Approches causales pour l'analyse de données observationnelles - application aux données longitudinales avec traitements multiples

François Bettega

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : ISCE - Ingénierie pour la Santé la Cognition et l'Environnement
Spécialité : MBS - Modèles, méthodes et algorithmes en biologie, santé et environnement
Unité de recherche : Hypoxie et Physiopathologie

Approches causales pour l'analyse de données observationnelles application aux données longitudinales avec traitements multiples

Causal inference for longitudinal observationnal data - application to multinomial longitudinal data

Présentée par :

François BETTEGA

Direction de thèse :

Sébastien BAILLY CHARGE DE RECHERCHE, Université Grenoble Alpes	Directeur de thèse

Rapporteurs :

Matthieu RESCHE-RIGON
PROFESSEUR DES UNIVERSITES - PRATICIEN HOSPITALIER, Université Paris Cité
David HAJAGE
PROFESSEUR DES UNIVERSITES - PRATICIEN HOSPITALIER, Faculté de Médecine Sorbonne Université

Thèse soutenue publiquement le **12 décembre 2023**, devant le jury composé de :

Sébastien BAILLY CHARGE DE RECHERCHE, INSERM	Directeur de thèse
Emilie DEVIJVER MAITRESSE DE CONFERENCES, Université Grenoble Alpes	Examinatrice
Delphine MAUCORT-BOULCH PROFESSEURE DES UNIVERSITES - PRATICIENNE HOSPITALIERE, Université Claude Bernard Lyon 1	Présidente
Matthieu RESCHE-RIGON PROFESSEUR DES UNIVERSITES - PRATICIEN HOSPITALIER, Université Paris Cité	Rapporteur
David HAJAGE PROFESSEUR DES UNIVERSITES - PRATICIEN HOSPITALIER, Faculté de Médecine Sorbonne Université	Rapporteur
Guillaume DUMAS PROFESSEUR DES UNIVERSITES - PRATICIEN HOSPITALIER, Université Grenoble Alpes	Examinateur

Invités :

Clemence Leyrat
MAITRESSE DE CONFERENCES, The London School of Hygiene & Tropical Medicine

Table des matières

Liste des figures

Glossaire

AIPTW : Augmented Inverse Probability of Treatment Weighting

ATE : Average Treatment Effect

ATT : Average Treatment Effect on the Treated

CBPS : Covariate Balancing Propensity Score

DAG : Directed Acyclic Graph

ECR : Essai Contrôlé Randomisé

GBM : Generalized Boosted Models

IAH : Index Apnées-Hypopnées

ESS : Epworth Sleepiness Scale

IPW-RA : Inverse Propensity Weighted Regression Adjustment

IC : Intervalle de Confiance

IMC : Indice de Masse Corporelle

IPTW : Inverse Probability of Treatment Weighting

IQR : Inter Quartile Range

MsM : Marginal structural Model

OSFP : Observatoire Sommeil de la Fédération de Pneumologie

Outcome : Variable réponse

PPC : Pression Positive Continue

PSAD : Prestataires de Soins à Domicile

SAOS : Syndrome d'Apnées Obstructives du Sommeil

SMD : Standardized Mean Difference

SP : Scores de Propensions

TMLE : Targeted Maximum Likelihood Estimation

VAS : Voies Aériennes Supérieures

Financement

Pour ce travail de thèse j'ai été financé par l'Agence Nationale de la Recherche dans le cadre du programme "Investissements d'avenir" (ANR-15-IDEX-02) et les chaires d'excellence "e-santé et soins intégrés et médecine des trajectoires et intelligence artificielle MIAI" de l'Université Grenoble Alpes.

Remerciement

Je souhaite remercier les membres de mon jury Emilie DEVIJVER, Delphine MAUCORT-BOULCH et Guillaume DUMAS pour l'intérêt que vous avez porté à mon travail de thèse, votre présence et nos échanges lors de ma soutenance.

Merci à Matthieu RESCHE-RIGON et David HAJAGE d'avoir accepté de rapporter ce travail de thèse, pour vos relectures avisées et pour vos retours constructifs et extrêmement intéressants à l'écrit et lors de ma soutenance.

Je tiens à remercier mes directeurs de thèse Clemence Leyrat et Sébastien BAILLY pour votre aide tout au long de ce travail de thèse, merci à vous deux pour votre mentorat complémentaire. Clémence, merci d'avoir pris de ton temps pour m'avoir partagé ton expertise en matière d'inférence causale, merci encore d'avoir accepté de continuer a répondre à mes questions même durant tes congés. Sébastien, merci pour ton soutien sans failles pour me permettre de réaliser ce travail de thèse et ta disponibilité. Merci d'avoir pris de ton temps pour m'initier au travail de chercheur et à la rédaction d'un travail de recherche.

Merci à Monique Mendelson pour avoir accepté de relire mes travaux en anglais afin d'en améliorer grandement la rédaction.

Merci à Christelle Gonindard pour m'avoir permis d'enseigner et m'avoir soutenu dans les difficultés liées a l'enseignement.

Je remercie également le Professeur Jean-Louis Pépin pour votre accueil au sein du laboratoire HP2 et m'avoir accompagné dans la mise en place de mes projets d'avenir après thèse.

Merci à Paul et Nicolas de m'avoir permis de savoir ce qu'impliquait d'être doctorant et d'avoir accepté de m'écouter parler de mon travail de thèse pendant 3 ans.

Merci Clélia d'avoir supporté de vivre avec moi pendant la dernière année de ma thèse et de m'avoir soutenu sans faille, merci pour tout.

Merci à ma famille de m'avoir soutenu pendant toutes mes études.

Merci a tous mes amis de m'avoir permis de survivre à ces 3 ans en me permettant de décompresser.

I. Introduction

1.1 Syndrome d'apnées obstructives du sommeil

Le syndrome d'apnées obstructives du sommeil (SAOS) est une maladie respiratoire chronique qui se caractérise par des interruptions involontaires et répétitives de la respiration (apnées) et/ou des réductions significatives du flux d'air respiratoire (hypopnées) pendant le sommeil. Ces apnées-hypopnées sont principalement causées par des obstructions des voies aériennes supérieures (VAS) résultant de l'affaissement des tissus mous des VAS du fait d'anomalies anatomiques ou fonctionnelles. Les facteurs de risque majeurs de cette pathologie sont l'obésité, l'âge avancé et le sexe masculin [45]. En Figure 1.1 le mécanisme physiopathologique du SAOS.

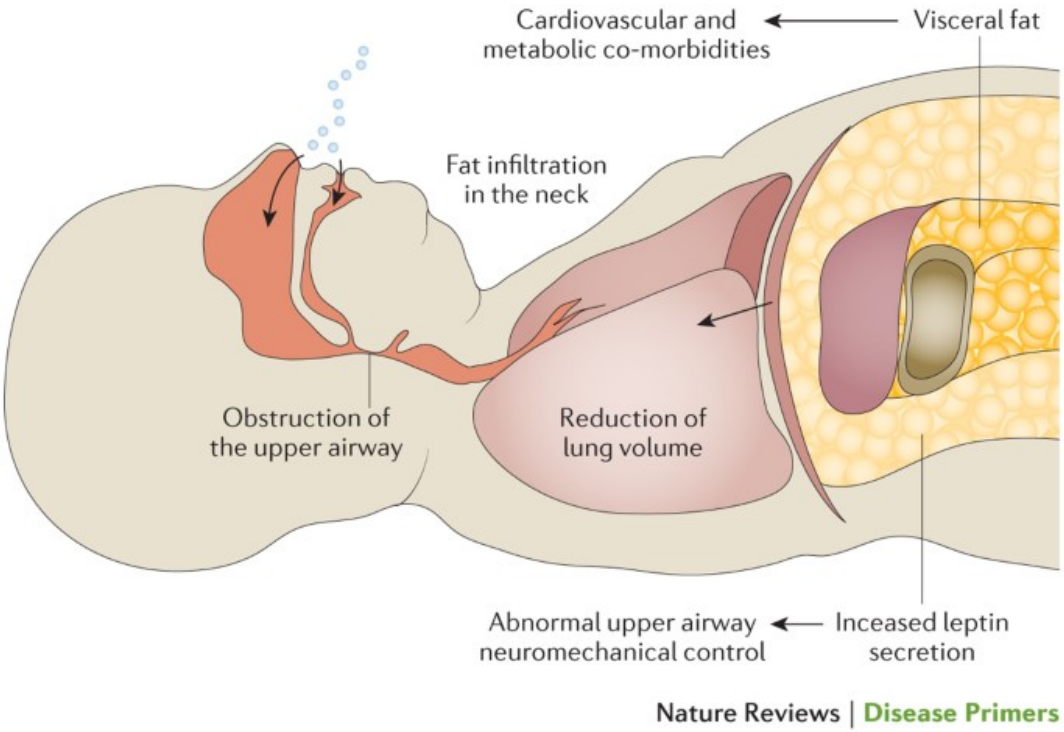

Figure 1.1: Obésité et syndrome d'apnées obstructives du sommeil.
L'obésité prédispose au SAOS par l'infiltration de graisses dans le cou conduisant à des affaissements des VAS. Elle augmente également la pression abdominale, réduisant le volume pulmonaire. L'accumulation de tissus adipeux pourrait aussi affecter le contrôle neuromécanique des VAS par les effets spécifiques de la leptine. La graisse viscérale favorise les comorbidités cardiométaboliques. Les "bulles" représentent les gaz inhalés et expirés. Légende et figure tirées de [45], la figure est adaptée de Drager L, Togeiro S, Polotsky V, et al. Obstructive Sleep Apnea. J Am Coll Cardiol. 2013 Aug, 62 (7) 569–576, Elsevier..

Le SAOS engendre plusieurs effets néfastes directs sur l'organisme tels qu'une augmentation de l'effort respiratoire avec des micro-éveils, des pressions intra-thoraciques modifiées qui exercent des contraintes mécaniques sur le cœur et les vaisseaux, ainsi qu'une exposition intermittente à l'hypoxie qui correspond à une diminution temporaire de l'apport en oxygène dans l'organisme avant de retrouver un niveau normal. Il en résulte divers symptômes qui altèrent la qualité de

vie, notamment une somnolence diurne, une détérioration des capacités cognitives comme la difficulté de concentration, la dépression et un risque accru d'accidents vasculaires [45]. En outre, le SAOS entraîne une baisse de productivité et une augmentation de l'absentéisme professionnel [3]. De nombreux phénomènes interdépendants entraînent une perturbation systémique de l'organisme, provoquant une altération des fonctions métaboliques, hépatiques et cardiovasculaires. Ces phénomènes interdépendants sont présentés en Figure 1.2. Le SAOS peut également entraîner une hypertension artérielle et il est associé à des comorbidités cardiométaboliques telles que le diabète, l'insuffisance cardiaque et les accidents vasculaires cérébraux [45, 35].

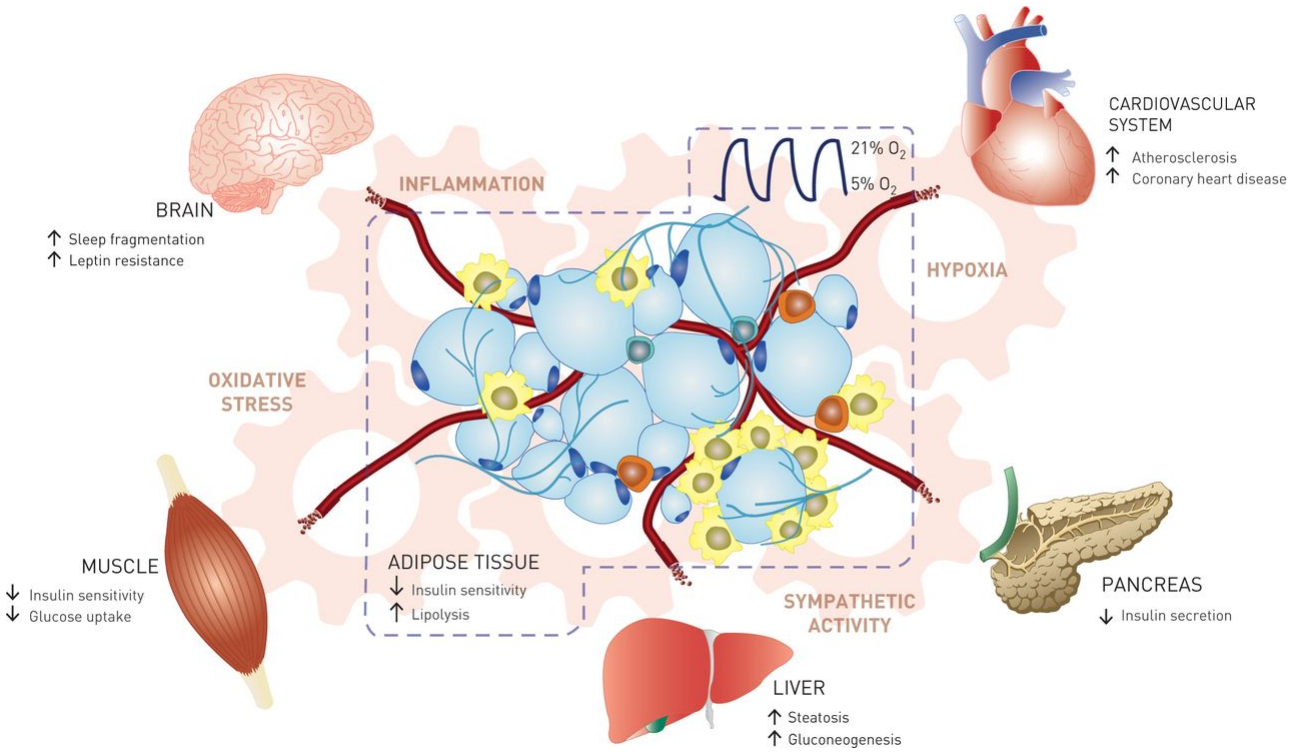

Figure 1.2: Le tissu adipeux vu comme un acteur majeur dans les conséquences systémiques de l'hypoxie intermittente. Source : [71]

Le SAOS est une pathologie complexe avec différents phénotypes identifiables résultant de plusieurs caractéristiques liées : 1) à l'individu et ses facteurs de risque personnels qui peuvent contribuer à son développement ; 2) aux différentes manifestations possibles, à différents stades de sommeil et différents niveaux de gravité en fonction de la fréquence des micro-éveils, de l'ampleur et de la fréquence des épisodes d'hypoxie ; 3) à une variété de symptômes, en dehors des manifestations pendant le sommeil, qui peuvent affecter les patients atteints de SAOS ; et 4) aux différentes comorbidités des patients entraînant des interactions spécifiques avec le SAOS et donc des conséquences différentes à long terme [89].

Selon les sociétés savantes, le diagnostic et la gravité du SAOS sont évalués de diverses manières [21, 52, 19]. La société de pneumologie de langue française prend en compte les symptômes cliniques du patient ainsi que l'index d'apnées-hypopnées (IAH). L'IAH est un score objectif qui calcule la fréquence des événements d'apnées et d'hypopnées observés lors d'un examen du sommeil [11]. Pour diagnostiquer un SAOS, il faut au moins cinq événements d'apnées ou d'hypopnées par heure de sommeil. Une apnée obstructive est caractérisée par une réduction du flux respiratoire d'au moins 90 % pendant plus de 10 secondes malgré la présence de mouvements respiratoires. Une hypopnée est définie comme une réduction du flux respiratoire d'au moins 30 % pendant plus de 10 secondes, accompagnée soit d'une diminution de la teneur en oxygène du sang artériel au-delà de 3 % (désaturation artérielle en oxygène), soit d'un micro-éveil. Le SAOS est considéré comme

sévère lorsque l'IAH est supérieur à 30 événements par heure.

Selon Benjafield et al. [9], le SAOS pourrait être largement sous-diagnostiqué à l'échelle mondiale. Ils estiment à plus de 900 millions les individus, âgés de 30 à 69 ans, ayant un IAH supérieur à 5 événements par heure. Parmi eux, plus de 400 millions ont un IAH supérieur à 15 événements par heure. Ces estimations prennent en compte les variations dans les taux de prévalence selon les pays et les régions du monde. Les 10 pays avec la plus grande prévalence sont présentés en Figure 1.3.

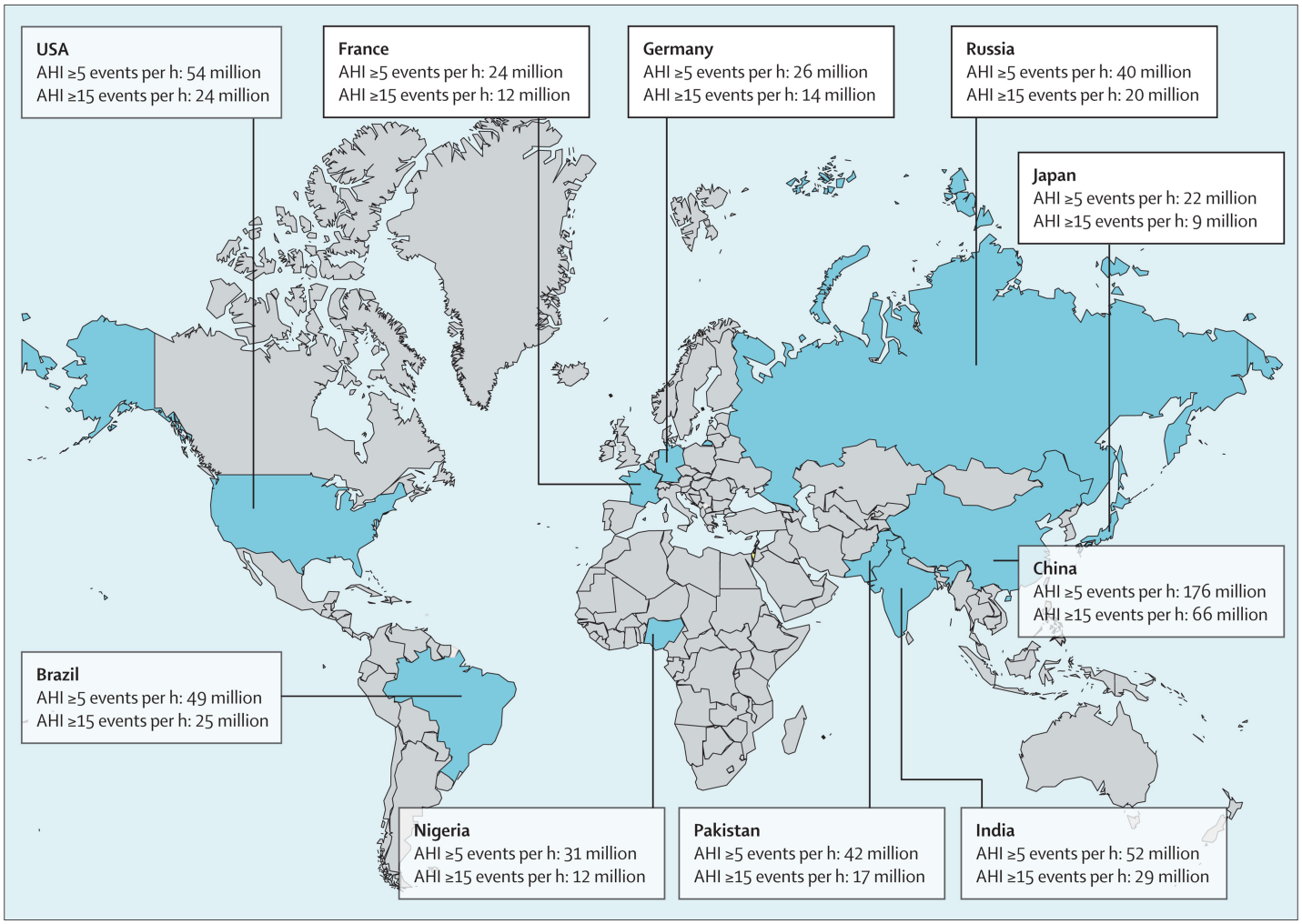

Figure 1.3: Les 10 pays avec les plus grandes prévalences estimées du syndrome d'apnées obstructives du sommeil selon les critères 2012 de l'American Academy of Sleep Medicine [11].
Source : [9]

Le SAOS a des implications économiques significatives en termes de prise en charge qui nécessite une gestion des coûts. Un rapport de Frost & Sullivan de 2016, mandaté par l'American Academy of Sleep Medicine, a évalué que les coûts de prise en charge du SAOS s'élevaient à 12,4 milliards de dollars américains pour 5,9 millions de patients diagnostiqués, tandis que le coût du sous-diagnostic était estimé à 149,6 milliards de dollars pour 23,5 millions de personnes supposées non diagnostiquées. Les coûts pris en compte dans cette étude comprenaient les comorbidités, les accidents de véhicules motorisés, les accidents de travail et la perte de productivité. Les auteurs de cette étude ont recommandé une amélioration du diagnostic et du traitement du SAOS. Une étude ultérieure menée par Wickwire et al. [86] en 2020 a estimé le surcoût annuel d'un SAOS non traité aux États-Unis chez les personnes de plus de 65 ans entre 13 000 et 26 000 dollars.

En France, la ventilation en pression positive continue (PPC) est le traitement de première intention pour les cas modérés à sévères du SAOS. Elle est prescrite principalement en fonction de l'IAH et dans une moindre mesure en fonction

des symptômes, des risques d'accidents et des comorbidités [62]. La Figure 1.4 présente les effets sur l'obstruction des voies aériennes supérieures d'un appareil de PPC avec masque bucco-nasal.

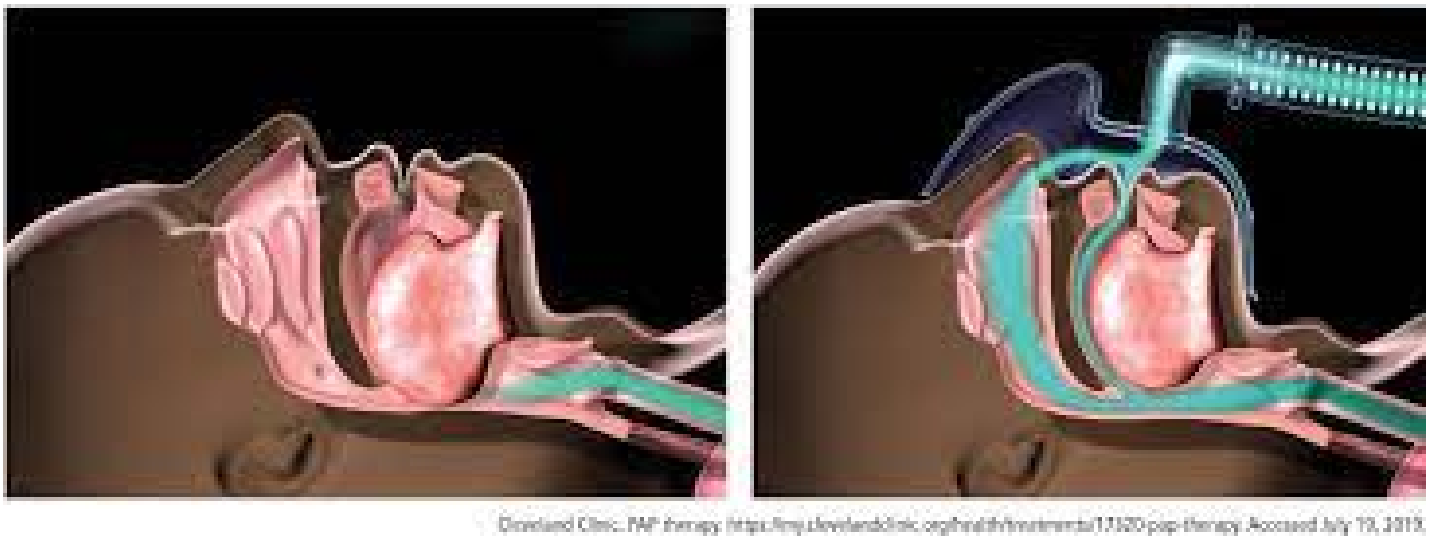

Figure 1.4: Représentation schématique de l'utilisation d'un appareil de pression positive continue par l'intermédiaire d'un masque bucco-nasal et de son effet sur les voies aériennes supérieures.
Source : Cleveland Clinic. cPAP therapy. https://my.clevelandclinic.org/health/treatments/17320-pap-therapy Accessed april 2023.

Cette thérapie consiste à normaliser les événements respiratoires en maintenant les voies aériennes supérieures ouvertes pendant le sommeil à l'aide d'une pression d'air positive continue dans le pharynx. Début 2021, environ 1,2 million de personnes en France utilisaient ce traitement.

Le traitement par PPC a montré un effet bénéfique sur l'atténuation des symptômes diurnes, l'amélioration de la qualité de vie [45], des fonctions neurocognitives [16] et de la productivité subjective au travail [3]. Elle peut également avoir un impact bénéfique sur certains troubles cardiométaboliques tels que la diminution de l'hypertension artérielle [54, 34] et la résistance à l'insuline [33]. Une étude observationnelle conduite par Marin et al. [47] suggère que la PPC peut protéger contre les événements cardiovasculaires et la mortalité cardiovasculaire mais l'étude randomisée SAVE [50] n'a pas confirmé cet effet. Cependant, cette dernière avait des limites notamment le temps de traitement quotidien moyen et le niveau effectif d'utilisation des machines par les patients [61]. Une analyse des données sur les patients en France a montré que l'interruption du traitement par PPC était associée à un risque plus élevé de décès et d'insuffisance cardiaque [79]. Cependant, l'observance à la PPC est souvent irrégulière et d'une durée insuffisante [69] ce qui limite son efficacité. De plus, l'acceptation du traitement et l'adhésion à long terme sont des obstacles importants [42, 81]. De fait, près de la moitié des prises en charge par la PPC, initiées en France entre 2015 et 2016, ont été interrompues dans les trois premières années [62] après exclusion des patients décédés ou ayant eu recours à un traitement alternatif. Ces problèmes d'acceptation, d'adhésion et d'observance à la PPC s'ajoutent au sous-diagnostic du SAOS. Il reste donc à investiguer l'effet bénéfique de la PPC en prenant en compte les différents phénotypes du SAOS, les causes de mortalité et le niveau d'observance des patients [61, 52, 40, 73].

En France, les prestataires de soins à domicile (PSAD) sont chargés de fournir, régler et entretenir les machines de PPC chez les patients. Augmenter l'observance semble justifié à trois niveaux. Tout d'abord pour les patients, afin d'améliorer leur réponse au traitement. Ensuite pour l'État, pour maintenir les avantages de l'investissement dans la prise en charge du SAOS car certains effets de la PPC sont réversibles dès la première nuit d'interruption [16, 39], pour réduire

les coûts de santé [86] et sociétaux associés à la non prise en charge du SAOS. Enfin pour les PSAD, afin d'obtenir un remboursement plus élevé par l'assurance maladie car le remboursement est déterminé par plusieurs forfaits en fonction de l'observance quotidienne moyenne par périodes de 28 jours consécutifs. Les PSAD doivent donc rendre compte de l'utilisation des machines pour chacun de leurs patients. Certains modèles de machine enregistrent des données d'utilisation supplémentaires utiles pour mesurer des indicateurs quotidiens. Ces indicateurs peuvent être des objectifs pertinents pour le suivi de la thérapie. Ces données peuvent inclure : 1) la durée pendant laquelle le patient porte le masque lorsque la machine délivre une pression thérapeutique qui mesure l'observance au traitement, appelée ici la "durée de port du masque" ; 2) les pressions et fuites d'air au niveau du masque, qui permettent de contrôler l'usure des consommables et/ou les réglages de l'appareil et 3) l'IAH résiduel, qui correspond au nombre d'événements d'apnées ou d'hypopnées observées sous traitement qui indique l'efficacité curative de la PPC. Depuis le 1er janvier 2018, les PSAD ont l'obligation légale de proposer aux patients le télé-suivi de leur traitement qui permet la transmission automatique de ces données au PSAD. Ce dernier les met à disposition des médecins, des patients et fournit les données d'observance à l'assurance maladie.

