.3.3.4 Adaptation du modèle pour la censure par intervalle

Une fois le modèle multi-états estimé, les probabilités de survi de chaque profil potentiel de variable d'exposition $P(Z(x, D(x')) > t)$ peut être calculé à partir :

— des paramètres estimés β

— des risques instantanés prédit $\hat{\lambda}^{a\to a'}$

— des estimateurs de Nelson–Aalen des risques cumulés prédit $\hat{\Lambda}_{aa'}$

où a et a' correspondent à deux états. Les intervalles de confiance peuvent être obtenus par bootstrap non-paramétrique.

4.3.4 Adaptation aux données censurées par intervalle

Dans le cas d'un événement diagnostiqué au cours d'une visite de cohorte, le temps exact est inconnu. Il est censuré par intervalle entre la dernière visite avec un diagnostic négatif D_L et la première visite avec un diagnostic positif D_U.

La définition des estimandes causaux ne change pas, ni les hypothèses tant que la censure reste non informative. Cependant, l'estimation par le modèle multi-états change, faisant ainsi intervenir deux complexités. Premièrement, le temps intermédiaire ne peut

plus être directement inclus dans la transition 12 de l'équation 4.9. Deuxièmement la vraisemblance doit tenir compte de la censure par intervalle.

Une façon d'approcher le temps exact de l'événement intermédiaire est de considérer la moitié de l'intervalle, telle que $\hat{D} = \frac{D_L + D_U}{2}$ au moment de l'estimation.

Concernant la vraisemblance, dès lors que le temps intermédiaire est observé de façon censurée par intervalle, la contribution d'un individu à la vraisemblance change :

— pour un sujet diagnostiqué dément, son risque de démence est intégré dans l'intervalle D_L et D_U.

— pour un sujet décédé avant le diagnostic de démence, la contribution doit inclure le fait que le sujet a pu développer la démence entre D_L et Z^*

Pour plus de détails sur la vraisemblance d'un modèle multi-états pour données censurées par intervalle sont présentés dans Touraine et al. (2016). Ce modèle peut être estimé par le package **SmoothHazard** en supposant des risques instantanés approchés par splines pénalisés. Actuellement ce package ne permet pas de prendre en compte des variables explicatives dépendantes du temps. Or, afin de pouvoir considérer la censure par intervalle dans nos modèles, il faudrait dans un des modèles (transition de l'état 1 à l'état 2) ajustée sur le temps de démence.

4.4 Application

Pour illustrer notre méthodologie, nous avons utilisé les données de la cohorte $3C$, pour étudier le mécanisme causal sous-jacent entre le score de santé cardiovasculaire, la démence et le décès. Plus précisement, nous cherchons à quantifier la part d'effet du score de santé cardiovasculaire à l'inclusion sur la survenue de décès, qui passe par la démence. Le score de santé cardiovasculaire, créé en 2010 par la American Heart Association (AHA), repose sur 7 items différents : 4 concernent les habitudes de vie (tabagisme, poids du corps, activité physique et régime alimentaire) et 3 items sont sur les mesures biologiques (cholestérol, glycémie et pression artérielle) (Lloyd-Jones (2010))).

4.4.1 Sélection de l'échantillon

Notre échantillon d'étude inclut les 7978 participants de la cohorte 3C ayant eu une évaluation de leur santé cardio. Dans notre étude, nous allons plus spécifiquement nous intéresser à 6197 participants de la cohorte $3C$, en ne considérant que ceux ayant eu un score de santé cardiovasculaire à l'inclusion et ne manifestant aucun antécédent de démence ou de maladie cardiovasculaire au début de l'étude, comme représenté sur la figure 4.3.

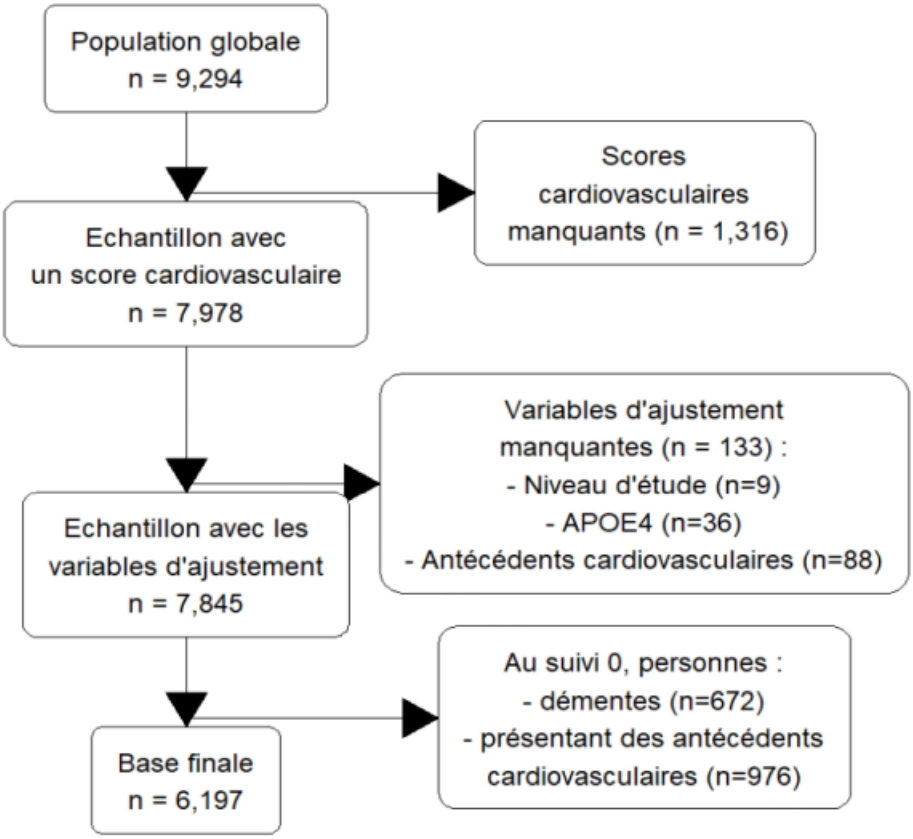


FIGURE 4.3 – Diagramme de flux pour la selection de notre échantillon dans la cohorte 3C

4.4.2 Description

Parmi les 6197 participants de notre échantillon, 3918 (63%) étaient des femmes, 2476 (40%) avaient un niveau d'éducation supérieur au secondaire et 1227 (19,8%) étaient porteurs de l'APOE4 (Tableau 4.1). Les participants avaient en moyenne 73,5 ans (sd = 5,8) au début de l'étude. Parmi les 6197 participants 751 ont fait une démence au cours de leurs suivi, avec un âge diagnostic moyen de 83 ans.

TABLE 4.1 – Caractéristiques de 6197 participants de la cohorte 3C

Variable	**Moyenne (écart-type)**	**Effectif (%)**
Score cardiovasculaire (sur 14)	8, 42 (2, 05)	
Sexe (femme)		3918 (63, 0%)
Niveau d'étude (Supérieur)		2476 (40, 0%)
APOE4		1227 (19, 8%)
Age à l'entrée de l'étude	73, 52 (5, 80)	
Age en fin d'étude (ou censuré)	84, 91 (6, 18)	
Démence		751 (12, 1%)
Age au diagnostic	83, 08 (6, 48)	
Décès		1887 (30, 5%)
Age du décès	85, 63 (7, 18)	

En l'absence de programme disponible pour traiter la censure par intervalle dans le modèle multi-états considéré (avec la transition démence-décès dépendante du temps de démence), nous avons imputé le temps de démence. Les sujets ayant été dépistés pour la démence tous les deux/trois ans après leur entrée dans l'étude, à savoir au moment des visites, l'âge de démence a été imputé comme le milieu de l'intervalle entre la dernière visite avec un diagnostic négatif et la première visite avec un diagnostic positif. En ce qui concerne les sujets décédés, leur âge de décès est connu exactement. Parmi les 1887 participants décédés, 1532 n'ont pas été diagnostiquées avec une démence. En définissant le temps de démence ainsi, nous ignorons le risque d'avoir été atteint de la maladie entre leur dernière visite et leur décès.

4.4.3 Estimation du modèle multi-états avec temps exact de démence

4.4.3.1 Estimation des modèles de survie

Trois modèles de Cox ont été effectués, chacun étudiant soit : le risque de développer une démence, le risque de décès, le risque de décès sachant que l'on est atteint de démence.

Dans ces modèles nous avons considéré le délai depuis l'inclusion et le délai depuis la démence pour le modèle de décès avec démence. Pour chacun des modèles nous avons ajusté sur le sexe, le niveau d'étude, l'apoe4, l'âge à baseline et le centre. Le modèle de survie après démence a aussi été ajusté sur le délai de démence. Les résultats des différents modèles sont resumés dans le tableau 4.2.

TABLE 4.2 – Estimation du modèle multi-états dans l'échantillon issu de la cohorte 3C : risques relatifs (RR) et p-valeurs des tests de Wald pour chaque modèle

Modèle	Modèle de démence		Modèle de décès sans démence		Modèle de décès avec démence	
	$\hat{RR}$	P-valeur	$\hat{RR}$	P-valeur	$\hat{RR}$	P-valeur
Score cardiovasculaire	0,94	<0.05	0,91	<0.05	0,97	0,30
Femme (ref. homme)	1,00	0,99	0,51	<0.05	0,61	<0.05
Âge au début de l'étude (pour 10 ans)	2,05	<0.05	1,95	<0.05	0,75	<0.05
Présence Apoe4 (ref. absence)			1,07	0,30	1.15	0.26
Haut niveau d'étude (ref. bas)	0,81	<0.05	0,94	0,28	1,20	0,10
Temps de démence					1.13	<0.05

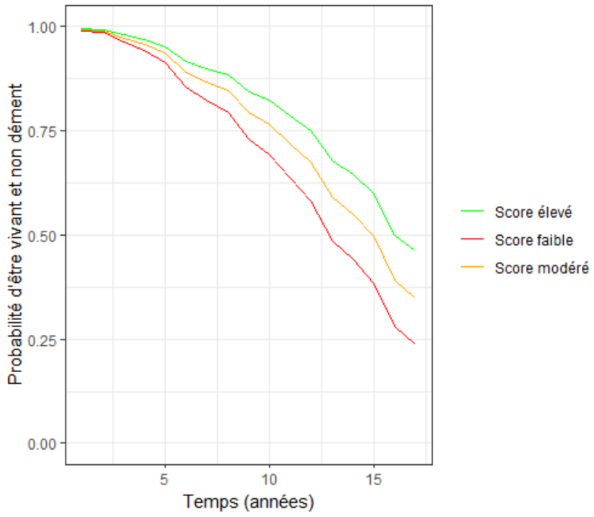
Ces modèles montrent qu'un score de santé cardiovasculaire élevé est un facteur protecteur du décès mais aussi de la démence après ajustement sur les autres variables (p-value < 0.05 pour les modèles de démence et décès sans démence). Le risque de développer une démence est réduit de 6 % pour un individu par rapport à un autre, à caractéristiques égales.

4.4.4 Estimation des probabilités de survie selon le niveau de santé cardiovasculaire

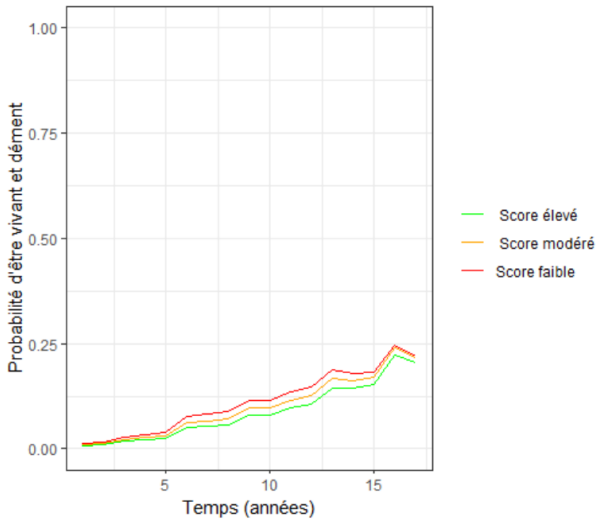
Les probabilités de survie selon le niveau de santé cardiovasculaire sont représentées sur la Figure 4.4. Il s'agit respectivement de la probabilité d'être vivant et non dément, de la probabilité de survie et de la probabilité d'être vivant et dément, en (a), (b) et (c).

Nous pouvons remarquer sur ces graphiques que les personnes avec un score de santé cardiovasculaire plus hauts ont une probabilité plus grande d'être vivants en général et vivants non-déments (Figures 4.4 a et c.). Sur la figure 4.4 b, les différences entre les courbes sont très faibles, ainsi il semblerait que le score cardiovasculaire ait moins d'importance sur la probabilité d'être dément et vivant.

(a)



(b)



(c)

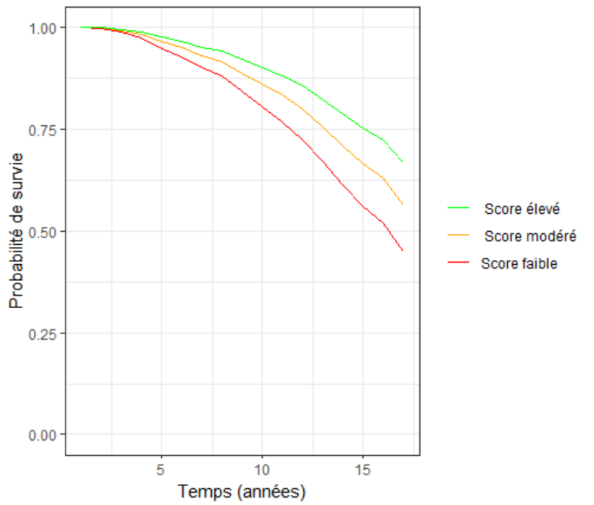


FIGURE 4.4 – Evolution au cours du temps avec différents scores cardiovasculaires (4, 8 et 12) des différentes probabilités (a). d'être non dément et vivant, (b). d'être dément et vivant et (c). d'être vivant

La figure 4.5 représente les effets total et indirect du score cardiovasculaire sur la démence passant par le décès. Nous pouvons constater sur cette figure, que les courbes de l'effet total (en rouge) et de l'effet indirect (en bleu) se superposent, indiquant que jusqu'à 5ans de suivi, l'effet du score cardiovaculaire sur le décès serait totalement médié par la démence. Néanmoins, cela est à mettre en relation avec le faible risque de faire l'événement jusqu'à 5 ans.

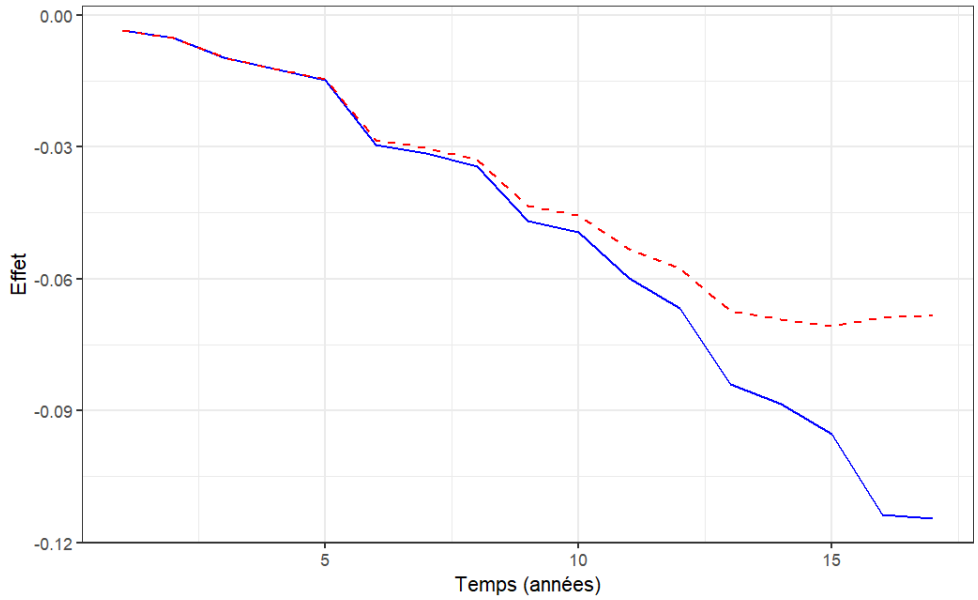


FIGURE 4.5 – Effet total (en rouge) versus effet indirect (en bleu) d'une intervention sur le score cardiovasculaire à 12 versus à 4 au cours du temps

4.5 Discussion

Dans ce travail, nous avons adapté et appliqué l'approche d'analyse de médiation proposée par Valeri et al. (2023), pour des effets naturels en présence d'une variable d'exposition fixe, et d'un médiateur et d'un outcome tous deux des temps d'événements censurés à droite, l'outcome étant un événement terminal. Nous avons défini nos effets causaux, les hypothèses d'identifiabilité associées, et nous avons utilisé comme modèle de travail un modèle multi-états qui permet la modélisation simultanée de deux événements.

Bien que la plupart des modèles multi-états utilisent l'hypothèse Markovienne, il est généralement plus raisonnable, dans certains cas, d'adopter l'hypothèse Semi-Markovienne, notamment lorsque l'événement intermédiaire entraîne un risque plus important d'événe-

ments terminaux. C'est le cas pour la démence. Plusieurs travaux ont rapporté des risques accrus de décès après la démence (Agüero-Torres et al. (1999)).

Néanmoins, le passage de l'hypothèse semi-markovienne à l'hypothèse markovienne est très simple et intervient uniquement dans la spécification du modèle de travail. Cela n'altère donc pas l'approche d'analyse de médiation que nous proposons.

Dans le cadre de la thèse, l'approche tenant compte de la censure par intervalle n'a malheureusement pas pu être mise en œuvre. En effet, le logiciel d'estimation du modèle multi-états pour censure par intervalle ne permettait pas de dériver des probabilités de décès lorsque le temps intermédiaire est inclus en variable explicative, ce qui est requis en médiation.

Notre hypothèse est que l'impact est signifiant. En effet, il a été montré par des simulations que les coefficients estimés dans un modèle multi-états négligeant la censure par intervalle étaient biaisés. Ce biais est particulièrement notable dans le cas de facteur de risque ayant des effets dans la même direction pour les deux événements (Leffondré et al. (2013)). C'est précisement ce que nous observons lorsque nous étudions le lien entre le score cardiovasculaire et le décès, passant par la démence, voire plus généralement lorsque nous effectuons des analyses de médiation.

Pour illustrer ce travail, nous avons quand même réalisé l'analyse sur les données de la cohorte $3C$. Du fait de ce biais potentiel, les résultats restent à interpréter avec précaution.

Chapitre 5

Approche par variables instrumentales

Motivé à nouveau par l'étude des facteurs de risque du vieillissement cognitif, ce chapitre vise à proposer une approche statistique pour étudier le lien causal existant entre un facteur de risque et un processus de santé dans les études de cohortes observationnelles, en présence de facteurs de confusion non mesurés.

Plus précisément, nous nous intéressons à la compréhension du lien causal entre le diabète et le déclin cognitif. Dans cette situation, nous sommes confrontés aux problèmes de la multifactorialité du vieillissement, avec de multiples facteurs de risque qui agissent sur le diabète et de multiples facteurs de risque qui agissent sur le déclin cognitif. Ces facteurs sont parfois communs à la fois au diabète et au déclin cognitif, sans pour autant en être la cause directe. C'est ce que nous avons défini dans l'état de l'art comme étant des facteurs de confusion. La non-prise en compte de ces facteurs peut mener à des biais lorsque l'on étudie le lien entre le facteur de risque et le processus de santé, et ne permet donc pas une interprétation causale. Pourtant, lorsqu'une étude de cohorte est effectuée, elle ne peut être exhaustive, et considérons qu'elle le soit, les facteurs de confusion ne sont pas tous connus.

La méthode par variables instrumentales permet de pallier le biais induit par les facteurs de confusion non mesurés en introduisant une variable exogène, c'est-à-dire une

variable qui n'est pas influencée par les autres variables du modèle.

Lorsque l'on souhaite étudier le rôle du diabète à l'entrée dans l'étude sur le déclin cognitif (possédant des mesures répétées dans le temps), il est courant de réaliser un modèle mixte. Néanmoins, en présence de facteurs de confusion omis, il n'est pas possible d'interpréter causalement les résultats obtenus.

Dans ce chapitre, nous proposons une approche par variables instrumentales. Elle considère un modèle mixte pour la variable d'intérêt Y, répétée. Elle est basée sur l'estimation classique des méthodes par variables instrumentales, à savoir une estimation en deux étapes (modèle pour X en fonction de Z, prédiction de X, modèle pour Y en fonction de la prédiction de X). Au travers de simulations, la méthode est évaluée et comparée aux méthodes naïves qui ne tiennent pas compte de la confusion non observée.

Ce travail a fait l'objet d'une publication dans le journal : *Biometrical Journal*

Received: 19 December 2022 | Revised: 12 June 2023 | Accepted: 11 August 2023

DOI: 10.1002/bimj.202200358

Biometrical Journal

RESEARCH ARTICLE

Addressing unmeasured confounders in cohort studies: Instrumental variable method for a time-fixed exposure on an outcome trajectory

Kateline Le Bourdonnec[1] | **Cécilia Samieri**[1] | **Christophe Tzourio**[1] | **Thibault Mura**[2] | **Aniket Mishra**[1] | **David-Alexandre Trégouët**[1] | **Cécile Proust-Lima**[1]

[1]Inserm, BPH, U1219, University of Bordeaux, Bordeaux, France

[2]Institute for Neurosciences of Montpellier INM, University of Montpellier, INSERM, Montpellier, France

Correspondence
Kateline Le Bourdonnec, Univ. Bordeaux, Inserm, BPH, U1219, F-33000 Bordeaux, France.
Email: kateline.le-bourdonnec@u-bordeaux.fr

Funding information
INSERM GOLD Cross-Cutting program; Agence Nationale de la Recherche, Grant/Award Number: ANR-18-CE36-0004-01

Abstract

Instrumental variable methods, which handle unmeasured confounding by targeting the part of the exposure explained by an exogenous variable not subject to confounding, have gained much interest in observational studies. We consider the very frequent setting of estimating the unconfounded effect of an exposure measured at baseline on the subsequent trajectory of an outcome repeatedly measured over time. We didactically explain how to apply the instrumental variable method in such setting by adapting the two-stage classical methodology with (1) the prediction of the exposure according to the instrumental variable, (2) its inclusion into a mixed model to quantify the exposure association with the subsequent outcome trajectory, and (3) the computation of the estimated total variance. A simulation study illustrates the consequences of unmeasured confounding in classical analyses and the usefulness of the instrumental variable approach. The methodology is then applied to 6224 participants of the 3C cohort to estimate the association of type-2 diabetes with subsequent cognitive trajectory, using 42 genetic polymorphisms as instrumental variables. This contribution shows how to handle endogeneity when interested in repeated outcomes, along with a R implementation. However, it should still be used with caution as it relies on instrumental variable assumptions hardly testable in practice.

KEYWORDS
causality, cohort study, instrumental variable, mixed model, repeated data

1 | INTRODUCTION

Observational studies are widely used in epidemiology to assess the relation between an exposure X and an outcome Y, with the perspective to identify the causal effect of X on Y. Statistical techniques (Ertefaie et al., 2017; Hernan & Robins, 2020) have been used to derive causal interpretations in the presence of confounding. However, they rely on the assumption that all the sources of confounding have been observed and controlled for. Yet, in many contexts, the assumption that all the confounders are observed is unrealistic, and statistical analyses are likely to provide biased estimates of causal associations (Fewell et al., 2007). For instance, when studying the relation between cardiometabolic factors on cognitive aging, so many confounders may intervene (Rawlings et al., 2014) that residual unobserved confounding is very likely. The issue of unmeasured confounding relates to the more general problem of endogeneity that occurs when the covariate is partly explained by the system under study. Beyond confounding, endogeneity also encompasses reverse causation that occurs when the outcome or its underlying process may cause a change in the exposure (Wagner, 2018).

To handle endogeneity, instrumental variable (IV) analysis, first developed in Economics (Wright, 1928), was applied in Public Health from the early 2000s (Greenland, 2000). This method consists in using an exogenous variable, the "IV", which is not subject to unmeasured confounding and recreates the randomization framework. The principle of the IV methodology can be illustrated in the cross-sectional framework (Figure 1A). Let us denote Z the IV, X the endogenous exposure variable, Y the outcome, and U the unobserved confounders. To be considered as valid, the IV needs to satisfy three assumptions (Greenland, 2000): (1) Z is strongly associated with X; (2) Z is associated with Y only through X; and (3) Z is independent of U conditionally on X. Under these assumptions, Z can be used to retrieve the causal association between X and Y. In epidemiology, genetic data have been considered as promising IV because genes are determined from birth, thus not subject to confounding; in this context, IV methodology is called Mendelian randomization (MR) (Davies et al., 2018). Finally, to be interpreted as causal effects, IV analyses require a fourth assumption of homogeneity for the average causal effect or monotonicity for the local average causal effect (Hernán & Robins, 2006; Swanson & Hernán, 2018).

The most widely used estimation technique in IV methodology is the two-stage approach, called two-stage least square (2SLS) method (Burgess et al., 2017): (1) the endogenous exposure is regressed on the IV and (2) the derived prediction, which is independent of the unmeasured confounders due to the assumptions of Z, substitutes the exposure in the regression of the outcome to quantify the causal relation between X and Y. First proposed in the cross-sectional framework where X and Y were continuous variables measured at a single time point (Burgess et al., 2017), it was adapted to handle binary exposures and/or binary outcomes (Li et al., 2022; Terza et al., 2008), and to treat grouped data (Li et al., 2015, 2020).

Recently, the methodology was extended to handle longitudinal data. Two settings were explored: (i) an exposure repeatedly measured over time and its effect on the concomitant level of a repeatedly measured outcome (Hogan & Lancaster, 2004; O'Malley, 2012) and (ii) a time-fixed exposure and its effect on the subsequent risk of an event (Li et al., 2015; Martínez-Camblor et al., 2019; Tchetgen Tchetgen et al., 2015). Yet, another frequent setting encountered in longitudinal studies concerns a time-fixed exposure and its effect on the subsequent trajectory of an outcome repeatedly measured over time.

In the present contribution, we aim to didactically explain how the IV methodology can be used in observational cohort studies to assess the association between an exposure collected at baseline and the trajectory of an outcome repeatedly measured over follow-up in the presence of potential unmeasured confounding. Our solution consists in considering

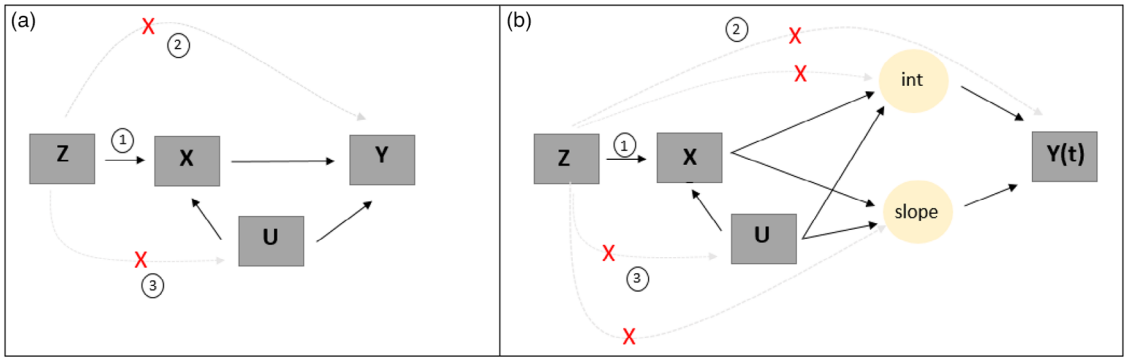


FIGURE 1 Directed acyclic graph for the IV methodology with a cross-sectional outcome Y (Panel A) or a longitudinal continuous outcome Y (Panel B). X is the exposure, Z the instrumental variable (with 1, 2, 3 the corresponding IV assumptions), and U the (partially) unobserved confounders. Int and slope represent the underlying latent level of Y at baseline and the latent slope of Y over time, respectively.

a mixed model for the repeated marker in the second step of the two-stage IV approach. We show how this can solve situations of unmeasured confounding and endogeneity, and we illustrate it in a simulation study considering both a binary and a continuous exposure, and a continuous outcome. We finally apply the methodology to assess the association between type-2 diabetes and cognitive aging in the French cohort "Three city" (3C) (Alperovitch, 2003), by using genetic polymorphisms as the exogenous variable.

2 | METHODS

2.1 | Framework

Let us consider a classical longitudinal framework (Figure 1B) where X is the time-fixed exposure, $\mathbf{U}$ is a r-vector of confounders, and $\mathbf{Z}$ is a p-vector of exogenous (instrumental) variables, all defined and measured at entry in the cohort while the continuous outcome Y is repeatedly measured over time t after baseline. Without loss of generality, we assume $\mathbb{E}(\mathbf{U}) = 0$.

To ease the problem description, we first consider the case of a continuous exposure, and we assume that Y evolves linearly over time and can be summarized by its latent level at baseline and its latent slope over time, on which the other variables can have an effect. The generalization to a nonlinear trajectory over time is straightforward by considering a more flexible basis of time functions instead of only intercept and slope.

Let us assume that the true relations schematized in Figure 1(B) translate for each subject i ($i = 1, \dots, N$) of a sample and each occasion j ($j = 1, \dots n_i$) in a linear regression for the continuous exposure (1) and a linear mixed model for the outcome (2):

$$X_i = \alpha_0^* + \mathbf{Z_i}^\top \boldsymbol{\alpha}_\mathbf{Z}^* + \mathbf{U_i}^\top \boldsymbol{\alpha}_\mathbf{U}^* + \epsilon_i^{X*}, \tag{1}$$

$$\begin{aligned} Y_{ij} = & \underbrace{\beta_0^* + X_i\beta_e^* + \mathbf{U_i}^\top \boldsymbol{\beta}_\mathbf{U}^* + b_{0i}^*}_{Int_i} \\ & + \underbrace{\left(\beta_t^* + X_i\beta_{te}^* + \mathbf{U_i}^\top \boldsymbol{\beta}_\mathbf{tU}^* + b_{1i}^*\right)}_{Slope_i} t_{ij} + \epsilon_{ij}^{Y*}. \end{aligned} \tag{2}$$

For the sake of readability, conditioning on covariates and random effects, although systematic, is not made explicit in any of the linear regressions throughout the manuscript.

Following classical definitions of the linear mixed model (Commenges & Jacqmin-Gadda, 2015; Laird & Ware, 1982), $\mathbf{b_i^*} = (b_{0i}^*, b_{1i}^*)^\top \sim \mathcal{N}(0, \mathbf{B}^*)$ is the vector of individual random effects that accounts for the intraindividual correlation within the repeated Y measures. The measurement error in the exposure regression ϵ_i^{X*} is independent of Z_i and $\mathbf{U_i}$ and the measurement error at time t_{ij} in the outcome regression $\epsilon_{ij}^{Y*} \sim \mathcal{N}(0, \sigma_Y)$ is independent of all the other measurement errors at different times $\epsilon_{ij'}^{Y*}$ with $j' \neq j$, and of X_i, $\mathbf{U_i}$, and $\mathbf{b_i^*}$. The random effects $\mathbf{b_i^*}$ are also independent of X_i and $\mathbf{U_i}$. In Equations (1) and (2), superscript * refers to the parameters and latent variables under the true model.

The parameters of interest are β_e^* and β_{te}^* corresponding to the effect of X on the level of Y at inclusion and the effect of X on the subsequent change of Y over time, respectively. Since all confounders are included through U in model (2), we can interpret these parameters in a causal way. The fundamental problem is that this model and these parameters cannot be directly estimated when some of the confounders $\mathbf{U}$ are not observed. Let us split $\mathbf{U} = (\mathbf{U^o}, \mathbf{U^m})$ with $\mathbf{U^o}$ the observed confounders and $\mathbf{U^m}$ the unobserved confounders.

2.2 | Naive approach neglecting unobserved confounding

In the presence of unobserved confounding, a naive solution consists in estimating the association between X and the trajectory of Y by considering the model that includes $\mathbf{U^o}$ but omits $\mathbf{U^m}$:

$$\begin{aligned} Y_{ij} = & \beta_0^N + \beta_e^N X_i + b_{0i}^N + \mathbf{U_i^{o\top}} \boldsymbol{\beta}_\mathbf{Uo}^\mathbf{N} \\ & + (\beta_t^N + \beta_{te}^N X_i + \mathbf{U_i^{o\top}} \boldsymbol{\beta}_\mathbf{tUo}^\mathbf{N} + b_{1i}^N) t_{ij} + \epsilon_{ij}^{NY}. \end{aligned} \tag{3}$$

The estimation of this model relies on the same distributions and independence assumptions as defined for model (2). Yet, those are not satisfied anymore in the presence of unobserved confounding: the neglected confounders $\mathbf{U^m}$ are absorbed by the individual random effects: $b_{0i}^N = b_{0i}^* + \mathbf{U_i^{m\top}}\boldsymbol{\beta}_{\mathbf{Um}}^*$ and $b_{1i}^N = b_{1i}^* + \mathbf{U_i^{m\top}}\boldsymbol{\beta}_{\mathbf{tUm}}^*$, so that $\mathbf{b_i^N} = (b_{0i}^N, b_{1i}^N)^\top$ is not independent of X_i anymore, and is not homoscedastic anymore. Of note, $\mathbf{U_i^m}$ induces a correlation between b_{0i}^N and b_{1i}^N even when b_{0i}^* and b_{1i}^* were initially independent.

When $\mathbf{U^m}$ is not a confounder, $(\hat{\beta}_e^N, \hat{\beta}_{te}^N)$ is an unbiased estimate of $(\beta_e^*, \beta_t e^*)$ from Equation (2), and under the assumption that $\mathbb{E}(\mathbf{U^m}) = 0$, $E(Y_{ij}|X_i, \mathbf{Z_i}, \mathbf{U_i}, t_{ij}) = E(Y_{ij}|X_i, \mathbf{Z_i}, \mathbf{U_i^o}, t_{ij})$. However, when $\mathbf{U^m}$ includes confounders, $E(Y_{ij}|X_i, \mathbf{Z_i}, \mathbf{U_i}, t_{ij}) \neq E(Y_{ij}|X_i, \mathbf{Z_i}, \mathbf{U_i^o}, t_{ij})$ since $E(b_{0i}^N|X_i, \mathbf{Z_i}, \mathbf{U_i^o}, t_{ij}) \neq 0$ and $E(b_{1i}^N|X_i, \mathbf{Z_i}, \mathbf{U_i^o}, t_{ij}) \neq 0$, and $(\hat{\beta}_e^N, \hat{\beta}_{te}^N)$ is not an unbiased estimator of $(\beta_e^*, \beta_{te}^*)$ anymore.

2.3 | Instrumental variable approach

The two-stage IV methodology aims at correcting the bias due to residual unmeasured confounding. We show here how it can be adapted to the longitudinal framework described above by replacing the second-stage least-square regression by a second-stage linear mixed model.

For clarity, we distinguish below the case of a continuous endogenous exposure from the case of a binary endogenous exposure. The method relies on the independence between the regressors $(\mathbf{Z}, \mathbf{U^o})$ and the unobserved variables $\mathbf{U^m}$. As this assumption may likely be violated between $\mathbf{U^m}$ and $\mathbf{U^o}$, we consider below the total vector $\mathbf{U} = (\mathbf{U^m}, \mathbf{U^o})$ as being unobserved to ensure independence.

2.3.1 | X continuous

With a continuous endogenous exposure, the two-stage methodology is defined as follows:

$$X_i = \alpha_0 + \mathbf{Z_i^\top}\boldsymbol{\alpha}_{\mathbf{Z}} + e_i^X, \tag{4}$$

$$\begin{aligned} Y_{ij} = {} & \beta_0 + E(X_i|\mathbf{Z_i})\beta_e + b_{0i} \\ & + (\beta_t + E(X_i|\mathbf{Z_i})\beta_{te} + b_{1i})t_{ij} + \varepsilon_{ij}^Y. \end{aligned} \tag{5}$$

This model relies on the same distributions and independence assumptions as model (2).

From the IV conditional independence assumption (3), the conditional expectation $E(X_i|\mathbf{Z_i}) = \tilde{X}_i = \alpha_0^* + \mathbf{Z_i^\top}\boldsymbol{\alpha}_{\mathbf{Z}}^*$ and the residual $X_i - \mathbb{E}(X_i|\mathbf{Z_i}) = \mathbf{U_i^\top}\boldsymbol{\alpha}_{\mathbf{U}}^* + \epsilon_i^{X*}$.

When rewriting Equation (2) according to $\mathbb{E}(X_i|\mathbf{Z_i})$, one obtains:

$$\begin{aligned} Y_{ij} &= \beta_0^* + X_i\beta_e^* + \mathbf{U_i^\top}\boldsymbol{\beta}_{\mathbf{U}}^* + b_{0i}^* \\ &\quad + \left(\beta_t^* + X_i\beta_{te}^* + \mathbf{U_i^\top}\boldsymbol{\beta}_{\mathbf{tU}}^* + b_{1i}^*\right)t_{ij} + \epsilon_{ij}^{Y*} \\ &= \beta_0^* + \mathbb{E}(X_i|\mathbf{Z_i})\beta_e^* + (X_i - \mathbb{E}(X_i|\mathbf{Z_i}))\beta_e^* + \mathbf{U_i^\top}\boldsymbol{\beta}_{\mathbf{U}}^* + b_{0i}^* \\ &\quad + \left(\beta_t^* + \mathbb{E}(X_i|\mathbf{Z_i})\beta_{te}^* + (X_i - \mathbb{E}(X_i|\mathbf{Z_i}))\beta_{te}^* + \mathbf{U_i^\top}\boldsymbol{\beta}_{\mathbf{tU}}^* + b_{1i}^*\right)t_{ij} + \epsilon_{ij}^{Y*}. \end{aligned} \tag{6}$$

And using that $X_i - \mathbb{E}(X_i|\mathbf{Z_i}) = \mathbf{U_i^\top}\boldsymbol{\alpha}_{\mathbf{U}}^* + \epsilon_i^{X*}$ from model (1),

$$\begin{aligned} Y_{ij} = {} & \beta_0^* + \mathbb{E}(X_i|\mathbf{Z_i})\beta_e^* + (\mathbf{U_i^\top}\boldsymbol{\alpha}_{\mathbf{U}}^* + \epsilon_i^{X*})\beta_e^* + \mathbf{U_i^\top}\boldsymbol{\beta}_{\mathbf{U}}^* + b_{0i}^* \\ & + \left(\beta_t^* + \mathbb{E}(X_i|\mathbf{Z_i})\beta_{te}^* + (\mathbf{U_i^\top}\boldsymbol{\alpha}_{\mathbf{U}}^* + \epsilon_i^{X*})\beta_{te}^* + \mathbf{U_i^\top}\boldsymbol{\beta}_{\mathbf{tU}}^* + b_{1i}^*\right)t_{ij} + \epsilon_{ij}^{Y*}, \end{aligned} \tag{7}$$

which reduces to:

$$Y_{ij} = \beta_0^* + \mathbb{E}(X_i|\mathbf{Z_i})\beta_e^* + b_{0i} + \left(\beta_t^* + \mathbb{E}(X_i|\mathbf{Z_i})\beta_{te}^* + b_{1i}\right)t_{ij} + \epsilon_{ij}^{Y*} \tag{8}$$

with $b_{0i} = \mathbf{U_i}^\top(\boldsymbol{\alpha}^*_{\mathbf{U}}\boldsymbol{\beta}^*_{\mathbf{e}} + \boldsymbol{\beta}^*_{\mathbf{U}}) + \epsilon_i^{X*}\beta_e^* + b_{0i}^*$ and $b_{1i} = \mathbf{U_i}^\top(\boldsymbol{\alpha}^*_{\mathbf{U}}\boldsymbol{\beta}^*_{\mathbf{te}} + \boldsymbol{\beta}^*_{\mathbf{tU}}) + \epsilon_i^{X*}\beta_e^* + b_{1i}^*$. By definition, $\mathbb{E}(X_i|\mathbf{Z_i})$ and $\mathbf{U_i}$ are independent, so $\mathbf{b_i} = (b_{0i}, b_{1i})^\top$ is independent of the covariates in the model, as required in a linear mixed model. The model defined in Equation (5) is thus equivalent to the target model in Equation (2), except that the variance of the random effects is not homoskedastic anymore.