Comme nous avons pu le voir, le SAOS est une pathologie complexe avec de nombreuses conséquences sur la santé à long terme des patients atteints. Le télé-suivi de l'observance à la PPC offrant des données précises et en quantité, ces données offrent une opportunité importante de mieux comprendre l'impact de l'observance à la PPC sur les conséquences du SAOS.

1.2 Les facteurs associés à l'observance à la PPC

Les seuils de référence de l'observance à la PPC visés par les soignants et les patients sont relativement arbitraires et la qualité des données est insuffisante pour établir une durée optimale d'utilisation de la PPC. Il n'est pas clair si les améliorations liées au traitement par PPC sont dues à un effet de seuil (par exemple, lorsque l'utilisation est $\geq$ 4 h/nuit) ou montrent une relation dose-réponse. La situation est d'autant plus complexe, que la durée optimale de l'observance à la PPC peut varier en fonction de l'outcome étudié par exemple le risque d'événements cardiovasculaires, le risque de trouble métabolique ou la somnolence diurne.

Le taux d'arrêt de la PPC reste élevé, moins de 50% des patients continuent à utiliser la PPC plus de 4h par nuit après plusieurs années [50, 17]. Au cours des 20 dernières années, aucune amélioration significative de l'observance à la PPC n'a été observée malgré des améliorations évidentes des aspects techniques des appareils [53]. La gravité du SAOS, évaluée par l'indice d'apnées-hypopnées, et les aspects techniques relatifs aux appareils de PPC semblent être des facteurs mineurs dans l'explication de l'adhésion à la PPC. Les autres facteurs (comorbidités, facteurs psychologiques, profil des couples, statut socio-économique, accès aux soins et diversité culturelle) devraient être mieux reconnus et inclus dans des interventions sur mesure [32, 56, 63]. Le manque d'observance à la PPC n'est pas sans conséquence, l'étude SAVE a montré que le sous-groupe dont l'observance à la PPC était >4 h par nuit présentait une réduction du risque d'accident vasculaire cérébral et d'hypertension artérielle [50]. Dans l'essai RICCADSA [60], le sous-groupe utilisant la PPC >4 h par nuit a montré un taux plus faible d'événements cardiovasculaires. Il existe également des données concrètes démontrant un lien entre l'observance à la PPC et les événements cardiovasculaires et de mortalité prématurée [20]. Les coûts économiques associés au SAOS sont considérables pour l'individu et pour la société [20]. À titre d'exemple, le coût économique des accidents de la route liés au SAOS a été estimé à 810 000 collisions et 1 400 décès aux États-Unis en 2000 [20]. Le SAOS non traité entraîne l'agrégation et la progression des comorbidités qui peuvent potentiellement augmenter la consommation

de soins de santé. Nous abordons également les défis et les pièges pour la visualisation et l'analyse des plateformes de surveillance à distance de la PPC. Les stratégies pour améliorer l'observance devraient être adaptées individuellement et viser à améliorer les habitudes de vie, notamment l'activité physique et l'alimentation. L'accès à ces stratégies devrait être soutenu en améliorant les tableaux de bord de visualisation des plateformes de surveillance à distance de la PPC et en diffusant la télé-santé et les analyses innovantes, y compris l'intelligence artificielle. Ces éléments ont été publiés dans une revue dont je suis co-auteur présenté en annexe 1. La figure 1.5 est un résumé graphique.

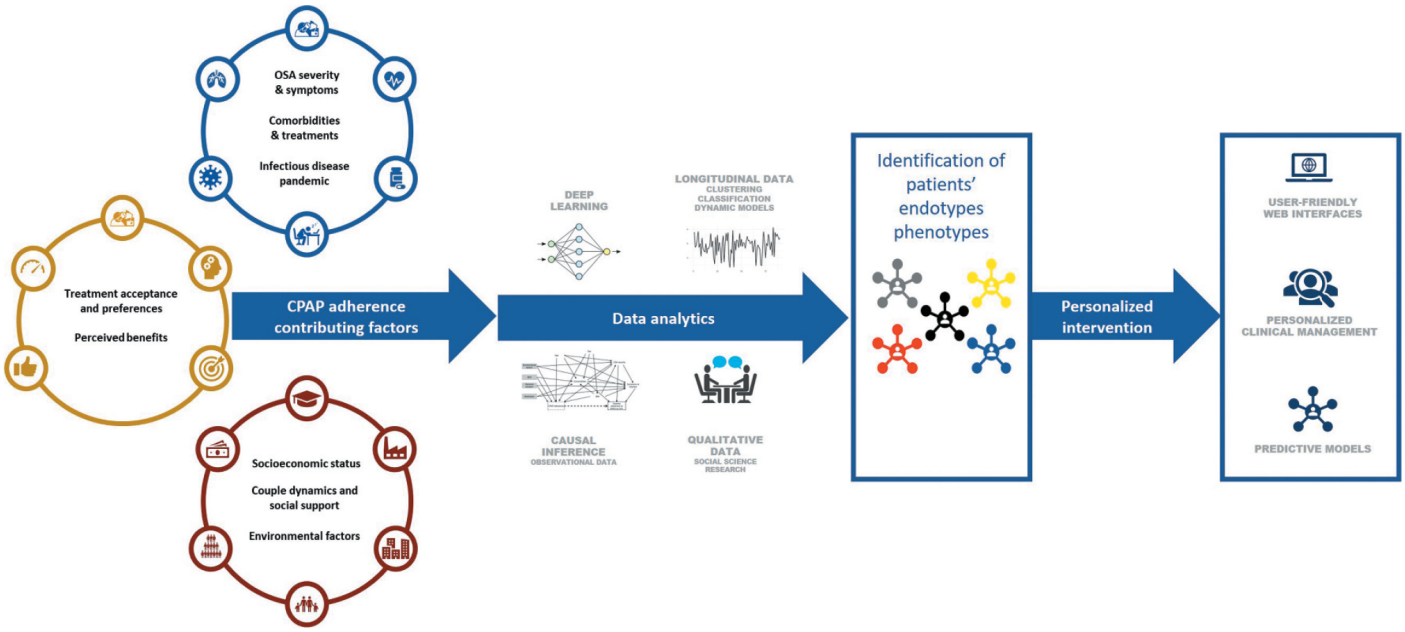

Figure 1.5: Facteurs influençant l'adhésion à la PPC, outils/méthodes d'analyse des données, identification des phénotypes de patients et développement d'interventions personnalisées.
Source : [53]

1.3 Evaluation de la PPC en vie réelle

En recherche, et en particulier en recherche clinique, il est fréquent de chercher à évaluer l'effet d'un traitement ou d'un facteur d'exposition sur une variable réponse. Les essais contrôlés randomisés (ECRs) sont les études qui, bien menées, permettent d'obtenir le plus haut niveau de preuve. Dans ce type d'étude, l'assignation du traitement est aléatoire, chaque patient ayant la même probabilité de recevoir chaque traitement et les caractéristiques de chaque groupe ne diffèrent que par le traitement. Les essais contrôlés randomisés présentent aussi des limitations et ne sont pas toujours réalisables, en fonction de contraintes éthiques, logistiques ou financières. [85, 10]. Par exemple, il serait non acceptable d'un point de vue éthique de demander à des patients non-fumeurs de fumer afin d'évaluer l'effet du tabac[57, 58, 51]. Dans ce cas, une analyse de données observationnelles, comparant les fumeurs et non-fumeurs est une solution pragmatique.

Les données recueillies en vie réelle possèdent des avantages multiples. Elles ont généralement des effectifs beaucoup plus importants que dans les ECRs et elles sont souvent plus représentatives de la vie réelle (les patients des ECRs étant généralement plus observants, mieux suivis et en meilleure santé que la population générale[38]). Il existe, de plus, une très grande variété de sources de données observationnelles. Ainsi, chaque base de données complétée régulièrement constitue un potentiel jeu de données [28].

Les études observationnelles sont, par nature, exposées à des biais systématiques. Bien que ces biais puissent aussi se produire dans les ERCs mal menés, ils sont omniprésents dans les études observationnelles, si ces dernières ne sont pas

planifiées et analysées de manière adéquate. Les biais les plus fréquents sont :

- Le biais de confusion. Dans une étude observationnelle, les traitements ne sont pas prescrits aléatoirement à chaque patient. Un patient aura donc plus ou moins de chance de recevoir un des traitements selon ses caractéristiques. Si certaines de ces caractéristiques ont aussi un effet sur le critère de jugement, alors l'estimation non ajustée de l'effet causal du traitement sera biaisée. On appelle de telles variables des facteurs de confusion. Plus généralement, le biais de confusion existe lorsque l'effet observé d'un traitement est expliqué partiellement ou complètement par une ou plusieurs variables non prises en compte dans l'analyse.

- Le biais de sélection, quant à lui, est le biais induit par le processus de sélection des individus dans l'étude (par exemple, pré-sélection de patients ayant un meilleur état de santé général) ou dans l'analyse [30]. S'il s'agit d'un problème de sélection des patients dans l'étude, les résultats ne peuvent être généralisés a la population dans son ensemble. S'il s'agit d'un problème de sélection des patients pour l'analyse, l'effet estimé du traitement est biaisé par rapport a l'échantillon d'origine.

- Le biais de mesure ou de classification se produit lorsque la quantité mesurée (exposition ou critère de jugement) est différente de sa valeur réelle. Cela peut être le cas pour les self-reported outcomes, pour lesquels les patients peuvent avoir un biais de rappel. [18]. Le biais de classification peut prendre différentes formes. Par exemple, le biais de temps immortel, qui peut apparaître quand le début du suivi et le début du traitement ne coïncident pas [44, 48] peut être vu comme une mauvaise classification du traitement, et conduire a une sur-estimation de l'effet causal du traitement.

Il existe des méthodes statistiques permettant d'établir des relations causales à partir de données observationnelles. Ces méthodes reposent, comme dans le cas des ECRs, sur des hypothèses strictes et un plan d'étude spécifiques. Les méthodes reposant sur les scores de propension (SP) ont été proposées pour corriger le biais de confusion présent dans l'analyse des données observationnelles. La prise en compte des autres biais nécessite d'autres outils, qui ne font pas l'objet de cette thèse. La théorie des scores de propension a été initialement développée pour des expositions binaires et sa mise en oeuvre pour des expositions à plusieurs niveaux n'a reçu que peu d'attention, ce qui peut expliquer son utilisation limitée dans la pratique [88].

Ces méthodes sont de plus en plus explorées dans le domaine médical. Elles semblent particulièrement bien adaptées au SAOS qui est une pathologie où il existe de nombreuses bases de données de suivi de patients. On peut citer par exemple, la cohorte European Sleep Apnoea Database (ESADA), la base de données MARS développée par le CHU de Grenoble, pour Multimorbidity Apnea Respiratory failure Sleep ou l'Observatoire Sommeil de la Fédération de Pneumologie (OSFP). Il est donc possible d'utiliser ces bases de données et les méthodes basées sur les scores de propension pour évaluer l'effet causal marginal de l'adhésion à la PPC sur différentes variables réponses (outcomes).

L'observance à la PPC est un processus complexe avec de nombreux facteurs comme : le sexe [62], l'IMC, la gravité du SAOS [75] et l'âge [72] qui ont un effet causal direct à la fois sur l'observance à la PPC mais aussi sur différents outcomes. Afin d'évaluer l'effet de l'observance à la PPC, il est important de prendre en compte ces facteurs potentiels de confusion. L'observance à la PPC étant une durée, la résumer à un seuil binaire durée d'utilisation supérieure ou inférieure à 4h par exemple, est une perte d'information importante[4]. Il semble donc utile d'évaluer l'effet de plusieurs seuils d'observance afin de mesurer l'effet causal de la PPC.

L'évaluation de l'effet de l'observance à la PPC semble donc être une problématique dans laquelle les SP, appliqués aux traitements à plusieurs niveaux, sont un outil adapté afin de déterminer si l'effet du traitement varie avec le niveau

d'observance.

1.4 Objectifs de la thèse

Cette thèse a pour objectif d'approfondir l'application des approches SP multi-niveaux dans le domaine clinique, notamment dans le domaine du syndrome d'apnées du sommeil. Pour cela, les travaux réalisés se sont décomposés de la façon suivante :

1. Proposer une exploration systématique de l'usage des méthodes d'inférence causale avec traitement multi-niveaux dans la littérature médicale afin d'évaluer les spécialités médicales dans lesquelles ces méthodes sont les plus appliquées.

   Et en particulier de :

   (a) Évaluer avec quel modèle ces méthodes sont implémentées et dans quel langage de programmation.

   (b) Comparer pour combien de groupes de traitement, de quelle taille et pour quel type d'outcomes ces méthodes sont utilisées.

   (c) Évaluer la qualité de l'usage rapporté des ces méthodes.

2. Proposer des applications des méthodes basées sur les scores de propension appliquées aux traitements multi-niveaux, afin de sensibiliser à l'usage de ces méthodes et des hypothèses dont elles dépendent.

3. Évaluer les performances et les limites de modèles autres que les régressions multinomiales pour l'estimation des scores de propension dans le cas des traitements multi-niveaux.

Le chapitre 2 pose les bases des méthodes de l'inférence causale. Le chapitre 3 développe la méthode et les résultats d'une revue systématique de l'usage des "Inverse Probability of Treatment Weighting" (IPTW) multi-niveaux dans la recherche médicale. Le chapitre 4 développe une application des IPTW multi-niveaux pour évaluer l'effet causal de l'observance à la PPC sur la somnolence diurne. Le chapitre 5 compare les performances d'une méthode de machine learning utilisée pour l'inférence avec d'autres méthodes couramment utilisées. Enfin le chapitre 6 conclut mon travail de thèse.

II. Inférence Causale

La théorie présentée ici est relativement succincte, les différents éléments sont détaillés plus longuement dans les sections suivantes et les articles associés inclus dans le manuscrit.

2.1 Cadre théorique des outcomes potentiels

Contrairement aux statistiques traditionnelles qui visent à évaluer les associations entre une exposition et un outcome, l'inférence causale fait référence à des hypothèses et à un plan d'étude spécifique pour pouvoir tirer des conclusions causales à partir des données [59]. Un cadre théorique, le cadre théorique des outcomes [70], a été proposé pour définir un langage permettant d'exprimer des quantités causales et définissant des hypothèses nécessaires à l'identification de ces effets causaux.

Dans cette section, les concepts principaux de l'inférence causale sont présentés dans le cadre d'un outcome continu ou binaire Y, un traitement binaire A, avec $A = 1$ correspond au traitement d'intérêt et $A = 0$ correspond à l'absence de traitement, et un ensemble $\mathbf{C}$ de covariables mesurées au niveau individuel. Cette théorie sera étendue aux traitements à plus de deux modalités dans les sections suivantes.

La théorie des outcomes potentiels, aussi appelée théorie des contrefactuels, postule l'existence hypothétique, pour chaque individu, d'un outcome correspondant à ce qui aurait été observé à chaque niveau de traitement. Par exemple, dans notre cas d'un traitement A binaire, chaque individu i ($i = 1, ..., n$) a deux outcomes potentiels, $Y_i(0)$ et $Y_i(1)$ capturant la réponse de l'individu i en l'absence de traitement, ou recevant le traitement dont on essaye d'estimer l'effet causal. [82]. Le contraste entre $Y_i(0)$ et $Y_i(1)$ correspond à l'effet causal du traitement A pour l'individu i. A l'échelle de la population, le contraste entre $Y_(0)$ et $Y_(1)$ correspond à l'effet causal marginal.

En pratique, pour chaque individu, seulement un seul outcome est observé (l'outcome factuel) et l'autre outcome est appelé contrefactuel. Dans ce contexte, l'inférence causale peut donc être vu comme un problème de données manquantes. Sous un ensemble d'hypothèses décrites dans la sous-section suivante, et grâce à des outils statistiques spécifiques, il est néanmoins possible d'identifier et estimer des effets causaux à l'échelle de la population (mais pas au niveau individuel).

Hypothèses permettant l'identification de l'effet causal

Les hypothèses nécessaires à l'identification d'effets causaux dans le cadre théorique des outcomes potentiels sont présentées ci-dessous. Il est important de noter qu'elles sont empiriquement non vérifiables, à l'exception de l'hypothèse de positivité.

- L'hypothèse de positivité stipule qu'il existe une probabilité non nulle de recevoir chaque niveau de traitement à chaque niveau de combinaison des covariables parmi les individus de la population cible [14], c'est à dire $0 < P(A = a|\mathbf{C}) < 1, \text{for a=0,1}$. Par exemple, l'existence de contre-indications formelles à l'un des traitements évalués dans la

population observée constitue une violation de l'hypothèse de positivité car les patients présentant la contre-indication ne pourraient pas être exposés au traitement contre-indiqué.

- L'hypothèse de consistance postule que l'outcome observé d'un individu sous son exposition observée est identique à l'outcome potentiel de cet individu sous l'intervention hypothétique correspondant à ce niveau d'exposition [67, 27], c'est-à-dire: pour $A = a$, $Y(a) = Y$ avec Y représentant l'outcome observé. Cela suppose que l'observation est la même chose que l'intervention. La plausibilité de cette hypothèse dépend de la précision avec laquelle le traitement est défini. Par exemple, pour un traitement pharmacologique, une définition précise de la molécule, sa posologie, le mode d'administration, etc. est nécessaire pour garantir la consistance de l'effet du traitement.

- L'hypothèse de non-interférence stipule que le traitement d'un individu n'a aucune influence sur les outcomes potentiels d'autres individus: $A_i \perp Y_j(0), Y_j(1)\ \forall i, j = 1, ..., n\ i \neq j$. Un exemple de violation de cette hypothèse est de considérer les vaccins car la vaccination d'un individu peut affecter le statut de la maladie d'autres individus.

- L'échangeabilité conditionnelle fait référence à l'hypothèse de l'absence de facteurs de confusion non mesurés: $Y(0), Y(1) \perp A|\mathbf{C}$. En inférence causale, tous les prédicteurs conjoints de l'exposition et des outcomes doivent être pris en compte. Ainsi, toutes les variables liées aux traitements et aux outcomes doivent être incluses [14]

- L'hypothèse d'une spécification correcte du modèle postule que les probabilités inconnues pour un patient d'appartenir à chaque groupe de traitement, connaissant tous les facteurs de confusion, sont modélisées par un modèle correctement spécifié. Par exemple, la modélisation d'une relation exponentielle à l'aide d'un modèle linéaire constitue une violation de cette hypothèse. Cette hypothèse stipule également que toutes les variables confondantes et leurs formes fonctionnelles réelles sont utilisées pour ajuster le modèle.

Les hypothèses 1 à 4 sont nécessaires pour l'identification de l'effet causal et l'hypothèse 5 est nécessaire pour son estimation.

2.2 Méthode d'inférence causale sur données observationnelles

Estimands d'intérêt en inférence causale

Plusieurs estimands sont souvent définis dans le cadre de l'inférence causale :

- l'Average Treatment Effect (ATE) correspond à l'effet moyen du traitement en comparant deux mondes hypothétiques dans lesquels tous les individus de la population sont traités versus non traités. Dans le cas d'un traitement binaire avec un outcome Y et un traitement A, les outcomes potentiels obtenus avec les 2 modalités du traitement sont notés $Y(0)$ et $Y(1)$. $ATE = \mathbb{E}[Y(A_1)] - \mathbb{E}[Y(A_0)]$.

- L'Average Treatment effect on the Treated (ATT) correspond à l'effet moyen du traitement chez les traités. Y $ATT = \mathbb{E}[Y(1)|A = 1] - \mathbb{E}[Y(0)|A = 1]$

- L'Average Treatment effect on the Untreated (ATU) correspond à l'effet moyen du traitement chez les non-traités. Y $ATU = \mathbb{E}[Y(1)|A = 0] - \mathbb{E}[Y(0)|A = 0]$

- L'Average Treatment effect on in the Overlap population (ATO), a été proposé plus récemment et correspond à l'effet moyen du traitement dans une population similaire à ce qui aurait été observé dans un essai randomisé.

L'ATT est un estimand particulièrement intéressant quand on essaye d'évaluer l'effet bénéfique de supprimer une exposition à risque comme le tabac par exemple chez les individus exposés. L'ATU, à l'inverse est utile lorsqu'on cherche à évaluer l'effet qu'aurait la prescription du traitement à une population plus large que celle définie par les recommandations cliniques. Quand il n'y a pas d'interaction entre l'effet traitement et les variables associées avec l'attribution du traitement, l'ATE, l'ATT, l'ATU et l'ATO sont égaux. Dans le cadre de cette thèse, nous nous focaliserons uniquement sur l'ATE.

Estimation ajustée par une régression

Une première manière de tenir compte des déséquilibres sur les facteurs de confusion est d'ajuster sur les caractéristiques déséquilibrées. Pour cela, on modélise l'outcome en fonction du traitement ainsi que des facteurs de confusion. Puis, il est possible d'utiliser le modèle ainsi obtenu pour prédire l'outcome qu'aurait obtenu chaque patient s'il avait reçu chacun des différents traitements.

Il est alors possible de faire la différence entre les outcomes obtenus pour chaque traitement et ainsi calculer les écarts d'outcomes entre les groupes de traitement, cette méthode est la g-computation [66].

Score de propension

Le score de propension est la probabilité pour chaque patient $i, i = 1, ..., n$ d'appartenir à son propre groupe de traitement, connaissant ses caractéristiques individuelles $\mathbf{C} = (C_1, ..., C_k)$, avec k nombre de caractéristiques individuelles $SP = P(A = a|C_1 = c_1, ..., C_k = c_k)$. Les scores de propension sont des scores d'équilibre, tels que définis par Rosenbaum et Rubin [68]. Un score d'équilibre, b(x), est une fonction des covariables observées x telle que la distribution conditionnelle de x compte tenu de b(x) est la même pour l'unité traitée (Z = 1) et l'unité témoin (Z = 0). Cela signifie qu'à chaque valeur du score de propension, les individus ont, en moyenne, des distributions similaires de leurs caractéristiques. Par conséquent, les scores de propension sont utilisés pour équilibrer les covariables entre les groupes pour l'estimation de l'ATE. Cela nécessite, outre les quatre hypothèses d'identification, que le modèle de SP soit correctement spécifié, c'est-à-dire que le modèle pour l'affectation du traitement comprenne tous les facteurs de confusion dans leur forme fonctionnelle correcte et qu'il inclut les interactions potentielles. Les lignes directrices actuelles recommandent d'inclure tous les facteurs de risque (confondants ou non) et d'éviter d'inclure les prédicteurs du traitement (variables instrumentales) dans le modèle de score de propension [43].

Appariement et stratification

Une fois estimé et après avoir obtenu un bon équilibre des covariables, le score de propension peut être utilisé de différentes manières pour l'estimation de l'effet du traitement. Les approches les plus courantes sont l'appariement, la stratification et la pondération par l'IPTW. L'appariement, qui implique la création de paires de personnes traitées et non traitées ayant des valeurs de score de propension similaires, permet principalement d'estimer l'ATT, même si des méthodes d'appariement bidirectionnel permettent d'estimer l'ATE. Cependant, l'appariement entraîne généralement une perte de taille de l'échantillon, ce qui affecte à la fois la puissance et la généralisabilité de l'étude. Ce phénomène est amplifié lorsque le nombre de niveaux de traitements augmente. La stratification crée des sous-classes de patients ayant des scores de propension similaires dans lesquelles les effets du traitement sont estimés et mis en commun entre les strates. La stratification entraîne souvent des déséquilibres résiduels, en particulier si le nombre de strates est trop faible. Cependant, cette approche peut être utile pour explorer la présence d'interaction entre le SP et le traitement.

Les limitations de l'appariement et de la stratification m'ont incité à concentrer mon travail de thèse sur les IPTWs.

Inverse Probability-of-Treatment Weighting

L'IPTW est une méthode basée sur les scores de propension issus des travaux de Rosenbaum and Rubin [68]. La paternité de l'IPTW revient à Robins [67]. Dans un IPTW, il est nécessaire de pondérer chaque individu par l'inverse de leur score de propension.

$$IPW_i = \frac{1}{P(A_i = a_i | c_1 = c_{1i}, ..., c_k = c_{ki})}$$

Quand les effectifs sont assez grands, les IPTWs et la g-computation, présentée précédemment, conduisent à des résultats similaires [76]. L'avantage de l'IPTW, par rapport à un ajustement classique sur les covariables est qu'il conduit à un résultat marginal alors que les résultats de l'ajustement sont conditionnels.

Estimateur doublement robuste

Les méthodes doublement robustes combinent les deux approches présentées ci-dessus. En utilisant à la fois un modèle basé sur les scores de propension et une modélisation de la distribution de l'outcome, ces 2 modèles fournissent une estimation asymptotiquement non biaisée et dont la variance tend vers zéro quand la taille de l'échantillon tend vers l'infini si un des deux modèles est spécifié correctement.[36]. Des exemples de méthodes doublement robustes sont :

- l'Inverse Probability Weighted Regression Adjustment (IPW-RA) qui consiste à effectuer une régression multivariée pondérée par l'inverse du score de propension pour estimer l'effet causal (conditionel) du traitement.
- l'Augmented Inverse Probability of Treatment Weighted estimator (AIPTW). Les estimateurs IPTW ont souvent une imprécision majorée par rapport à l'ajustement par régression multivariable. En combinant les deux approches, il est possible de gagner en efficacité. L'AIPTW est une méthode doublement robuste qui consiste à ajouter à l'estimateur IPTW un terme de moyenne nulle basé sur la régression de l'outcome. L'estimation des poids se fait par modélisation de la régression, comme dans l'estimation IPTW standard. L'estimateur AIPTW est un estimateur doublement robuste, ce qui signifie qu'il reste cohérent même si l'un des modèles de score de propension ou de résultat est mal spécifié [80].

$$\hat{ATE}_{AIPW} = \frac{1}{n}\sum_{i=1}^{n}\left\{\left[\frac{X_i Y_i}{\hat{\pi}(Z_i)} - \frac{(1-X_i)Y_i}{1-\hat{\pi}(Z_i)}\right] - \frac{(X_i - \hat{\pi}(Z_i))}{\hat{\pi}(Z_i)(1-\hat{\pi}(Z_i))}\right.$$
$$\left.[(1-\hat{\pi}(Z_i))\hat{\mathbb{E}}(Y_i \| X_i = 1, Z_i) + \hat{\pi}(Z_i)\hat{\mathbb{E}}(Y_i \| X_i = 0, Z_i)]\right\}$$

Contrôle de l'équilibre

La validité des estimateurs de SP pour l'estimation de l'effet du traitement repose sur leur capacité à équilibrer les covariables entre les groupes de traitement. Cette capacité peut être évaluée à l'aide de mesures telles que la différence moyenne standardisée (SMD). La SMD, pour une covariable C et un traitement binaire, est définie comme suit : $SMD = \frac{(\bar{C}_1 - \bar{C}_0)}{\sqrt{\sigma_1^2 + \sigma_0^2}}$ avec $\bar{C}_0$ et $\bar{C}_1$ la moyenne de l'échantillon des variables C et σ_0^2 et σ_1^2 la variance de l'échantillon de la variable C dans les groupes de traitement 0 et 1, respectivement. Les SMD peuvent être calculées pour chaque variable avant et après la pondération, en utilisant les moyennes et les variances pondérées. En règle générale, des SMD $< 10\%$ sont considérées comme un seuil satisfaisant [77].