Maximum likelihood estimates of the fixed effects in a mixed model being unbiased even when the covariance structure is misspecified (following the same principle as with generalized estimating equations, Liang & Zeger, 1986), $\hat{\beta_e}$ and $\hat{\beta_{te}}$ are unbiased estimators of β_e^* and β_{te}^*; they may be used to quantify the causal relation between X and Y. However, their variance needs to be corrected for the heteroskedasticity and the use of an IV. By applying the same principle of robust variances (Royall, 1986; White, 1980) as in IV methods for cross-sectional studies (e.g., in ivtools R package, Sjolander & Martinussen, 2019), we define the following sandwich estimator:

$$V_{2\text{-S}}(\hat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{N} \hat{\mathbf{W}}_\mathbf{i}^\mathbf{T}\hat{\mathbf{V}}_\mathbf{i}^{-\mathbf{1}}\hat{\mathbf{W}}_\mathbf{i}\right)^{-1}\left(\sum_{i=1}^{N} \hat{\mathbf{W}}_\mathbf{i}^\mathbf{T}\hat{\mathbf{V}}_\mathbf{i}^{-\mathbf{1}}\mathbf{V_i}\hat{\mathbf{V}}_\mathbf{i}^{-\mathbf{1}}\hat{\mathbf{W}}_\mathbf{i}\right)\left(\sum_{i=1}^{N} \hat{\mathbf{W}}_\mathbf{i}^\mathbf{T}\hat{\mathbf{V}}_\mathbf{i}^{-\mathbf{1}}\hat{\mathbf{W}}_\mathbf{i}\right)^{-1}, \tag{9}$$

where $\hat{\mathbf{W}}_\mathbf{i}$ is the matrix of variables associated with the vector of fixed effects $\boldsymbol{\beta}$ (in our example in Equation (5), $\hat{\mathbf{W}}_\mathbf{i}$ is a $n_i \times 4$-matrix with intercept, time, $\mathbb{E}(X_i|\mathbf{Z_i})$ and its interaction with time, and $\boldsymbol{\beta} = (\beta_0, \beta_t, \beta_e, \beta_{te})^\top$), $\hat{\mathbf{V}}_\mathbf{i} = M_i\hat{\mathbf{B}}M_i^T + \hat{\sigma}_y^2 I_{n_i}$ with M_i the matrix of variables related to the random effects (in our example, an $n_i \times 2$ with intercept and time), I_{n_i} is the identity matrix, and $\hat{\boldsymbol{\beta}}$, $\hat{\mathbf{B}}$, $\hat{\sigma}$ are the estimates obtained in the second-stage model (5). Finally, $\mathbf{V_i}$ is the empirical covariance matrix of Y, that is, $\mathbf{V_i} = \text{Cov}(\mathbf{Y_i} - \mathbf{W_i}^\top\hat{\boldsymbol{\beta}}, \mathbf{Y_i} - \mathbf{W_i}^\top\hat{\boldsymbol{\beta}})$ where $\mathbf{W_i}$ is the $n_i \times 4$ matrix with intercept, time, X_i, and its interaction with time.

The robust variance $V_{2\text{-S}}(\hat{\beta})$ quantifies the second-stage variability in the estimates, but it neglects the first-stage uncertainty. To compute the total variance that accounts for the variability in the two stages, we use a parametric bootstrap (Efron & Tibshirani, 1993): instead of running the second-stage analysis once from the maximum likelihood estimates $\hat{\alpha}$, the second stage is replicated M times from first-stage parameters α_m ($m = 1, .., M$) randomly drawn from their asymptotic normal distribution with mean $\hat{\alpha}$ and variance $\widehat{V(\hat{\alpha})}$. The total variance estimate of $\hat{\boldsymbol{\beta}}$ can then be derived with the Rubin's rule (Little & Rubin, 2019) from the M second-stage estimates $\hat{\boldsymbol{\beta}}_m$ as:

$$V_{\text{tot}}(\hat{\boldsymbol{\beta}}) = \frac{1}{M}\sum_{m=1}^{M}\widehat{V_{2\text{-S}}(\hat{\boldsymbol{\beta}}_m)} + \frac{(M+1)}{M(M-1)}\sum_{m=1}^{M}\left(\hat{\boldsymbol{\beta}}_m - \overline{\hat{\boldsymbol{\beta}}_m}\right)\left(\hat{\boldsymbol{\beta}}_m - \overline{\hat{\boldsymbol{\beta}}_m}\right)^\top.$$

2.3.2 | X binary

The absence of bias demonstrated for the continuous exposure comes from the use of additive models in both stages. Although not frequent, a linear model could also be considered for a binary exposure. Called linear probability model (Li et al., 2022), it translates into the exact same inference technique as described for the continuous exposure with $E(X_i|Z_i)$ derived from a linear model for X and included into the second-stage linear mixed model, and the same variance estimator.

Alternatively, the more classical logistic model can also be considered:

$$\text{logit}(\mathbb{E}(X_i|\mathbf{Z_i})) = \alpha_0 + \mathbf{Z_i}^\top\boldsymbol{\alpha}_\mathbf{Z} \tag{10}$$

with the derived $E(X_i|Z_i)$ included in the second-stage linear mixed model in (5), and the same total variance estimator used. However, due to the nonlinear nature of the logistic regression, $E(X_i|Z_i, U_i)$ does no longer equal $E(X_i|Z_i)$, and the convergence of the estimates of β_e and β_{te} to β_e^* and β_{te}^* in (2) is not ensured anymore. To further account for the residual effect of the unmeasured confounders, some authors recommended to replace the substitution of X by $E(X_i|Z_i)$ by the combination of X and the residual $X - E(X_i|Z_i)$ in the second stage. We call these three options linear/substitution, logistic/substitution, and logistic/residual inclusion, respectively.

2.4 | Software

The IV estimation technique for a binary or continuous time-fixed exposure and a continuous repeatedly measured outcome is implemented in the R package **IVmm** available at *url of the package —blinded version*. It relies on the hlme function of lcmm R package for the linear mixed model estimation (Proust-Lima et al., 2017).

3 | SIMULATION STUDY

We ran a simulation study to illustrate the behavior of the naive approach and of the IV methods in the presence of unmeasured confounding.

3.1 | Simulation design

The simulation setting followed the DAG of Figure 1(B). The procedure of data generation including parameters values considered is fully summarized in Table S1. For each individual i in a sample of size N, we first generated an exogenous IV Z_i and an unobserved confounder U_i according to standard Gaussian distributions, and random visit times $t_{ij} = j + u_{ij}$ around theoretical annual visits j (with $j = 1, .., 6$) with u_{ij} a visit-and-subject-specific random Gaussian departure ($\mathcal{N}(0, 0.05)$). We then generated the endogenous continuous exposure X_i according to model (4) (for a binary, a logistic version of (4) was considered) and the repeated measures of the outcome Y_i according to model (2).

We considered scenarios with different sample sizes (N=2000, 6000, or 20,000) and different strengths of association between the IV and the exposure α_z resulting in different strengths of the IV. As common in the IV literature, the strength of association between the IV and the exposure was quantified with the F-statistic (ratio of the explained variance and the residual variance) (Andrews et al., 2019) and the Nagelkerke R^2 for a continuous and binary exposure, respectively. For each scenario, 500 datasets were simulated.

3.2 | Simulation results

The results of the naive and the IV approaches are reported in Tables 1 and 2; they are also displayed in Figure 2 for the slope with time (and in Figure S1 for the initial level).

As expected, whatever the sample size and the strength of the IV association with the exposure, the naive method showed very large bias and null coverage rate for the association between the exposure and the change over time in all cases. In contrast, the two-stage IV methods retrieved the true causal association without any bias for the continuous

TABLE 1 Simulation results for continuous exposure (over 500 replicates) for the association between the exposure and the trajectory of Y (summarized by the effect on the baseline level and the slope over time) according to the sample size, and strength of the instrumental variable (α_Z).

			$\alpha_Z = 0.5$					$\alpha_Z = 1$			
			Baseline Level		Slope Over time			Baseline Level		Slope Over time	
N	**Methods**	**Strength[a]**	**RB**	**CR**	**RB**	**CR**	**Strength***	**RB**	**CR**	**RB**	**CR**
2000	Naive	–	44.3	0.0	44.3	0.0	–	33.3	0.0	33.2	0.0
	IV	251	−0.1	93.6	0.3	95.6	1003	0.1	96.8	0.1	95.6
6000	Naive	–	44.5	0.0	44.5	0.0	–	33.4	0.0	33.3	0.0
	IV	757	0.9	95.4	0.4	95.0	3003	−0.1	96.8	−0.1	96.2
20,000	Naive	–	44.4	0.0	44.5	0.0	–	33.3	0.0	33.3	0.0
	IV	2503	0.08	96.2	−0.0	94.6	10,009	−0.0	95.2	0.0	93.4

[a]Strength of association is assessed with the F-statistic for continuous X.

Abbreviations: CR, coverage rate of the 95% confidence interval; N, sample size; RB, relative bias (defined as the average percentage of difference between the estimate and the true parameter value).

TABLE 2 Simulation results for binary exposure with naive method, linear/substitution, and logistic/substitution IV methods (over 500 replicates) for the association between the exposure and the trajectory of Y (summarized by the effect on the baseline level and the slope over time) according to the type of exposure, the sample size, and strength of the instrumental variable (α_Z).

			$\alpha_Z = 2$					$\alpha_Z = 3$					$\alpha_Z = 4$			
			Baseline level		Slope over time			Baseline level		Slope over time			Baseline level		Slope over time	
N	Methods	Str[a]	RB	CR	RB	CR	Str[a]	RB	CR	RB	CR	Str[a]	RB	CR	RB	CR
2000	Naive	-	135.9	0.0	135.5	0.0	-	106.9	0.0	106.7	0.0	-	67.6	0.0	67.7	0.0
	Log/Res	14.3	100.3	0.0	100.2	0.0	35.0	82.7	0.0	82.5	0.0	58.6	67.9	0.0	67.7	0.0
	Log/Sub	14.3	−1.6	94.6	−2.0	95.2	35.0	−0.8	94.8	−1.4	95.4	58.6	−0.4	94.6	−1.0	95
	Lin/Sub	10.3 (229)	−1.0	95.4	−1.4	95.4	25.1 (676)	−0.1	96.0	−0.1	93.8	41.6 (1406)	0.0	94.0	0.2	94.0
6000	Naive	–	135.9	0.0	135.5	0.0	–	106.7	0.0	106.3	0.0	–	68.0	0.0	67.8	0.0
	Log/Res	14.3	100.4	0.0	100.2	0.0	35.0	82.4	0.0	81.8	0.0	58.6	−21.6	0.0	16.2	0.0
	Log/Sub	14.3	−1.3	94.6	−1.2	93.8	35.4	−1.0	94.6	−0.9	94.0	58.6	−0.7	94.0	−0.7	94.4
	Lin/Sub	10.3 (692)	−1.0	94.8	−0.1	95.4	25.1 (2025)	−0.6	96.8	−0.4	96.4	41.6 (4218)	−0.1	93.0	0.2	96.0
20,000	Naive	-	135.7	0.	135.7	0.0	–	106.7	0.0	106.8	0.0	–	67.9	0.0	67.9	0.0
	Log/Res	14.3	100.4	0.0	100.4	0.0	35.0	82.2	0.0	82.3	0.0	58.6	67.4	0.0	67.4	0.0
	Log/Sub	14.3	−0.3	93.8	0.0	95.6	35.4	−0.6	93.8	−0.3	95.6	58.6	−0.5	94.0	−0.4	95.4
	Lin/Sub	10.3 (2301)	−0.6	94.0	−0.2	95.0	25.1 (6763)	0.2	93.8	0.2	94.6	41.6 (14,037)	−0.2	94.6	−0.1	94.6

[a]Strength of association is assessed with the R^2 expressed in % (and *F*-statistic) for the linear regression, and the R^2 of Nagelkerke for the logistic regression also expressed in %.

Abbreviations: CR, coverage rate expressed in % of the 95% confidence interval;Log/Sub, logistic/substitution method; Lin/Sub = linear/substitution method; *N*, sample size; RB, relative bias expressed in % (defined as the average percentage of difference between the estimate and the true parameter value); Str, strength.

exposure, and for the binary exposure when using the linear/substitution and logistic/substitution methods, even for the scenarios with a weak instrument. In contrast, the logistic/residual methodology for a binary exposure showed large bias and null coverage rate. In the following, we thus did not investigate this method further. The simulation study also validated the proposed estimate of variance with reported coverage rate of the 95% confidence interval very close the nominal value in both the continuous and binary cases. However, although correct, the two-stage IV method showed substantial variability in the estimates when the IV was weaker.

4 | APPLICATION

We aimed to assess the relation between type-2 diabetes measured at baseline and subsequent cognitive trajectory in the elderly population. Indeed, biological mechanisms suggest an implication of type-2 diabetes on cognitive aging (Frison, 2019), but unmeasured confounders can interfere with this process. To handle this, we used a genetic IV defined by the 42 single nucleotide polymorphisms (SNPs) (listed in the Supporting Information) that were previously identified in genome-wide association studies of type-2 diabetes (Morris et al., 2012; Tchetgen Tchetgen et al., 2015).

4.1 | The Three-city study

The 3C study is a population-based prospective cohort that aimed at assessing the relation between vascular diseases and dementia in the elderly (Alperovitch, 2003). Participants, aged 65 years and older, were randomly selected in 1999 from the electoral lists of three French cities. In total, 9294 participants underwent an in-depth examination of their health and risk factors at baseline, and were then followed every 2–3 years for up to 20 years with an extensive interview and a neuropsychological battery. Among them, 6948 participants have been typed on genome-wide genotyping arrays and further imputed from Haplotype Reference Consortium panel (Lambert et al., 2009). Genotype data that were retained

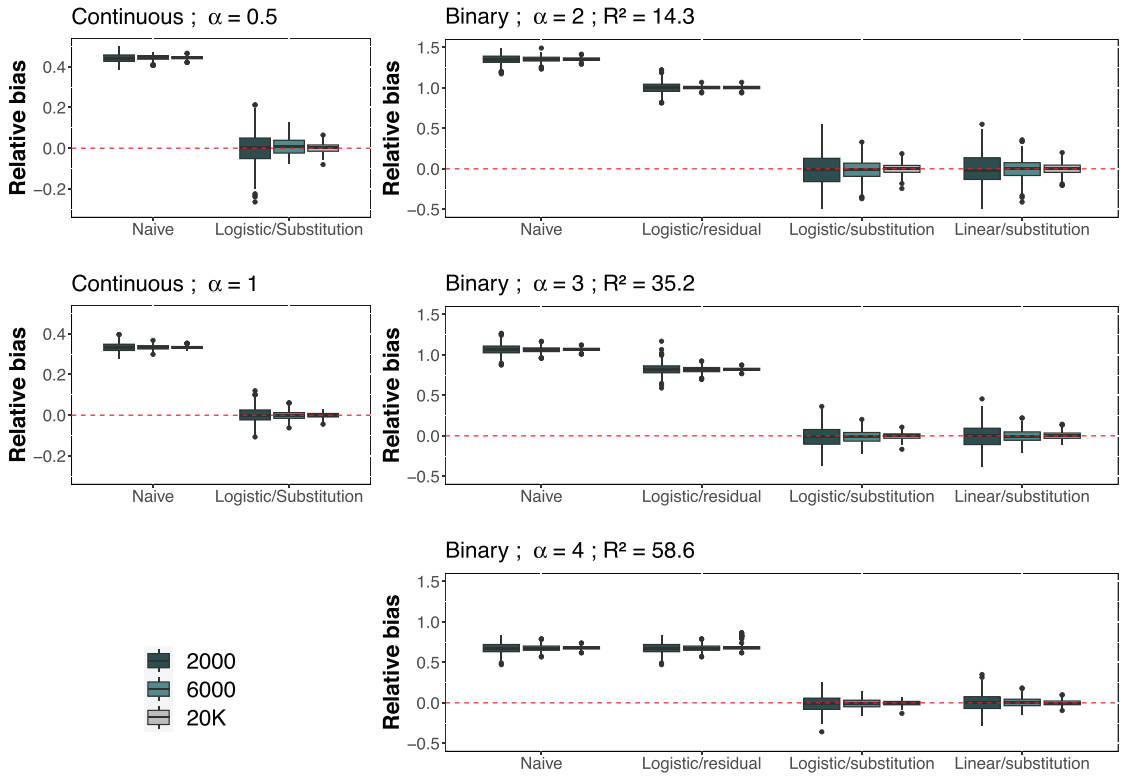


FIGURE 2 Association estimates (over 500 replicates) of the continuous exposure or the binary exposure with the change of the outcome over time using the naive or the IV approaches (logistic/residual, linear/substitution, and logistic/substitution in the binary case) for different sample sizes (N) and different intensities of association (through the regression coefficient α). In the binary case only, the Nagelkerke R^2 is also reported to further illustrate the strength of the IV in comparison with the application setting.

in the study are those with an imputation quality greater than 0.70. Type-2 diabetes were determined from blood glucose level (fasting glucose level≥ 7.0 mmol/L) or the use of antidiabetic treatment at baseline. We studied the cognitive trajectory through the Isaacs set test (IST), which measures verbal fluency and has been shown to differentiate early in the pathological process toward dementia (Amieva et al., 2014). The score is the total number of words given in four semantic categories in 15 s.

The final sample size included 6224 participants whose type-2 diabetes were ascertained at baseline, who were genotyped, and had at least one IST measure during the follow-up. Participants were 74 years old at baseline on average, 61 % were women, and 38% had an educational level higher than secondary school (Table 3). Among them, 598 (9.6 %) were ascertained with diabetes at baseline; those with diabetes were more often male, more likely to have a low educational level. Participants were followed up for 8 years on average with a mean of four repeated measures of IST.

4.2 | The IV analysis

We primarily used the logistic/substitution method. The R^2 of 4.8% showed a weak association between type-2 diabetes and genetic polymorphisms. The linear mixed model for the IST trajectory included a basis of four natural cubic splines on the time from baseline to account for the nonlinear trajectories over time. Diabetic status (in the naive model) or its expectation based on the 42 polymorphisms (in the IV model) was included in interaction with each spline function. For the naive model, we considered both no adjustment or adjustment on measured potential confounders (educational level, age at baseline). Parameter estimates are given in Table S2. Predicted trajectories of IST according to diabetic status are displayed in Figure 3(A) (corresponding differences over time between groups in Figure 3B).

The naive method, whether it was adjusted or not for potential confounders, highlighted a difference at inclusion according to the type-2 diabetes but no differential change over time. At any time, the mean IST score was lower for

TABLE 3 Characteristics of the 6224 participants of 3C sample according to their type-2 diabetes and overall.

Characteristics	Diabetics ($N = 598$)		No diabetics ($N = 5626$)		Overall ($N = 6224$)	
	Number (%)	Mean (SD)	Number (%)	Mean (SD)	Number (%)	Mean (SD)
Sex						
female	285 (47.7)		3498 (62.2)		3783 (60.8)	
male	313 (52.3)		2128 (37.8)		2441 (39.2)	
Education level						
no education	78 (13.0)		458 (8.1)		536 (8.6)	
primary school	112 (18.7)		924 (16.4)		1036 (16.7)	
secondary school	218 (36.5)		2086 (37.1)		2304 (37.0)	
high school	99 (16.6)		1138 (20.2)		1237 (19.9)	
university	91 (15.2)		1020 (18.1)		1111 (17.9)	
Age at entry		74.44 (5.4)		74.29 (5.5)		74.31 (5.5)
IST score at baseline		30.48 (6.8)		32.24 (7.0)		32.08 (7.0)
Number of IST		4.06 (1.8)		4.47 (1.9)		4.42 (1.9)
measures/subject						
Years of follow-up		7.08 (4.6)		8.12 (4.8)		8.02 (4.7)

IST, Isaacs set test; *N*, sample size; SD, standard deviation.

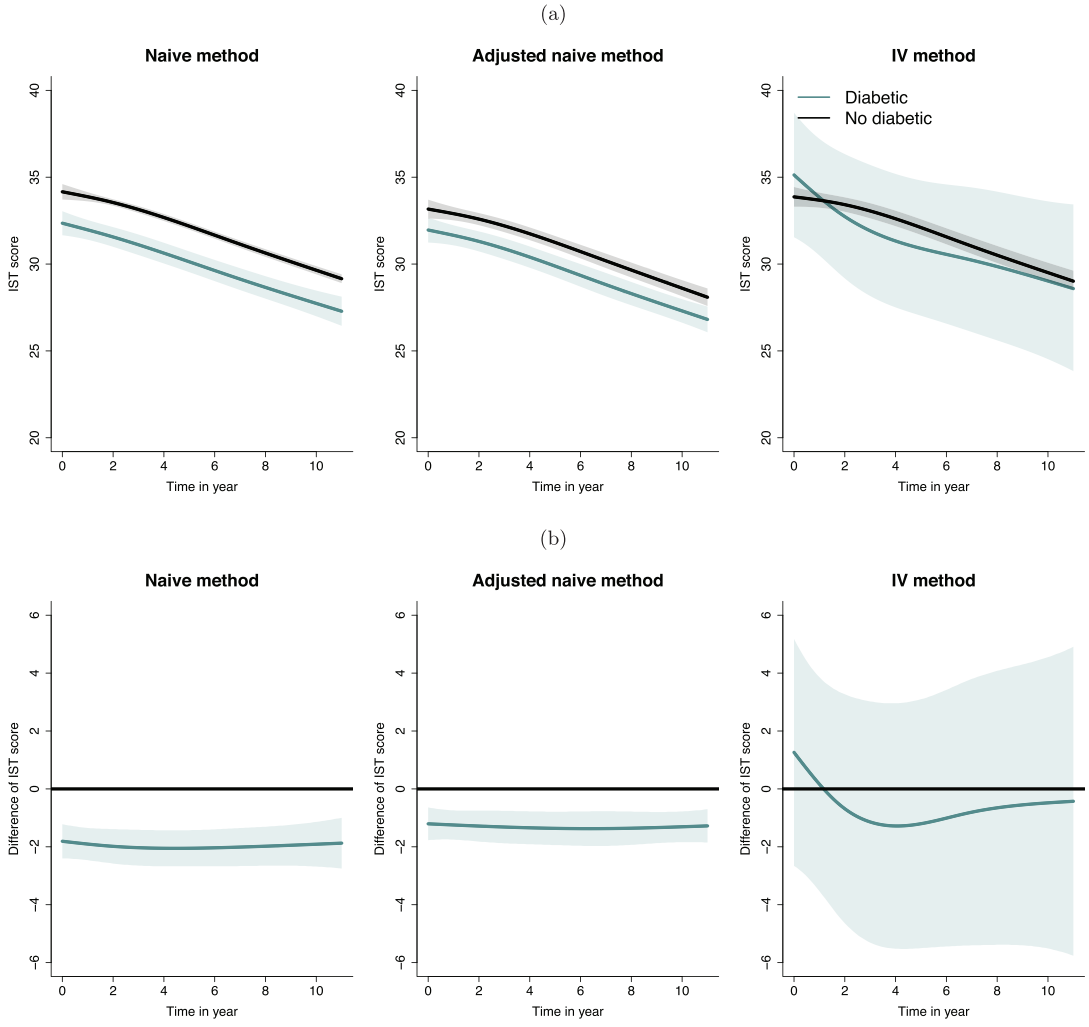


FIGURE 3 (A) Predicted trajectories of IST score according to type-2 diabetes at baseline and associated 95% confidence interval. (B) Estimated difference in IST score over time for diabetic compared to nondiabetic using the naive method (not adjusted or adjusted on gender, educational level, and age) and the logistic/substitution instrumental variable method.

participants with type-2 diabetes than for those without type-2 diabetes (mean difference in the adjusted model of −1.20 [−1.77;-0.64], −1.36 [−1.94;−0.79], −1.31 [−1.84;−0.78] points at 0, 5, and 10 years). In contrast, the logistic/substitution IV method did not show evidence of substantial difference in cognitive trajectory according to the type-2 diabetes although the point estimates suggested a higher level at baseline for participants with type-2 diabetes (mean difference of 1.26 [−2.66;5.18] points at baseline) and a steeper cognitive decline in the first years for participants with type-2 diabetes (mean difference of −1.20 [−5.50;3.10], −0.48 [−5.51;4.55] points at 5 and 10 years, respectively). Results were similar when using the linear/substitution IV model (see Figure S3).

5 | DISCUSSION

The IV method has gained interest in observational studies to address unmeasured confounding. Yet, although the framework is very common in observational longitudinal studies, an IV solution for the assessment of an exposure collected at baseline on the subsequent trajectory of a repeated outcome had not been previously described in the medical statistics literature. We showed in this work how the two-stage approach frequently used in IV methodology for cross-sectional or survival outcomes (Burgess et al., 2017; Tchetgen Tchetgen et al., 2015) could be adapted to study the association between a time-fixed exposure and the subsequent trajectory of an outcome using the mixed model theory. Previous contributions dealing with repeated data over time had systematically focused on time-dependent exposures (rather than time-fixed) and associations with either the level of a time-fixed outcome (Sánchez et al., 2017) or the level of a repeated outcome at a given time using distributed lag models (Hogan & Lancaster, 2004; O'Malley, 2012). To our knowledge, the use of a mixed model with an IV approach in epidemiology was limited to the analysis of a complex clinical trial to treat noncompliance over time (Bond et al., 2007), the issue of measurement error of time-dependent exposures with regression calibration (Strand et al., 2014), and the issue of between/within unmeasured confounding in cross-sectional grouped data (Li et al., 2015).

The conducted simulation study emphasized the highly biased estimations obtained when ignoring unmeasured confounding. They also showed the correct inference that our IV solution could provide for assessing the causal association between a time-fixed continuous or binary exposure and a continuous longitudinal outcome in the presence of endogeneity. However, we noticed a very high variance for moderate sample sizes (a few thousand subjects) when the IV was weakly associated with the exposure. For simplicity of result reporting, we focused in the methodology and in the simulations on scenarios with a linear trajectory for the outcome. However, the methodology applies equivalently to any scenario with a nonlinear trajectory, provided that the mixed model remains linear in the fixed and random effects, and random effects are included for each time function. This is what was done in the application considering natural splines to approximate the nonlinear cognitive trajectory.

The IV methodology highly relies on additive model properties to eliminate the association with the unmeasured confounders. The use of nonlinear models may prevent from a total elimination of this association and induce biased estimates. When considering a binary exposure, we explored linear and nonlinear regressions. Our simulations showed that the causal association could be correctly retrieved when using the linear probability model for the binary exposure but also when using the nonlinear logistic model combined with a substitution method in the second stage. In the application, both methods also gave the same results. In contrast, the logistic regression combined with the residual inclusion in the second stage (Terza et al., 2008) showed large bias in our simulation setting with a linear mixed model in the second stage and was not further investigated. Regarding the outcome, we restricted our framework to continuous longitudinal outcomes with linear mixed models and leave extensions to other types of outcomes to future research.

Our motivating application aimed at evaluating the causal association between type-2 diabetes and cognitive decline by using 42 genetic polymorphisms associated with type-2 diabetes as IV. While the classical (naive) regression ignoring unmeasured confounders highlighted a lower cognitive level for type-2 diabetics at all times, the IV methodology that handles unobserved confounding suggested a different and time-varying association. However, the analysis by IV does not allow to reach a conclusion as the confidence intervals were excessively large because of the limited sample size for an IV application with a binary exposure ($N = 6224$), and the weakness of the association between genetic polymorphisms and type-2 diabetes ($R^2 = 4.8\%$). These results were similar when considering logistic and linear models in first step.

MR studies had already been conducted to assess the causal association between type-2 diabetes and cerebral aging. Cross-sectional studies had focused on cognitive level (Ware et al., 2021) and dementia risk (Østergaard et al., 2015; Walter et al., 2016), and one longitudinal survival study had investigated the association with dementia risk (Tchetgen Tchetgen et al., 2015). None had identified a causal association between genetically predicted type-2 diabetes and cerebral aging.

Our work goes one step further by considering the association with prospective cognitive decline. Although in accordance with the literature, the highly variable results call for a replication in a much larger sample to overcome a potential lack of power. Additional simulations based on a similar instrument as in our application (Figure S2) showed the substantial gain in accuracy when considering, for instance, 20,000 subjects rather than 6000 subjects.

The method we proposed relies on assumptions coming from both the IV theory and the mixed model theory. First, the method is based on the fundamental assumptions that define valid instruments: (1) Z is strongly associated with X; (2) Z is associated with Y only through X; and (3) Z is independent of U conditionally on X (Figure 1). In our application as in many MR analyses, the genetic IV explains only a small part of the exposure (assumption (1)) leading to a weak instrument, high variances, and need for very large sample sizes. The simulation study did not reveal any issue of bias or coverage rate with weak instruments. However, it showed a huge variability that can make the IV method inconclusive, except when carried out on very large samples (20,000 subjects, e.g., in our case). To better address assumption (1) and not rely on a predetermined set of IVs, Fan and Zhong (2018) proposed an adaptive lasso technique that simultaneously selects the IV variables from a high-dimensional set of candidates. Developed for cross-sectional data, an extension to longitudinal outcome data using our mixed modeling strategy could be possible.

As fixed at birth, the genetic IV cannot be affected by the confounders (Assumption 3). However, to guarantee assumptions (2) and (3), we further need to assume that the SNPs associated with type-2 diabetes are not associated with other diseases (pleiotropy). Moreover, the use of genetic variants as an IV for a later in life study relies on the implicit assumption that the genetic variants are not associated with the probability to be alive at the timing of eligibility definition, exposure, and outcome collection (Swanson, 2019; Vansteelandt at al., 2018). Our application was performed under the assumption that genetic polymorphisms and type-2 diabetes were not associated with mortality prior to cohort entry. Finally, causal interpretation of the IV analysis requires a fourth assumption, either the homogeneity for the average causal effect or monotonicity for the local average causal effect (Hernán & Robins, 2006; Swanson & Hernán, 2018).

Note that with binary exposures, the interpretation of IV analyses may not be straightforward, especially when the binary exposure reflects an underlying continuous process that should be considered instead (Burgess & Labrecque, 2018). This is, however, unlikely the case with diabetes. In particular, its definition differs from blood glucose because a diabetic person under treatment may be controlled for hyperglycemia.

Our methodology also relies on classical assumptions of longitudinal analyses. We considered the linear mixed model theory rather than marginal models as they better handle selection over time for etiological studies (Rouanet et al., 2022). Our methodology is robust to missing data under the missing at random mechanism (i.e., missingness can be fully determined by the observations) (Little & Rubin, 1987) for both the intermittent missing outcome and study dropout. In case of informative dropout linked to the outcome process, the methodology can be easily extended by jointly modeling the risk of dropout according to the trajectory of the outcome (Rizopoulos, 2012). In the application, we performed such a sensitivity analysis where death and dropout from the study were modeled along with the cognitive decline (Table S3); it showed concordant results.

To conclude, we provided a full methodology and associated software solution to apply the IV technique to the frequent framework of an exposure measured at baseline and the subsequent trajectory of a continuous marker. It must be used with caution due to the strong and hardly controllable assumptions IV methods must satisfy. However, as illustrated with the causal association between type-2 diabetes and cognitive decline, it constitutes a useful statistical tool to take into account unobserved confounders in prospective cohort studies.

ACKNOWLEDGMENTS

Computer time was provided by the computing facilities MCIA (Mesocentre de Calcul Intensif Aquitain) at the University of Bordeaux and the University of Pau and Pays de l'Adour. This work was funded by the French National Research Agency (Project DyMES - ANR-18-CE36-0004-01) and was carried out in the framework of the INSERM GOLD Cross-Cutting program. D.-A.T. is supported, in part, by the EPIDEMIOM-VT Senior Chair from the University of Bordeaux initiative of excellence Initiative d'Excellence. This study was carried out in the framework of the University of Bordeaux's France 2030 program/RRI PHDS.

CONFLICTS OF INTEREST STATEMENT

The authors declare that they have no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are not available and are not publicly available due to privacy or ethical restrictions.

ORCID

Kateline Le Bourdonnec https://orcid.org/0000-0002-7732-5642

REFERENCES

Alperovitch, A. (2003). Vascular factors and risk of dementia: Design of the three-city study and baseline characteristics of the study population. *Neuroepidemiology*, *22*, 316–325.

Amieva, H., Mokri, H., Le Goff, M., Meillon, C., Jacqmin-Gadda, H., Foubert-Samier, A., Orgogozo, J.-M., Stern, Y., & Dartigues, J.-F. (2014). Compensatory mechanisms in higher-educated subjects with Alzheimer's disease: A study of 20 years of cognitive decline. *Brain: A Journal of Neurology*, *137*(Pt 4), 1167–1175.

Andrews, I., Stock, J. H., & Sun, L. (2019). Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics*, *11*(1), 727–753.

Bond, S. J., White, I. R., & Sarah Walker, A. (2007). Instrumental variables and interactions in the causal analysis of a complex clinical trial. *Statistics in Medicine*, *26*(7), 1473–1496.

Burgess, S., & Labrecque, J. A. (2018). Mendelian randomization with a binary exposure variable: Interpretation and presentation of causal estimates. arXiv:1804.05545 [stat].

Burgess, S., Small, D. S., & Thompson, S. G. (2017). A review of instrumental variable estimators for Mendelian randomization. *Statistical Methods in Medical Research*, *26*(5), 2333–2355.

Commenges, D., & Jacqmin-Gadda, H. (2015). *Modèles biostatistiques pour l'épidémiologie*. De Boeck Superieur. Google-Books-ID: twEtDwAAQBAJ.

Davies, N. M., Holmes, M. V., & Smith, G. D. (2018). Reading Mendelian randomisation studies: A guide, glossary, and checklist for clinicians. *BMJ (Clinical research ed.)*, *362*, k601.

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. No. 57 in Monographs on Statistics and Applied Probability. Chapman & Hall.

Ertefaie, A., Small, D. S., Flory, J. H., & Hennessy, S. (2017). A tutorial on the use of instrumental variables in pharmacoepidemiology. *Pharmacoepidemiology and Drug Safety*, *26*(4), 357–367.

Fan, Q., & Zhong, W. (2018). Nonparametric additive instrumental variable estimator: A group shrinkage estimation perspective. *Journal of Business & Economic Statistics*, *36*(3), 388–399.

Fewell, Z., Smith, G. D., & Sterne, J. A. C. (2007). The impact of residual and unmeasured confounding in epidemiologic studies: A simulation study. *American Journal of Epidemiology*, *166*(6), 646–655.

Frison. (2019). Diabèete et risque de démence. Thèse de doctorat, Spécialité Santé Publique, option épidémiologie Université de Bordeaux.

Greenland, S. (2000). An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology*, *29*(4), 722–729.

Hernán, M. A., & Robins, J. M. (2006). Instruments for causal inference: An epidemiologist's dream? *Epidemiology*, *17*(4), 360–372.

Hernan, M. A., & Robins, J. M. (2020). *Causal inference: What if*. Chapmann & Hall/CRC.

Hogan, J. W., & Lancaster, T. (2004). Instrumental variables and inverse probability weighting for causal inference from longitudinal observational studies. *Statistical Methods in Medical Research*, *13*(1), 17–48.

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, *38*(4), 963–974.

Lambert, J. C., Heath, S., Even, G., Campion, D., Sleegers, K., Hiltunen, M., Combarros, O., Zelenika, D., Bullido, M. J., Tavernier, B., Letenneur, L., Bettens, K., Berr, C., Pasquier, F., Fiévet, N., Barberger-Gateau, P., Engelborghs, S., De Deyn, P., Mateo, I., & Amouyel, P. (2009). Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. *Nature Genetics*, *41*(10), 1094–1099.

Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13–22.

Li, C., Poskitt, D. S., Windmeijer, F., & Zhao, X. (2022). Binary outcomes, OLS, 2SLS and IV probit. *Econometric Reviews*, *41*(8), 859–876.

Li, J., Fine, J., & Brookhart, A. (2015). Instrumental variable additive hazards models. *Biometrics*, *71*(1), 122–130.

Li, Y., Lee, Y., Port, F. K., & Robinson, B. M. (2020). The impact of unmeasured within- and between-cluster confounding on the bias of effect estimators of a continuous exposure. *Statistical Methods in Medical Research*, *29*, 2119.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. Wiley.

Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data*. John Wiley & Sons. Google-Books-ID: BemMDwAAQBAJ.

Martínez-Camblor, P., MacKenzie, T. A., Staiger, D. O., Goodney, P. P., & James O'Malley, A. (2019). An instrumental variable procedure for estimating Cox models with non-proportional hazards in the presence of unmeasured confounding. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *68*(4), 985–1005.