Problématiques des traitements multi-niveaux

La théorie IPTW a été développée à l'origine pour les traitements binaires et sa mise en œuvre pour les traitements multi-niveaux n'a reçu que peu d'attention, ce qui peut expliquer son utilisation limitée dans la pratique [88]. Quelques exemples d'application des SP pour les traitements multi-niveaux peuvent toutefois être trouvés dans [2] qui a étudié l'association entre les bêta-agonistes à longue durée d'action, les corticostéroïdes oraux et la combinaison de ces deux produits sur l'exacerbation de l'asthme et [1] qui a étudié l'association entre plusieurs niveaux d'exercice physique et les accidents vasculaires cérébraux, l'insuffisance cardiaque, et la mortalité. Dans le cas de traitements multiples ou multi-niveaux, chaque individu possède plusieurs scores de propension. Pour un traitement binaire, chaque patient aura deux scores de propension, mais un seul est directement estimé, puisque l'autre est son complément. Pour les traitements multi-niveaux, il y a autant de scores que de groupes de traitement. Ceux-ci peuvent être obtenus à partir des prédictions d'un modèle de régression multinomiale, mais d'autres méthodes sont présentées dans les sous-sections suivantes. Il est possible d'utiliser des méthodes spécifiques aux traitements ordonnés quand on rencontre ce type de traitement, mais l'estimation des PS est plus complexe dans ce cas. Même dans la situation des traitements ordonnés, une approche multi-niveaux reste une approche valable permettant une estimation plus simple des PS au prix d'une légère perte de puissance. Lors de l'estimation de l'ATT, les traitements multi-niveaux offrent une grande variété de choix d'estimateurs causaux. Le choix du comparateur doit être guidé par l'objectif de l'étude et la question clinique sous-jacente. Pour l'ATE, il y a autant de contrastes que de comparaisons deux à deux, même s'ils ne sont pas tous pertinents pour la question de recherche.

D'autre part, l'estimation de l'équilibre des covariables dans ce cas est plus complexe, car il y a autant de SMDs que de paires de traitement. En pratique, il est possible pour chaque covariable de présenter la moyenne des SMDs sur toutes les paires pour cette covariable, ou la SMD la plus grande parmi toutes les paires.

Dans cette thèse je me concentrerai sur le cas particulier des traitements multi-niveaux en commençant par étudier comment ils sont employés dans la littérature médicale puis en appliquant ces méthodes à une problématique de recherche clinique et pour finir en comparant différentes méthodes pour leur estimations.

III. Revue systématique de la littérature sur l'application des IPTWs multi-niveaux

3.1 Présentation du travail

Comme nous l'avons vu précédemment, les méthodes d'inférence causale basées sur des données observationnelles représentent une alternative aux ECR lorsque ces derniers ne sont pas réalisables ou lorsque l'on recherche des preuves en situation réelle. L'IPTW est l'une des approches marginales les plus populaires pour tenir compte des facteurs de confusion dans les études observationnelles. La figure 3.1 présente le nombre de publications dans Pubmed faisant mention des IPTW.

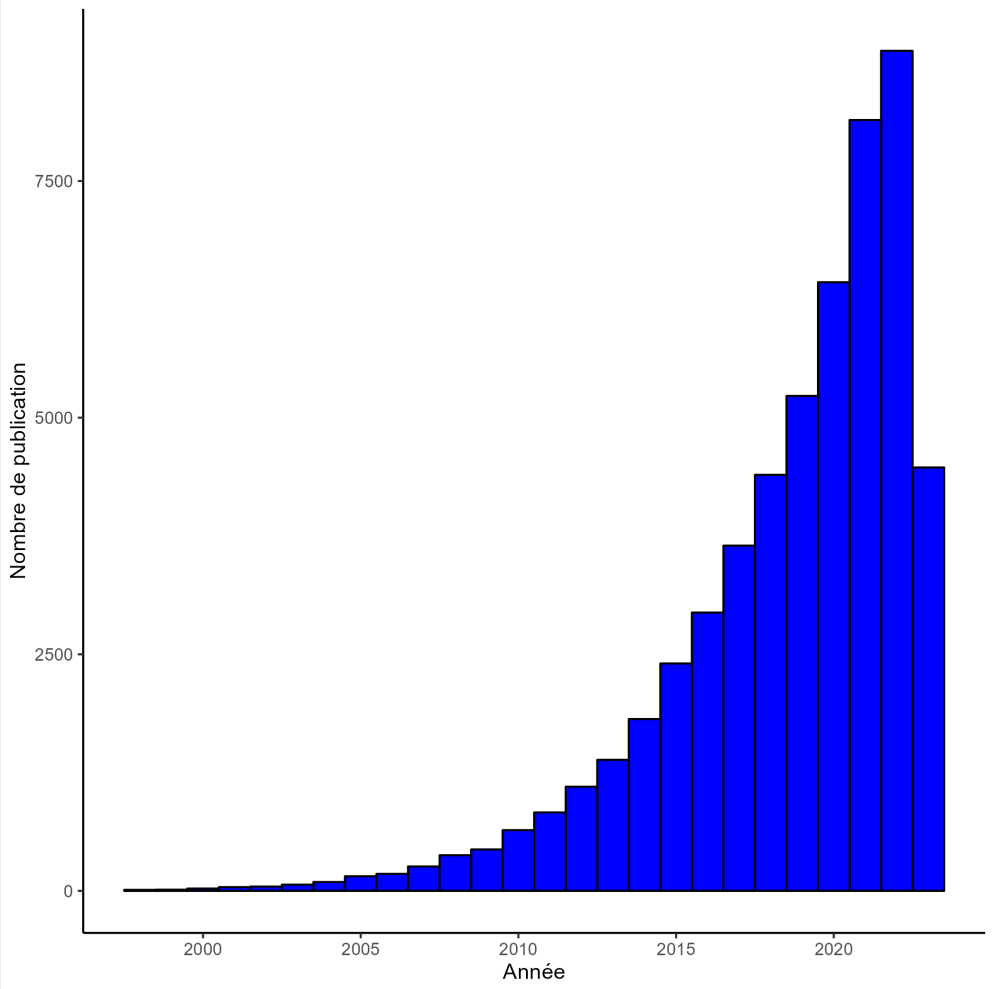

Figure 3.1: Nombre de publications dans Pubmed faisant mention des IPTW (mots clef : propensity score, Inverse Probability of Treatment Weighting, IPTWs) (01/06/2023)

Dans la recherche médicale, l'estimateur IPTW est principalement utilisé pour estimer l'effet causal moyen d'un traitement binaire, même lorsque le traitement comporte en réalité plusieurs niveaux, malgré la disponibilité d'estimateurs IPTW pour des niveaux de traitements multiples. Cependant on peut retrouver des exemples d'IPTW avec traitement multi-niveaux comme [2] ou [1]. Cela soulève des questions quant à la pertinence de l'utilisation de l'IPTW dans ce contexte. Nous avons donc procédé à une revue systématique des publications médicales faisant état de l'utilisation de l'IPTW en présence d'un traitement à plusieurs niveaux. Nos objectifs étaient d'étudier la fréquence d'utilisation et la mise en œuvre de ces méthodes dans la pratique et d'évaluer la qualité de leurs rapports. Cette revue systématique est enregistrée sur PROSPERO (CRD42022352669). Afin de dédupliquer et de sélectionner les résumés correspondant au traitement multi-niveaux nous avons utilisé Rayyan. C'est une application open-source permettant le travail collaboratif avec un outil de recherche de duplicata et qui permet d'organiser les résumés et de mettre en évidence les mots clefs d'inclusion ou d'exclusion. Sur les 5 299 articles sélectionnés dans Pubmed, Embase et Web of Science, 106 articles provenant de 17 domaines médicaux différents ont été retenus pour l'analyse finale. Le nombre de groupes de traitement variait entre 3 et 9, la grande majorité des articles (90 (84.9 %)) comportant 3 ou 4 groupes. La méthode la plus couramment utilisée pour estimer les scores de propension était la régression multinomiale (51 (48,1 %)) et les "Generalized Boosted Models" (GBMs) (48 (45,3 %)). Les covariables du modèle de pondération été rapportées dans 91 articles (85,9 %). Vingt-six articles (24,5 %) ne discutaient pas de l'équilibre des covariables après la pondération et seize articles (15,1 %) faisaient référence aux hypothèses nécessaires pour obtenir des déductions correctes. Les résultats de cette revue systématique montrent que les publications médicales utilisent rarement les méthodes IPTW pour plus de deux niveaux de traitement. Parmi les publications qui l'ont fait, la qualité des rapports n'était pas optimale, en particulier en ce qui concerne les hypothèses et la construction du modèle. L'IPTW pour les traitements multi-niveaux pourrait être appliquée plus largement dans la recherche médicale et l'élaboration de lignes directrices pratiques dans ce contexte pourrait être nécessaire pour aider les chercheurs dans leur analyse et améliorer la qualité des rapports. Cette revue de la littérature est en cours de soumission dans le Journal of Clinical Epidemiology. Dans ce travail j'ai contribué à chaque étape : l'élaboration de la requête d'extraction des abstracts, la sélection manuelle des résumés, l'élaboration et la complétion de la grille d'extraction des résultats, l'analyse des résultats ainsi que la rédaction et la soumission de l'article. La figure 3.2 est un résumé graphique.

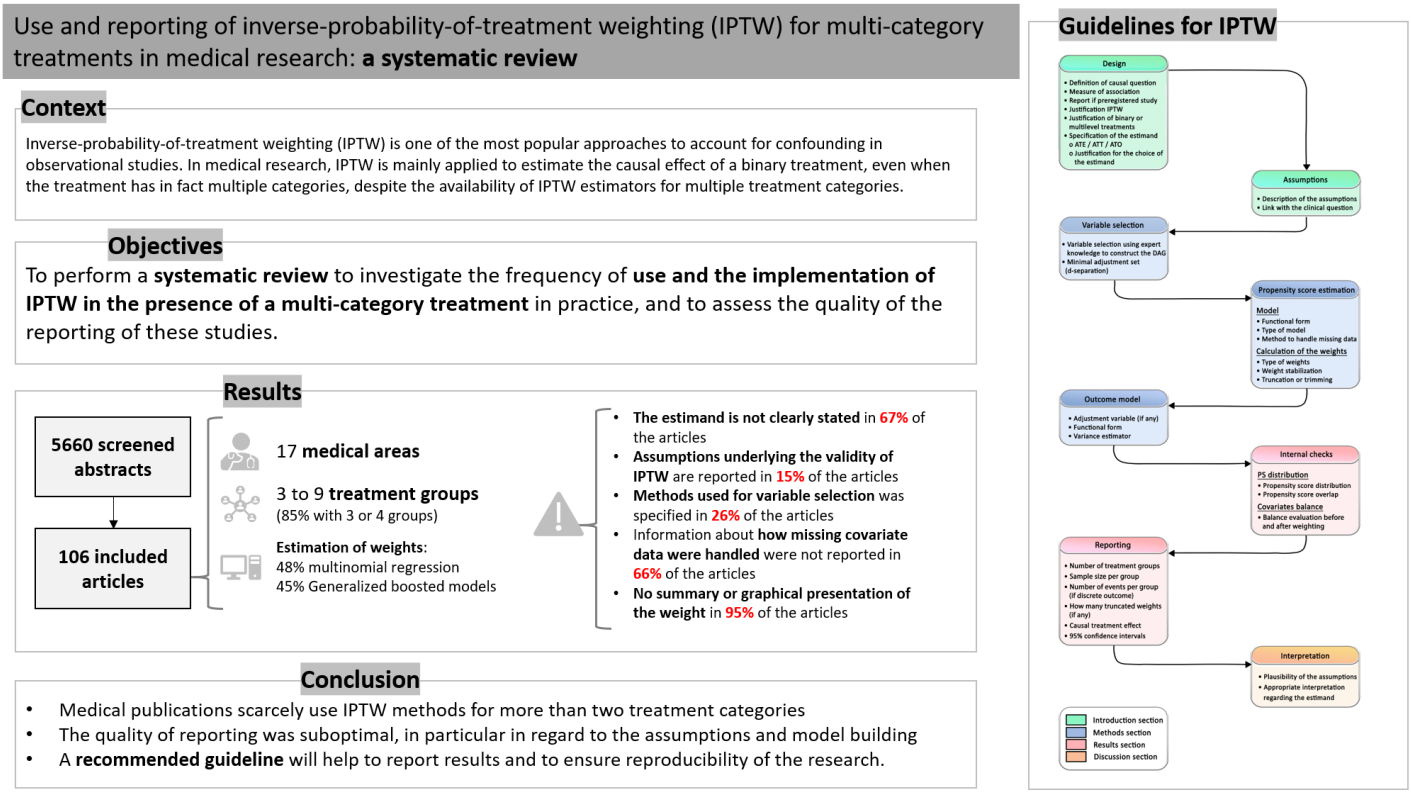

Figure 3.2: Résumé graphique des résultats de la revue systématique de la littérature sur l'application des IPTWs multi-niveaux dans la littérature médicale

3.2 Manuscrit en révision (Journal of Clinical Epidemiology)

Use and reporting of inverse-probability-of-treatment weighting (IPTW) for multi-category treatments in medical research: a systematic review.

Authors

François Bettega[1], Monique Mendelson[1], Clémence Leyrat[2], Sébastien Bailly[1]

[1]University Grenoble Alpes, Inserm, Grenoble Alpes University Hospital, HP2, 38000 Grenoble, France
[2]Department of Medical Statistics, Inequalities in Cancer Outcomes Network, London School of Hygiene and Tropical Medicine, London, UK

Fundings

SB, and FB are supported by the French National Research Agency in the framework of the "Investissements d'avenir" program (ANR-15-IDEX-02) and the "e-health and integrated care and trajectories medicine and MIAI artificial intelligence" Chairs of excellence from the Grenoble Alpes University Foundation. CL is supported by the UK Medical Research Council (MRC Skills Development Fellowship MR/T032448/1). This work has been partially supported by MIAI University Grenoble Alpes, (ANR-19-P3IA-0003)

Declarations of interest

none

Abstract

Causal inference methods for observational data represent an alternative to randomised controlled trials when they are not feasible or when real-world evidence is sought. Inverse-probability-of-treatment weighting (IPTW) is one of the most popular approaches to account for confounding in observational studies. In medical research, IPTW is mainly applied to estimate the causal effect of a binary treatment, even when the treatment has in fact multiple categories, despite the availability of IPTW estimators for multiple treatment categories. This raises questions about the appropriateness of the use of IPTW in this context. Therefore, we conducted a systematic review of medical publications reporting the use of IPTW in the presence of a multi-category treatment. Our objectives were to investigate the frequency of use and the implementation of these methods in practice, and to assess the quality of their reporting. This systematic review is registered on PROSPERO (CRD42022352669).

Using Pubmed, Embase and Web of Science, we screened 5660 articles and retained 106 articles in the final analysis that were from 17 different medical areas. The number of treatment groups varied

between 3 and 9, with a large majority of articles (90 (84.9%)) including 3 or 4 groups. The most commonly used method for estimating the weights was multinomial regression (51 (48.1%)) and generalized boosted models (48 (45.3%)). The covariates of the weight model were reported in 91 articles (85.9 %). Twenty-six articles (24.5 %) did not discuss the balance of covariates after weighting, and only 16 articles (15.1 %) referred to the assumptions needed to obtain correct inferences.

The results of this systematic review illustrate that medical publications scarcely use IPTW methods for more than two treatment categories. Among the publications that did, the quality of reporting was suboptimal, in particular in regard to the assumptions and model building. IPTW for multi-category treatments could be applied more broadly in medical research, and the application of the proposed guidelines in this context will help researchers to report their results and to ensure reproducibility of their research.

1. Introduction

Randomised controlled trials (RCTs) typically provide the highest level of evidence for causal inference. In RCTs, participants are randomised to treatment groups, which ensures that, on average, observed and unobserved participants' characteristics are balanced across groups(1). This balance is key to make causal inferences about the treatment(s) being studied. However, RCTs are not always feasible for ethical reasons (e.g. when the exposure of interest is harmful) (2,3), are very costly, and sometimes lack generalisability and transportability because of stringent inclusion criteria. Real-world data have been increasingly used as an alternative or complement to RCTs(4). Real-world observational studies present multiple advantages over RCTs: they usually include a wider diversity of patients (e.g., older patients with comorbidities are often excluded from trials). Furthermore, the follow-up period is usually longer in retrospective studies and the outcome can be directly available. Regarding prospective studies, the follow-up period can also be longer when compared to RCTs.Thus, observational studies are useful for investigating the long-term intervention effects as well as adverse events. Comparative effectiveness research using observational data has received increasing attention, and methodological work has been conducted to propose innovative designs, statistical tools and strategies for their analysis(5). In recent years, accreditation bodies have been giving more credit to such study designs, as evidenced by the validation of Covid-19 vaccines by the FDA (6).

Nevertheless, unlike RCTs, observational studies are prone to confounding. When estimating the causal effect of a treatment, estimates from unadjusted analyses are biased if risk factors differ between groups. Several causal inference methods have been proposed to account for observed confounding, and they are split into two categories: those modelling the confounders-outcome relationships (e.g. g-computation), and those modelling the confounders-treatment relationships (e.g. propensity score methods) (7). The latter has the advantage of mimicking RCTs, by recovering balance between groups on observed covariates using multiple balancing scores and allows researchers to define population based on conterfactuals(8). Inverse-probability-of-treatment weighting (IPTW), in which patients are re-weighted according to the inverse of their propensity of receiving the treatment actually received, creates a pseudo-population in which covariate distributions are similar between treatment groups. Because of its similarity with the philosophy of RCTs, IPTW is widely used for comparative effectiveness research (9).

However, IPTW is mostly implemented to estimate the causal effect of binary treatments(10), although researchers may be interested in the evaluation of several treatments (or several categories of treatment). This is the case when several treatments exist for the same indication, or when researchers want to compare the effect of two treatments and their combination (11–13). IPTW estimators have been proposed in this context (14). A review presented a methodological description of causal inference for multiple treatments (15) and suggested how to apply these methods, but it is unclear how often and how well they are used in practice. In particular, models for categorical outcomes are not always used for the estimation of the weights, but instead a series of models for binary outcomes are used. In addition, several types of modelling

strategies can be used to estimate the weights, including parametric and non-parametric approaches, but it is unclear which approaches are commonly implemented.

Therefore, we conducted a systematic review of the medical literature to describe current practice in the use of IPTW for a multi-category treatment, and the quality of reporting of these studies. Based on the findings, we propose recommendations that we hope may contribute to a better transparency in the use and reporting of IPTW in this setting.

2. IPTW and causal inference on observational data

Unlike traditional statistics, estimating associations between exposures and outcomes, causal inference refers to specific hypotheses, study designs and statistical methods to draw causal conclusions from the data (16). A specific framework, based on the concept of potential outcomes, has been developed to propose a causal language allowing the mathematical representation of causal questions. The potential outcome is what would have happened had the patient received a particular treatment (17). Patients have as many potential outcomes as there are treatment categories. In this framework, the main issue is that only the effect of one treatment on a given patient can be observed, and other potential outcomes must be estimated from the data.

Within this framework, the causal effect can be identified from the data under the assumptions of consistency, no interference, positivity, and conditional exchangeability. Under consistency, the outcome of an individual under their observed exposure is the same as their potential outcome had they received their observed intervention via the hypothetical intervention (5,18). The no interference assumption states that the treatment received by an individual has no influence on the potential outcomes of the other individuals. The positivity assumption states that, given their own characteristics, every individual has a non-zero probability of receiving any exposure categories (19). Finally, the conditional exchangeability states that, given the measured variables, the exposure and potential outcomes are independent. The validity of these assumptions is required to be able to identify the causal effect from the data, but the additional assumption of a correct specification of the analysis model(s) is needed to ensure the validity of the causal effect estimate. These assumptions, together with the assumed causal relationships between confounding factors, treatments and outcomes are at the heart of the reasoning necessary for the application of causal inference methods, and their plausibility should always be discussed when reporting results.

The IPTW is a weighting propensity score-based method (18). Regarding Rosenbaum and Rubin's definition, the propensity score is "the conditional probability of assignment to a particular treatment given a vector of observed covariates"(8). Thus, the propensity score is the probability, given the individuals' characteristics, to receive a specific treatment. When the treatment is binary, the probability of receiving the control treatment (or no treatment) is 1-PS. When the treatment has multiple categories, each individual has a propensity score for each treatment category.(15,20) The ATE can then be estimated using the IPTW estimator, in which individuals are weighted by the inverse of the treatment they actually received. Other methods, such as gradient boosting could be used (22). However, the weights can be modified to target other estimands, such as the average treatment effect on the treated (ATT) or the average treatment effect in the overlapping population (ATO (23)). The balancing ability of the weights can be checked by comparing covariate distributions between treatment groups, using, for instance,

standardised mean differences. IPTW estimates are unbiased if a good balance is achieved between groups, but residual imbalance can be addressed with augmented IPTW (AIPTW)(24,25). AIPTW combines multivariable regression and IPTW in a way that only one of the two models needs to be correctly specified to obtain unbiased estimates of the causal effect. Checking for the absence of extreme weights is also key to ensure the validity of the estimation. Weight truncation or trimming(26) is sometimes used to limit the of large weights to the analysis (19). Large weights may increase the variance of the estimates and lead to biased estimates in some instance. Such methods are used to increase the precision of the estimate, but may lead to residual confounding and change the target population, making the causal interpretation more difficult. Another important consideration when using IPTW estimators is variance estimation which must account for two aspects of the estimation. The uncertainty in propensity score estimation and the intraindividual correlation introduced via weighting should be captured in the outcome model to avoid misestimating the variance. Estimators based on the delta method (27) and non-parametric bootstrap have been proposed.

In summary, for the validity of IPTW and AIPW estimates we must ensure that: (i) the identification assumptions for causal inference are plausible, (ii) the estimator targets the correct estimands, (iii) the propensity score model is correctly specified and (iv) appropriate variance estimators are used. It is therefore very important for these elements to be reported when publishing the findings of a study analysed using IPTW.

Methods using IPTW for more than two treatment categories face additional technical challenges. Because of data scarcity or strong indication bias, the plausibility of the positivity assumption may be less likely when the number of treatment categories increases. With multiple treatment categories, it is necessary to question the choice of treatment reference for the estimand. Therefore, our systematic review focused on this setting, where a correct implementation and a clear reporting are required to ensure validity and reproducibility.

3. Methods

3.1. Inclusion and exclusion criteria

We performed literature searches on PubMed, Web of Science and Embase from 01/01/2011 to 27/06/2021, for peer-reviewed articles published in English. The systematic review included all publications in medical research involving human participants using an IPTW estimator with multiple treatment categories for the primary analysis. The review was limited to applied research and did not focus on methological papers. The study is registered on PROSPERO (CRD42022352669).

There was no restriction in terms of research area, study design, type of intervention or outcome. Exclusion criteria were: non-medical research, methodological studies (e.g. simulation study, reviews.), non-original research articles (e.g. letters), articles using IPTW for subgroup or sensitivity analyses.

3.2. Search strategy

The search strategy screened articles whose abstracts, title or keywords contained the followings: inverse probability weight, inverse probability of treatment weight, augmented inverse propensity

weight, as well as the associated acronyms (IPW, IPTW, AIPW, AIPTW, AIPWE). The generic term “propensity score” was not considered to improve the specificity of the algorithm, as previously done (9). We also conducted a reverse search for articles citing McCaffrey et al. (2013)(14), Yoshida et al. (2018)(28), or Li and Li (2019)(29). The search strategy is given in Appendix 1. The abstracts were then manually and independently screened for eligibility by FB and SB.

3.3. Extracted information

The extracted information was divided into nine fields: 1) description of the studies, 2) estimand and measure of association, 3) assumptions, 4) covariate selection, 5) propensity score estimation, 6) covariate balance, 7) analysis model, 8) software and statistical packages and 9) good research practice (Table 1).

3.4. Data extraction procedure

A standardised, pre-piloted form was used to extract data from the included studies and tested on 10 randomly selected studies. The full text of the eligible studies identified after screening were retrieved and the data extracted by FB and SB. Any disagreement over the eligibility or extracted items was resolved through discussion with CL if an agreement could not be reached.

4. Results

4.1. Screening and inclusion

The search yielded a total of 5,299 articles (after the removal of duplicates), which were screened based on abstracts. From these, 303 were identified for full-text screening and 106 articles fulfilled the inclusion criteria and were included (complete list given in Appendix 2). The selection process is summarised in Figure 1.

4.2. Description of the included studies

Multi-category treatments were observed in 17 different medical fields, but three medical specialties accounted for half of included studies: 35 studies (47.3%) were either in cardiology (32 studies: 30.2%), 9 in nephrology (8.5%) and 7 in Gastroenterology (6.6%). Almost all the included articles (103, 97.2 %) were cohort studies.

A variety of wording was used to refer to the method applied. IPTW and weighted regression were the two most common wording encountered (n=73 (68.9%) of and n=15 studies (14.2%), respectively). The majority of studies (n=100(94.3%)) justified the use of IPTW, the main reason being confounding adjustment.

The number of treatment groups ranged between 3 and 9 with a majority of studies comparing three groups (n=59, 56.7%). Forty-Four articles (41.5%) had between 4 and 6 groups, and 1 (1%) article included 9 groups.

The total sample sizes ranged from 65 to 12,700,000 with a median of 161,583 participants. The minimum total sample size ranged from 12 to 638,905 with a median of480.A summary of the results is presented in Figure 2 and Table 3.

4.3. Estimand and measure of association

In two-thirds of the articles (71 (67%)), the estimand was not clearly stated and had to be determined from the calculation of the weights, when this was available. The majority of the studies (90 (85%))

focused on estimating the ATE, (5 (4.7%)) focused on estimating the ATT and 1 (0.9%) article estimated the ATO. It was impossible to identify the estimand in 10 (9.4%) articles.

For the measure of association, 58 studies (54.7%) reported hazard ratios, 22 (20.8%) reported odds ratios, 8 (7.6 %) reported risk ratios, 7 (6.6%) reported difference and 9 (8.5%) articles used other measures. It was impossible to determine the measure of association in 2 (1.9%) articles.

4.4. Assumptions

Only sixteen articles (15.1%) explicitly mentioned the assumptions underlying the validity of IPTW and discussed their context-specific plausibility.

4.5. Covariate selection and handling of missing data

The variables included in the weight model were mentioned in 91 (85.9%) articles. The method used for variable selection was specified in 28 (26.4%) of the articles: 14 (50%) used evidence from the literature, 11 (40.3%) articles used automated selection ~~9 (40.9%) used evidence from the literature~~ and 3 (10.7%) used a DAG to inform the selection. In most articles (70 (66%)), the way missing covariate data was handled was not reported. In the articles which did report this, 16 (40%) used a complete case analysis, 8 (20%) used multiple imputation and 12 (30%) used other ad hoc methods.

4.6. Propensity score estimation

A majority of the studies (51 (48.1%)) used multinomial regression to estimate the weights, followed by Generalized Boosted Models (GBMs) (48 (45.3%)), and one (0.9%) used covariate balancing propensity score. The remaining 6 studies (5.7%) did not specify the method. One hundred and one (95.3%) of the articles did not present any summary or graphical representation of the weights, 4 (3.8%) articles presented a histogram of the weights and 1 (0.9%) another representation. Among the included articles, 23 (21.7%) articles explicitly stated whether or not they stabilised weights. Among these articles, 16 used (69.6%) stabilised weights. Trimming or weight truncation were mentioned in 24 (22.6%) articles. Of these 24 articles, only 18 studies actually did perform trimming or truncation and the other 6 studies just mentioned trimming or weight truncation without applying it. One potential explanation for not applying weight or trimming in these 6 papers is that extreme weights were under 10.

4.7. Assessing covariate balance

Among the reviewed articles, 26 (24.5%) did not mention whether the covariate balance after weighting was investigated. The most frequent method for estimating equilibrium was the SMD (i.e., 60 articles, 56.6%), 4 (3.8%) the Kolmogorov-Smirnov distance, 12 (11.3%) used p-values, 3 (2.8%) used graphs and 1 (0.9%) reported the population standard bias.

4.8. Analysis model

IPTW was implemented in 83 (78.3%) articles, and AIPTW in 23 (21.7%). These estimators were applied to a wide range of outcomes: time-to-event 64 (60.4%), binary 20 (18.9%), continuous 10 (9.4%),

count 6 (5.6%), categorical 5 (4.7%) and ordinal 1 (0.9%). The weighted outcome models used to estimate the causal effect of the treatments were diverse and depended mainly on the type of outcomes. The results are summarised in Table 3 and in Figure 2. Methods for variance estimation were reported in 18 (17%) studies, 13 (72.2%) studies used a robust estimator, 4 (22.2%) used non-parametric bootstrap and 1 used uncorrected variance.

4.9. Software and statistical packages

The three main software packages were: R 39 (37.1%), SAS 35 (33.3%) and STATA 18 (17.1%). Four (3.8%) of the articles used a combination of R, SAS and STATA and in 4 (3.8%) of the articles the statistical software was not clearly identified. Among the articles using a programming language other than R, 7 (6.6%) used R as a secondary programming software for the "TWANG" package in order to estimate the weights using GBMs.

4.10. Good research practice

Only 8 (7.6%) studies had previously registered a protocol. Only 4 (3.8%) study reported that the code was available and 3 (2.8%) proposed an access to all or part of the data. Finally, 9 studies (8.5 %) referred to the STROBE statement.

5. Discussion

This systematic review aimed to collect detailed information from published observational studies to assess how IPTW methods with a multi-category treatment are applied in medical research. As we focused the review on practical implementation in applied studies, we excluded methodological papers. From 5,660 screened articles, only 106 (3.4%) focussed on a multi-category treatment and the reasons for choosing IPTW over other approaches were rarely given. Moreover, the plausibility assumptions underpinning the validity of IPTW were discussed in very few studies. Overall, the quality of the reporting was poor, with key elements missing, thus compromising the interpretability and generalizability of the results.

In the majority of the studies, the estimand was not reported. This is a concern as the estimand determines the way results are interpreted. In addition, estimation of the ATT relies on less stringent assumptions.

The implementation and reporting were also often inadequate. Indeed, one of the most striking results from this review is the low frequency of studies reporting the assumptions for the identification of causal effects, and their plausibility, which is however crucial to make causal claims. This result was already observed in a previous study focusing on binary treatments (9). Interestingly, a few studies discussed the plausibility of modelling assumptions (e.g., proportionality of hazards), but failed to report the assumptions for causal inference.

This could be explained by a lack of practical guidelines for the reporting of these studies. Although these assumptions are not empirically verifiable, the plausibility of the assumption of no interference can be determined based on the knowledge of the clinical setting. The plausibility of the consistency assumption may be ensured with a precise definition of the exposure of interest. While the assumption of conditional exchangeability is often questionable in observational studies, the elaboration of a DAG from expert knowledge followed by an application of d-separation rules may dramatically reduce the risk of confounding. Finally, the plausibility of the positivity assumption can be explored from the distribution of the propensity score and the absence of extreme weights (24) , although the absence of extreme weights does not guarantee that the positivity assumption holds.".

Empirical positivity may be a challenge when estimating causal effects for more than two treatment groups as indication guidelines may be more specific when multiple treatments are available for the same condition and the sample size may be smaller in each group increasing the chance of violation of the positivity assumption in the sample. In this systematic review some studies analysed up to 9 groups, and did not investigate violations of the positivity assumption.).

The propensity score model was most often a multinomial regression model. This can probably be explained by the fact that this method is a direct extension of the logistic regression model used in IPTW for binary treatments, is simple to implement in standard statistical software, and relatively inexpensive in terms of computational power. Generalized Boosted Models (GBMs), a machine learning method based on regression trees, is featured prominently in this literature review. Although these methods are more computationally expensive, there is an easy-to-use implementation in the "TWANG" package (30) with a tutorial for causal effect estimation in the case of multi-category treatments (14). An advantage of GBMs is that they are non-parametric and therefore do not require

the specification of a functional form. Furthermore, in the TWANG implementation, the stopping rule for the GBM algorithm is based on a balance metrics for the covariates, thus maximising the balance across treatment groups. However, the method to compute the weights was not always reported, which compromises the transparency and reproducibility of the results.

The validity of propensity score methods depends on the ability of the scores to balance treatment groups with respect to the covariates (24). In our review, balance was assessed in most studies, but a few used p-values, that are not recommended because they strongly depend on the sample size. Balance should be assessed before and after weighting for instance by presenting standardized mean differences for each covariate and for each pair of treatments or by presenting the mean or maximum SMD per variable across all treatment comparisons. However, there is currently no consensus on the way to assess balance for multiple treatments and further work is needed to provide practical guidelines.

In terms of analysis model, the type of model was generally well reported, but not the variance estimator. This is very important because the estimated variance must account for (i) the correlation introduced via weighting (ii) the uncertainty around the propensity score estimates. In practice, many authors used sandwich estimators for (i) but issue (ii) is often overlooked, despite available estimators (27) including for multi-category treatments (29) and the validity of non-parametric bootstrap.

Guidelines for the application of IPTW for binary treatments exist (9,31), and we would like to propose steps for their reporting in the case of multi-category treatments. These recommendations are summarised in Figure 3.

6. Conclusion

Causal inference approaches using IPTW are largely applied in medical research but multi-category treatment remains scarcely used. This systematic review highlighted the suboptimal reporting quality of studies in this context, in particular for assumptions and model building. The application of practical guidelines, as proposed here, is needed to help researchers improve the presentation of their results to ensure a better understanding of their methods and the reproducibility of their results.

References

1. Senn S. Seven myths of randomisation in clinical trials. Stat Med. 2012 Dec;32(9):1439–50.

2. Ware JH, Hamel MB. Pragmatic Trials — Guides to Better Patient Care? N Engl J Med. 2011 May;364(18):1685–7.

3. Benson K, Hartz AJ. A Comparison of Observational Studies and Randomized, Controlled Trials. N Engl J Med. 2000 Jun;342(25):1878–86.

4. Kennedy-Martin T, Curtis S, Faries D, Robinson S, Johnston J. A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. Trials. 2015 Nov;16(1).

5. Hernan MA. A definition of causal effect for epidemiological research. J Epidemiol Ampmathsemicolon Community Health. 2004 Apr;58(4):265–71.

6. Pawlowski C, Lenehan P, Puranik A, Agarwal V, Venkatakrishnan AJ, Niesen MJM, et al. FDA-authorized mRNA COVID-19 vaccines are effective per real-world evidence synthesized across a multi-state health system. Med. 2021 Aug;2(8):979-992.e8.

7. Smith MJ, Mansournia MA, Maringe C, Zivich PN, Cole SR, Leyrat C, et al. Introduction to computational causal inference using reproducible Stata, R, and Python code: A tutorial. Stat Med. 2021 Oct;41(2):407–32.

8. ROSENBAUM PR, RUBIN DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41–55.

9. Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Stat Med. 2015 Aug;34(28):3661–79.

10. Ali MS, Groenwold RHH, Belitser SV, Pestman WR, Hoes AW, Roes KCB, et al. Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: a systematic review. J Clin Epidemiol. 2015 Feb;68(2):122–31.

11. Bettega F, Leyrat C, Tamisier R, Mendelson M, Grillet Y, Sapène M, et al. Application of Inverse-Probability-of-Treatment Weighting to Estimate the Effect of Daytime Sleepiness in Patients with Obstructive Sleep Apnea. Ann Am Thorac Soc. 2022 Sep;19(9):1570–80.

12. Carr DC, Willis R, Kail BL, Carstensen LL. Alternative Retirement Paths and Cognitive Performance: Exploring the Role of Preretirement Job Complexity. Meeks S, editor. The Gerontologist. 2019 Jul;60(3):460–71.

13. Rannanheimo PK, Tiittanen P, Hartikainen J, Helin-Salmivaara A, Huupponen R, Vahtera J, et al. Impact of Statin Adherence on Cardiovascular Morbidity and All-Cause Mortality in the Primary Prevention of Cardiovascular Disease: A Population-Based Cohort Study in Finland. Value Health. 2015 Sep;18(6):896–905.

14. McCaffrey DF, Griffin BA, Almirall D, Slaughter ME, Ramchand R, Burgette LF. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. Stat Med. 2013 Mar;32(19):3388–414.

15. Lopez MJ, Gutman R. Estimation of Causal Effects with Multiple Treatments: A Review and New Ideas. Stat Sci. 2017 Aug;32(3):432–54.

16. Pearl J. An Introduction to Causal Inference. Int J Biostat. 2010 Jan;6(2).

17. Vandenbroucke JP, Broadbent A, Pearce N. Causality and causal inference in epidemiology: the need for a pluralistic approach. Int J Epidemiol. 2016 Jan;45(6):1776–86.

18. Robins JM, Hernán MÁ, Brumback B. Marginal Structural Models and Causal Inference in Epidemiology. Epidemiology. 2000 Sep;11(5):550–60.

19. Cole SR, Hernan MA. Constructing Inverse Probability Weights for Marginal Structural Models. Am J Epidemiol. 2008 Jul;168(6):656–64.

20. Imbens G. The Role of the Propensity Score in Estimating Dose-Response Functions. National Bureau of Economic Research; 1999.

21. Friedman J. The Elements of Statistical Learning Data Mining, Inference, and Prediction: Data Mining, Inference, and Prediction. Springer-Verlag New York; 2009. 745 p.

22. Hastie T, Tibshirani R, Friedman J. Boosting and Additive Trees. In: The Elements of Statistical Learning. New York, NY: Springer New York; 2009. p. 337–87. (Springer Series in Statistics).

23. Greifer N, Stuart EA. Choosing the Causal Estimand for Propensity Score Analysis of Observational Studies. 2021 Jun;

24. Hernan MA, Robins JM. Causal Inference: What If. Taylor & Francis Group; 2019. 352 p.

25. Funk MJ, Westreich D, Wiesen C, Stürmer T, Brookhart MA, Davidian M. Doubly Robust Estimation of Causal Effects. Am J Epidemiol. 2011 Mar;173(7):761–7.

26. Crump RK, Hotz VJ, Imbens GW, Mitnik OA. Dealing with limited overlap in estimation of average treatment effects. Biometrika. 2009 Jan;96(1):187–99.

27. Williamson EJ, Forbes A, White IR. Variance reduction in randomised trials by inverse probability weighting using the propensity score. Stat Med. 2013 Sep;33(5):721–37.

28. Yoshida K, Solomon DH, Haneuse S, Kim SC, Patorno E, Tedeschi SK, et al. Multinomial Extension of Propensity Score Trimming Methods: A Simulation Study. Am J Epidemiol. 2018 Dec;188(3):609–16.

29. Li F, Li F. Propensity score weighting for causal inference with multiple treatments. Ann Appl Stat. 2019 Dec;13(4):2389–415.

30. Ridgeway G, McCaffrey DF, Morral AR, Cefalu M, Burgette LF, Pane JD, et al. Toolkit for Weighting and Analysis of Nonequivalent Groups: A Tutorial for the R TWANG Package. RAND Corporation; 2022.

31. Williamson E, Morley R, Lucas A, Carpenter J. Propensity scores: From naïve enthusiasm to intuitive understanding. Stat Methods Med Res. 2011 Jan;21(3):273–93.

Table 1: presentation of extracted information

Fields	Extracted information
Description of the included studies	Area of research Study registration number Study design Wording used to refer to IPTW Justification of the method Presence and appropriateness of sample size calculation Type of analysis model Sample size in each treatment groups Number of treatment groups Nature of the comparator Nature of the outcome
Estimand and measure of association	Estimand (ATE, ATT, ATO) Measure of association (HR, OR, RR, Other)
Assumptions	Mention of assumptions Mention of use of STROBE checklist
Covariate selection	Presence of DAG Variable include in weight model Method used for variable selection Method used for missing values
Propensity score estimation	Model used for propensity scores Type of weight Summary of weights Weight stabilisation Weight trimming or truncation
Assessing covariate balance	Covariate balance Methods used for assessing balance
Analysis model	Method used (IPTW, AIPTW) Variance estimation method
Software and statistical packages	Name of softwares and packages used
Good research practice	Protocol Open data Open code

Table 2: Type of outcome reported and outcome models

Outcome type	Outcome model	N (%)
Continuous	Linear	9 (8.5%)
Binary	Negative binomial	1 (0.9%)
	Logistic	18 (17%)
Time-to-event	Cox	63 (59.4%)
	Linear	1 (0.9%)
Categorical	Multinomial	5 (4.7%)
Count	Poisson	6 (5.7%)
Ordinal	Multinomial	1 (0.9%)

Table 3: Summary of the main results

Elements of IPW method		N (%)
Estimand	ATE	90(85.0%)
	ATT	5(4.7%)
	ATO	1(0.9%)
	Unknown	10(9.4%)
Estimand definition	Guessed from the weights	71(67%)
	Explicitly written	25 (23.6%)
	Not reported	10(9.4%)
Measure of association	HR	58(54.7%)
	OR	22(20.8%)
	RR	8(7.6%)
	Other	9(8.5%)
	Unknown	2(1.9%)
Assumptions	Mention of assumptions (yes)	16(15.1%)
	Mention of STROBE (yes)	9(8.5%)
Covariate selection	Variables included in the weight model	91(85.9%)
Method used for variable selection	From the literature	14(13.2%)
	Automated selection	11 (11.4%)
	From the DAG	3(2.8%)
	Not specified	78(73.6%)
Method used for missing values	Complete-case	16(15.1%)
	Multiple imputation	8(7.6%)
	Adjustment	1(0.9%)
	Group mean	1(0.9%)
	Other	9(8.5%)
	Unknown	70(66%)
Summary of weights	Histograms	4(3.8%)
	Other	1(0.9%)
	Unknown	101(95.2%)
Weight stabilisation	Yes	16(15.1%)
	No	7(6.6%)
	Unknown	83(78.3%)
Model used for propensity scores	Multinomial	51(48.1%)
	GBM	48(45.3%)
	Other	1(0.9%)
	Unclear	6(5.7%)
Methods used for assessing balance	SMD	60(56.6%)
	KS	4(3.8%)
	Graph	3(2.8%)
	P-values	12(11.3%)
	Other	1(0.9%)
	Unknown	26(24.5%)
Analysis model	IPTW	83(78.3%)
	AIPTW	23 (21.7%)
Variance estimation method	Robust	13(12.3%)
	Bootstrap	4 (3.8%)
	Uncorrected	1 (0.9%)
	Unknown	88(83%)

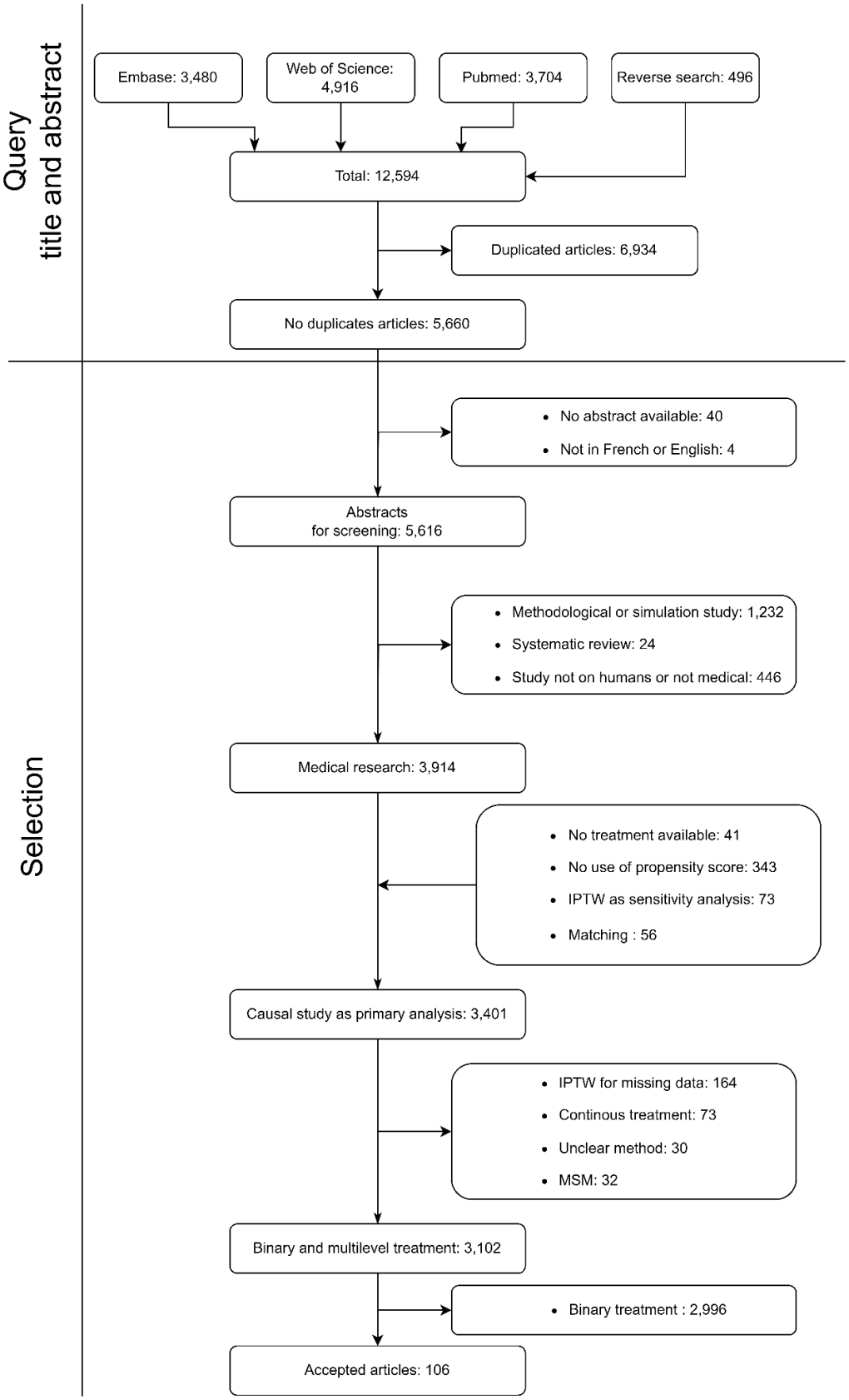

Figure 1: Flow chart of the systematic review process.

MSM: Marginal structural model, IPTW: inverse probability of treatment weight.

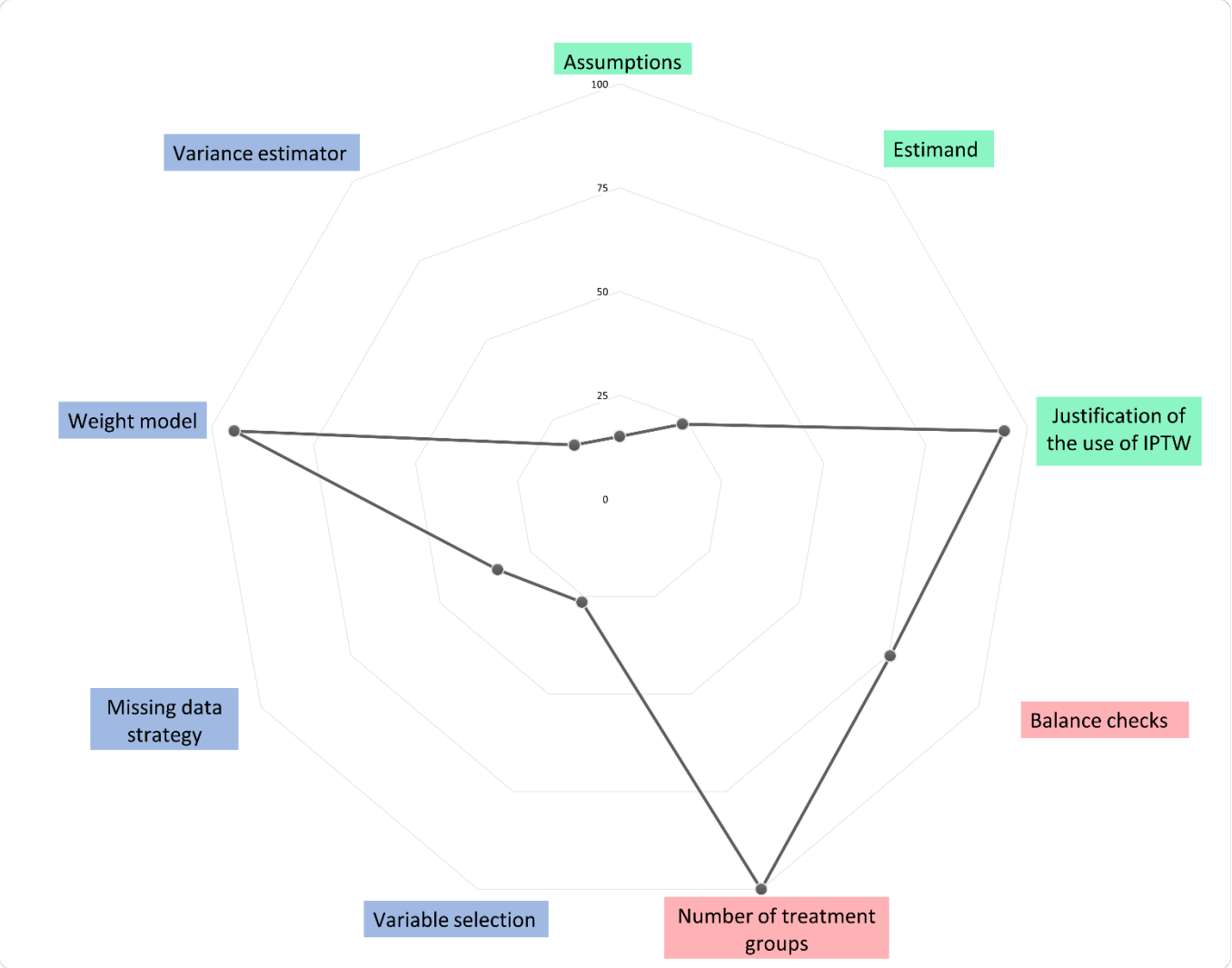

Figure 2: Summary of the main results

Results are presented in percentage. In green: point related to the introduction section, in blue: points related to the methods section and in red points related to the results.

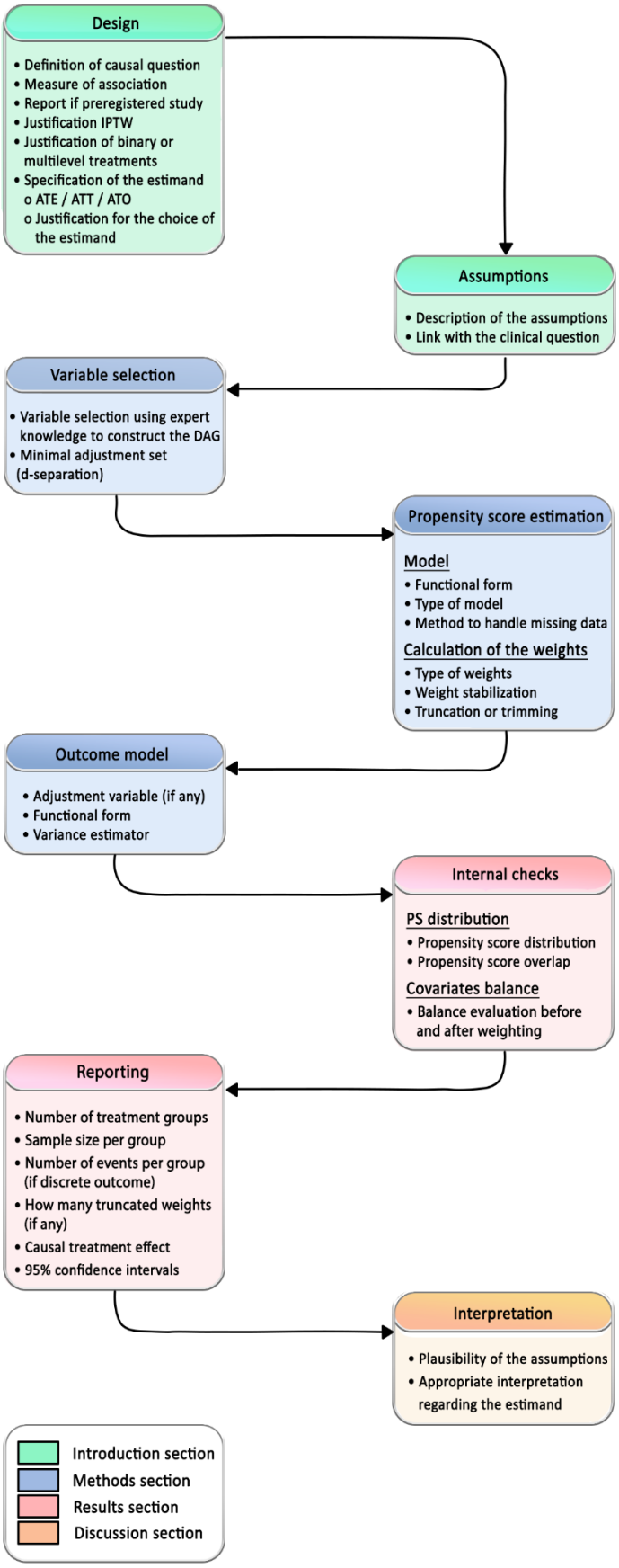
Design
• Definition of causal question
• Measure of association
• Report if preregistered study
• Justification IPTW
• Justification of binary or multilevel treatments
• Specification of the estimand
o ATE / ATT / ATO
o Justification for the choice of the estimand
Assumptions
• Description of the assumptions
• Link with the clinical question
Variable selection
• Variable selection using expert knowledge to construct the DAG
• Minimal adjustment set (d-separation)
Propensity score estimation
Model
• Functional form
• Type of model
• Method to handle missing data
Calculation of the weights
• Type of weights
• Weight stabilization
• Truncation or trimming
Outcome model
• Adjustment variable (if any)
• Functional form
• Variance estimator
Internal checks
PS distribution
• Propensity score distribution
• Propensity score overlap
Covariates balance
• Balance evaluation before and after weighting
Reporting
• Number of treatment groups
• Sample size per group
• Number of events per group (if discrete outcome)
• How many truncated weights (if any)
• Causal treatment effect
• 95% confidence intervals
Interpretation
• Plausibility of the assumptions
• Appropriate interpretation regarding the estimand
Introduction section
Methods section
Results section
Discussion section

Figure 3: Guideline for causal inference approaches

This guideline proposes a list of the main points to follow regarding the introduction (green windows), methods (blue windows), results (red windows) and discussion (red windows) sections to report a study based on weighted approaches in general and using multilevel treatment in particular.