Morris, A. P., Voight, B. F., Teslovich, T. M., Ferreira, T., Segrè, A. V., Steinthorsdottir, V., Strawbridge, R. J., Khan, H., Grallert, H., Mahajan, A., Prokopenko, I., Kang, H. M., Dina, C., Esko, T., Fraser, R. M., Kanoni, S., Kumar, A., Lagou, V., Langenberg, C., & McCarthy, M. I. (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genetics*, *44*(9), 981–990.

O'Malley, A. J. (2012). Instrumental variable specifications and assumptions for longitudinal analysis of mental health cost offsets. *Health Services and Outcomes Research Methodology*, *12*(4), 254–272.

Østergaard, S. D., Mukherjee, S., Sharp, S. J., Proitsi, P., Lotta, L. A., Day, F., Perry, J. R. B., Boehme, K. L., Walter, S., Kauwe, J. S., Gibbons, L. E., Consortium, A. D. G., Larson, E. B., Powell, J. F., Langenberg, C., Crane, P. K., & Scott, R. A., Consortium, GERAD1, Consortium, EPIC-InterAct. (2015). Associations between potentially modifiable risk factors and Alzheimer disease: A Mendelian randomization study. *PLoS Medicine*, *12*(6), e1001841.

Proust-Lima, C., Philipps, V., & Liquet, B. (2017). Estimation of extended mixed models using latent classes and latent processes: The R Package lcmm. *Journal of Statistical Software*, *78*, 1–56.

Rawlings, A. M., Sharrett, A. R., Schneider, A. L., Coresh, J., Albert, M., Couper, D., Griswold, M., Gottesman, R. F., Wagenknecht, L. E., Windham, B. G., & Selvin, E. (2014). Diabetes in midlife and cognitive change over 20 years: The Atherosclerosis Risk in Communities Neurocognitive Study. *Annals of Internal Medicine*, *161*(11), 785–793.

Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data* (1st ed.). Routledge.

Rouanet, A., Avila-Rieger, J., Dugravot, A., Lespinasse, J., Stuckwisch, R., Merrick, R., Anderson, E., Long, L., Helmer, C., Jacqmin-Gadda, H., Dufouil, C., Judd, S., Manly, J., Sabia, S., Gross, A., & Proust-Lima, C. (2022). How selection over time contributes to the inconsistency of the association between sex/gender and cognitive decline across cognitive aging cohorts. *American Journal of Epidemiology*, *191*(3), 441–452.

Royall, R. M. (1986). The prediction approach to robust variance estimation in two-stage cluster sampling. *Journal of the American Statistical Association*, *81*(393), 119–123.

Sánchez, B. N., Kim, S., & Sammel, M. D. (2017). Estimators for longitudinal latent exposure models: Examining measurement model assumptions. *Statistics in Medicine*, *36*(13), 2048–2066.

Sjolander, A., & Martinussen, T. (2019). Instrumental variable estimation with the R package ivtools. *Epidemiologic Methods*, *8*(1), 20180024.

Strand, M., Sillau, S., Grunwald, G. K., & Rabinovitch, N. (2014). Regression calibration for models with two predictor variables measured with error and their interaction, using instrumental variables and longitudinal data. *Statistics in Medicine*, *33*(3), 470–487.

Swanson, S. A., & Hernán, M. A. (2018). The challenging interpretation of instrumental variable estimates under monotonicity. *International Journal of Epidemiology*, *47*(4), 1289–1297.

Swanson, S. A. (2019). A practical guide to selection bias in instrumental variable analyses. *Epidemiology*, *30*(3), 345–349.

Tchetgen Tchetgen, E. J., Walter, S., Vansteelandt, S., Martinussen, T., & Glymour, M. (2015). Instrumental variable estimation in a survival context. *Epidemiology (Cambridge, Mass.)*, *26*(3), 402–410.

Terza, J. V., Basu, A., & Rathouz, P. J. (2008). Two-stage residual inclusion estimation: Addressing endogeneity in health econometric modeling. *Journal of Health Economics*, *27*(3), 531–543.

Wagner, M., Dartigues, J. F., Samieri, C., & Proust-Lima, C. (2018). Modeling risk-factor trajectories when measurement tools change sequentially during follow-up in cohort studies: Application to dietary habits in prodromal dementia. *American Journal of Epidemiology*, *187*(4), 845–854.

Walter, S., Marden, J. R., Kubzansky, L. D., Kubzansky, L. D., Mayeda, E. R., Crane, P. K., Chang, S.-C., Cornelis, M., Rehkopf, D. H., Mukherjee, S., & Glymour, M. M. (2016). Diabetic phenotypes and late-life dementia risk: A mechanism-specific Mendelian randomization study. *Alzheimer Disease & Associated Disorders*, *30*(1), 15–20.

Ware, E. B., Morataya, C., Fu, M., & Bakulski, K. M. (2021). Type 2 diabetes and cognitive status in the health and retirement study: A Mendelian randomization approach. *Frontiers in Genetics*, *12*, 634767.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*(4), 817–838.

Wright, P. G. (1928). *The tariff on animal and vegetable oils*. Macmillan.

Vansteelandt, S., Dukes, O., & Martinussen, T. (2018). Survivor bias in Mendelian randomization analysis. *Biostatistics (Oxford, England)*, *19*(4), 426–443.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Le Bourdonnec, K., Samieri, C., Tzourio, C., Mura, T., Mishra, A., Trégouët, D.-A., & Proust-Lima, C. (2024). Addressing unmeasured confounders in cohort studies: Instrumental variable method for a time-fixed exposure on an outcome trajectory. *Biometrical Journal*, *66,* 2200358.
https://doi.org/10.1002/bimj.202200358

Supplementary materials

List of the 42 polymorphisms used as instrumental variables:
rs10203174, rs243088, rs1801282, rs1801214, rs459193, rs849135, rs516946, rs3802177, rs12571751, rs1111875, rs163184, rs10830963, rs11063069, rs10842994, rs7955901, rs7177055, rs9936385, rs7202877, rs12970134, rs8108269, rs2075423, rs3923113, rs2943640, rs1496653, rs6795735, rs11708067, rs4402960, rs6769511, rs6878122, rs7756992, rs17168486, rs10965250, rs2796441, rs7903146, rs5215, rs1552224, rs2261181, rs12427353, rs1359790, rs12899811, rs11651052, rs10401969

Table S1: Summary of the procedure of generation in the simulation study including the data generation model and the parameters considered

Sample size (N)	N = 2000; 6000; 20000	
Instrumental variable (Z)	$Z_i \sim N(0,1)$	
Unobserved confounder (U)	$U_i \sim \mathcal{N}(0,1)$	
Time (t)	$t_{ij} \sim \mathcal{N}(j, 0.05)$ with $j = 1, ..., 6$	
Random effects (b_i)	$b_i = (b_{0i}, b_{1i})^\top \sim \mathcal{N}\left(\left(\begin{smallmatrix}0\\0\end{smallmatrix}\right), \left(\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right)\right)$	
Measurement errors (ϵ)	$\epsilon_{ij} \sim \mathcal{N}(0,1)$ with $j = 1, ..., 6$	
Exposure (X)	Generation model	Generated parameters
	Continuous	
	$X = \alpha_{int} + \alpha_Z Z + \alpha_U U + \epsilon_X$ $\epsilon_X \sim \mathcal{N}(0,1)$	with $\alpha_{int} = 10, \alpha_Z = 0.5$ or 1, $\alpha_U = 1$,
	Binary	
	$X \sim$ Bernoulli(p) with $p = \text{expit}(\alpha_{int} + \alpha_Z Z + \alpha_U U)$	with $\alpha_{int} = 3, \alpha_Z = 0.5, \alpha_U = 1$ or $\alpha_{int} = -2, \alpha_Z = 2, \alpha_U = 4$ or $\alpha_{int} = -2, \alpha_Z = 3, \alpha_U = 3$ or $\alpha_{int} = -2, \alpha_Z = 4, \alpha_U = 2$
Outcome (Y)	Generation model	Generated parameters
	$Y_{ij} = \beta_{int} + \beta_e X_i + \beta_U U_i + \beta_t + b_{0i} + t_{ij}(\beta_{te} X_i + \beta_{tU} U + b_{1i}) + \epsilon_{ij}$	with $\beta_{int} = 5, \beta_e = 1, \beta_U = 1, \beta_t = 1, \beta_{te} = 1, \beta_{tU} = 1$

Table S2: Estimates of the association between type-2 diabetes and cognitive trajectory (approximated by natural cubic splines) using the naive method and the logistic/substitution IV method

	Naive method			**Adjusted naive method***			**Logistic/Substitution IV method**		
	Coef	SE	$CI_{95\%}$	Coef	SE	$CI_{95\%}$	Coef	SE**	$CI_{95\%}$
Diabete	-1.81	0.30	[-2.40 ; -1.22]	-1.21	0.29	[-1.77; -0.64]	1.26	2.00	[-2.66 ; 5.18]
Spline1	-2.46	0.37	[-3.19 ; -1.73]	-2.46	0.85	[-4.12 ; -0.79]	-4.58	0.25	[-3.01 ; -2.04]
Spline2	-4.62	0.71	[-6.02 ; -3.22]	-7.76	1.00	[-9.72 ; -5.80]	-4.56	0.57	[-5.69 ; -3.47]
Spline3	-4.65	0.40	[-5.43 ; -3.87]	-4.72	0.59	[-5.88 ; -3.55]	-2.52	0.22	[-5.01 ; -4.12]
Diabete x Spline1	-0.16	0.32	[-0.78 ; 0.47]	-0.18	0.35	[-0.88 ; 0.51]	-1.09	2.10	[-5.21 ; 3.04]
Diabete x Spline2	-0.35	0.49	[-1.32 ; 0.62]	-0.21	0.53	[-1.25 ; 0.83]	-4.41	3.41	[-11.10 ; 2.28]
Diabete x Spline3	0.10	0.36	[-0.60 ; 0.79]	-0.01	0.09	[-0.17 ; 0.16]	0.13	2.12	[-4.03 ; 4.29]

All method are adjusted on primo-passation

*: with adjustment on age at baseline, sex and education level

**: Standard error after variance correction

Table S3: Estimation the association between type-2 diabetes and cognitive trajectory (approximated by natural cubic splines) using the Logistic/substitution IV method with either a linear mixed model (LMM) or a joint model (JM) to account for informative dropout in the second stage

		LMM			**JM**		
		Coef	SE*	$CI_{95\%}$	Coef	SE	$CI_{95\%}$
Longitudinal Process	Diabete	1.26	2.00	[-2.66 ; 5.18]	1.31	1.95	[-2.51 ; 5.13]
	Spline1	-4.58	0.25	[-3.01 ; - 2.04]	-2.65	0.25	[-3.15 ; -2.15]
	Spline2	-4.56	0.57	[-5.69 ; -3.47]	-5.13	0.56	[-6.23 ; -4.03]
	Spline3	-2.52	0.22	[-5.01 ; -4.12]	-4.98	0.23	[-5.43 ; -4.52]
	Diabete x Spline1	-1.09	2.10	[-5.21 ; 3.04]	-1.06	2.23	[-5.43 ; 3.31]
	Diabete x Spline2	-4.41	3.41	[-11.10 ; 2.28]	-4.64	3.19	[-10.89 ; 1.61]
	Diabete x Spline3	0.13	2.12	[-4.03 ; 4.29]	0.20	2.15	[-4.03 ; 4.42]
Survival Process	Diabete				0.16	0.34	[-0.51 ; 0.83]
	ISA association				-0.07	0.01	[-0.07 ; -0.06]

All method are adjusted on primo-passation

*: Standard error after variance correction

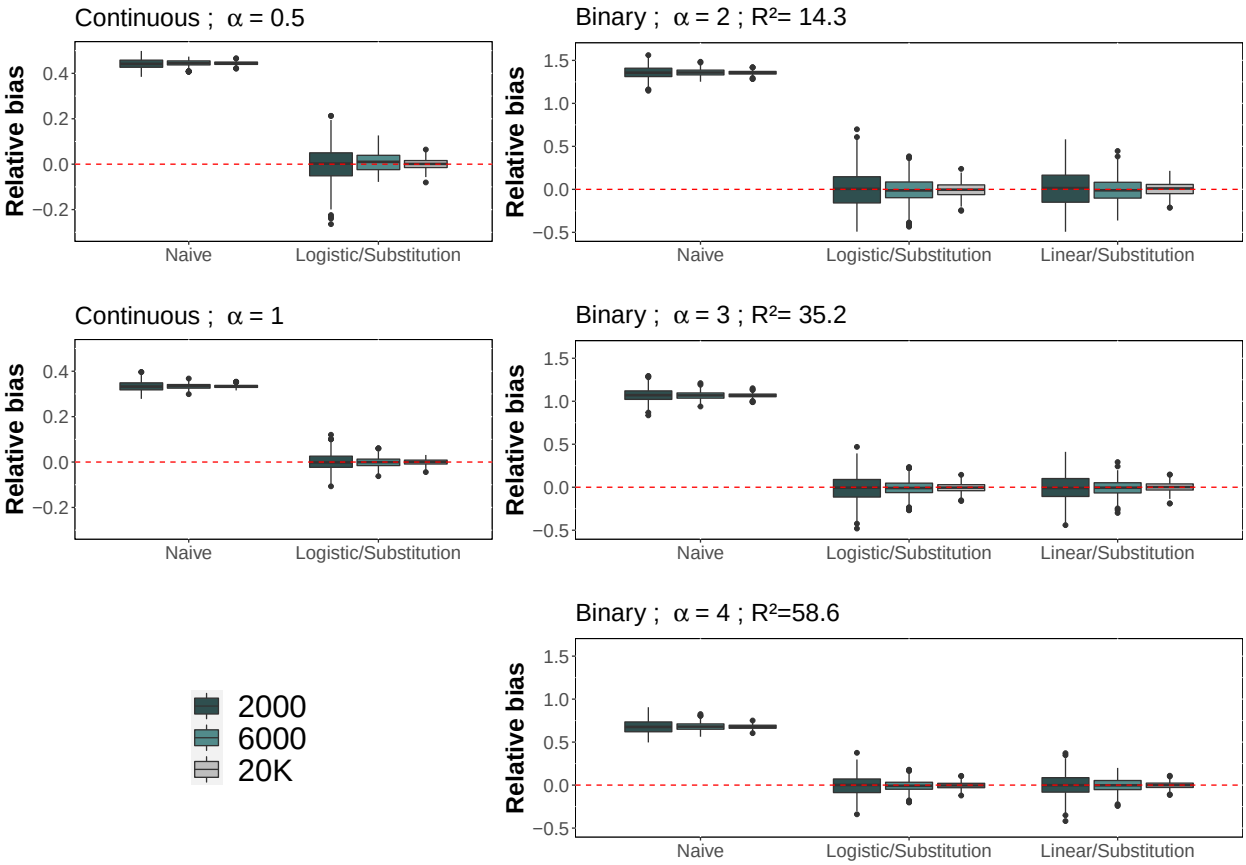


Figure S1: Association estimates (over 500 replicates) of the continuous exposure and binary exposure with the outcome at baseline in the naive approach and in the IV approach for different sample sizes (N) and strengths of the association (i.e., different regression coefficient α). In the binary case only, the Nagelkerke R^2 is also reported to further illustrate the strength of the IV in comparison with the application setting.

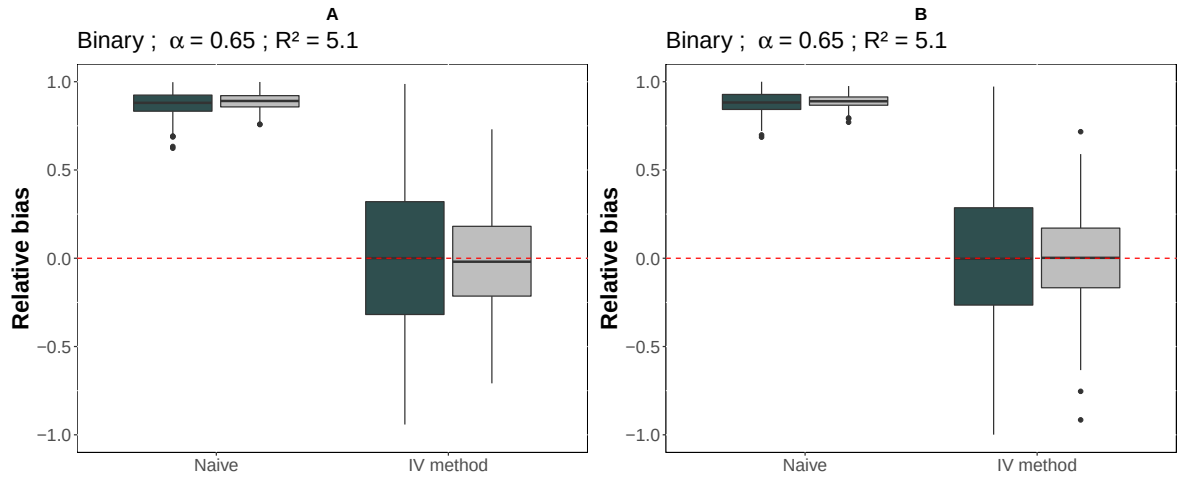


Figure S2: Simulation results (over 500 replicates) for the association between the binary exposure and the outcome at baseline (A) or its change over time (B) for two sample sizes N=6000 and N=20000 and for an intensity of association α_Z=0.65 close to the one observed in the application.

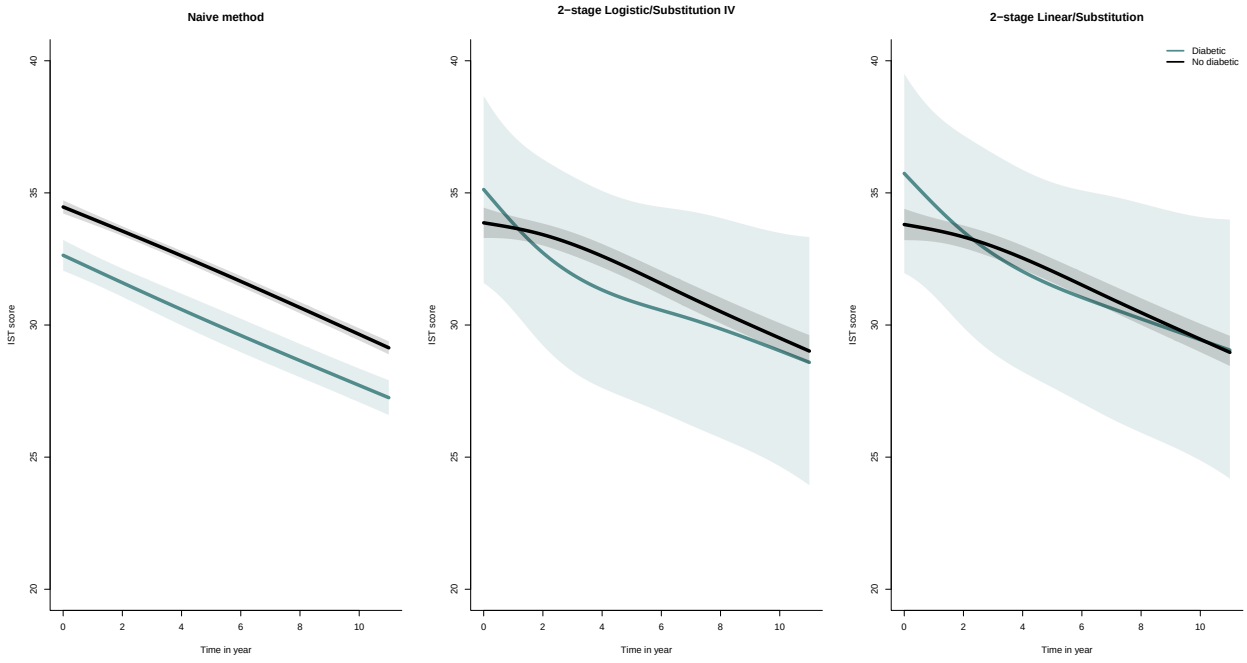


Figure S3: Predicted trajectories (and associated 95% confidence interval) of IST score according to type-2 diabetes at baseline using different methods: Naive model, 2-stage Logistic/Substitution IV method and 2-stage Linear/Substitution IV method

Chapitre 6

Discussion et perspectives

Nous avons étendu dans cette thèse des approches statistiques permettant de faire de l'inférence causale à partir de données de cohortes observationnelles, en tenant compte de leur nature longitudinale. Chacune des approches proposées était motivée par une application épidémiologique dans l'étude du vieillissement cognitif.