IPTW: inverse probability of treatment weight, ATE: Average treatment effect, ATT: Average treatment effects on the treated, ATO: average treatment effect in the overlap population, DAG: directed acyclic graph, PS: propensity score.

The algorithm below will be used to search PubMed, Web of sciences and Embase from 01/01/2001 to 28/02/2018 for papers published in English only.

Pubmed :

(((("IPW"[Title/Abstract] NOT ("IPW"[Title/Abstract] AND ("rmrp"[Title/Abstract] OR "rmst"[Title/Abstract] OR "ftx"[Title/Abstract] OR "lncrna"[Title/Abstract])))OR "AIPW"[Title/Abstract] OR "IPTW"[Title/Abstract] OR "AIPTW"[Title/Abstract] OR "AIPWE"[Title/Abstract] OR "Inverse probability weight*"[Title/Abstract] OR "inverse propensity weight*" [Title/Abstract] "inverse probability of treatment weight*"[Title/Abstract] OR "augmented inverse propensity weight*"[Title/Abstract] OR "Inverse-probability weight*"[Title/Abstract] OR "inverse-probability of treatment weight*"[Title/Abstract] OR "augmented inverse-propensity weight*"[Title/Abstract] OR "Marginal Structural Model*"[Title/Abstract] OR "inverse-propensity weight*" [Title/Abstract]) NOT ("matching*"[Title/Abstract]) AND (2011:2021[pdat])) AND (English[Language])

Web of sciences:

(TS=((IPW NOT (IPW AND (rmrp OR rmst OR ftx OR lncrna))) OR AIPW OR IPTW OR AIPTW OR AIPWE OR Inverse probability weight* OR inverse probability of treatment weight* OR augmented inverse propensity weight* OR Inverse$probability weight* OR inverse$probability of treatment weight* OR augmented inverse$propensity weight* OR Marginal Structural Model* OR inverse$propensity weight*) NOT TS=(matching*)) AND LA=(English) AND DT=(Article) AND PY=(2011-2021) NOT WC= (GEOCHEMISTRY GEOPHYSICS OR PHYSICS MATHEMATICAL OR ENERGY FUELS OR TRANSPORTATION SCIENCE TECHNOLOGY OR ASTRONOMY ASTROPHYSICS OR OCEANOGRAPHY OR AGRICULTURAL ECONOMICS POLICY OR GREEN SUSTAINABLE SCIENCE TECHNOLOGY OR ENGINEERING ELECTRICAL ELECTRONIC OR PHYSICS APPLIED OR PHYSICS MULTIDISCIPLINARY OR REPRODUCTIVE BIOLOGY OR AUTOMATION CONTROL SYSTEMS OR MECHANICS OR COMPUTER SCIENCE THEORY METHODS OR ENGINEERING MECHANICAL OR MATHEMATICS OR ECOLOGY OR TRANSPORTATION OR MATERIALS SCIENCE MULTIDISCIPLINARY OR CHEMISTRY MULTIDISCIPLINARY OR OPERATIONS RESEARCH MANAGEMENT SCIENCE OR REHABILITATION OR WATER RESOURCES OR ENGINEERING CHEMICAL OR METEOROLOGY ATMOSPHERIC SCIENCES OR ENGINEERING ENVIRONMENTAL OR ENGINEERING MULTIDISCIPLINARY OR INSTRUMENTS INSTRUMENTATION OR ECONOMICS OR BIOPHYSICS OR BIOCHEMISTRY MOLECULAR BIOLOGY OR SOCIAL SCIENCES MATHEMATICAL METHODS OR CELL BIOLOGY OR GEOSCIENCES MULTIDISCIPLINARY OR TELECOMMUNICATIONS OR ENVIRONMENTAL STUDIES OR COMPUTER SCIENCE INTERDISCIPLINARY APPLICATIONS OR GEOLOGY OR FORESTRY OR CHEMISTRY PHYSICAL OR BIOCHEMICAL RESEARCH METHODS OR ENGINEERING INDUSTRIAL OR GEOGRAPHY PHYSICAL)

Embase :

('ipw':ab,ti NOT ('ipw':ab,ti AND ('rmrp':ab,ti OR 'rmst':ab,ti OR 'ftx':ab,ti OR 'lncrna':ab,ti)) OR 'aipw':ab,ti OR 'iptw':ab,ti OR 'aiptw':ab,ti OR 'aipwe':ab,ti OR 'inverse probability weight*':ab,ti OR 'inverse probability of treatment weight*':ab,ti OR 'augmented inverse propensity weight*':ab,ti OR "inverse$probability weight*":ab,ti OR "inverse$probability of treatment weight*":ab,ti OR "augmented inverse$propensity weight*":ab,ti OR 'marginal structural model*':ab,ti OR "inverse$propensity weight*":ab,ti) NOT 'matching*':ab,ti AND ([article]/lim OR [article in press]/lim) AND [2011-2021]/py

IV. Application de l'IPTW multi-niveaux pour l'évaluation de l'effet de l'observance à la PPC sur la somnolence diurne

4.1 Présentation du travail

Dans cette étude, nous montrons comment deux méthodes d'inférence causale peuvent être appliquées à des données observationnelles pour l'estimation de l'effet de différentes plages d'adhésion à la PPC sur la somnolence diurne mesurée par le score de somnolence d'Epworth (ESS). L'ESS est un score obtenu à partir de 8 questions dont les réponses sont cotées de 0 à 3. Le score final, compris entre 0 et 24 évalue leur niveau de somnolence, une somnolence diurne normale conduit à un score entre 0 et 6. Les données ont été collectées à partir d'une vaste cohorte prospective d'observation française de patients souffrant de SAOS. Quatre groupes d'observance à la PPC ont été pris en compte (0-4 ; 4-6 ; 6-7 et 7-10 heures par nuit). La régression multivariable, l'IPTW et l'IPTW-RA ont été utilisées pour évaluer l'impact du niveau d'adhésion à la PPC sur la somnolence diurne. Cette étude a porté sur 9 244 patients souffrant de SAOS et traités par PPC. Un Graphe Orienté Acyclique (DAG) a été développé pour représenter les liens causaux supposés entre l'observance à la PPC et la somnolence diurneFigure 4.1.

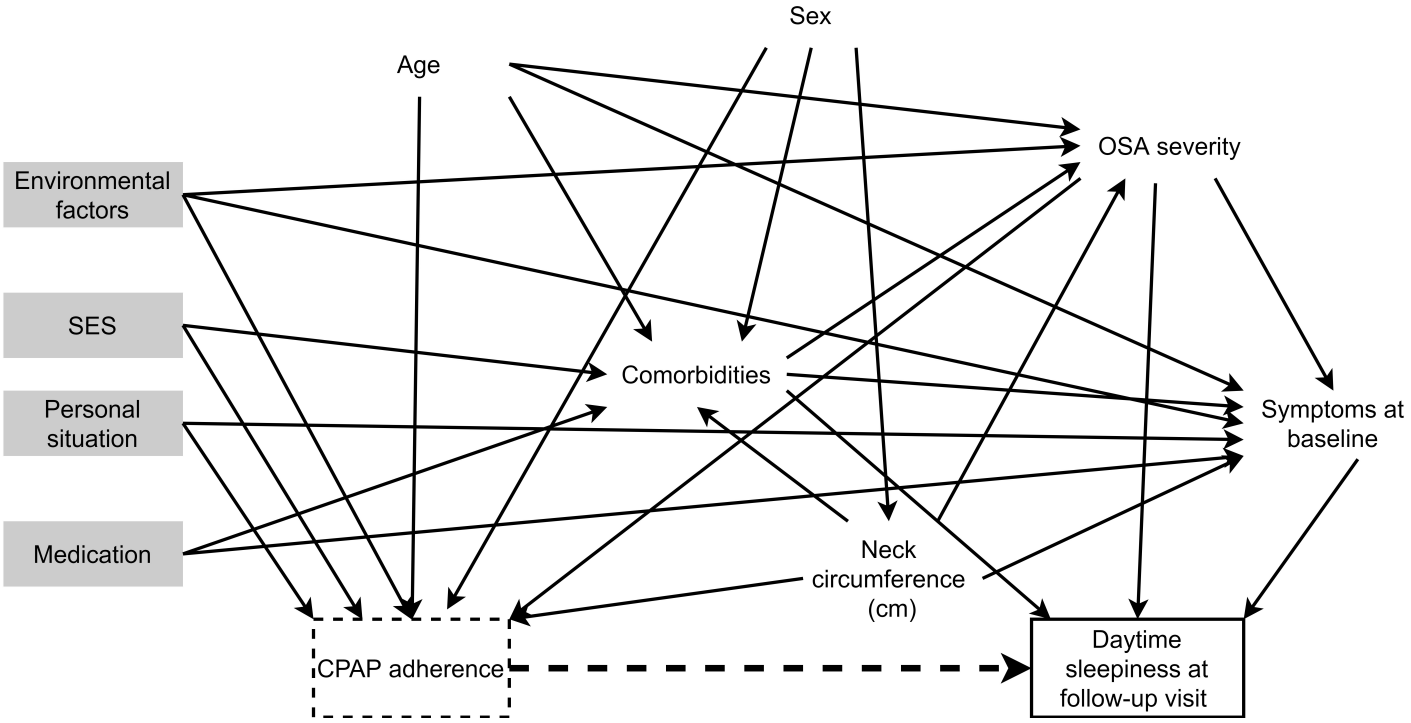

Figure 4.1: Graphe orienté acyclique

Graphe orienté acyclique pour la relation causale entre l'adhésion à la PPC à plusieurs niveaux et la somnolence diurne résiduelle. La flèche droite en pointillé indique la relation causale à l'étude ; les flèches pleines indiquent les relations connues. PPC : Pression positive continue ; SAOS : Apnées obstructives du sommeil. ; SES : Statut socio-économique. La situation personnelle regroupe : le mode de vie, la situation matrimoniale, les enfants. Fond gris : facteurs de confusion non observés Cadre pointillé : exposition ; Cadre plein : résultat Symptômes au départ : somnolence au volant, fatigue matinale, maux de tête matinaux, troubles de la libido, transpiration nocturne, fatigue mesurée par l'échelle de Pichot et SaO2 nocturne moyenne Comorbidités : dépression mesurée par l'échelle de dépression de Pichot et syndrome des jambes sans repos. source : [12]

L'ESS initial moyen était de 11 ($\pm$5,2) avec une réduction moyenne de 4 points ($\pm$5,1) sous traitement par PPC. Dans l'ensemble, l'effet causal de l'observance à la PPC sur la somnolence diurne a été mis en évidence principalement dans le groupe ayant une faible observance à la PPC (0-4h) par rapport au groupe ayant une forte observance à la PPC (7-10h). Il n'y a pas de différence si l'on considère un niveau plus élevé d'adhésion à la PPC (>4h). Nous avons montré que l'IPTW et l'IPTW-RA peuvent être facilement mises en œuvre pour répondre aux questions concernant les effets causaux en utilisant des données d'observation lorsque des essais randomisés ne peuvent pas être menés. Les deux méthodes donnent une interprétation causale directe au niveau de la population et permettent d'évaluer la prise en compte appropriée des facteurs de confusion mesurés. La figure Figure 4.2 issue de l'article présente les différents résultats obtenus avec chaque méthode.

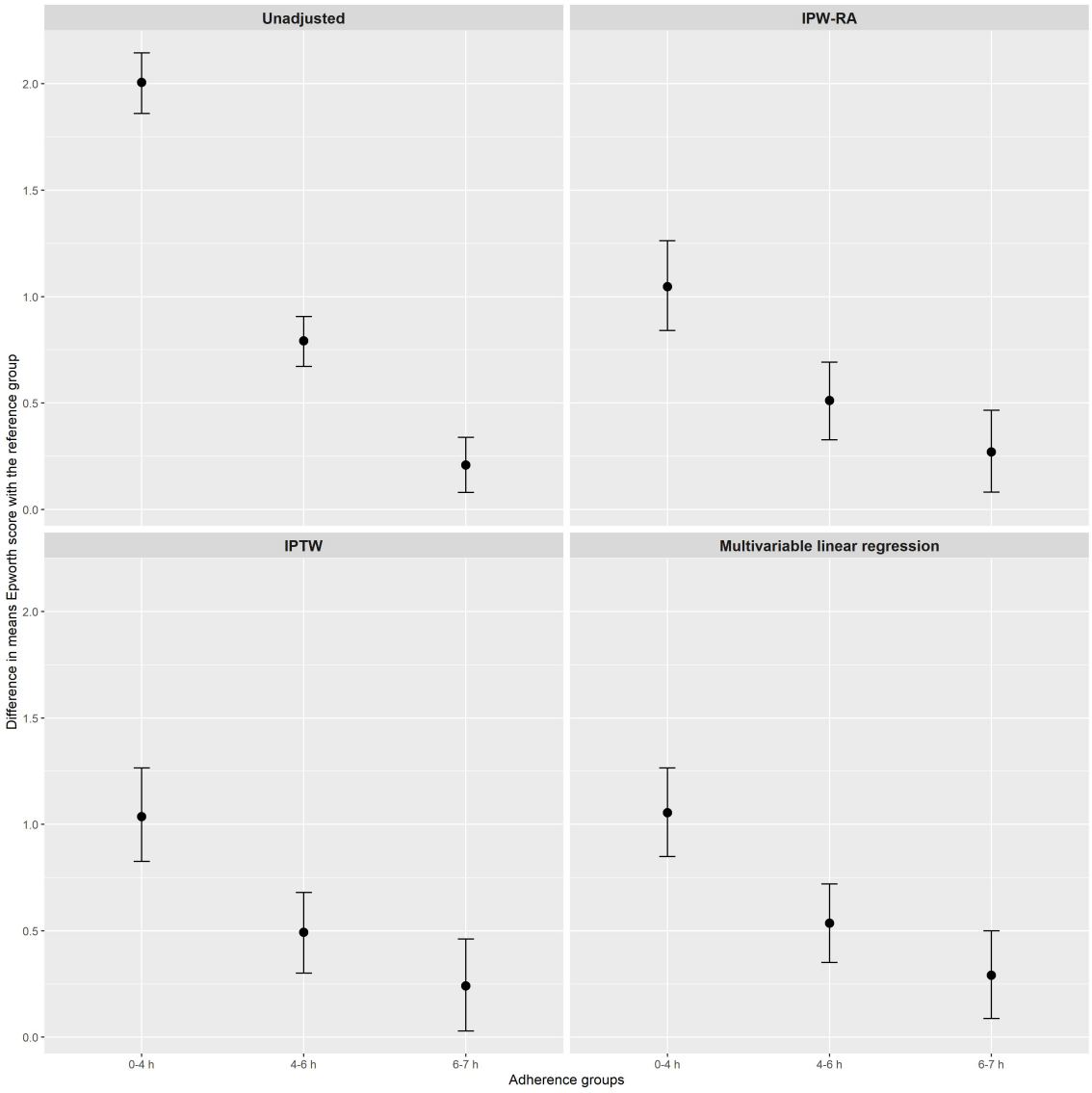

Figure 4.2: Différence moyenne du score d'Epworth entre chaque groupe d'observance et le groupe de référence utilisant différentes méthodes

Chaque point représente la différence moyenne du score d'Epworth entre chaque groupe d'observance et le groupe de référence (7-10h). Les barres verticales représentent les intervalles de confiance à 95 % de ces estimations. source : [12]

Ce travail a fait l'objet d'une publication dans le journal Annals of american thoracic society. Le choix d'un article applicatif dans une revue clinique a été motivé par le souhait de rendre les IPTWs multi-niveaux plus largement accessibles aux cliniciens. J'ai contribué à chaque étape de ce travail : la définition de la question de recherche, le nettoyage des données, l'analyse des données et la rédaction du manuscrit.

4.2 Manuscrit publié (annals of american thoracic society)

ORIGINAL RESEARCH

Application of Inverse-Probability-of-Treatment Weighting to Estimate the Effect of Daytime Sleepiness in Patients with Obstructive Sleep Apnea

François Bettega[1], Clémence Leyrat[2], Renaud Tamisier[1], Monique Mendelson[1], Yves Grillet[3], Marc Sapène[4], Maria R. Bonsignore[5], Jean Louis Pépin[1], Michael W. Kattan[6], and Sébastien Bailly[1]

[1]Grenoble Alpes University, Grenoble Alpes University Hospital, HP2, Inserm, Grenoble, France; [2]Department of Medical Statistics, Inequalities in Cancer Outcomes Network, London School of Hygiene and Tropical Medicine, London, United Kingdom; [3]Private Practice Sleep and Respiratory Disease Centre, Nouvelle Clinique Bel Air, Bordeaux, France; [4]Private Practice Sleep and Respiratory Disease Centre, Valence, France; [5]Respiratory Medicine, PROMISE Department, University of Palermo and Istituto per la Ricerca e l'Innovazione Biomedica Consiglio Nazionale delle Ricerche (IRIB-CNR), Palermo, Italy; and [6]Department of Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic, Cleveland, Ohio

ORCID IDs: 0000-0002-9736-5289 (F.B.); 0000-0002-4097-4577 (C.L.); 0000-0002-2179-4650 (S.B.).

Abstract

Rationale: Continuous positive airway pressure (CPAP), the first line therapy for obstructive sleep apnea (OSA), is considered effective in reducing daytime sleepiness. Its efficacy relies on adequate adherence, often defined as >4 hours per night. However, this binary threshold may limit our understanding of the causal effect of CPAP adherence and daytime sleepiness, and a multilevel approach for CPAP adherence can be more appropriate.

Objectives: In this study, we show how two causal inference methods can be applied on observational data for the estimation of the effect of different ranges of CPAP adherence on daytime sleepiness as measured by the Epworth Sleepiness Scale (ESS).

Methods: Data were collected from a large prospective observational French cohort for patients with OSA. Four groups of CPAP adherence were considered (0–4, 4–6, 6–7, and 7–10 h per night). Multivariable regression, inverse-probability-of-treatment weighting (IPTW), and inverse propensity weighting with regression adjustment (IPW-RA) were used to assess the impact of CPAP adherence level on daytime sleepiness.

Results: In this study, 9,244 patients with OSA treated by CPAP were included. The mean initial ESS score was 11 (±5.2), with a mean reduction of 4 points (±5.1). Overall, there was evidence of the causal effect of CPAP adherence on daytime sleepiness which was mainly observed between the lower CPAP adherence group (0–4 h) compared with the higher CPAP adherence group (7–10 h). There are no differences by considering higher level of CPAP adherence (>4 h).

Conclusions: We showed that IPTW and IPW-RA can be easily implemented to answer questions regarding causal effects using observational data when randomized trials cannot be conducted. Both methods give a direct causal interpretation at the population level and allow the assessment of the appropriate consideration of measured confounders.

Keywords: causal inference; inverse probability weight; daytime sleepiness; sleep apnea

(Received in original form September 9, 2021; accepted in final form April 4, 2022)

Supported by grants from the French National Research Agency in the framework of the "Investissements d'avenir" program (ANR-15-IDEX-02; J.L.P., S.B., and F.B.) and the "e-health and integrated care and trajectories medicine and MIAI artificial intelligence" Chairs of Excellence from the Grenoble Alpes University Foundation; the United Kingdom Medical Research Council (MRC Skills Development Fellowship MR/T032448/1; C.L.), and MIAI University Grenoble Alpes (ANR-19-P3IA-0003).

Author Contributions: F.B., C.L., and S.B. contributed to the study design, analysis, and interpretation of the data. F.B., C.L., S.B., M.M., R.T., M.R.B., J.L.P, and M.W.K. contributed substantially to the study design, data interpretation, and writing of the manuscript. R.T., Y.G., and M.S. were responsible for acquisition of data, contributed to the discussion, and reviewed the manuscript. S.B., C.L., and J.L.P. were responsible for the study concept and design, supervised the study, and critically revised the manuscript. S.B. is the guarantor of this work and, as such, had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Correspondence and requests for reprints should be addressed to Sébastien Bailly, Pharm.D., Ph.D., Laboratoire EFCR - CHU de Grenoble Rondpoint de la Chantourne - CS10217, Cedex 9, 38043 Grenoble, France. E-mail: sbailly@chu-grenoble.fr.

This article has an online supplement, which is accessible from this issue's table of contents at www.atsjournals.org.

Ann Am Thorac Soc Vol 19, No 9, pp 1570–1580, Sep 2022

DOI: 10.1513/AnnalsATS.202109-1036OC
Internet address: www.atsjournals.org

Obstructive sleep apnea (OSA) is a major health concern with multiorgan consequences and significant economic cost and social burdens (1). OSA is defined by recurrent complete or partial obstruction of the upper airway during sleep. It has been estimated that more than 1 billion men and women aged 30–69 years worldwide suffer from moderate to severe OSA (2). OSA frequently co-occurs with comorbidities such as obesity, diabetes, hypertension, or other cardiovascular and metabolic diseases and has a major impact on quality of life (1, 3–5). Continuous positive airway pressure (CPAP), the first-line therapy for OSA, is highly effective in terms of symptom improvement, even in minimally symptomatic patients who initially complain of fatigue and nonrestorative sleep (6). Previous studies have demonstrated that adequate adherence to CPAP treatment is the prerequisite for reducing symptoms. CPAP also has an effect on quality of life by improving daytime sleepiness. Indeed, CPAP use is associated with a significant decrease in the Epworth Sleepiness Scale (ESS) score (7), and a majority of initially sleepy patients (with ESS scores >10) experience a significant improvement in their ESS scores after CPAP initiation (8). The effect size of the response and the dose–response relationship have mainly been established by meta-analyses summarizing existing randomized clinical trials (9, 10).

Randomized controlled trials (RCTs) are considered as the gold standard for causal inference in medical research, providing the highest level of evidence. Unfortunately, RCTs are not always feasible, for ethical, logistical, or financial reasons (11, 12). Furthermore, RCTs often have strict inclusion criteria and typically include younger patients with few or no comorbidities, thus limiting the generalizability of the findings (13). Current data emerging from RCTs might not represent the true impact of CPAP for reducing subjective sleepiness in unbiased real-life populations (14–16).

When RCTs cannot be implemented or when real-world evidence is needed, observational studies, such as cohorts or registries, contain a wealth of data for causal inference. However, unlike RCTs, observational studies are prone to confounding bias due to the absence of randomization, meaning that treatment groups might be unbalanced. Therefore, specific statistical methods have been proposed to address this issue to target the causal nature of the relationship between multilevel exposures and outcomes. There are several ways to account for the effect of measured confounding factors. In medical research, the most standard approaches are multivariable regression and standardization for the estimation of marginal effects. In medical research, the most standard approaches are multivariable regression and standardization for the estimation of marginal effects. Propensity score (PS) approaches are a first methodological way to replicate covariate balance associated with randomized trials (with the difference that PSs only achieve balance on measured variables) and minimize selection bias. This was explored by Keenan and colleagues (17) for balance CPAP adherence. However, PS approaches result in a decrease in the overall sample size. Finally, the inverse-probability-of-treatment weighting (IPTW) estimator developed within the counterfactual theory has been increasingly used (18).

When a study population is large enough, PS-based methods, such as IPTW, and multivariable regression lead to similar results (19). Unlike multivariable regression, IPTW allows the comparison and evaluation of covariate balance after weighting and leads to directly interpretable marginal effects (and not conditional effects). Most of the IPTW theory has been developed for binary exposures, and its implementation for multilevel exposures has been given little attention, which can explain its limited use in practice (20).

In the present study, we aim to describe two weighted methods for the estimation of causal effects—namely, IPTW and inverse propensity weighting with regression adjustment (IPW-RA)—and to illustrate their implementation and interpretation for the analysis of the causal effect of CPAP adherence on the change in ESS score from baseline in a large national prospective cohort.

Methods

Patients with a diagnosis of sleep apnea by polygraphy or polysomnography, who were older than 18 years of age, and who were treated by CPAP were included from the "Observatoire Sommeil de la Fédération de Pneumologie" (OSFP) database, a national French registry for sleep apnea. Patients with missing values for CPAP adherence or ESS scores either at the diagnostic visit or at the first follow-up visit were excluded from our study.

We used a multiple imputation by chained equations method to replace missing data values in the dataset under certain assumptions about the data missingness mechanism (i.e., assuming that the data are missing at random). Details on imputation are available (*see* Supplementary Material 1, as well as Table E1 for the number of missing values, in the online supplement).

The exposure (i.e., average objective adherence) came from CPAP device downloads during the first follow-up visit by the pulmonologist. To evaluate adherence as a multilevel treatment, patients were divided into four equally sized adherence groups on the basis of average nightly CPAP adherence as follows: CPAP use *1*) between 0 and 4 hours by night; *2*) between 4 and 6 hours by night; *3*) between 6 and 7 hours by night; and *4*) between 7 and 10 hours by night. The last group was used as the reference.

The outcome, daytime sleepiness, was assessed using the self-administered ESS questionnaire, which leads to a score between 0 and 24, with 24 being the maximum drowsiness. The patient's reported ESS score was considered at two time points: at the diagnostic visit and at the first follow-up visit.

To assess the impact of CPAP adherence on the ESS score, all potential confounders must be accounted for (i.e., all variables that can be related with the ESS score and CPAP adherence must be considered). For example, among patients with OSA, those who were younger, very elderly, or female were much more likely to present lower adherence to CPAP (21). Thus, age, obesity, and sex have an impact on the outcome (daytime sleepiness) and, therefore, must be considered as confounding factors; otherwise, the estimate of the causal effect between the exposure and response variables would be biased. To ensure that these potential confounding factors are accounted for, the IPTW estimator can be applied as explained in the next section of this article. Moreover, the IPW-RA can be used to increase the robustness of standard IPTW to a misspecification of the weight model. This method is a double robust estimator of average treatment effect (ATE) (22). In both methods, two steps are considered: *1*) A weight is computed for each patient; and *2*) these weights are subsequently used in a regression model to predict the ATE on the ESS score. The ATE is defined as the average difference

between the potential outcomes for every individual in the population. It is the contrast between two hypothetical worlds. When the exposure has multiple levels, there are as many ATEs as there are possible contrasts.

Counterfactual Theory

Contrary to traditional statistics, which aim to assess associations between an exposure and an outcome, causal inference refers to specific assumptions and study design to be able to draw causal conclusions from the data (23). In the potential outcome framework (the framework developed for causal inference), the potential outcome refers to what would have happened if a patient had received a treatment (24). To evaluate the effect of several treatments, it is necessary to establish the effect of each treatment on each patient. There are as many potential outcomes as there are treatments. However, the observation of each potential outcome is nearly impossible, because each patient usually receives only one treatment.

One of the main problems with observational data is the fact that the exposure is not independent of the other variables. To address this issue, the IPTW is based on the creation of a pseudopopulation in which the exposure variable (i.e., CPAP adherence) becomes independent of the potential outcomes given the covariates.

The pseudopopulation is the result of assigning a weight to each participant that is, informally, proportional to the participant's probability of receiving his or her own exposure.