Dans ce chapitre, après avoir résumé les approches que nous avons proposées, nous discuterons de leurs avantages et limites, ainsi que des perspectives futures, avant de conclure.

6.1 Résumé des travaux réalisés

Dans cette thèse, deux approches majeures de causalité ont été explorées pour étudier des données longitudinales issues de cohortes observationnelles : l'analyse de médiation et la méthode des variables instrumentales. Ces travaux étaient motivés par la volonté d'étudier des facteurs de risque de vieillissement cérébral et de mieux comprendre les mécanismes sous-jacents. Ils font suite à la constatation de la complexité d'interpréter de façon causale les associations identifiées dans les cohortes observationnelles, et le manque de méthodologies statistiques adaptées aux données répétées dans la littérature.

Tout d'abord, nous avons pu proposer (c.f. section 4.1) une approche d'analyse de

médiation pour traiter des données longitudinales mesurées en temps continu. Plus précisément, la méthodologie proposée permet de décortiquer le mécanisme sous-jacent de la relation entre une variable d'exposition fixe dans le temps et d'un processus d'intérêt, en présence de processus intermédiaires de médiation et de confusion.

Nous avons également étendu l'analyse des variables instrumentales, permettant de pallier le biais lié aux facteurs de confusion non mesurés pour une variable d'exposition indépendante du temps et un outcome longitudinal. Ces nouvelles approches ont été validées dans le cadre d'une étude de simulation pour vérifier leurs performances.

Nous avons également approfondi (c.f. section 4.2) une méthodologie d'analyse de médiation similaire à Valeri et al. (2023) pour analyser l'impact d'un facteur de risque indépendant du temps sur un événement terminal, en présence d'un événement intermédiaire, tous deux sujets à de la censure à droite.

6.2 Avantages des méthodes proposées

Bien que de nombreux modèles existent actuellement pour étudier les données longitudinales (e.g. modèles mixtes multivariés) ou les données de survie (e.g. modèles multi-états), leurs interprétations se limitent bien souvent à des associations. Or, pour pouvoir comprendre l'étiologie d'une maladie et les mécanismes sous-tendant sa progression, il est nécessaire d'en connaître les causes. Ainsi, les méthodes proposées dans cette thèse permettent de répondre à un besoin réel pour l'analyse causale de données longitudinales ou de type temps d'événement. Ces travaux sont motivés par l'étude du vieillissement cérébral, néanmoins ils sont d'une utilité beaucoup plus large et peuvent être appliqués à de nombreuses autres études de cohortes, que ce soit pour étudier une maladie spécifique ou s'intéresser à d'autres processus de santé.

Même si ces deux méthodes ont pour objectif commun d'interpréter la relation causale entre un facteur de risque et une maladie, elles ne jouent pas le même rôle. Tandis que les analyses de médiation s'attachent à comprendre les mécanismes sous-jacents de la

relation entre le facteur de risque et la maladie (en utilisant des variables intermédiaires), la méthode des variables instrumentales vise à corriger un biais causé par l'omission de facteurs de confusion dans la relation entre le facteur de risque et la maladie.

De plus, les méthodes proposées permettent de traiter des complexités méthodologiques fréquentes dans les études de cohorte, telles que des phénomènes mesurés en temps continu avec l'approche d'analyse de médiation proposée en section 4.1, l'existence de facteurs de confusion non mesurés qui sont abordés avec l'analyse de variables instrumentales, ou la problématique de la censure par intervalle dans l'approche d'analyse de médiation proposée en section 4.2.

6.3 Limites des méthodes proposées

Outres les limites très spécifiques à chaque approche, évoquées dans les articles, les méthodes que nous avons proposé sont sujettes à différentes limites.

Premièrement, les trois approches proposées dans cette thèse reposent toutes sur une variable d'exposition indépendante du temps. Pourtant, il est fréquent que les facteurs de risque soient des variables variant dans le temps comme cela a été illustré dans l'introduction avec les facteurs de risque de démence modifiables (Figure 1.1). Pour l'approche de médiation longitudinale proposée en section 4.1, nous avons considéré deux variables réellement indépendantes du temps dans une population âgée avec le facteur génétique de l'APOE4 et le niveau d'étude. Cependant, il n'en est pas de même avec les deux autres travaux qui portaient sur le diabète ou la santé cardiovasculaire. Ces variables peuvent évoluer dans le temps et nous avons uniquement considéré le statut à l'inclusion dans la cohorte.

D'autre part, nous avons démontré dans le chapitre 5 que l'interprétation des modèles mixtes est biaisée lorsqu'on cherche à l'analyser de manière causale l'effet d'une exposition en présence de facteurs de confusion non mesurés. Les méthodes d'analyse de médiation décrites dans ce travail ne traitent pas cette confusion non-observée. Elles offrent une in-

terprétation causale uniquement en absence de tout facteur de confusion non mesuré. Or, cette hypothèse est difficile à vérifier en pratique, voire impossible à vérifier, et pourrait entraîner des biais dans les applications aux données de la cohorte 3C.

De plus, nous avons travaillé sur des données de cohorte très riches avec un suivi standardisé. Cela a offert des applications pertinentes. Cependant, le projet autour de l'analyse de médiation pour données répétées avait initialement pour perspective d'explorer le rôle de l'imagerie dans le vieillissement cérébral. Or nous ne disposions que de deux à trois mesures pour ces marqueurs. Nous avons découvert que les méthodes de médiation en temps continu ne pouvaient pas appréhender de façon correcte des questions de médiation avec aussi peu de données répétées. Nous avons dû réorienté notre application vers des processus mesurés de façon plus fréquente. Cette limite est un vrai frein pour l'étude de mécanismes causaux dans les études observationnelles. En effet, du fait du coût, et de la pénibilité pour le participant, il n'est pas envisageable de collecter plus de mesures d'imagerie. Cela empêche l'investigation des chemins sous-jacents aux manifestations cliniques du vieillissement cérébral.

Par ailleurs, la censure par intervalle était un des objectifs de la thèse. Malheureusement, ce travail n'a pas pu être abouti du fait de l'absence de modèles multi-états traitant à la fois la censure par intervalle et permettant d'inclure le temps d'événement intermédiaire dans le modèle de l'événement final.

Enfin, les applications dans ce travail sont à visée d'illustration des méthodes statistiques. Certaines applications pourraient être plus poussées avec des analyses de complémentaires et de sensibilité. Notamment pour le score de santé cardiovasculaire, où nous n'avons pas exploré des associations non linéaires entre le score et les risques de démence et de décès.

6.4 Perspectives

Plusieurs perspectives sont à envisager dans nos travaux.
À court terme, un package R implémentant les méthodes d'analyse de médiation sera proposé, avec une mise à disposition sur un site dédié afin de rendre les méthodes développées plus accessibles. De plus, le travail sur la médiation pour données censurées par intervalles sera finalisé en implémentant le modèle multi-états nécessaire à son estimation.

Dans une perspective à plus long terme, il serait intéressant d'étudier le comportement des méthodologies proposées en analyse de médiation en utilisant différents modèles de travail. Par exemple, pour l'analyse de médiation avec des données longitudinales, nous avons choisi d'utiliser le modèle dynamique proposé par Taddé et al. (2020). Nous pourrions envisager la même méthodologie en utilisant d'autres approches de modélisation. Comme vu dans l'article, la contrainte est de pouvoir prédire des distributions conditionnelles. Nous pourrions ainsi envisager l'approche par modèle mixte multivarié considérant l'influence d'un marqueur sur l'autre (présentée en section 3.1.2.2), ou bien utiliser des techniques issues de l'apprentissage automatique. La difficulté pour cela est d'avoir à disposition des outils adaptés aux données répétées bruitées et irrégulièrement mesurées. Des études de simulation pourraient ensuite être réalisées pour étudier l'impact du choix du modèle de travail sur les résultats obtenus.
De la même manière, la méthodologie d'analyse de médiation pour des temps d'événements s'appuie actuellement sur un modèle multi-états faisant intervenir des modèles à risques proportionnels de Cox. Cette hypothèse, supposant un effet constant au cours du temps, est idéale pour l'interprétation des résultats. Néanmoins, il s'agit d'une hypothèse forte et pas toujours vérifiée (Kragh Andersen et al. (2021)). Il pourrait être intéressant d'étudier des modèles de travail à risques non proportionnels, soit en travaillant toujours à partir de modèles multiplicatifs, soit en abordant des modèles à risques additifs. De nouveau, des études de simulation pourraient être réalisées afin de comparer la méthodologie en présence de différents modèles de travail.

Comme évoqué en limite de nos travaux, les méthodologies proposées se focalisent uniquement sur des variables d'exposition mesurées à un temps fixe. Prendre en compte l'aspect temporel de ces variables permettrait de rendre les méthodes proposées plus flexibles (c'est-à-dire adaptables à un plus grand nombre d'études). Néanmoins, obtenir cette flexibilité n'est pas évident au vu des hypothèses d'identifiabilité des approches actuelles.

Une extension intéressante des approches d'analyse de médiation que nous avons proposées serait d'étudier un médiateur répété dans le temps sur un événement d'intérêt et inversement (i.e. médiateur de type survie - outcome répété). Cela permettrait par exemple de pouvoir étudier le rôle de l'entrée en institution dans le déclin cognitif chez la personne âgée, ou bien de traiter le rôle des fonctions cognitives dans le risque de démence.

De plus, l'analyse de médiation et l'approche par variables instrumentales sont deux méthodes qui permettent de traiter la causalité. Cependant, tandis que dans l'approche par variables instrumentales, nous pallions un possible biais dû à des facteurs de confusion non mesurés, il n'en est pas de même pour l'analyse de médiation qui suppose simplement l'absence de ces facteurs de confusion entre l'exposition et l'outcome, entre l'exposition et le médiateur, et entre le médiateur et l'outcome. Or, comme évoqué en introduction de cette thèse, dans les études de cohorte, en particulier lorsque l'on étudie le vieillissement, l'omission de facteurs de confusion est inévitable. C'est pourquoi, ces dernières années, différents auteurs se sont intéressé à combiner l'approche par variables instrumentales et l'approche d'analyse de médiation dans un cadre de données fixes dans le temps (Chen et al. (2023), Rudolph et al. (2024)). Ainsi, il serait intéressant d'étendre ces méthodes à un cadre plus général permettant de pallier le biais des facteurs de confusion non mesurés dans les analyses de médiation en présence de données longitudinales et/ou de type survie.

Enfin, ma thèse visant à comprendre le vieillissement cognitif et ayant montré en introduction qu'il s'agit d'un processus multidimensionnel, lui-même composé de processus multidimensionnels, une extension possible des travaux effectués serait de prendre en

considération des médiateurs et/ou des outcomes multidimensionnels. Premièrement, il s'agirait de définir des quantités causales adaptées ainsi que les hypothèses d'identifiabilité associées. Ensuite, à l'aide de méthodes adaptées en grande dimension telles que les forêts aléatoires de survie avec des marqueurs longitudinaux (Devaux et al. (2022)), il serait possible d'explorer ces aspects multidimensionnels pour une meilleure compréhension du vieillissement cognitif.

6.5 Conclusion générale

Dans cette thèse, nous nous sommes intéressés à des méthodes d'inférence causale pour des données observationnelles issues d'étude de cohorte, à savoir des données répétées au cours du temps ou bien encore des temps d'événement.

Cette thèse a un fort impact tant d'un point de vue statistique qu'épidémiologique. En effet, de par le développement de méthodes pour l'inférence causale dans les études longitudinales, les travaux effectués durant cette thèse permettent de pallier un manque actuel de méthodologies pour traiter la causalité dans les cohortes observationnelles prospectives alors que ces cohortes sont extrêmement courantes en épidémiologie.
D'un point de vue épidémiologique, cette thèse présente de fortes retombées. Le vieillissement est actuellement une préoccupation majeure en Santé Publique, et l'identification de facteurs causaux sur lesquels il est possible d'agir est un élément essentiel pour la prévention. Les méthodologies appliquées aux données très riches de plusieurs cohortes permettent ainsi de fournir un nouveau regard sur les relations causales entre des facteurs environnementaux et plusieurs composantes du vieillissement, notamment le vieillissement cérébral. Mais elles apporteront aussi des éléments clés pour la compréhension des mécanismes causaux sous-jacents au vieillissement, général ou plus spécifique comme le vieillissement cérébral.

Bibliographie

O. O. Aalen. A linear regression model for the analysis of life times. *Stat Med*, 8(8) : 907–925, Aug. 1989. ISSN 0277-6715. doi : 10.1002/sim.4780080803.

O. O. Aalen and A. Frigessi. What can statistics contribute to a causal understanding ? *Scandinavian Journal of Statistics*, 34(1) :155–168, 2007. ISSN 1467-9469. doi : 10.1111/j.1467-9469.2006.00549.x. Place : United Kingdom Publisher : Blackwell Publishing.

O. O. Aalen, M. J. Stensrud, V. Didelez, R. Daniel, K. Røysland, and S. Strohmaier. Time-dependent mediators in survival analysis : Modeling direct and indirect effects with the additive hazards model. *Biom J*, 62(3) :532–549, May 2020. ISSN 1521-4036. doi : 10.1002/bimj.201800263.

Y. Agid. Vieillissement cérébral ou maladie dégénérative. In *Chimie et cerveau*, pages 75–86. EDP Sciences, 2016. ISBN 978-2-7598-1790-0. URL https://www.cairn-sciences.info/chimie-et-cerveau--9782759817900-page-75.htm.

H. Agüero-Torres, L. Fratiglioni, Z. Guo, M. Viitanen, and B. Winblad. Mortality from Dementia in Advanced Age : A 5-Year Follow-Up Study of Incident Dementia Cases. *Journal of Clinical Epidemiology*, 52(8) :737–743, Aug. 1999. ISSN 0895-4356. doi : 10.1016/S0895-4356(99)00067-0. URL https://www.sciencedirect.com/science/article/pii/S0895435699000670.

J. M. Albert, Y. Li, J. Sun, W. A. Woyczynski, and S. Nelson. Continuous-time causal mediation analysis. *Statistics in Medicine*, 38(22) :4334–4347, 2019. ISSN 1097-0258. doi : 10.1002/sim.8300. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.8300.

A. Alpérovitch, P. Amouyel, J.-F. Dartigues, P. Ducimetière, B. Mazoyer, K. Ritchie, and C. Tzourio. [Epidemiological studies on aging in France : from the PAQUID study to the Three-City study]. *C R Biol*, 325(6) :665–672, June 2002. ISSN 1631-0691. doi : 10.1016/s1631-0691(02)01476-2.

L. Angel and M. Isingrini. Le vieillissement neurocognitif : entre pertes et compensation. *L'Année psychologique*, 115(2) :289–324, 2015. ISSN 0003-5033. doi : 10.3917/anpsy.152.0289. URL https://www.cairn.info/revue-l-annee-psychologique1-2015-2-page-289.htm. Place : Paris Publisher : NecPlus.

J. D. Angrist, G. W. Imbens, and D. B. Rubin. Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association*, 91(434) :444–455, 1996. ISSN 0162-1459. doi : 10.2307/2291629. URL https://www.jstor.org/stable/2291629. Publisher : [American Statistical Association, Taylor & Francis, Ltd.].

N. Awad, M. Gagnon, and C. Messier. The Relationship between Impaired Glucose Tolerance, Type 2 Diabetes, and Cognitive Function. *Journal of Clinical and Experimental Neuropsychology*, 26(8) :1044–1080, Nov. 2004. ISSN 1380-3395. doi : 10.1080/13803390490514875. URL https://doi.org/10.1080/13803390490514875. Publisher : Routledge _eprint : https ://doi.org/10.1080/13803390490514875.

M. Baiocchi, J. Cheng, and D. S. Small. Instrumental variable methods for causal inference. *Stat Med*, 33(13) :2297–2340, June 2014. ISSN 1097-0258. doi : 10.1002/sim.6128.

R. M. Baron and D. A. Kenny. The moderator–mediator variable distinction in social psychological research : Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6) :1173–1182, 1986. ISSN 1939-1315. doi : 10.1037/0022-3514.51.6.1173. Place : US Publisher : American Psychological Association.

A. L. Benton. A visual retention test for clinical use. *Arch Neurol Psychiatry*, 54 :212–216, Sept. 1945. ISSN 0096-6754. doi : 10.1001/archneurpsyc.1945.02300090051008.

G. J. Biessels, S. Staekenborg, E. Brunner, C. Brayne, and P. Scheltens. Risk of dementia in diabetes mellitus : a systematic review. *The Lancet Neurology*, 5(1) :64–

74, Jan. 2006. ISSN 1474-4422, 1474-4465. doi : 10.1016/S1474-4422(05)70284-2. URL https://www.thelancet.com/journals/lancet/article/PIIS1474-4422(05)70284-2/abstract. Publisher : Elsevier.

M.-a. C. Bind, T. J. Vanderweele, B. A. Coull, and J. D. Schwartz. Causal mediation analysis for longitudinal data with exogenous exposure. *Biostatistics*, 17(1) :122–134, Jan. 2016. ISSN 1468-4357. doi : 10.1093/biostatistics/kxv029.

K. A. Bollen and J. Pearl. Eight Myths About Causality and Structural Equation Models. In S. L. Morgan, editor, *Handbook of Causal Analysis for Social Research*, pages 301–328. Springer Netherlands, Dordrecht, 2013. ISBN 978-94-007-6093-6 978-94-007-6094-3. doi : 10.1007/978-94-007-6094-3_15. URL http://link.springer.com/10.1007/978-94-007-6094-3_15. Series Title : Handbooks of Sociology and Social Research.

S. J. Bond, I. R. White, and A. Sarah Walker. Instrumental variables and interactions in the causal analysis of a complex clinical trial. *Statistics in Medicine*, 26(7) :1473–1496, Mar. 2007. ISSN 0277-6715. doi : 10.1002/sim.2644.

R. L. Buckner. Memory and executive function in aging and AD : multiple factors that cause decline and reserve factors that compensate. *Neuron*, 44(1) :195–208, Sept. 2004. ISSN 0896-6273. doi : 10.1016/j.neuron.2004.09.006.

S. Burgess, D. S. Small, and S. G. Thompson. A review of instrumental variable estimators for Mendelian randomization. *Stat Methods Med Res*, 26(5) :2333–2355, Oct. 2017. ISSN 0962-2802. doi : 10.1177/0962280215597579. URL https://doi.org/10.1177/0962280215597579. Publisher : SAGE Publications Ltd STM.