Assumptions

To use a pseudopopulation to measure the effect of CPAP adherence on daytime sleepiness without bias, we need to verify four assumptions to identify the causal effect—*1*) consistency, *2*) noninterference, *3*) conditional exchangeability, and *4*) positivity—and one assumption about the estimation of the causal effect: no model misspecification. The four assumptions are summarized in Table 1.

The consistency assumption is often stated such that an individual's potential outcome under his or her observed exposure is exactly the same outcome as it would have been if the patient had received his or her observed intervention by means of the hypothetical intervention (18, 25). This assumes that observing is the same as intervening. The treatment needs to be precisely defined to ensure that observed treatment and hypothetical treatment use in causal framework lead to the same outcomes for a given patient.

Noninterference assumption states that an individual's treatment has no influence on the potential outcomes of other individuals. An example of a violation of this assumption is to consider vaccines because vaccinating one individual may affect the disease status of other individuals.

Conditional exchangeability refers to the assumption of no unmeasured confounders. In causal inference, all joint predictors of exposure and outcomes must be accounted for. Thus, all variables related with treatments and outcomes, (i.e., in this case, all variables linked with CPAP adherence and daytime sleepiness variations) have to be included. In our study, this is illustrated by causal directed acyclic graphs (DAG) (Figure 1) (26). Information on the design of the DAG and the links between the variables is provided in Supplementary Material 2 in the online supplement. This assumption is empirically untestable and can only be verified through expert knowledge (27). However, we can assume that most important confounders have been properly included in the OSFP database because of expert medical knowledge and the selection of variables.

Positivity states that, given their own characteristics, every individual has a nonzero probability of receiving any exposure level (27). For example, the existence of formal contraindications to one of the treatments evaluated among the observed population is a violation of the positivity assumption, because the patients with the contraindication could not be exposed to the contraindicated treatment.

In addition to these four assumptions, we need a correct model specification, i.e., the unknown probabilities for a patient to belong to each treatment group knowing all confounders is modeled through a correctly specified model. For instance, modeling an exponential relationship using a linear model is a violation of this assumption. This assumption also states that all confounding variables and their real functional forms are used to fit the model. Double robust estimators such as IPW-RA can help address this assumption.

Weight Estimation and Balancing Properties

In our study, we predicted for the patient, *i*, the probability of belonging to the adherence group noted *A*, given their confounding factors *AGE* and *BMI*, and their observed values (*age* and *bmi*). We used age and body mass index, BMI, for illustrative purposes in this example:

$$\mathrm{P}(\mathrm{A_i} = \mathrm{a_i} \mid AGE_i = age_i, BMI_i = bmi_i, \ldots).$$

Then, each individual was weighted according to the inverse of the probability of

Table 1. Table of causality assumptions

Assumptions	Definitions	Can It Be Tested from the Data?
Consistency	The outcome of an individual under their observed exposure is the same as their potential outcome had they received their observed intervention by means of the hypothetical intervention.	No
Non-interference	The treatment received by an individual has no influence on the potential outcomes of the other individuals.	No
Conditional exchangeability	Given the measured variables, the exposure and potential outcomes are independent; i.e., all joint predictors of exposure and outcomes are accounted for.	No, but the investigation of the balance between exposure groups after weighting may give an indication of the plausibility of this assumption for the measured variables (but not the unmeasured variables)
Positivity	Given their own characteristics, every individual has a nonzero probability of receiving any exposure level.	Yes, by investigating the range of the estimated propensity score values

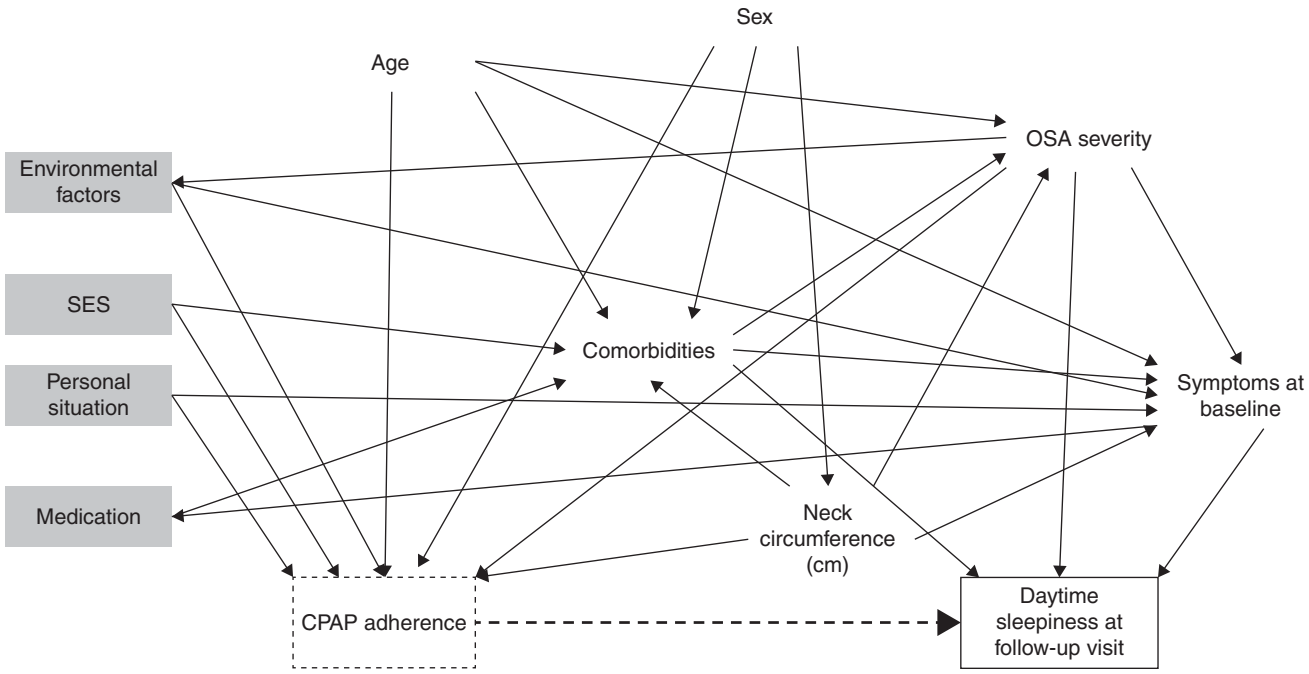

Figure 1. Causal directed acyclic graph for the relation between multilevel continuous positive airway pressure (CPAP) adherence and residual daytime sleepiness under CPAP. Dotted arrow indicates causal relation under investigation. Solid arrows indicate known relations. Personal situation regroups: lifestyle, marital status, children. Gray background indicates unobserved confounders. Dotted frame indicates exposure. Solid frame indicates outcome. Symptoms at baseline: sleepiness at the wheel, morning tiredness, morning headaches, libido disorder, night sweating, fatigue measured by Pichot's depression scale, and mean nocturnal arterial oxygen saturation (SaO_2). Comorbidities: depression measured by Pichot's depression scale and restless legs syndrome. OSA = obstructive sleep apnea; SES = socioeconomic status.

receiving the treatment they actually received (i.e., their adherence group). The probability, for each individual, of belonging to their treatment group was computed using a multinomial logistic regression:

$$IPW_i = \frac{1}{\mathrm{P}(\mathrm{A_i} = \mathrm{a_i} \mid AGE_i = age_i, BMI_i = bmi_i, \ldots)}.$$

An example of weight assessment is illustrated in Figure 2. To minimize the bias–variance compromise, a weight truncation was performed. All weights that exceed a specified threshold were each set to that threshold. Several thresholds for weight truncation were investigated from the 1st–99th to the 25th–75th percentiles, and the threshold that offered the best bias–variance ratio was chosen.

To create a model able to estimate the causal effect, we needed to select a set of variables. We use the recommendations of Lefebvre and colleagues on model specification (28). These authors recommended the inclusion of all risk factors (confounders or not) and avoid including pure predictors of exposure, also known as instrumental variables, in the treatment model. According to these rules, for this study, candidate variables are all variables that are not instrumental variables, selected by using univariable linear regressions for the outcome, with less than 60% of missing values, without collinearity. To control for the type I error rate, selection was performed on a subgroup consisting of 20% patients stratified by adherence group. These patients were used only to choose the variables and were removed for weight and final models. To ensure that no major confounding factors have been overlooked by our procedure, the choice of variables to be included, and their relationship to each other, was made in collaboration with OSA clinical experts (R.T., J.L.P., and M.R.B.).

Four approaches were implemented and compared for the estimation of the outcome: *1*) mean comparison using a *t* test; *2*) a multivariable regression; *3*) using the IPTW estimator; and *4*) using the double robust IPW-RA. The simple *t* test does not account for confounding and, therefore, is not appropriate for the analysis of nonrandomized studies. We report these results as a benchmark for adjusted methods. Final results are expressed as ATE (95% confidence interval) from the reference group (7–10 h).

For the IPTW, the final weighted model was adjusted for the exposure. For the IPW-RA, the final weighted model was adjusted for the exposure and all the confounders included in the weight model, to account for potential remaining imbalances in confounders between groups. The IPW-RA combines the strengths of the IPTW and multivariable regression: Confounders are adjusted for in a multiple regression model, also weighted by the inverse of the PS. By doing so, the causal effect estimate will be unbiased if either the weight model or the outcome model is correctly specified. The weights are the IPTW weights, and the covariates in the outcome model can be any covariate still unbalanced despite the weighting. A simple IPW-RA allows the estimation of a conditional causal effect, but it is possible to marginalize on the distribution of the covariates to estimate a marginal effect (the ATE).

As model-based standard errors are incorrect because they do not account for the uncertainty in PS estimation, we chose to bootstrap weight and treatment effect estimations. This allowed us to estimate confidence intervals based on percentiles for the treatment effect without making

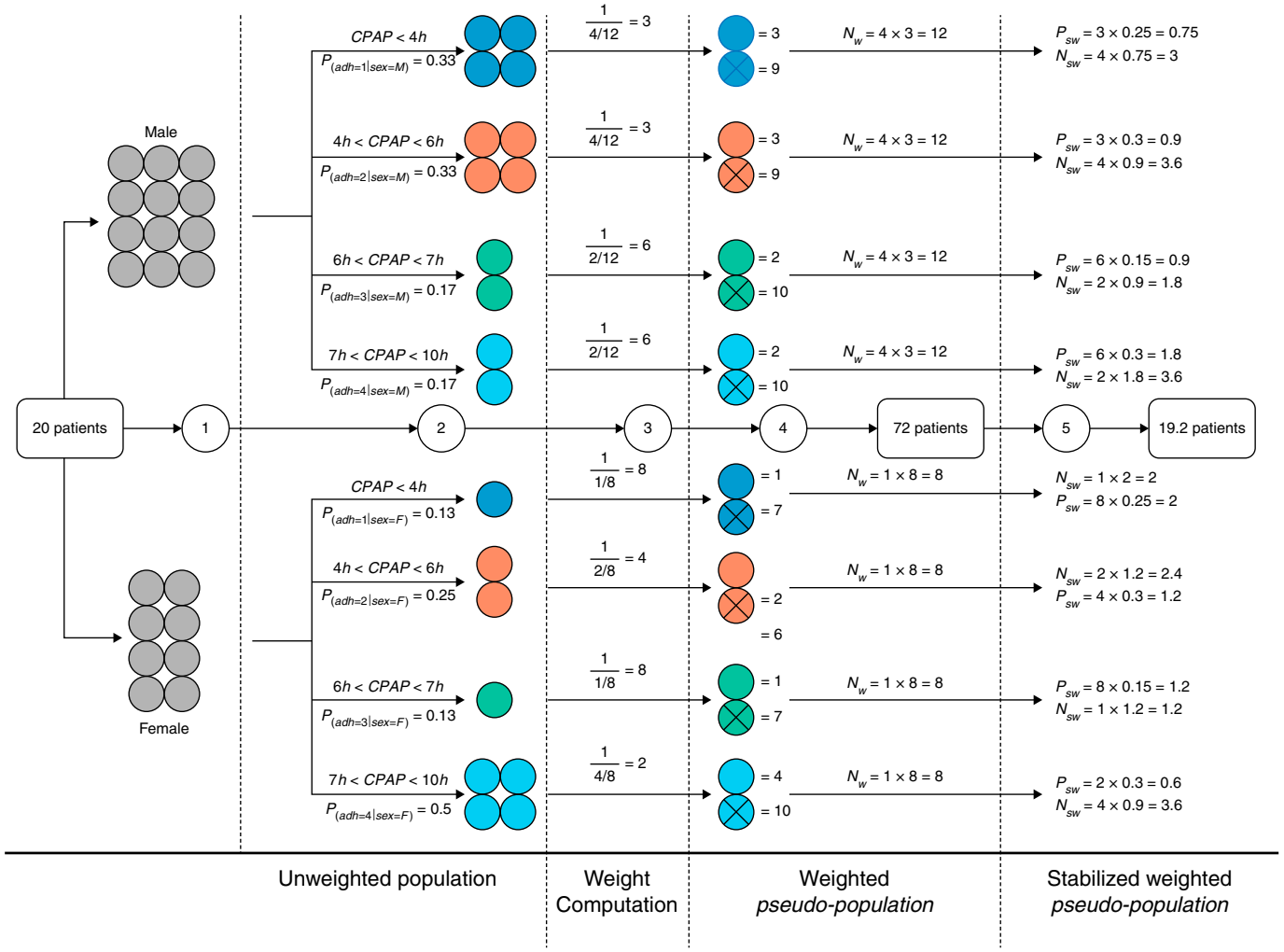

Figure 2. Illustration of the estimation the inverse-probability-of-treatment weight (IPTW) with a categorical exposure. *1*) The population may be divided into two groups according to sex (12 females and 8 males). *2*) Within each group, patients are divided according to their adherence to continuous positive airway pressure (CPAP). For each individual in each subgroup, the probability of belonging to their actual adherence group ($P_{CPAP|sex}$; i.e., probability of treatment given the sex) may be estimated from empirical proportions. *3*) From this probability, we compute the IPTW: $\frac{1}{P_{CPAP|sex}}$. *4*) We use this weighting to create a pseudopopulation. In this pseudopopulation, individuals with a high probability of belonging to a treatment group are downweighted; in contrast, individuals with a low probability of belonging to a treatment group are upweighted. The pseudopopulation encompasses both factual and counterfactual observations. In this pseudopopulation, all adherence groups are exchangeable, and it is possible to compute directly the difference for a specific outcome. *5*) A common issue with pseudopopulation is that individuals with a very low propensity score (very close to 0) will end up with a huge weight, resulting in extremely large pseudopopulation and potentially making the weighted estimation unstable. A common way to address this issue is the stabilized weight, which uses the marginal probability of treatment instead of 1 in the weight numerator resulting for a patient belonging in the first adherence group in $\frac{P_{CPAP=1}}{P_{CPAP|sex}}$.

assumptions about the distribution of the parameters. We evaluated the possibility of integrating missing value imputation into the bootstrap, but this procedure was time consuming and increased the time needed far too much. To keep a reasonable execution time, we therefore chose to impute the missing values before the bootstrap (29).

We performed a complementary analysis using the same model and method with morning fatigue as outcome. This analysis is presented in Supplementary Material 3 in the online supplement.

All statistical analyses were performed using R Statistical Software (version 4.0.2). The tests were performed at a 5% significant level.

Results

Population

From the OSFP database, 9,244 patients were included in the study. The included patients were mainly men ($n = 6,492$; 70.2%) with a mean age of 57 years (SD ± 12.4) and mean body mass index was 32 kg/m^2 (SD ± 6.9). The mean apnea–hypopnea index (AHI) was 41 events per hour (±20.4), and 6,510 (70%) patients had severe OSA.

The mean observance was 5 hours, 35 min by night, and patients were divided in four groups according to their average CPAP use by night as follows: *1*) CPAP use between 0 hours and 4 hours, $n = 1,977$ (21.4%); *2*) CPAP use between 4 hours and

6 hours, $n = 3{,}519$ (38.1%); *3*) CPAP use between 6 hours and 7 hours, $n = 2{,}023$ (21.9%); and *4*) CPAP use between 7 hours and 10 hours, $n = 1{,}725$ (18.7%). For more information, differences across all variables and subgroups are presented in Table 2.

Overall, the unadjusted mean initial ESS was 11 (±5.2). There was a mean reduction in the ESS score of 4 (±5.1) at the follow-up visit under CPAP treatment. This reduction was different according to the CPAP adherence groups: The smallest difference was observed in the adherence group with CPAP use for 0–4 hours, with a mean ESS score that varied from 11 (±5.3) to 8 (±4.7), resulting in a difference of 3 (±5), which was lower compared with the three other adherence groups. In the adherence group using CPAP for 4–6 hours, the mean reduction was 5 (±5.2) and was similar to that of the higher level of CPAP adherence groups (6–7 hours and 7–10 hours), who had a mean reduction of 5 (±5.2) (Figure 3).

Comparison of Statistical Approaches

First, we performed a weighted regression analysis using the IPTW estimator. The positivity assumption was verified by making sure that the mean of the maximum IPTW weights obtained by bootstrap was reasonable. To investigate the presence of outliers, we verified the distribution of the weights by adherence group (Table E2). The mean of truncated weights was 4.0 (4.0; 4.0) (Table E3).

After weighting, the standardized mean differences of all variables for each adherence group showed no imbalance on confounders (Figure 4). Adjustment in the final model was performed to correct for potential remaining imbalance in confounders. The coefficients of the multinomial logistic regression are available in Table E4. The second modeling approach was based on the double robust approach, IPW-RA. Coefficients of the IPW-RA are available in Table E5.

Confounders selected are the following variables at diagnosis: age (in years), neck circumference (in centimeters), sleepiness at the wheel, morning tiredness, morning headaches, libido disorder, dysfunction, night sweating, daytime sleepiness as measured by ESS score, fatigue as measured by Pichot's scale score, depression as measured by Pichot's depression scale, apnea–hypopnea index, sex, restless legs syndrome, morning tiredness, morning headaches, night sweating, and fatigue as measured by Pichot's scale.

When the adherence groups are compared with an unweighted mean difference, patients in the adherence group using CPAP for 0–4 hours had an average ESS score of 2 points higher than the reference group (95% bootstrap confidence intervals based on percentiles 1.9; 2.1). Patients using CPAP for 4–6 hours had an average ESS score of 0.8 (0.7; 0.9) points higher than patients using CPAP for 7–10 hours. Patients using CPAP for 6–7 hours had an average ESS score of 0.2 points (0; 0.3) higher than patients using CPAP for 7–10 hours, and there is evidence of a difference between groups (overall CPAP adherence effect, $p < 0.001$). By using multivariable regression, IPTW, or IPW-RA estimators, the results are attenuated as compared with the unadjusted analysis but similar to each other: Patients using CPAP for 0–4 hours had an average ESS score of 1.1 points (0.8; 1.3) higher than patients using CPAP for 7–10 hours with IPTW. Patients using CPAP for 4–6 hours had an

Table 2. Patient characteristics according to the adherence group

Variables	All Groups, (*N* = *9,244*)	Group 1: 0–4 h (*n* = *1,977*, 21.4%)	Group 2: 4–6 h (*n* = *3,519*, 38.1%)	Group 3: 6–7 h (*n* = *2,023*, 21.9%)	Group 4: 7–10 h (*n* = *1,725*, 18.7%)
Variables at diagnosis					
ESS score	10.9 (5.2)	11.0 (5.3)	10.9 (5.2)	10.8 (5.2)	10.8 (5.3)
Gender, male	6,485 (70.2%)	1,337 (67.6%)$_3$	2,470 (70.2%)	1,457 (72%)$_1$	1,221 (70.8%)
Age, yr	57.3 (12.4)$_{1,4}$	55.5 (12.8)$_{2,3,4}$	56.7 (12.1)$_{1,3,4}$	57.9 (11.9)$_{1,2,4}$	59.7 (12.4)$_{1,2,3}$
Body mass index, kg/m^2	32.0 (7.1)	31.9 (6.8)	31.9 (7.4)	32.0 (7.2)	32.4 (6.8)
Apnea–hypopnea index, events/h	40.7 (20.5)$_{1,3,4}$	37.8 (19.6)$_{2,3,4}$	39.7 (19.7)$_{1,3,4}$	42.7 (21.5)$_{1,2}$	43.6 (21.2)$_{1,2}$
Tobacco status	0.6 (0.7)	0.6 (0.8)$_4$	0.6 (0.7)$_4$	0.6 (0.7)	0.5 (0.7)$_{1,2}$
Depression scale	4.0 (3.8)	4.3 (3.9)$_2$	3.9 (3.8)$_1$	3.9 (3.7)	4.1 (3.8)
Pichot's fatigue scale	13.3 (8.2)	13.8 (8.6)$_2$	12.9 (8.1)$_{1,4}$	13.1 (8.1)	13.6 (8.2)$_2$
Morning headaches	3,697 (40%)	820 (41.5%)	1,423 (40.4%)	813 (40.2%)	641 (37.2%)
Morning tiredness	7,178 (77.7%)	1,546 (78.2%)	2,743 (77.9%)	1,560 (77.1%)	1,329 (77%)
Diabetes	2,492 (27%)	568 (28.7%)	906 (25.7%)	518 (25.6%)	500 (29%)
Variables at follow-up					
ESS score	6.3 (4.2)$_{1,3,4}$	7.5 (4.8)$_{2,3,4}$	6.3 (4.1)$_{1,3,4}$	5.7 (3.9)$_{1,2}$	5.5 (3.9)$_{1,2}$
Residual apnea–hypopnea index under CPAP	4.1 (4.7)$_1$	4.5 (5.5)$_{2,3}$	3.8 (4.2)$_1$	3.9 (4.5)$_1$	4.2 (4.9)
Pichot's fatigue scale	7.7 (7.2)$_{1,3}$	9.4 (7.7)$_{2,3,4}$	7.4 (6.9)$_{1,3}$	6.8 (6.7)$_{1,2,4}$	7.4 (7.3)$_{1,3}$
Morning headaches	2,479 (26.8%)$_4$	579 (29.3%)$_4$	982 (27.9%)$_4$	533 (26.3%)$_4$	385 (22.3%)$_{1,2,3}$
Morning tiredness	4,954 (53.6%)$_{1,4}$	1,166 (59%)$_{2,3,4}$	1,906 (54.2%)$_{1,4}$	1,055 (52.2%)$_1$	827 (47.9%)$_{1,2}$
Number of ADR types under CPAP	0.7 (1.1)$_{1,3,4}$	1.0 (1.3)$_{2,3,4}$	0.6 (1.0)$_{1,3,4}$	0.5 (1.0)$_{1,2}$	0.5 (1.0)$_{1,2}$
Duration since diagnosis, yr	0.9 (1.5)$_{1,4}$	0.8 (1.2)$_{3,4}$	0.9 (1.4)$_4$	1.0 (1.7)$_1$	1.1 (1.9)$_{1,2}$

Definition of abbreviations: ADR = adverse drug reaction; CPAP = continuous positive airway pressure; ESS = Epworth Sleepiness Scale.
Values are presented as mean (SD) for quantitative variables and as number (%) of individuals for *qualitative* variables.
A *t* test was performed for the quantitative variables and a Pearson's chi-squared test for the categorical variables after application of a Bonferroni correction for multiple testing.
Subscript numbers refer to columns statistically different at the 5% threshold; for example, 1 means that there is a statistically significant difference between the adherence group of that column and adherence group 1 (0–4 h) for the variable in question.

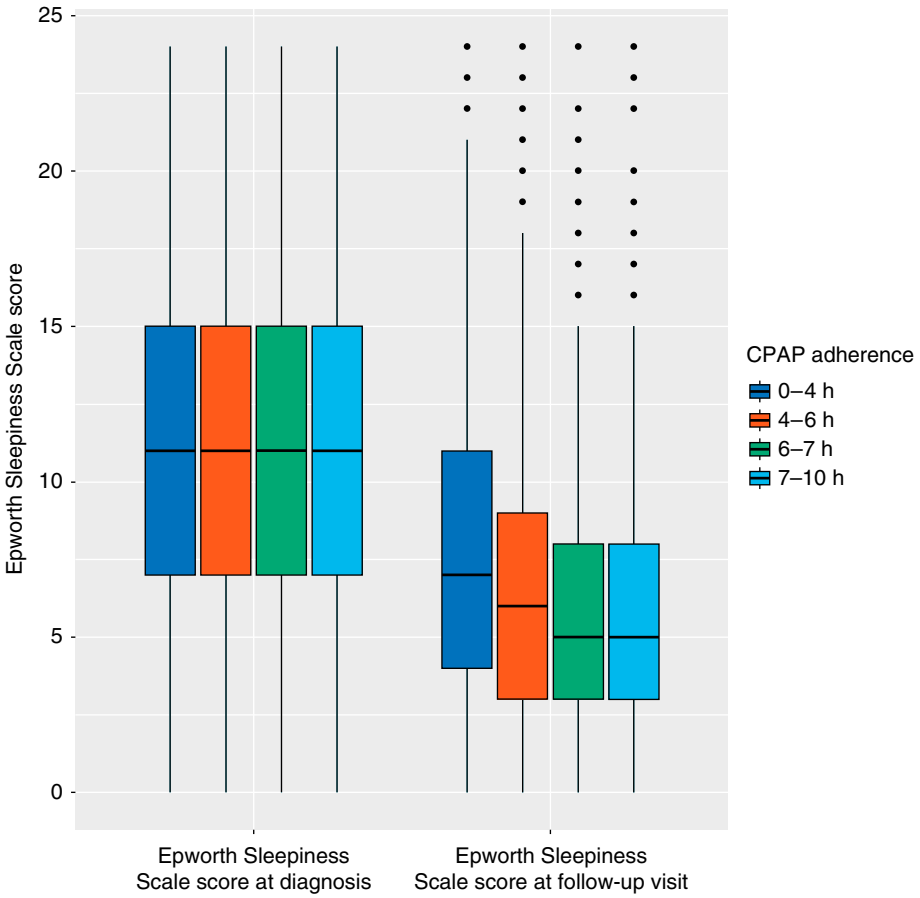

Figure 3. Box plot of raw Epworth Sleepiness Scale scores according to the visit by adherence group. CPAP = continuous positive airway pressure.

average ESS score of 0.5 points (0.3; 0.7) higher than patients using CPAP for 7–10 hours. Patients using CPAP for 6–7 hours had an average ESS score of 0.2 points (0; 0.5) higher than patients using CPAP for 7–10 hours. (Results for the four methods are presented in Figure 5.) IPTW-RA was the most efficient approach, leading to narrower 95% confidence intervals.

Discussion

In this study, we illustrated the application of an inverse probability of treatment weighting estimator by assessing the impact of CPAP adherence levels on ESS score in a large population ($n = 9{,}244$) of CPAP-treated patients. Our results suggested evidence of a difference in the ESS score at the follow-up visit between the low adherence groups and the high adherence group (>7 h). However, there was no evidence of a difference for patients with a high CPAP adherence level (6–7 h vs. 7–10 h). This result is consistent with other studies (30, 31). Further applications should be performed to investigate the use of such methods on other symptoms and signs of daytime sleepiness.