W. J. Busby, A. J. Campbell, and M. C. Robertsone. Is Low Blood Pressure in Elderly People just a Consequence of Heart Disease and Frailty ? *Age and Ageing*, 23(1) :69–74, Jan. 1994. ISSN 0002-0729. doi : 10.1093/ageing/23.1.69. URL https://doi.org/10.1093/ageing/23.1.69.

D. Canoy, J. Tran, M. Zottoli, R. Ramakrishnan, A. Hassaine, S. Rao, Y. Li, G. Salimi-Khorshidi, R. Norton, and K. Rahimi. Association between cardiometabolic disease multimorbidity and all-cause mortality in 2 million women and men registered in

UK general practices. *BMC Med*, 19 :258, Oct. 2021. ISSN 1741-7015. doi : 10.1186/s12916-021-02126-x. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8555122/.

F. Chen, W. Hu, J. Cai, S. Chen, and W. Liu. Instrumental variable-based high-dimensional mediation analysis with unmeasured confounders for survival data in the observational epigenetic study. *Front. Genet.*, 14, Feb. 2023. ISSN 1664-8021. doi : 10.3389/fgene.2023.1092489. URL https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2023.1092489/full. Publisher : Frontiers.

R. R. Clayton, A. M. Cattarello, and B. M. Johnstone. The effectiveness of Drug Abuse Resistance Education (project DARE) : 5-year follow-up results. *Prev Med*, 25(3) : 307–318, 1996. ISSN 0091-7435. doi : 10.1006/pmed.1996.0061.

T. D. Cosco, A. M. Prina, J. Perales, B. C. M. Stephan, and C. Brayne. Lay perspectives of successful ageing : a systematic review and meta-ethnography. *BMJ Open*, 3(6) : e002710, June 2013. ISSN 2044-6055. doi : 10.1136/bmjopen-2013-002710.

D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2) :187–220, 1972. ISSN 0035-9246. URL https://www.jstor.org/stable/2985181. Publisher : [Royal Statistical Society, Wiley].

R. M. Daniel, B. L. De Stavola, S. N. Cousens, and S. Vansteelandt. Causal mediation analysis with multiple mediators. *Biometrics*, 71(1) :1–14, 2015. ISSN 1541-0420. doi : 10.1111/biom.12248. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.12248.

J.-F. Dartigues and C. Helmer. Comment expliquer le retard au diagnostic de maladie d'Alzheimer en France ? *Gérontologie et société*, 32 / 128-129(1-2) :183–193, 2009. ISSN 0151-0193. doi : 10.3917/gs.128.0183. URL https://www.cairn.info/revue-gerontologie-et-societe1-2009-1-2-page-183.htm. Place : Paris Publisher : Fondation Nationale de Gérontologie.

A. Devaux, C. Helmer, C. Dufouil, R. Genuer, and C. Proust-Lima. *Random survival*

forests for competing risks with multivariate longitudinal endogenous covariates. Aug. 2022. doi : 10.48550/arXiv.2208.05801.

V. Didelez. Graphical models for marked point processes based on local independence. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 70(1) : 245–264, 2008. ISSN 1467-9868. doi : 10.1111/j.1467-9868.2007.00634.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2007.00634.x.

V. Didelez. Defining causal mediation with a longitudinal mediator and a survival outcome. *Lifetime Data Anal*, 25(4) :593–610, Oct. 2019. ISSN 1572-9249. doi : 10.1007/s10985-018-9449-0.

V. Didelez, A. P. Dawid, and S. Geneletti. Direct and Indirect Effects of Sequential Treatments. 2006.

J. C. Digitale, J. N. Martin, and M. M. Glymour. Tutorial on directed acyclic graphs. *J Clin Epidemiol*, 142 :264–267, Feb. 2022. ISSN 1878-5921. doi : 10.1016/j.jclinepi.2021.08.001.

D. A. Drachman. Aging of the brain, entropy, and Alzheimer disease. *Neurology*, 67(8) : 1340–1352, Oct. 2006. ISSN 1526-632X. doi : 10.1212/01.wnl.0000240127.89601.83.

A. Duray, S. Demoulin, J. Petermans, M. Moutschen, S. Saussez, G. Jerusalem, and P. Delvenne. Vieillissement et cancer : coincidence ou relation etiologique ? *Rev Med Liège*, 2014.

A. Edjolo. L'épidémiologie de la dépendance du sujet âgé. Histoire naturelle, tendances évolutives et déterminants - theses.fr, Nov. 2014. URL https://www.theses.fr/2014BORD0422.

J. P. Fine and R. J. Gray. A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Journal of the American Statistical Association*, 94(446) :496–509, 1999. ISSN 0162-1459. doi : 10.2307/2670170. URL https://www.jstor.org/stable/2670170. Publisher : [American Statistical Association, Taylor & Francis, Ltd.].

G. G. Fisher and L. H. Ryan. Overview of the Health and Retirement Study and Introduction to the Special Issue. *Work, Aging and Retirement*, 4(1) :1–9, Jan. 2018. ISSN 2054-4642, 2054-4650. doi : 10.1093/workar/wax032. URL http://academic.oup.com/workar/article/4/1/1/4762672.

M. F. Folstein, S. E. Folstein, and P. R. McHugh. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res*, 12(3) : 189–198, Nov. 1975. ISSN 0022-3956. doi : 10.1016/0022-3956(75)90026-6.

L. Gelfand, D. MacKinnon, R. DeRubeis, and A. Baraldi. Mediation Analysis with Survival Outcomes : Accelerated Failure Time vs. Proportional Hazards Models. *Frontiers in Psychology*, 7, Mar. 2016. doi : 10.3389/fpsyg.2016.00423.

B. Giffard, B. Desgranges, and F. Eustache. Le vieillissement de la mémoire : vieillissement normal et pathologique. *Gérontologie et société*, 24 / 97(2) :33–47, 2001. ISSN 0151-0193. doi : 10.3917/gs.097.0033. URL https://www.cairn.info/revue-gerontologie-et-societe1-2001-2-page-33.htm. Place : Paris Publisher : Fondation Nationale de Gérontologie.

E. Goetghebeur, S. le Cessie, B. De Stavola, E. E. Moodie, I. Waernbaum, and b. o. t. t. g. C. I. T. o. t. S. Initiative. Formulating causal questions and principled statistical answers. *Statistics in Medicine*, 39(30) :4922–4948, 2020. ISSN 1097-0258. doi : 10.1002/sim.8741. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.8741.

G. Grande, P. L. S. Ljungman, K. Eneroth, T. Bellander, and D. Rizzuto. Association Between Cardiovascular Disease and Long-term Exposure to Air Pollution With the Risk of Dementia. *JAMA Neurology*, 77(7) :801–809, July 2020. ISSN 2168-6149. doi : 10.1001/jamaneurol.2019.4914. URL https://doi.org/10.1001/jamaneurol.2019.4914.

Z. Guo, M. Viitanen, L. Fratiglioni, and B. Winblad. Low blood pressure and dementia in elderly people : the Kungsholmen project. *BMJ*, 312(7034) :805–808, Mar. 1996. ISSN 0959-8138, 1468-5833. doi : 10.1136/bmj.312.7034.805. URL https://www.bmj.

com/content/312/7034/805. Publisher : British Medical Journal Publishing Group Section : Paper.

S. B. Guze. Diagnostic and Statistical Manual of Mental Disorders, 4th ed. (DSM-IV). *AJP*, 152(8) :1228–1228, Aug. 1995. ISSN 0002-953X. doi : 10.1176/ajp.152.8.1228. URL https://ajp.psychiatryonline.org/doi/10.1176/ajp.152.8.1228. Publisher : American Psychiatric Publishing.

A. Gégout-Petit and D. Commenges. A general definition of influence between stochastic processes. *Lifetime Data Anal*, 16(1) :33–44, Jan. 2010. ISSN 1380-7870, 1572-9249. doi : 10.1007/s10985-009-9131-7. URL http://link.springer.com/10.1007/s10985-009-9131-7.

C. N. Harada, M. C. Natelson Love, and K. Triebel. Normal Cognitive Aging. *Clin Geriatr Med*, 29(4) :737–752, Nov. 2013. ISSN 0749-0690. doi : 10.1016/j.cger.2013.07.002. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4015335/.

D. Harman. The aging process. *Proc Natl Acad Sci U S A*, 78(11) :7124–7128, Nov. 1981. ISSN 0027-8424. doi : 10.1073/pnas.78.11.7124.

R. Henderson and A. Milner. Aalen Plots Under Proportional Hazards. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 40(3) :401–409, 1991. ISSN 0035-9254. doi : 10.2307/2347520. URL https://www.jstor.org/stable/2347520. Publisher : [Wiley, Royal Statistical Society].

M. A. Hernan and J. M. Robins. Causal Inference : What If. 2020.

M. A. Hernán, D. Clayton, and N. Keiding. The Simpson's paradox unraveled. *International Journal of Epidemiology*, 40(3) :780–785, June 2011. ISSN 0300-5771. doi : 10.1093/ije/dyr041. URL https://doi.org/10.1093/ije/dyr041.

P. W. Holland. Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396) :945–960, 1986. ISSN 0162-1459. doi : 10.2307/2289064. URL https://www.jstor.org/stable/2289064. Publisher : [American Statistical Association, Taylor & Francis, Ltd.].

P. Hougaard. Multi-state models : a review. *Lifetime Data Anal*, 5(3) :239–264, Sept. 1999. ISSN 1380-7870. doi : 10.1023/a:1009672031531.

Y.-T. Huang and H.-I. Yang. Causal Mediation Analysis of Survival Outcome with Multiple Mediators. *Epidemiology*, 28(3) :370, May 2017. ISSN 1044-3983. doi : 10.1097/EDE.0000000000000651. URL https://journals.lww.com/epidem/fulltext/2017/05000/Daylight_Savings_Time_Transitions_and_the.11.aspx.

K. Imai, L. Keele, and D. Tingley. A general approach to causal mediation analysis. *Psychological Methods*, 15(4) :309–334, 2010a. ISSN 1939-1463, 1082-989X. doi : 10.1037/a0020761. URL https://doi.apa.org/doi/10.1037/a0020761.

K. Imai, L. Keele, and T. Yamamoto. Identification, Inference and Sensitivity Analysis for Causal Mediation Effects. *Statist. Sci.*, 25(1), Feb. 2010b. ISSN 0883-4237. doi : 10.1214/10-STS321.

B. Isaacs and A. J. Akhtar. The set test : a rapid test of mental function in old people. *Age Ageing*, 1(4) :222–226, Nov. 1972. ISSN 0002-0729. doi : 10.1093/ageing/1.4.222.

C. R. Jack, D. S. Knopman, W. J. Jagust, L. M. Shaw, P. S. Aisen, M. W. Weiner, R. C. Petersen, and J. Q. Trojanowski. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *Lancet Neurol*, 9(1) :119–128, Jan. 2010. ISSN 1474-4465. doi : 10.1016/S1474-4422(09)70299-6.

A. Jerolon, L. Baglietto, E. Birmele, V. Perduca, and F. Alarcon. Causal mediation analysis in presence of multiple mediators uncausally related, Aug. 2019. URL http://arxiv.org/abs/1809.08018. arXiv :1809.08018 [stat].

A. F. Jorm. History of Depression as a Risk Factor for Dementia : An Updated Review. *Aust N Z J Psychiatry*, 35(6) :776–781, Dec. 2001. ISSN 0004-8674, 1440-1614. doi : 10.1046/j.1440-1614.2001.00967.x. URL http://journals.sagepub.com/doi/10.1046/j.1440-1614.2001.00967.x.

C. M. Judd and D. A. Kenny. Process analysis : Estimating mediation in treatment evaluations. *Evaluation Review*, 5(5) :602–619, 1981. ISSN 1552-3926. doi : 10.1177/0193841X8100500502. Place : US Publisher : Sage Publications.

S. P. Kennelly, B. A. Lawlor, and R. A. Kenny. Blood Pressure and Dementia – a Comprehensive Review. *Ther Adv Neurol Disord*, 2(4) :241–260, July 2009. ISSN 1756-2856. doi : 10.1177/1756285609103483. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3002634/.

R. P. Kloppenborg, E. van den Berg, L. J. Kappelle, and G. J. Biessels. Diabetes and other vascular risk factors for dementia : Which factor matters most ? A systematic review. *European Journal of Pharmacology*, 585(1) :97–108, May 2008. ISSN 0014-2999. doi : 10.1016/j.ejphar.2008.02.049. URL https://www.sciencedirect.com/science/article/pii/S0014299908002276.

P. Kragh Andersen, M. Pohar Perme, H. C. van Houwelingen, R. J. Cook, P. Joly, T. Martinussen, J. M. G. Taylor, M. Abrahamowicz, and T. M. Therneau. Analysis of time-to-event for observational studies : Guidance to the use of intensity models. *Statistics in Medicine*, 40(1) :185–211, 2021. ISSN 1097-0258. doi : 10.1002/sim.8757. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.8757.

W. S. Kremen, A. Beck, J. A. Elman, D. E. Gustavson, and C. A. Reynolds. Influence of young adult cognitive ability and additional education on later-life cognition. *Proc Natl Acad Sci U S A*, 116(6) :2021–2026, Feb. 2019. ISSN 1091-6490. doi : 10.1073/pnas.1811537116.

N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4) :963–974, Dec. 1982. ISSN 0006-341X.

J.-C. Lambert, C. A. Ibrahim-Verbaas, D. Harold, A. C. Naj, R. Sims, C. Bellenguez, and G. Jun. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet*, 45(12) :1452–1458, Dec. 2013. ISSN 1546-1718. doi : 10.1038/ng.2802. URL https://www.nature.com/articles/ng.2802. Publisher : Nature Publishing Group.

T. Lange and J. V. Hansen. Direct and Indirect Effects in a Survival Context. *Epidemiology*, 22(4) :575–581, July 2011. ISSN 1044-3983. doi : 10.1097/EDE.0b013e31821c680c. URL https://journals.lww.com/00001648-201107000-00024.

T. Lange, M. Rasmussen, and L. C. Thygesen. Assessing Natural Direct and Indirect Effects Through Multiple Pathways. *American Journal of Epidemiology*, 179(4) :513–518, Feb. 2014. ISSN 0002-9262. doi : 10.1093/aje/kwt270. URL https://doi.org/10.1093/aje/kwt270.

P.-L. Lee, W. Lan, and T.-W. Yen. Aging Successfully : A Four-Factor Model. *Educational Gerontology*, 37(3) :210–227, Feb. 2011. ISSN 0360-1277. doi : 10.1080/03601277.2010.487759. URL https://doi.org/10.1080/03601277.2010.487759.

K. Leffondré, C. Touraine, C. Helmer, and P. Joly. Interval-censored time-to-event and competing risk with death : is the illness-death model more accurate than the Cox model ? *International Journal of Epidemiology*, 42(4) :1177–1186, Aug. 2013. ISSN 1464-3685, 0300-5771. doi : 10.1093/ije/dyt126. URL https://academic.oup.com/ije/article-lookup/doi/10.1093/ije/dyt126.

K. Levenberg. A Method for the Solution of Certain Non-Linear Problems in Least Squares. *Quarterly of Applied Mathematics*, 2(2) :164–168, 1944. ISSN 0033-569X. URL https://www.jstor.org/stable/43633451. Publisher : Brown University.

J. Li, J. Fine, and A. Brookhart. Instrumental variable additive hazards models. *Biometrics*, 71(1) :122–130, Mar. 2015. ISSN 1541-0420. doi : 10.1111/biom.12244.

S.-H. Lin, J. G. Young, R. Logan, and T. J. VanderWeele. Mediation analysis for a survival outcome with time-varying exposures, mediators, and confounders. *Stat Med*, 36(26) : 4153–4166, Nov. 2017. ISSN 1097-0258. doi : 10.1002/sim.7426.

J. C. Lindsey and L. M. Ryan. Methods for interval-censored data. *Statistics in Medicine*, 17(2) :219–238, 1998. ISSN 1097-0258. doi : 10.1002/(SICI)1097-0258(19980130)17:2<219::AID-SIM735>3.0.CO;2-O. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-0258%2819980130%2917%3A2%3C219%3A%3AAID-SIM735%3E3.0.CO%3B2-O.

L. Liu, C. Zheng, and J. Kang. Exploring causality mechanism in the joint analysis of longitudinal and survival data. *Stat Med*, 37(26) :3733–3744, Nov. 2018. ISSN 1097-0258. doi : 10.1002/sim.7838.

G. Livingston, A. Sommerlad, V. Orgeta, S. G. Costafreda, J. Huntley, and D. Ames. Dementia prevention, intervention, and care. *Lancet*, 390(10113) :2673–2734, Dec. 2017. ISSN 1474-547X. doi : 10.1016/S0140-6736(17)31363-6.

G. Livingston, J. Huntley, A. Sommerlad, D. Ames, C. Ballard, S. Banerjee, C. Brayne, and A. Burns. Dementia prevention, intervention, and care : 2020 report of the Lancet Commission. *Lancet*, 396(10248) :413–446, Aug. 2020. ISSN 1474-547X. doi : 10.1016/S0140-6736(20)30367-6.

D. M. Lloyd-Jones. Cardiovascular Risk Prediction. *Circulation*, 121(15) :1768–1777, Apr. 2010. URL https://www.ahajournals.org/doi/10.1161/circulationaha.109.849166. Publisher : American Heart Association.

C. Luo, B. Fa, Y. Yan, Y. Wang, Y. Zhou, Y. Zhang, and Z. Yu. High-dimensional mediation analysis in survival models. *PLOS Computational Biology*, 16(4) :e1007768, Apr. 2020. ISSN 1553-7358. doi : 10.1371/journal.pcbi.1007768. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007768. Publisher : Public Library of Science.

T. A. MacKenzie, T. D. Tosteson, N. E. Morden, T. A. Stukel, and A. J. O'Malley. Using instrumental variables to estimate a Cox's proportional hazards regression subject to additive confounding. *Health services & outcomes research methodology*, 14(1-2) : 54–68, June 2014. ISSN 1387-3741. doi : 10.1007/s10742-014-0117-x. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4261749/.

D. Mapelli, E. Di Rosa, R. Nocita, and D. Sava. Cognitive stimulation in patients with dementia : randomized controlled trial. *Dement Geriatr Cogn Dis Extra*, 3(1) :263–271, 2013. ISSN 1664-5464. doi : 10.1159/000353457.

D. W. Marquardt. An Algorithm for Least-Squares Estimation of Nonlinear Parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(2) :431–441, 1963. ISSN 0368-4245. URL https://www.jstor.org/stable/2098941. Publisher : Society for Industrial and Applied Mathematics.

T. Martinussen and S. Vansteelandt. Instrumental variables estimation with competing risk data. *Biostatistics (Oxford, England)*, 21(1) :158–171, Jan. 2020. ISSN 1468-4357. doi : 10.1093/biostatistics/kxy039.

C. D. McCullagh, D. Craig, S. P. McIlroy, and A. P. Passmore. Risk factors for dementia. *Adv. psychiatr. treat*, 7(1) :24–31, Jan. 2001. ISSN 1355-5146, 1472-1481. doi : 10.1192/apt.7.1.24. URL https://www.cambridge.org/core/product/identifier/S1355514600009391/type/journal_article.