The results presented in Figure 5 showed that the effect size (the mean difference in the ESS score between each adherence group and the reference) is of greater magnitude when confounding is not accounted for (unadjusted analysis) compared with adjusted multivariable regression and weighted methods. This is due to the presence of positive confounding when comparing groups; therefore, these estimates are strongly biased. To avoid inducing bias, multivariable regression and weighted methods accounting for all confounding factors should be preferred. In summary, unadjusted methods lead to bias estimates if groups being compared differ in terms of characteristics also associated with the outcome, which is almost always the case in nonrandomized studies. Multivariable regression models, under a correct specification of the model and the validity of the four identifiability assumptions, lead to the estimation of unbiased conditional causal effects. Although marginal and conditional effects are identical in linear models, it is not the case in nonlinear model, and conditional effects are harder to interpreting in a public health context. Using weighted approaches, an investigation of the weight distribution allows us to identify violations of the positivity assumption and potential lack of overlap. When these problems exist, standard regression methods rely on extrapolation, but an incorrect extrapolation will be more difficult to identify. Furthermore, the treatment allocation mechanism is sometimes easier to model than the outcome mechanism, thus reducing the risk of model misspecification using weighted estimators. In addition, weighted estimators lead to the estimation of marginal effects, which are often more easily interpretable and useful for policy making, as compared with conditional estimates from multivariable regression. Finally, when using multivariable regression, the model is built to

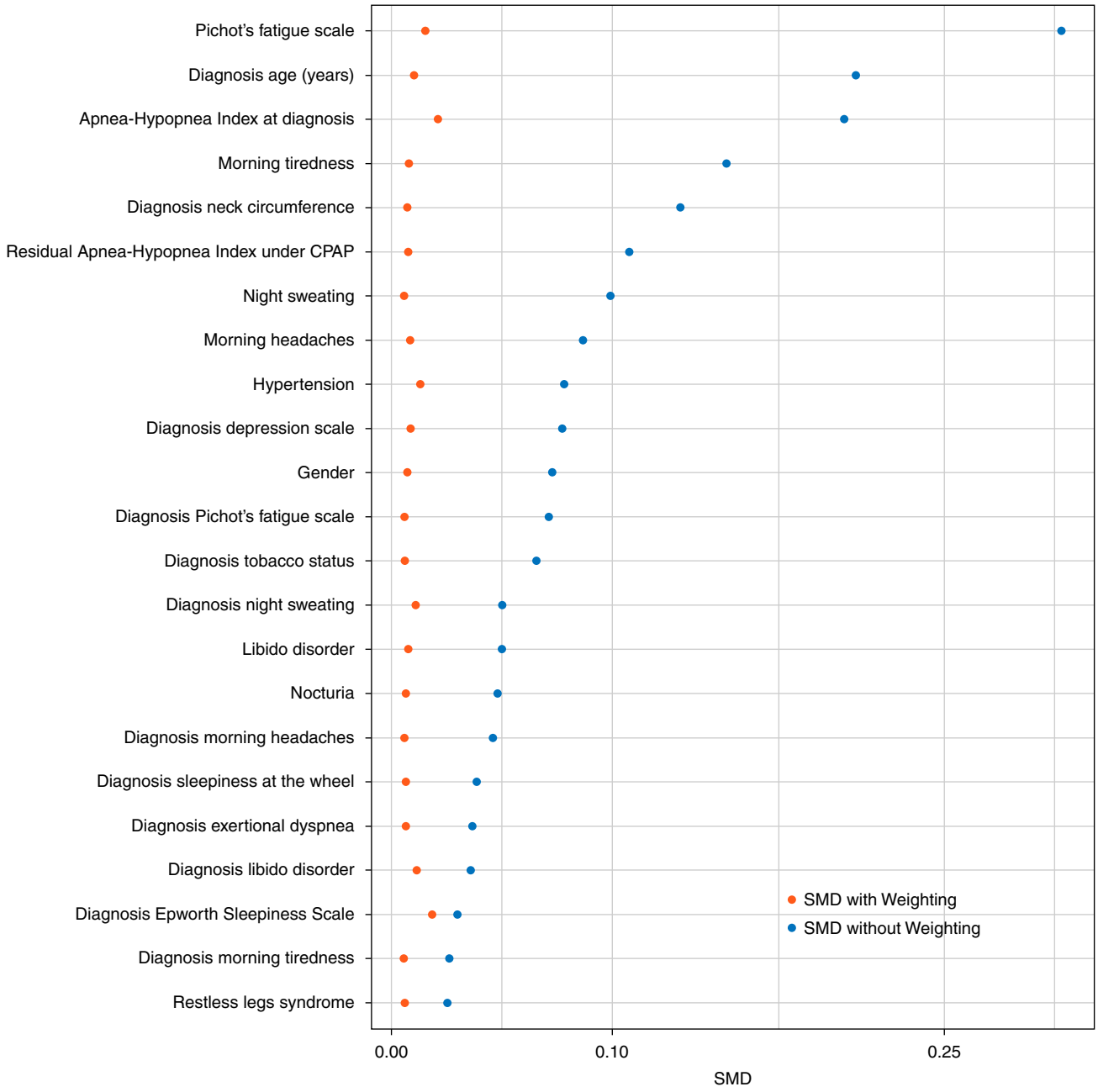

Figure 4. SMD before and after weighting. CPAP = continuous positive airway pressure; SMD = standardized mean difference.

estimate the causal effect of the intervention, but the other coefficients of the model do not have a causal interpretation. However, in practice, the other coefficients are reported and interpreted causally, whereas this is prevented with the use of weighted estimators, which provide a single estimate. In addition, IPW-RA has a double robust property, which minimizes the risk of bias due to model misspecification.

By comparing results of IPTW and IPW-RA, we found that the conclusions are similar; however, the use of a double robust IPW-RA estimator allows an increase in confidence in the consideration of possible risk of model misspecification. However, as in any observational study, unmeasured confounders cannot be ruled out, but it cannot be checked from the data.

From a methodological point of view, the present study highlights the benefits of applying IPW methods to estimate the effect of a multilevel exposure in assessing marginal causal effects. Under a set of assumptions, it is possible to estimate the causal effect of an exposure on an outcome with well-designed observational studies, which can be used as an alternative to randomized clinical trials (32). In this study, the ATE is the most relevant estimand for understanding the potential benefits if all the patients were adherent. However, the weights can be easily modified for the estimation of the ATE on the treated patients. IPTW and IPW-RA are examples of modern statistical methods developed over the past decades

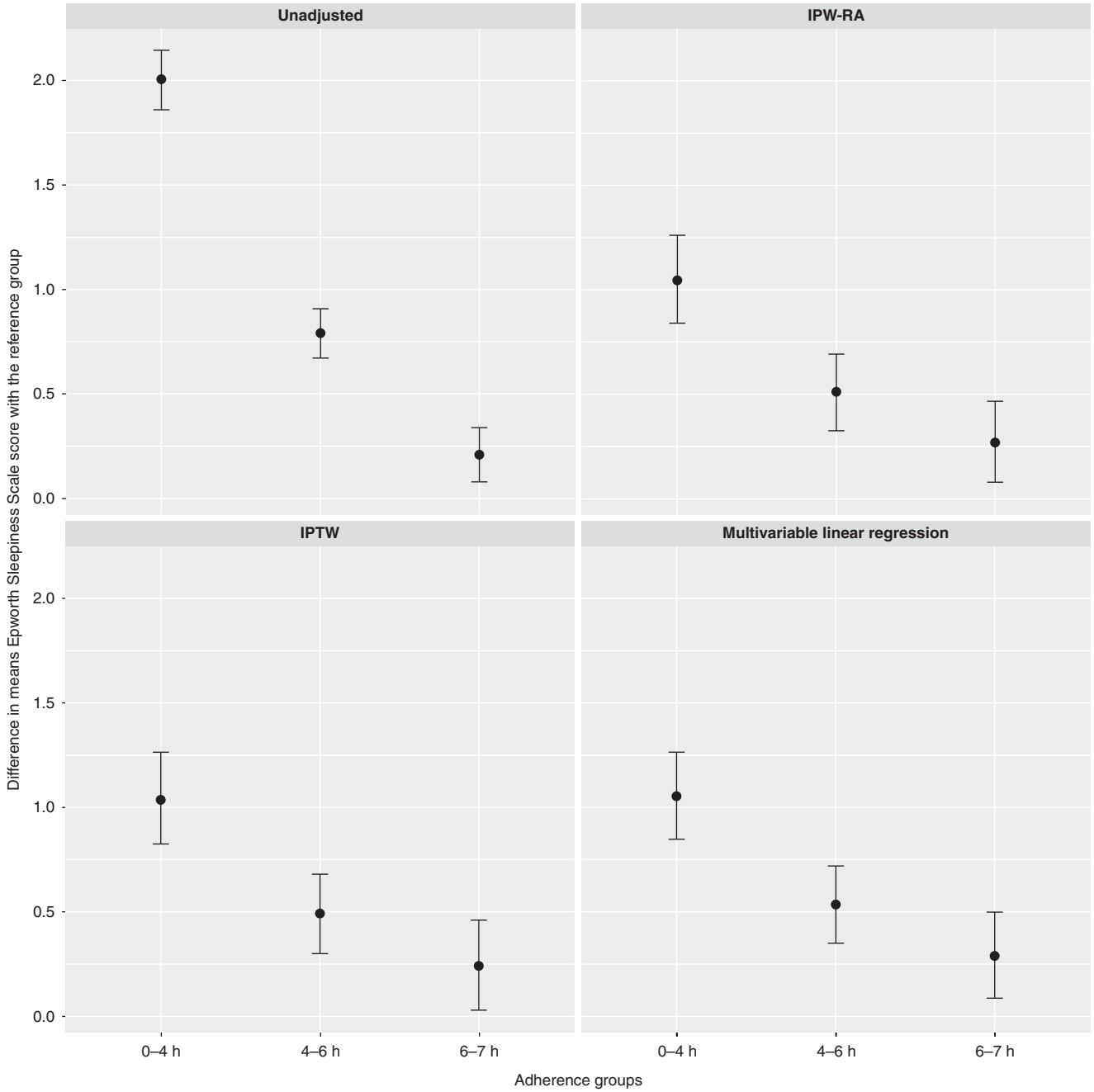

Figure 5. Difference in mean Epworth Sleepiness Scale scores between each adherence group and the reference group using different methods. Each point represents the difference in mean Epworth Sleepiness Scale score between each adherence group and the reference group (7–10 h). The vertical bars represent the 95% confidence intervals of these estimates. IPTW = inverse-probability-of-treatment weight; IPW-RA = inverse propensity weighting with regression adjustment.

that have improved health research by moving the interpretation from associational to causal (33).

To limit the risk of bias when using IPTW, it is important to assess the distribution of the weights and to consider whether the functional form of the weight model is correctly specified. In addition, a careful investigation of the plausibility of the assumptions required for causal inference is needed. This implies extensive discussion between the statistician and clinician to be vigilant with regard to variable selection and model validation. Moreover, if IPW estimators are now well known and extensively used, mainly for binary exposure, the application of such methods needs to be carefully performed, and the reporting of these methods in clinical research should be improved (34).

Unlike for binary exposures, a few published studies (35, 36) have applied IPW to multilevel exposure to assess the marginal causal effect on an outcome. However,

multilevel exposure is of great relevance in the medical field, as many treatments have several levels or need to be compared with other treatments or combination of treatments.

We proposed a method to consider CPAP adherence as a multilevel variable, in contrast to a binary one, and we have illustrated the application of the IPW method to reduce confounding bias. Further methodological research for causal inference applied to multivalued exposures could be proposed. Indeed, it could be of interest to compare these approaches to those based on machine learning algorithms to calculate weights, such as the gradient boosting algorithms that have already been used for this purpose (37). Beyond the trimming we used, which relies on the choice of an arbitrary cutoff, other methods, such as the overlap weighting methods (38), have been developed to minimize the risk of extreme PS.

In conclusion, IPW for multilevel exposures is a promising approach. This study showed that patients with different levels of CPAP adherence experienced a different reduction in their daytime sleepiness as measured by the ESS score. We also showed that patients who had high CPAP adherence experienced a greater reduction in daytime sleepiness than nonadherent patients at their first follow-up visit. ■

Author disclosures are available with the text of this article at www.atsjournals.org.

References

1 Lévy P, Kohler M, McNicholas WT, Barbé F, McEvoy RD, Somers VK, *et al.* Obstructive sleep apnoea syndrome. *Nat Rev Dis Primers* 2015;1: 15015.
2 Benjafield AV, Ayas NT, Eastwood PR, Heinzer R, Ip MSM, Morrell MJ, *et al.* Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. *Lancet Respir Med* 2019;7: 687–698.
3 Pépin JL, Timsit JF, Tamisier R, Borel JC, Lévy P, Jaber S. Prevention and care of respiratory failure in obese patients. *Lancet Respir Med* 2016;4:407–418.
4 Zinchuk AV, Jeon S, Koo BB, Yan X, Bravata DM, Qin L, *et al.* Polysomnographic phenotypes and their cardiovascular implications in obstructive sleep apnoea. *Thorax* 2018;73:472–480.
5 Pépin JL, Bailly S, Tamisier R. Incorporating polysomnography into obstructive sleep apnoea phenotyping: moving towards personalised medicine for OSA. *Thorax* 2018;73:409–411.
6 McEvoy RD, Antic NA, Heeley E, Luo Y, Ou Q, Zhang X*et al.*; SAVE Investigators and Coordinators. CPAP for prevention of cardiovascular events in obstructive sleep apnea. *N Engl J Med* 2016;375:919–931.
7 McDaid C, Durée KH, Griffin SC, Weatherly HLA, Stradling JR, Davies RJO, *et al.* A systematic review of continuous positive airway pressure for obstructive sleep apnoea-hypopnoea syndrome. *Sleep Med Rev* 2009;13:427–436.
8 Gasa M, Tamisier R, Launois SH, Sapene M, Martin F, Stach B, *et al.*; Scientific Council of the Sleep Registry of the French Federation of Pneumology-FFP. Residual sleepiness in sleep apnea patients treated by continuous positive airway pressure. *J Sleep Res* 2013;22:389–397.
9 Crook S, Sievi NA, Bloch KE, Stradling JR, Frei A, Puhan MA, *et al.* Minimum important difference of the Epworth Sleepiness Scale in obstructive sleep apnoea: estimation from three randomised controlled trials. *Thorax* 2019;74:390–396.
10 Patel S, Kon SSC, Nolan CM, Barker RE, Simonds AK, Morrell MJ, *et al.* The Epworth Sleepiness Scale: minimum clinically important difference in obstructive sleep apnea. *Am J Respir Crit Care Med* 2018;197:961–963.
11 Ware JH, Hamel MB. Pragmatic trials–guides to better patient care? *N Engl J Med* 2011;364:1685–1687.
12 Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *N Engl J Med* 2000;342:1878–1886.
13 Kennedy-Martin T, Curtis S, Faries D, Robinson S, Johnston J. A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. *Trials* 2015;16:495.
14 Pack AI, Magalang UJ, Singh B, Kuna ST, Keenan BT, Maislin G. To RCT or not to RCT? Depends on the question. A response to McEvoy et al. *Sleep* 2021;44:zsab042.
15 Pack AI, Magalang UJ, Singh B, Kuna ST, Keenan BT, Maislin G. Randomized clinical trials of cardiovascular disease in obstructive sleep apnea: understanding and overcoming bias. *Sleep* 2021;44:zsaa229 10.1093/sleep/zsaa229.
16 McEvoy RD, Sánchez-de-la-Torre M, Peker Y, Anderson CS, Redline S, Barbe F. Randomized clinical trials of cardiovascular disease in obstructive sleep apnea: understanding and overcoming bias. *Sleep* 2021;44:zsab019.
17 Keenan BT, Maislin G, Sunwoo BY, Arnardottir ES, Jackson N, Olafsson I, *et al.* Obstructive sleep apnoea treatment and fasting lipids: a comparative effectiveness study. *Eur Respir J* 2014;44:405–414.
18 Hernán MA. A definition of causal effect for epidemiological research. *J Epidemiol Community Health* 2004;58:265–271.
19 Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat Sci* 2010;25:1–21.
20 Yoshida K, Solomon DH, Haneuse S, Kim SC, Patorno E, Tedeschi SK, *et al.* Multinomial extension of propensity score trimming methods: a simulation study. *Am J Epidemiol* 2019;188:609–616.
21 Pépin J-L, Bailly S, Rinder P, Adler D, Szeftel D, Malhotra A, *et al.*; On Behalf of the medXcloud Group. CPAP therapy termination rates by OSA phenotype: a French nationwide database analysis. *J Clin Med* 2021;10:936.
22 Glynn AN, Quinn KM. An introduction to the augmented inverse propensity weighted estimator. *Polit Anal* 2010;18:36–56.
23 Pearl J. An introduction to causal inference. *Int J Biostat* 2010;6:7.
24 Vandenbroucke JP, Broadbent A, Pearce N. Causality and causal inference in epidemiology: the need for a pluralistic approach. *Int J Epidemiol* 2016;45:1776–1786.
25 Robins JM, Hernán MÁ, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000;11:550–560.
26 Etminan M, Collins GS, Mansournia MA. Using causal diagrams to improve the design and interpretation of medical research. *Chest* 2020; 158:S21–S28.
27 Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol* 2008;168:656–664.
28 Lefebvre G, Delaney JAC, Platt RW. Impact of mis-specification of the treatment model on estimates from a marginal structural model. *Stat Med* 2008;27:3629–3642.
29 Schomaker M, Heumann C. Bootstrap inference when using multiple imputation. *Stat Med* 2018;37:2252–2266.
30 Weaver TE, Maislin G, Dinges DF, Bloxham T, George CFP, Greenberg H, *et al.* Relationship between hours of CPAP use and achieving normal levels of sleepiness and daily functioning. *Sleep* 2007;30:711–719.
31 Stepnowsky CJ, Dimsdale JE. Dose-response relationship between CPAP compliance and measures of sleep apnea severity. *Sleep Med* 2002;3:329–334.
32 Frieden TR. Evidence for health decision making – beyond randomized, controlled trials. *N Engl J Med* 2017;377:465–475.
33 Glass TA, Goodman SN, Hernán MA, Samet JM. Causal inference in public health. *Annu Rev Public Health* 2013;34:61–75.
34 Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Stat Med* 2015;34:3661–3679 10.1002/sim.6607.
35 Ahn H-J, Lee S-R, Choi E-K, Han K-D, Jung J-H, Lim J-H, *et al.* Association between exercise habits and stroke, heart failure, and

mortality in Korean patients with incident atrial fibrillation: a nationwide population-based cohort study. *PLoS Med* 2021;18: e1003659.
36 Ali AK, Hartzema AG, Winterstein AG, Segal R, Lu X, Hendeles L. Application of multicategory exposure marginal structural models to investigate the association between long-acting beta-agonists and prescribing of oral corticosteroids for asthma exacerbations in the Clinical Practice Research Datalink. *Value Health* 2015;18:260–270.
37 Olmos A, Govindasamy P. A practical guide for using propensity score weighting in R. *Pract Assess, Res Eval* 2015;20:13.
38 Li F, Thomas LE, Li F. Addressing extreme propensity scores via the overlap weights. *Am J Epidemiol* 2019;188:250–257.

Causal inference with multiple exposures: application of inverse-probability-of-treatment weighting to estimate the effect of daytime sleepiness in obstructive sleep apnea patients. Supplementary Material

Author

François Bettega[1], Clémence Leyrat[2], Renaud Tamisier[1], Monique Mendelson[1],Yves Grillet[3], Marc Sapène[4], Maria R Bonsignore[5], Jean Louis Pépin[1], Michael W Kattan[6], Sébastien Bailly[1]

Affiliations

[1] University Grenoble Alpes, Inserm, Grenoble Alpes University Hospital, HP2, 38000 Grenoble, France

[2] Department of Medical Statistics, Inequalities in Cancer Outcomes Network, London School of Hygiene and Tropical Medicine, London, UK

[3] Private Practice Sleep and Respiratory Disease Centre, Nouvelle Clinique Bel Air, Bordeaux, France

[4] Private Practice Sleep and Respiratory Disease Centre, Valence, France

[5] Respiratory Medicine, PROMISE Department, University of Palermo and IRIB-CNR, Palermo, Italy

[6] The Department of Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic, Cleveland, Ohio, United States

Supplementary material 1 : Imputation

Variables contained at least one missing which needed to be imputed, table 1 summarized number of missing values by adherence group for each of those variables.

We performed 10 imputed data sets using predictive mean matching as imputation method. After that we verified that the 10 imputed data sets were consistent with each other by looking at the distributions of the imputed variables in the different data sets.

Supplementary material 2 : Directed acyclic graph

To construct the current DAG, we used existing knowledge from different original data publication and systematic reviews of expert consensus papers. We assumed that sex,[E1] body mass index (BMI), OSA severity[E2] and age[E3] have a direct causal effect on CPAP-adherence. This was reported from the analysis of a large cohort study including half a million participants. There are mechanisms explaining direct causal links between age, sex and BMI with cardiovascular and metabolic comorbidities. Obesity has an impact on OSA severity but is per se a major determinant for the occurrence of comorbidities. Some specific comorbidities such as cardiac failure,[E4] hypertension or stroke[E5] can induce rostral fluid shift during night that exacerbate or originate sleep apnea.[E6] Some of these comorbidities impact daytime sleepiness with a "protective effect" for heart failure[E7] or favoring daytime hypersomnia like obesity or diabetes. This is true both at baseline and under CPAP.[E8] There is a weak link between OSA severity and daytime sleepiness.[E9, E10] High OSA severity at baseline is a predictive factor for residual sleepiness under CPAP.[E11] Age and BMI are directly associated with OSA severity as it was shown in several publications using clustering approaches and there are distinct evolutions of symptoms from baseline under treatment according to age and sex.[E12–E15] Finally, severity of symptoms at baseline is a good predictor for residual daytime sleepiness under CPAP.[E8, E9, E16] Unmeasured confounders merit to be considered when assessing the link between CPAP adherence and OSAS severity: socio- economic status (SAS) with impact of deprivation index which has been shown associated to a lower CPAP adherence,[E17–E19] racial disparities,[E20] intimal context such as spousal involvement,[E21, E22] health literacy[E23] or personal perception of treatment efficacy.[E24] Environmental factors like pollution or temperature can impact also both CPAP adherence or OSAS severity.[E25, E26] Finally, medication can both impact CPAP adherence and daytime sleepiness.[E27] However, these variables do not need to be included in the minimal adjustment set for the estimation of the causal effect of CPAP adherence on daytime sleepiness. Indeed, by applying d-separation rules to this DAG, age, sex, comorbidities, BMI, OSAS severity were identified as the confounders to adjust for in our analysis to assess causal effect.

Supplementary material 3 : Additional analysis of causal effect of CPAP adherence with morning fatigue as outcome

An additional analysis was performed with morning fatigue as outcome and CPAP adherence as exposure. This analysis was performed on all patients without missing values for outcome, baseline outcome and exposure (n = 4833). Morning fatigue is self-reported by patients on a scale between 0 and 10.

The results are presented as the difference in morning fatigue scale compared to the most adherent group (7-10h) with 95% CI.

IPTW : 0-4h : 0.59 (0.43; 0.75), 4-6h: 0.23 (0.08; 0.37) and 6-7h: 0.00 (-0.17; 0.18),

IPWRA: 0-4h : 0.59 (0.43; 0.73), 4-6h: 0.23 (0.08; 0.37) and 6-7h: 0.04 (-0.12; 0.2),

Multivariable regression: 0-4h : 0.59 (0.44; 0.74), 4-6h: 0.23 (0.09; 0.37) and 6-7h: 0.03 (-0.13; 0.21)

Unadjusted mean comparison: 0-4h : 0.63 (0.43; 0.83), 4-6h: 0.4 (-0.14; 0.24) and 6-7h: -0.22 (-0.41; -0.02). The results are consistent with those observed with Epworth score.