D. H. Mellor. *The Facts of Causation*. Routledge, New York, 1995.

X. Meng and C. D'Arcy. Education and dementia in the context of the cognitive reserve hypothesis : a systematic review with meta-analyses and qualitative analyses. *PLoS One*, 7(6) :e38268, 2012. ISSN 1932-6203. doi : 10.1371/journal.pone.0038268.

M. N. Mittinty and S. Vansteelandt. Longitudinal Mediation Analysis Using Natural Effect Models. *Am J Epidemiol*, 189(11) :1427–1435, Nov. 2020. ISSN 1476-6256. doi : 10.1093/aje/kwaa092.

L. A. Morgan and S. R. Kunkel. *Aging, Society, and the Life Course*. Springer Publishing Company, 2007. ISBN 978-0-8261-0098-6.

D. L. Murman. The Impact of Age on Cognition. *Semin Hear*, 36(3) :111–121, Aug. 2015. ISSN 0734-0451. doi : 10.1055/s-0035-1555115. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4906299/.

F. Nourhashémi, V. Deschamps, S. Larrieu, L. Letenneur, J.-F. Dartigues, and P. Barberger-Gateau. Body mass index and incidence of dementia. *Neurology*, 60 (1) :117–119, Jan. 2003. doi : 10.1212/01.WNL.0000038910.46217.AA. URL https://www.neurology.org/doi/abs/10.1212/01.wnl.0000038910.46217.aa. Publisher : Wolters Kluwer.

L. Nyberg, M. Lövdén, K. Riklund, U. Lindenberger, and L. Bäckman. Memory aging and brain maintenance. *Trends Cogn Sci*, 16(5) :292–305, May 2012. ISSN 1879-307X. doi : 10.1016/j.tics.2012.04.005.

H.-R. Park and A. Uno. Investigation of Cognitive Abilities Related to Reading and Spelling in Korean : Readers with High, Average, and Low Skill Levels. *Dyslexia*, 18(4) :199–215, 2012. ISSN 1099-0909. doi : 10.1002/dys.1443. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/dys.1443.

J. Pearl. *Causality : Models, reasoning, and inference*. Causality : Models, reasoning, and inference. Cambridge University Press, New York, NY, US, 2000. ISBN 978-0-521-77362-1. Pages : xvi, 384.

J. Pearl. Direct and Indirect Effects. 2001.

J. Pearl. The Causal Foundations of Structural Equation Modeling :. Technical report, Defense Technical Information Center, Fort Belvoir, VA, Feb. 2012. URL http://www.dtic.mil/docs/citations/ADA557445.

V. Philipps, B. Hejblum, P., M. Prague, D. Commenges, and C. Proust-Lima. Robust and Efficient Optimization Using a Marquardt-Levenberg Algorithm with R Package marqLevAlg. *The R Journal*, 13(2) :273, 2021. ISSN 2073-4859. doi : 10.32614/RJ-2021-089. URL https://journal.r-project.org/archive/2021/RJ-2021-089/index.html.

M. Prague, D. Commenges, and R. Thiébaut. Dynamical models of biomarkers and clinical progression for personalized medicine : The HIV context. *Advanced Drug Delivery Reviews*, 65(7) :954–965, June 2013. ISSN 0169409X. doi : 10.1016/j.addr.2013.04.004. URL https://linkinghub.elsevier.com/retrieve/pii/S0169409X13000641.

M. Prague, J. M. Gerold, I. Balelli, C. Pasin, J. Z. Li, D. H. Barouch, J. B. Whitney, and A. L. Hill. Viral rebound kinetics following single and combination immunotherapy for HIV/SIV. preprint, Systems Biology, July 2019. URL http://biorxiv.org/lookup/doi/10.1101/700401.

K. Preacher and A. Hayes. Asymptotic and Resampling Strategies for Assessing and Comparing Indirect Effects in Multiple Mediator Models. *Behavior research methods*, 40 :879–91, Sept. 2008. doi : 10.3758/BRM.40.3.879.

C. Proust-Lima, J.-F. Dartigues, and H. Jacqmin-Gadda. Joint modeling of repeated multivariate cognitive measures and competing risks of dementia and death : a latent process and latent class approach. *Stat Med*, 35(3) :382–398, Feb. 2016. ISSN 1097-0258. doi : 10.1002/sim.6731.

R. Pérez-Ocón, J. E. Ruiz-Castro, and M. L. Gámiz-Pérez. Semi-Markov Models for Lifetime Data Analysis. In J. Janssen and N. Limnios, editors, *Semi-Markov Models and Applications*, pages 229–238. Springer US, Boston, MA, 1999. ISBN 978-1-4613-3288-6. doi : 10.1007/978-1-4613-3288-6_14. URL https://doi.org/10.1007/978-1-4613-3288-6_14.

K. Pérès, C. Helmer, H. Amieva, J.-M. Orgogozo, I. Rouch, J.-F. Dartigues, and P. Barberger-Gateau. Natural history of decline in instrumental activities of daily living performance over the 10 years preceding the clinical diagnosis of dementia : a prospective population-based study. *J Am Geriatr Soc*, 56(1) :37–44, Jan. 2008. ISSN 1532-5415. doi : 10.1111/j.1532-5415.2007.01499.x.

R. M. Reitan. The relation of the trail making test to organic brain damage. *J Consult Psychol*, 19(5) :393–394, Oct. 1955. ISSN 0095-8891. doi : 10.1037/h0044509.

M. Rey-Millet, M. Pousse, C. Soithong, J. Ye, A. Mendez-Bermudez, and E. Gilson. Senescence-associated transcriptional derepression in subtelomeres is determined in a chromosome-end-specific manner. *Aging Cell*, 22(5) :e13804, 2023. ISSN 1474-9726. doi : 10.1111/acel.13804. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/acel.13804.

J. M. Robins and S. Greenland. Identifiability and Exchangeability for Direct and Indirect Effects. *Epidemiology*, 3(2) :143–155, 1992. ISSN 1044-3983. URL https://www.jstor.org/stable/3702894. Publisher : Lippincott Williams & Wilkins.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5) :688–701, 1974. ISSN 1939-2176. doi : 10.1037/h0037350. Place : US Publisher : American Psychological Association.

K. E. Rudolph, J. Levy, and M. J. van der Laan. Transporting Stochastic Direct and Indirect Effects to New Populations. *Biometrics*, 77(1) :197–211, Mar. 2021. ISSN 0006-341X. doi : 10.1111/biom.13274. URL https://doi.org/10.1111/biom.13274.

K. E. Rudolph, N. Williams, and I. Díaz. Using instrumental variables to address unmeasured confounding in causal mediation analysis. *Biometrics*, 80(1) :ujad037, Mar. 2024. ISSN 0006-341X. doi : 10.1093/biomtc/ujad037. URL https://doi.org/10.1093/biomtc/ujad037.

T. A. Salthouse. The processing-speed theory of adult age differences in cognition. *Psychol Rev*, 103(3) :403–428, July 1996. ISSN 0033-295X. doi : 10.1037/0033-295x.103.3.403.

T. A. Salthouse. Interrelations of Aging, Knowledge, and Cognitive Performance. In U. M. Staudinger and U. Lindenberger, editors, *Understanding Human Development : Dialogues with Lifespan Psychology*, pages 265–287. Springer US, Boston, MA, 2003. ISBN 978-1-4615-0357-6. doi : 10.1007/978-1-4615-0357-6_12. URL https://doi.org/10.1007/978-1-4615-0357-6_12.

T. Schweder. Composable Markov Processes. *Journal of Applied Probability*, 7(2) :400–410, 1970. ISSN 0021-9002. doi : 10.2307/3211973. URL https://www.jstor.org/stable/3211973. Publisher : Applied Probability Trust.

J. Snowden, E. Tilden, and M. Odden. Formulating and Answering High-Impact Causal Questions in Physiologic Childbirth Science : Concepts and Assumptions. *Journal Of Midwifery & Women's Health*, 63(6) :721–730, Nov. 2018. doi : 10.1111/jmwh.12868. URL https://pdxscholar.library.pdx.edu/sph_facpub/167.

J. H. Stock. Nonparametric Policy Analysis. *Journal of the American Statistical Association*, 84(406) :567–575, June 1989. ISSN 0162-1459. doi : 10.1080/01621459.1989.10478805. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1989.10478805.

J. P. Sy, J. M. Taylor, and W. G. Cumberland. A stochastic model for the analysis of bivariate longitudinal AIDS data. *Biometrics*, 53(2) :542–555, June 1997. ISSN 0006-341X.

B. N. Sánchez, S. Kim, and M. D. Sammel. Estimators for longitudinal latent exposure models : examining measurement model assumptions. *Statistics in medicine*, 36(13) : 2048–2066, June 2017. ISSN 0277-6715. doi : 10.1002/sim.7268. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5418122/.

B. O. Taddé, H. Jacqmin-Gadda, J.-F. Dartigues, D. Commenges, and C. Proust-Lima. Dynamic modeling of multivariate dimensions and their temporal relationships using latent processes : Application to Alzheimer's disease. *Biometrics*, 76(3) :886–899, Sept. 2020. ISSN 1541-0420. doi : 10.1111/biom.13168.

A.-S. Tai, C.-A. Tsai, and S.-H. Lin. Survival mediation analysis with the death-truncated mediator : The completeness of the survival mediation parameter. *Stat Med*, 40(17) : 3953–3974, July 2021. ISSN 1097-0258. doi : 10.1002/sim.9008.

E. J. T. Tchetgen and I. Shpitser. Semiparametric theory for causal mediation analysis : Efficiency bounds, multiple robustness and sensitivity analysis. *The Annals of Statistics*, 40(3) :1816–1845, June 2012. ISSN 0090-5364, 2168-8966. doi : 10.1214/12-AOS990. Publisher : Institute of Mathematical Statistics.

E. J. Tchetgen Tchetgen, S. Walter, S. Vansteelandt, T. Martinussen, and M. Glymour. Instrumental variable estimation in a survival context. *Epidemiology (Cambridge, Mass.)*, 26(3) :402–410, May 2015. ISSN 1531-5487. doi : 10.1097/EDE.0000000000000262.

R. Thiébaut, H. Jacqmin-Gadda, G. Chêne, C. Leport, and D. Commenges. Bivariate linear mixed models using SAS proc MIXED. *Computer Methods and Programs in Biomedicine*, 69(3) :249–56, Nov. 2002. URL https://hal.science/hal-00143963. Publisher : Elsevier.

V. Tigano, G. L. Cascini, C. Sanchez-Castañeda, P. Péran, and U. Sabatini. Neuroimaging and Neurolaw : Drawing the Future of Aging. *Front Endocrinol (Lausanne)*, 10 :217, Apr. 2019. ISSN 1664-2392. doi : 10.3389/fendo.2019.00217. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6463811/.

C. Touraine, C. Helmer, and P. Joly. Predictions in an illness-death model. *Stat Methods Med Res*, 25(4) :1452–1470, Aug. 2016. ISSN 0962-2802. doi : 10.1177/0962280213489234. URL https://doi.org/10.1177/0962280213489234. Publisher : SAGE Publications Ltd STM.

M. J. Valenzuela and P. Sachdev. Brain reserve and cognitive decline : a non-parametric systematic review. *Psychol. Med.*, 36(8) :1065–1073, Aug. 2006. ISSN 0033-2917, 1469-8978. doi : 10.1017/S0033291706007744. URL https://www.cambridge.org/core/product/identifier/S0033291706007744/type/journal_article.

L. Valeri, C. Proust-Lima, W. Fan, J. T. Chen, and H. Jacqmin-Gadda. A multistate approach for the study of interventions on an intermediate time-to-event in health disparities research. *Stat Methods Med Res*, 32(8) :1445–1460, Aug. 2023. ISSN 0962-2802. doi : 10.1177/09622802231163331. URL https://doi.org/10.1177/09622802231163331. Publisher : SAGE Publications Ltd STM.

W. M. van der Flier and P. Scheltens. Epidemiology and risk factors of dementia. *J Neurol Neurosurg Psychiatry*, 76(Suppl 5) :v2–v7, Dec. 2005. ISSN 0022-3050. doi : 10.1136/jnnp.2005.082867. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1765715/.

F. M. van Oudenhoven, S. H. N. Swinkels, T. Hartmann, and D. Rizopoulos. Modeling the underlying biological processes in Alzheimer's disease using a multivariate competing risk joint model. *Stat Med*, 41(17) :3421–3433, July 2022. ISSN 1097-0258. doi : 10.1002/sim.9425.

T. J. VanderWeele. Marginal structural models for the estimation of direct and indirect effects. *Epidemiology*, 20(1) :18–26, Jan. 2009. ISSN 1531-5487. doi : 10.1097/EDE.0b013e31818f69ce.

T. J. VanderWeele. Causal mediation analysis with survival data. *Epidemiology*, 22(4) : 582–585, July 2011. ISSN 1044-3983. doi : 10.1097/EDE.0b013e31821db37e. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3109321/.

T. J. VanderWeele and Y. Chiba. Sensitivity analysis for direct and indirect effects in the presence of exposure-induced mediator-outcome confounders. *Epidemiol Biostat Public Health*, 11(2) :e9027, 2014. ISSN 2282-2305. doi : 10.2427/9027.

T. J. VanderWeele and S. Vansteelandt. Mediation Analysis with Multiple Mediators. *Epidemiol Methods*, 2(1) :95–115, Jan. 2014. ISSN 2194-9263. doi : 10.1515/em-2012-0010.

M. C. Voelkle, C. Gische, C. C. Driver, and U. Lindenberger. The Role of Time in the Quest for Understanding Psychological Mechanisms. *Multivariate Behavioral Research*, 53(6) :782–805, Nov. 2018. ISSN 0027-3171. doi : 10.1080/00273171.2018.1496813. URL https://doi.org/10.1080/00273171.2018.1496813. Publisher : Routledge _eprint : https ://doi.org/10.1080/00273171.2018.1496813.

M. Wagner, F. Grodstein, K. Leffondre, C. Samieri, and C. Proust-Lima. Time-varying associations between an exposure history and a subsequent health outcome : a landmark approach to identify critical windows. *BMC Medical Research Methodology*, 21(1) : 266, Nov. 2021. ISSN 1471-2288. doi : 10.1186/s12874-021-01403-w. URL https://doi.org/10.1186/s12874-021-01403-w.

R. L. West. An application of prefrontal cortex function theory to cognitive aging. *Psychol Bull*, 120(2) :272–292, Sept. 1996. ISSN 0033-2909. doi : 10.1037/0033-2909.120.2.272.

L. C. d. Wreede, M. Fiocco, and H. Putter. mstate : An R Package for the Analysis of Competing Risks and Multi-State Models. *Journal of Statistical Software*, 38 :1–30, Jan. 2011. ISSN 1548-7660. doi : 10.18637/jss.v038.i07. URL https://doi.org/10.18637/jss.v038.i07.

S. Wright. The Method of Path Coefficients. *The Annals of Mathematical Statistics*, 5 (3) :161–215, 1934. ISSN 0003-4851. URL https://www.jstor.org/stable/2957502. Publisher : Institute of Mathematical Statistics.

N. Zanjari, M. Sharifian Sani, M. H. Chavoshi, H. Rafiey, and F. Mohammadi Shahboulaghi. Successful aging as a multidimensional concept : An integrative review. *Med J Islam Repub Iran*, 31 :100, Dec. 2017. ISSN 1016-1430. doi : 10.14196/mjiri.31.100. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6014811/.

H. Zhang, Y. Zheng, L. Hou, C. Zheng, and L. Liu. Mediation analysis for survival data with high-dimensional mediators. *Bioinformatics*, 37(21) :3815–3821, Nov. 2021. ISSN 1367-4811. doi : 10.1093/bioinformatics/btab564.

C. Zheng and L. Liu. Quantifying direct and indirect effect for longitudinal mediator and survival outcome using joint modeling approach. *Biometrics*, 78(3) :1233–1243, Sept. 2022. ISSN 1541-0420. doi : 10.1111/biom.13475.

W. Zheng and M. van der Laan. Longitudinal Mediation Analysis with Time-varying Mediators and Exposures, with Application to Survival Outcomes. *J Causal Inference*, 5(2) :20160006, Sept. 2017. ISSN 2193-3677. doi : 10.1515/jci-2016-0006.

Résumé : L'épidémiologie du vieillissement pose de nombreux problèmes méthodologiques ayant mené au développement de modèles statistiques adaptés. Toutefois, la recherche de facteurs impactant de façon causale le processus de vieillissement dans les études de cohorte observationnelles ainsi que la compréhension des voies d'action causales de ces facteurs restent encore limitées par la rareté voire l'absence de méthodes d'inférence causale adaptées aux données longitudinales. Cette thèse vise à développer de nouveaux outils d'inférence causale pour l'étude des facteurs de risque du vieillissement et des mécanismes sous-jacents dans les études observationnelles longitudinales. Dans une première partie, nous nous sommes intéressés aux méthodes d'analyse de médiation permettant de décomposer les effets totaux entre un facteur de risque et une maladie, en un effet direct et des effets indirects passant par des variables médiatrices. Plus précisément, nous avons proposé une approche d'analyse de médiation pour étudier le lien causal entre une exposition fixe dans le temps et des processus de médiation, de confusion et d'outcome final, tous les trois définis en temps continu mais mesurés de façon irrégulière au cours du temps. pour le médiateur et l'outcome. Nous avons également discuté une approche d'analyse de médiation permettant d'étudier des variables intermédiaires et terminales de type temps d'événement, avec une possible censure par intervalle du temps d'événement intermédiaire. Dans la deuxième partie, nous avons étendu la méthode par variables instrumentales pour traiter la confusion non observée lorsque l'on étudie une exposition fixe dans le temps et un outcome mesuré de façon répétée dans le temps. Nous avons appliqué ces approches aux données de la cohorte populationnelle 3C, s'interessant au processus de vieillissement cérébral chez les personnes âgées. Les travaux présentés dans cette thèse ouvrent ainsi la voie à une meilleure compréhension des mécanismes causaux impliqués dans diverses pathologies, en présence de phénomènes d'intérêt évoluant au cours du temps.
Mots-clés : Causalité, Variable instrumentale, Analyse de médiation, Données longitudinales, Vieillissement cérébral

Abstract : The field of aging epidemiology poses numerous statistical challenges that have led to the development of adapted statistical models. However, the search for factors impacting the aging process causally in observational cohort studies, as well as the understanding of the causal pathways of these factors, is still limited by the rarity or even absence of causal inference methods adapted to longitudinal data. This thesis aims to develop new causal inference tools for studying aging risk factors and underlying mechanisms in longitudinal observational studies. In the first part, we focused on mediation analysis methods to decompose total effects between a risk factor and a disease into a direct effect and indirect effects through intermediate variables. More specifically, we proposed a mediation analysis approach to study the causal link between a fixed exposure over time and mediator, confounder and outcome all defined in continuous time but measured irregularly over time. We also discussed a mediation analysis approach to study intermediate and terminal time-to-event variables, with possible interval censoring of the intermediate time-to-event. In the second part, we extended the instrumental variables method to address unobserved confounding when studying a time-fixed exposure and an outcome measured repeatedly over time. We applied these approaches to data from the 3C population cohort, focusing on the process of cerebral aging in the elderly. The work presented in this thesis thus paves the way for a better understanding of the causal mechanisms involved in various pathologies, in presence of health phenomena evolving over time.
Keywords : Causality, Instrumental variable, Mediation analysis, Longitudinal data, Cerebral aging

Unité de recherche Inserm U1219,
Bordeaux Population Health,
Université de Bordeaux
146, rue Léo Saignat 33000 Bordeaux, France