E1 Table: Number of missing value by variables and group

	All groups 11553	0-4 h (1) 2471	4-6 h (2) 4398	6-7 h (3) 2528	7-10 h (4) 2156
Variables at diagnosis					
Body mass index (kg/m2)	118 (1.0%)	30 (1.2%)	46 (1.0%)	22 (0.9%)	20 (0.9%)
Neck circumference	4,148 (35.9%)	870 (35.2%)	1,535 (34.9%)	937 (37.1%)	806 (37.4%)
Tobacco status	30 (0.3%)	5 (0.2%)	15 (0.3%)	7 (0.3%)	3 (0.1%)
Nocturia	44 (0.4%)	9 (0.4%)	14 (0.3%)	15 (0.6%)	6 (0.3%)
Respiratory arrests	425 (3.7%)	94 (3.8%)	148 (3.4%)	98 (3.9%)	85 (3.9%)
Pichot's fatigue scale	941 (8.1%)	141 (5.7%)	365 (8.3%)	226 (8.9%)	209 (9.7%)
Depression scale	1,099 (9.5%)	180 (7.3%)	410 (9.3%)	262 (10.4%)	247 (11.5%)
Micro awake index	3,957 (34.3%)	694 (28.1%)	1,504 (34.2%)	939 (37.1%)	820 (38.0%)
Mean nocturnal SaO2	3,206 (27.8%)	691 (28.0%)	1,266 (28.8%)	717 (28.4%)	532 (24.7%)
Gender (male)	14 (0.1%)	2 (0.1%)	8 (0.2%)	3 (0.1%)	1 (0.0%)
Hypertension	50 (0.4%)	12 (0.5%)	21 (0.5%)	10 (0.4%)	7 (0.3%)
Hypercholesterolemia	48 (0.4%)	9 (0.4%)	19 (0.4%)	12 (0.5%)	8 (0.4%)
Restless legs syndrome	73 (0.6%)	19 (0.8%)	22 (0.5%)	20 (0.8%)	12 (0.6%)
Variables at follow-up					
Nocturia	83 (0.7%)	20 (0.8%)	32 (0.7%)	16 (0.6%)	15 (0.7%)
Pichot's fatigue scale	1,135 (9.8%)	182 (7.4%)	443 (10.1%)	256 (10.1%)	254 (11.8%)
Mask type	1,171 (10.1%)	261 (10.6%)	431 (9.8%)	263 (10.4%)	216 (10.0%)
Humidifier	28 (0.2%)	9 (0.4%)	8 (0.2%)	7 (0.3%)	4 (0.2%)
Residual apnea hypopnea index under CPAP	769 (6.7%)	206 (8.3%)	281 (6.4%)	167 (6.6%)	115 (5.3%)
Average pressure of CPAP	1,660 (14.4%)	370 (15.0%)	638 (14.5%)	369 (14.6%)	283 (13.1%)

ESS : Epworth Sleepiness Scale; CPAP : Continuous Positive Airway Pressure

E2 Table: Distribution of weights

Adherence group	Mean	Minimum	Maximum
0-4 h	4.7(4.5; 4.9)	1.5(1.2; 1.7)	20.1(13.2; 29.5)
4-6 h	2.6(2.6; 2.7)	1.8(1.7; 2.0)	5.9(4.4; 8.6)
6-7 h	4.6(4.4; 4.7)	2.5(2.2; 2.8)	13.2(10.4; 17.5)
7-10 h	5.4(5.2; 5.6)	2.3(2.0; 2.6)	20.1(14.7; 27.4)

Data are presented as mean (95% confidence interval) of bootstrap iterations

E3 Table: Weight truncations

Truncations	Mean	Minimum	Maximum
(0; 1)	4.0(4.0; 4.0)	1.5(1.3; 1.7)	22.4(17.1; 28.8)
(0.01; 0.99)	4.0(4.0; 4.0)	2.0(1.9; 2.1)	10.1(9.6; 10.7)
(0.05; 0.95)	3.9(3.9; 3.9)	2.2(2.2; 2.3)	7.3(7.1; 7.5)
(0.1; 0.9)	3.9(3.8; 3.9)	2.3(2.3; 2.4)	6.2(6.1; 6.3)
(0.25; 0.75)	3.7(3.6; 3.7)	2.7(2.6; 2.7)	4.8(4.8; 4.9)
(0.5; 0.5)	3.6(3.5; 3.7)	3.6(3.5; 3.7)	3.6(3.5; 3.7)

Data are presented as mean (95% confidence interval) of bootstrap iterations

E4 Table: IPWRA weight model coefficient to assess the probability of being in an adherence group

Variables	Coefficients of 4-6 h group	Coefficients of 6-7 h group	Coefficients of 7-10 h group
Diagnosis age (years)	0.013(0.008; 0.018)	0.022(0.016; 0.028)	0.030(0.024; 0.037)
Diagnosis neck circumference	0.010(-0.005; 0.027)	0.007(-0.011; 0.024)	0.021(0.004; 0.039)
Diagnosis sleepiness at the wheel	0.205(0.084; 0.338)	0.197(0.047; 0.348)	0.060(-0.088; 0.213)
Diagnosis morning tiredness	0.114(-0.037; 0.273)	0.045(-0.138; 0.222)	0.153(-0.033; 0.338)
Diagnosis morning headaches	0.012(-0.126; 0.153)	0.053(-0.112; 0.229)	0.005(-0.168; 0.162)
Diagnosis libido disorder	-0.096(-0.253; 0.071)	-0.015(-0.199; 0.184)	-0.056(-0.252; 0.149)
Diagnosis night sweating	-0.046(-0.191; 0.098)	-0.004(-0.173; 0.167)	-0.005(-0.176; 0.162)
Diagnosis exertional dyspnea	-0.111(-0.243; 0.019)	-0.129(-0.275; 0.013)	-0.033(-0.179; 0.119)
Diagnosis epworth sleepiness scale	0.001(-0.012; 0.013)	-0.010(-0.024; 0.005)	-0.009(-0.024; 0.006)
Diagnosis pichot's fatigue scale	0.005(-0.006; 0.016)	0.025(0.013; 0.037)	0.029(0.016; 0.041)
Diagnosis depression scale	0.003(-0.016; 0.022)	0.006(-0.014; 0.028)	0.004(-0.019; 0.028)
Apnea Hypopnea Index at diagnosis	0.004(0.002; 0.007)	0.012(0.008; 0.015)	0.013(0.009; 0.016)
Gender (male)	-0.002(-0.152; 0.145)	0.073(-0.090; 0.231)	-0.016(-0.194; 0.152)
Hypertension	-0.037(-0.155; 0.084)	0.001(-0.153; 0.132)	0.035(-0.105; 0.183)
Restless legs syndrome	0.146(0.005; 0.288)	0.069(-0.099; 0.243)	0.020(-0.155; 0.183)
Morning tiredness	-0.106(-0.254; 0.028)	-0.082(-0.244; 0.067)	-0.295(-0.459; -0.133)
Morning headaches	0.056(-0.106; 0.208)	0.052(-0.116; 0.228)	-0.093(-0.290; 0.108)
Libido disorder	0.304(0.121; 0.487)	0.301(0.092; 0.495)	0.398(0.187; 0.614)
Night sweating	-0.084(-0.236; 0.068)	-0.137(-0.321; 0.034)	-0.255(-0.446; -0.065)
Nocturia	-0.084(-0.199; 0.031)	-0.230(-0.360; -0.090)	-0.025(-0.165; 0.119)
Pichot's fatigue scale	-0.040(-0.050; -0.031)	-0.061(-0.073; -0.051)	-0.045(-0.057; -0.033)
Residual apnea hypopnea index under CPAP	-0.035(-0.047; -0.024)	-0.037(-0.052; -0.023)	-0.029(-0.043; -0.017)

Data are presented as mean (95% confidence interval) of bootstrap iterations

ESS : Epworth Sleepiness Scale; CPAP : Continuous Positive Airway Pressure; ADR : adverse drug reaction

E5 Table: IPWRA weight model coefficient to assess the probability of being in an adherence group

Variables	Coefficients of 4-6 h group	Coefficients of 6-7 h group	Coefficients of 7-10 h group
Diagnosis age (years)	0.013(0.008; 0.018)	0.022(0.016; 0.028)	0.030(0.024; 0.037)
Diagnosis neck circumference	0.010(-0.005; 0.027)	0.007(-0.011; 0.024)	0.021(0.004; 0.039)
Diagnosis sleepiness at the wheel	0.205(0.084; 0.338)	0.197(0.047; 0.348)	0.060(-0.088; 0.213)
Diagnosis morning tiredness	0.114(-0.037; 0.273)	0.045(-0.138; 0.222)	0.153(-0.033; 0.338)
Diagnosis morning headaches	0.012(-0.126; 0.153)	0.053(-0.112; 0.229)	0.005(-0.168; 0.162)
Diagnosis libido disorder	-0.096(-0.253; 0.071)	-0.015(-0.199; 0.184)	-0.056(-0.252; 0.149)
Diagnosis night sweating	-0.046(-0.191; 0.098)	-0.004(-0.173; 0.167)	-0.005(-0.176; 0.162)
Diagnosis exertional dyspnea	-0.111(-0.243; 0.019)	-0.129(-0.275; 0.013)	-0.033(-0.179; 0.119)
Diagnosis epworth sleepiness scale	0.001(-0.012; 0.013)	-0.010(-0.024; 0.005)	-0.009(-0.024; 0.006)
Diagnosis pichot's fatigue scale	0.005(-0.006; 0.016)	0.025(0.013; 0.037)	0.029(0.016; 0.041)
Diagnosis depression scale	0.003(-0.016; 0.022)	0.006(-0.014; 0.028)	0.004(-0.019; 0.028)
Apnea Hypopnea Index at diagnosis	0.004(0.002; 0.007)	0.012(0.008; 0.015)	0.013(0.009; 0.016)
Gender (male)	-0.002(-0.152; 0.145)	0.073(-0.090; 0.231)	-0.016(-0.194; 0.152)
Hypertension	-0.037(-0.155; 0.084)	0.001(-0.153; 0.132)	0.035(-0.105; 0.183)
Restless legs syndrome	0.146(0.005; 0.288)	0.069(-0.099; 0.243)	0.020(-0.155; 0.183)
Morning tiredness	-0.106(-0.254; 0.028)	-0.082(-0.244; 0.067)	-0.295(-0.459; -0.133)
Morning headaches	0.056(-0.106; 0.208)	0.052(-0.116; 0.228)	-0.093(-0.290; 0.108)
Libido disorder	0.304(0.121; 0.487)	0.301(0.092; 0.495)	0.398(0.187; 0.614)
Night sweating	-0.084(-0.236; 0.068)	-0.137(-0.321; 0.034)	-0.255(-0.446; -0.065)
Nocturia	-0.084(-0.199; 0.031)	-0.230(-0.360; -0.090)	-0.025(-0.165; 0.119)
Pichot's fatigue scale	-0.040(-0.050; -0.031)	-0.061(-0.073; -0.051)	-0.045(-0.057; -0.033)
Residual apnea hypopnea index under CPAP	-0.035(-0.047; -0.024)	-0.037(-0.052; -0.023)	-0.029(-0.043; -0.017)

Data are presented as mean (95% confidence interval) of bootstrap iterations

ESS : Epworth Sleepiness Scale; CPAP : Continuous Positive Airway Pressure; ADR : adverse drug reaction

E6 Table: Multivariable weighted linear regression to assess the impact of CPAP adherence group on ESS under CPAP

Label	Model coefficients
Diagnosis age (years)	-0.016(-0.022; -0.009)
Diagnosis neck circumference	-0.014(-0.032; 0.004)
Diagnosis tobacco status 1	0.055(-0.104; 0.210)
Diagnosis tobacco status 2	-0.166(-0.399; 0.041)
Diagnosis sleepiness at the wheel	0.328(0.170; 0.490)
Diagnosis morning tiredness	-0.356(-0.546; -0.163)
Diagnosis morning headaches	-0.038(-0.217; 0.147)
Diagnosis libido disorder	0.047(-0.148; 0.256)
Diagnosis night sweating	-0.028(-0.189; 0.131)
Diagnosis exertional dyspnea	-0.103(-0.254; 0.049)
Diagnosis epworth sleepiness scale	0.344(0.325; 0.362)
Diagnosis pichot's fatigue scale	-0.110(-0.124; -0.096)
Diagnosis depression scale	-0.028(-0.053; -0.002)
Apnea Hypopnea Index at diagnosis	-0.019(-0.023; -0.015)
Gender (male)	0.268(0.091; 0.435)
Hypertension	-0.012(-0.172; 0.140)
Restless legs syndrome	0.145(-0.045; 0.330)
Morning tiredness	0.421(0.259; 0.586)
Morning headaches	-0.010(-0.227; 0.208)
Libido disorder	0.124(-0.089; 0.341)
Night sweating	0.283(0.080; 0.478)
Nocturia	-0.199(-0.346; -0.054)
Pichot's fatigue scale	0.336(0.321; 0.351)
Residual apnea hypopnea index under CPAP	0.030(0.014; 0.048)
Adherence groups 1	1.054(0.848; 1.265)
Adherence groups 2	0.536(0.351; 0.720)
Adherence groups 3	0.291(0.087; 0.500)

Data are presented as mean (95% confidence interval) of bootstrap iterations

ESS : Epworth Sleepiness Scale; CPAP : Continuous Positive Airway Pressure; ADR : adverse drug reaction

References

[E1] Pépin JL, Bailly S, Rinder P, Adler D, Szeftel D, Malhotra A, et al. CPAP Therapy Termination Rates by OSA Phenotype: A French Nationwide Database Analysis. *J Clin Med* . 2021. 10(5):936.

[E2] Stepnowsky CJ, Dimsdale JE. Dose-response relationship between CPAP compliance and measures of sleep apnea severity. *Sleep Med* . 2002. 3(4):329–334.

[E3] Sabil A, Le Vaillant M, Stitt C, Goupil F, Pigeanne T, Leclair-Visonneau L, et al. A CPAP data-based algorithm for automatic early prediction of therapy adherence. *Sleep Breath* . 2021. 25(2):957–962.

[E4] Lévy P, Naughton MT, Tamisier R, Cowie MR, Bradley TD. Sleep Apnoea and Heart Failure. *Eur Respir J* . 2021. 2101640.

[E5] Brown DL, Yadollahi A, He K, Xu Y, Piper B, Case E, et al. Overnight Rostral Fluid Shifts Exacerbate Obstructive Sleep Apnea After Stroke. *Stroke* . 2021. 52(10):3176–3183.

[E6] Javaheri S, Barbe F, Campos-Rodriguez F, Dempsey JA, Khayat R, Javaheri S, et al. Sleep Apnea: Types, Mechanisms, and Clinical Cardiovascular Consequences. *J Am Coll Cardiol* . 2017. 69(7):841–858.

[E7] Arzt M, Young T, Finn L, Skatrud JB, Ryan CM, Newton GE, et al. Sleepiness and sleep in patients with both systolic heart failure and obstructive sleep apnea. *Arch Intern Med* . 2006. 166(16):1716–1722.

[E8] Koutsourelakis I, Perraki E, Economou NT, Dimitrokalli P, Vagiakis E, Roussos C, et al. Predictors of residual sleepiness in adequately treated obstructive sleep apnoea patients. *Eur Respir J* . 2009. 34(3):687–693.

[E9] Bonsignore MR, Pepin JL, Cibella F, Barbera CD, Marrone O, Verbraecken J, et al. Excessive Daytime Sleepiness in Obstructive Sleep Apnea Patients Treated With Continuous Positive

Airway Pressure: Data From the European Sleep Apnea Database. *Front Neurol* . 2021. 12:690.008.

[E10] Mitra AK, Bhuiyan AR, Jones EA. Association and Risk Factors for Obstructive Sleep Apnea and Cardiovascular Diseases: A Systematic Review. *Diseases* . 2021. 9(4):88.

[E11] Thorarinsdottir EH, Janson C, Aspelund T, Benediktsdottir B, Júlíusson S, Gislason T, et al. Different components of excessive daytime sleepiness and the change with positive airway pressure treatment in patients with obstructive sleep apnea: Results from the Icelandic Sleep Apnea Cohort (ISAC). *J Sleep Res* . 2021. e13528.

[E12] Bailly S, Grote L, Hedner J, Schiza S, McNicholas WT, Basoglu OK, et al. Clusters of sleep apnoea phenotypes: A large pan-European study from the European Sleep Apnoea Database (ESADA). *Respirology* . 2021. 26(4):378–387.

[E13] Gagnadoux F, Le Vaillant M, Paris A, Pigeanne T, Leclair-Visonneau L, Bizieux-Thaminy A, et al. Relationship Between OSA Clinical Phenotypes and CPAP Treatment Outcomes. *Chest* . 2016. 149(1):288–290.

[E14] Bailly S, Destors M, Grillet Y, Richard P, Stach B, Vivodtzev I, et al. Obstructive Sleep Apnea: A Cluster Analysis at Time of Diagnosis. *PLoS One* . 2016. 11(6):e0157.318.

[E15] Holfinger SJ, Lyons MM, Keenan BT, Mazzotti DR, Mindel J, Maislin G, et al. Diagnostic Performance of Machine Learning-Derived OSA Prediction Tools in Large Clinical and Community-Based Samples. *Chest* . 2021. S0012–3692(21)04.248–3.

[E16] Gasa M, Tamisier R, Launois SH, Sapene M, Martin F, Stach B, et al. Residual sleepiness in sleep apnea patients treated by continuous positive airway pressure. *J Sleep Res* . 2013. 22(4):389–397.

[E17] Daabek N, Tamisier R, Foote A, Revil H, Joyeux-Jaure M, Pépin JL, et al. Impact of Healthcare Non-Take-Up on Adherence to Long-Term Positive Airway Pressure Therapy. *Front Public Health* . 2021. 9:713.313.

[E18] Wickwire EM, Jobe SL, Oldstone LM, Scharf SM, Johnson AM, Albrecht JS. Lower socioeconomic status and co-morbid conditions are associated with reduced continuous positive airway pressure adherence among older adult medicare beneficiaries with obstructive sleep apnea. *Sleep* . 2020. 43(12):zsaa122.

[E19] Palm A, Grote L, Theorell-Haglöw J, Ljunggren M, Sundh J, Midgren B, et al. Socioeconomic Factors and Adherence to CPAP: The Population-Based Course of Disease in Patients Reported to the Swedish CPAP Oxygen and Ventilator Registry Study. *Chest* . 2021. 160(4):1481–1491.

[E20] Borker PV, Carmona E, Essien UR, Saeed GJ, Nouraie SM, Bakker JP, et al. Neighborhoods with Greater Prevalence of Minority Residents Have Lower Continuous Positive Airway Pressure Adherence. *Am J Respir Crit Care Med* . 2021. 204(3):339–346.

[E21] Gentina T, Bailly S, Jounieaux F, Verkindre C, Broussier PM, Guffroy D, et al. Marital quality, partner's engagement and continuous positive airway pressure adherence in obstructive sleep apnea. *Sleep Med* . 2019. 55:56–61.

[E22] Mendelson M, Gentina T, Gentina E, Tamisier R, Pépin JL, Bailly S. Multidimensional Evaluation of Continuous Positive Airway Pressure (CPAP) Treatment for Sleep Apnea in Different Clusters of Couples. *J Clin Med* . 2020. 9(6):E1658.

[E23] Bakker JP, O'Keeffe KM, Neill AM, Campbell AJ. Ethnic disparities in CPAP adherence in New Zealand: Effects of socioeconomic status, health literacy and self-efficacy. *Sleep* . 2011. 34(11):1595–1603.

[E24] Borriboon C, Chaiard J, Tachaudomdach C, Turale S. Continuous positive airway pressure adherence in people with obstructive sleep apnoea. *J Clin Nurs* . 2021.

[E25] Staats R, Bailly S, Bonsignore MR, Ryan S, Riha RL, Schiza S, et al. Impact of temperature on obstructive sleep apnoea in three different climate zones of Europe: Data from the European Sleep Apnoea Database (ESADA). *J Sleep Res* . 2021. 30(5):e13.315.

[E26] Rapelli G, Pietrabissa G, Manzoni GM, Bastoni I, Scarpina F, Tovaglieri I, et al. Improving CPAP Adherence in Adults With Obstructive Sleep Apnea Syndrome: A Scoping Review of Motivational Interventions. *Front Psychol* . 2021. 12:705.364.

[E27] Revol B, Joyeux-Faure M, Albahary MV, Gressin R, Mallaret M, Pepin JL, et al. Severe excessive daytime sleepiness induced by hydroxyurea. *Fundam Clin Pharmacol* . 2017. 31(3):367–368.

V. Performance des GBM pour l'IPTW multi-niveaux

5.1 Présentation du travail

Les IPTWs sur les données observationelles sont des méthodes statistiques bien identifiées et développées dans divers domaines d'application, en particulier en médecine clinique[22].

Dans le cas de traitements multiples, les régressions multinomiales sont, comme nous l'avons vu dans la revue de la littérature (section 4), la méthode la plus utilisée pour l'estimation des poids en IPTW ; elles ont été utilisées dans différentes applications [13, 12] dans le cadre d'un traitement multi-niveaux. Le calcul des scores de propension à l'aide de la régression multinomiale implique les mêmes hypothèses que celles utilisées pour la régression logistique.

Les GBMs sont des méthodes non paramétriques basées sur le boosting et les arbres de décision. Elles présentent l'avantage de ne pas faire d'hypothèse a priori sur la distribution des données. Les GBMs sont couramment utilisés pour estimer les scores de propension [74, 84], en particulier dans le cadre d'un traitement multi-niveaux. Leur introduction à cette fin a été proposée par McCaffrey [49], et les GBM pour les scores de propension sont mis en œuvre dans le package R *twang* [65].

Le boosting crée un classificateur à partir d'un ensemble de classificateurs faibles [26]. Pour l'estimation du score de propension avec les GBM, les classificateurs faibles sont un ensemble d'arbres de régression qui saisissent les relations complexes entre l'affectation du traitement et les facteurs de confusion, tout en évitant l'ajustement excessif [49, 64]. Cet ensemble de prédictions est combiné pour obtenir un classificateur fort.

Ces modèles présentent des propriétés intéressantes pour l'estimation des scores de propension. Plusieurs auteurs ont montré que dans le cas d'un traitement binaire, ils conduisent à un meilleur équilibre entre les covariables et à une erreur quadratique moyenne plus faible [24, 41] quand ils sont comparés avec les régressions multinomiales. Dans le contexte d'un traitement multi-niveaux, McCaffrey et al. ont montré que les GBMs permettent d'atteindre un meilleur équilibre entre les groupes et d'observer une variabilité réduite des poids par rapport à une régression multinomiale [49]. Comme nous l'avons montré dans notre revue systématique (Section 4), les GBMs sont la deuxième méthode d'estimation du score de propension la plus répandue dans la littérature médicale. Cela peut s'expliquer par l'existence du tutoriel de McCaffrey et al [49] ainsi que par l'existence du package R "TWANG" qui fournit une implémentation facile à utiliser des GBMs. Il semble donc important d'évaluer leurs performances en comparaison avec d'autres méthodes d'estimation classiques.

Les objectifs de ce travail étaient d'évaluer les différences en termes de capacité d'équilibrage et de biais dans l'estimation de l'ATE entre quatre méthodes: les GBMs, le Covariate balancing propensity score (CBPS), l'Augmented Inverse Probability of Treatment Weighting (AIPTW) et une méthode de référence, la régression logistique multinomiale, en présence d'un traitement à plusieurs niveaux. Pour cela nous avons procédé à une étude de simulation comprenant de multiples tailles d'échantillons. Elle était basée sur 3 scénarios : (i) un scénario où les covariables ont un effet additif sur l'assignation

du traitement (ii) un scénario avec des termes d'interactions supplémentaires (iii) un scénario avec du bruit et des variables instrumentales. Pour cette étude de simulation, nous avons suivi les bonnes pratiques recommandées par Morris et al [55]. Dans ce travail, j'ai montré que les GBMs ne conduisaient pas à un meilleur équilibre entre les covariables que les régressions multinomiales ou les CBPSs quand ceux-ci étaient correctement spécifiés. Les GBMs ont conduit à un meilleur équilibre que les CBPSs ou les régressions multinomiales quand celles-ci n'étaient pas correctement spécifiées comme dans les scénarios (i) et (ii). En revanche, même dans ces 2 scénarios où les GBMs conduisent à un meilleur équilibre, leur estimation de l'ATE était aussi biaisée que dans les modèles mal spécifiés (AIPTW, ajustement CBPS et régression multinomiale) contrairement aux versions correctement spécifiées de ces modèles qui conduisent à une estimation non biaisée de l'ATE.

Le manuscrit de ce travail est en cours de finalisation pour soumission. J'ai contribué à l'élaboration du plan de simulation, à l'écriture du script de simulation, à l'analyse des résultats et à la rédaction du manuscrit.

5.2 Manuscrit en cours de finition pour une soumission

IPTW for multilevel treatments : Comparison of multinomial logistic regression and generalized boosted models for propensity score estimation for multilevel treatments

François Bettega[1], Sébastien Bailly[1], Clémence Leyrat[2]

October 13, 2023

Affiliations

[1] University Grenoble Alpes, Inserm, Grenoble Alpes University Hospital, HP2, 38000 Grenoble, France

[2] Department of Medical Statistics, Inequalities in Cancer Outcomes Network, London School of Hygiene and Tropical Medicine, London, UK

1 Introduction

A wide range of statistical methods for causal inference from observational data has been developed in the last decades, and these methods have been applied in many fields, especially in medicine and epidemiology Granger et al. (2020). The potential outcomes framework has been proposed to provide a unified language for causal inference and facilitate the implementation and interpretation of causal inference methods Rubin (2005). It relies on the definition of potential outcomes under hypothetical interventions, which refers to what would have happened to each individual under assignment of each intervention. The causal effect is then expressed as a contrast of these potential outcomes, in the entire population (for the estimation of the average treatment effect - ATE) or among those actually exposed to the intervention (for the estimation of the average effect on the treated - ATT). The potential outcomes framework and statistical methods for causal inference are particularly useful to draw causal inferences from observational studies when controlled randomised trials are not feasible, or when one wants to estimate real-world effects Hernan (2004).

Propensity score (PS) methods were developed within the potential outcomes framework to address confounding bias ROSENBAUM and RUBIN (1983). Briefly, the propensity score is defined as the individual's probability of receiving the treatment, conditionally to their own characteristics. The propensity

score can be used to construct an inverse-probability-of-treatment (IPTW) estimator, in which individuals are weighted by the inverse of the propensity score of the treatment actually received. Under a set of assumptions, the comparison of outcomes between treatment groups in the weighted sample is the causal treatment effect. Observational studies using IPTW estimators are gaining increasing recognition, especially by regulatory agencies such as the Food and Drug Administration (FDA), for example in the evaluation of COVID-19 vaccine Pawlowski et al. (2021).

Most studies using IPTW focus on the estimation of the causal effect of a binary treatment, and fit a logistic regression model to estimate the propensity score and derive the weights. However, although the situation of a treatment with more than two levels is quite common, appropriate implementations of IPTW to this setting remain rare, despite the availability of statistical methods for this situation McCaffrey et al. (2013). Instead, it is frequent in the literature to observe that a series of binary logistic regressions are used rather than considering a categorical variable for the treatment, together with modelling strategies for categorical outcomes. Rodríguez-Bernal et al. (2020); Ali et al. (2015). For example, the effect of treatment adherence can be categorized to assess a dose-response effect of different adherence levels on an outcome. This is the case for instance when evaluating the causal effect of adherence to Continuous Positive Airway Pressure (CPAP) treatment – the first-line treatment for moderate to severe obstructive sleep apnea – which is often considered to be efficient with a use greater than 4 hours per night. However, the effect of different thresholds of CPAP adherence has not been extensively investigated.

When investigating the causal effect of multiple treatments or a treatment with more than two levels, a multinomial regression model is a standard approach for the estimation of the propensity score. This approach is commonly implemented in the literature Bozorgmehri et al. (2021); Bettega et al. (2022). Estimating the weights from the predictions of a multinomial regression model relies on the same assumption of correct specification of the propensity score model as with a binary treatment. The parametric form of this model may be difficult to specify, which motivates the use of alternative methods such as generalized boosted models (GBM), Covariate Balancing Propensity Score (CBPS) or Augmented inverse probability of treatment weighted estimator (AIPTW). For binary treatments, several authors showed that these models could lead to a better balance of the covariates between groups after weighting, and a lower mean squared error of the treatment effect estimate Harder et al. (2010); Lee et al. (2009). For multilevel treatments, McCaffrey *et al.* showed that GBM allowed a better balance between groups and reduced the variability of the weights compared to a multinomial regression McCaffrey et al. (2004).

GBMs are the second most common method for the estimation of the weights for IPTW with a multilevel tretament in the medical literature. This can be explained by the existence of the tutorial developed by McCaffrey *et al* McCaffrey et al. (2004), as well as the existence of the R package *twang* Ridgeway et al. (2022), which provides an off-the-shelf implementation of IPTW using GBMs. However, it is important to determine if the popularity of GBMs stems from

superior statistical properties, or convenience in the implementation.

Therefore, the aims of this paper is to evaluate the statistical performances of multinomial regression, CBPS and GBMs for the estimation of propensity score weights used in IPTW, and AIPW with a multi-level treatment using Monte-Carlo simulations. More specifically, we will explore the balancing abilities of each method, and the bias in the resulting ATE estimates. The paper is organised as follows. Section 2 introduces the potential outcomes framework, the IPTW and AIPW estimators, and methods for the estimation of the weights. Sections 3 and 4 present the design and the results of a Monte-Carlo simulation assessing the performances of the fours methods for the estimation of the ATE. Section 5 presents an application on these methods for the estimation of the causal effect of CPAP adherence on daytime sleepiness among individuals with obstructive sleep apnea. Section 6 provides a discussion of the results and avenues for future research.

2 Theoretical background

2.1 Potential outcome framework

The potential outcome framework was proposed by Rubin (1974) to provide a causal language for causal inference from observational data. With a binary outcome, the aim to estimate the causal effect of a treatment Z (Z=1 if treated, 0 if not) on an outcome Y. The potential outcomes framework relies on the concept of hypothetical interventions and hypothetical counterfactual worlds. We note $Y(0)$ the potential outcome under hypothetical treatment Z=0 and $Y(1)$ the potential outcome under hypothetical treatment Z=1. Then, the average treatment effect in the population (ATE) can be expressed as the contrast between these two potential outcomes $\mathbb{E}[Y(1) - Y(0)]$. Or the ATT is an estimate of the causal effect on the sub-population of treated individuals. It is defined as $\mathbb{E}[Y(1)|Z = 1] - \mathbb{E}[Y(0)|Z = 1]$. It differs from ATE when the treatment does not affect treated and untreated individuals in the same way.

In practice, an issue is that we cannot observe both Y_1 and Y_0 at the individual level, but instead, we only observe the effect on the outcome under one of a treatment levels. This is called the fundamental problem of causal inference (Rubin, 1974). However, under a set of four assumptions, causal effects can be identifiable from the data at hand. These assumptions are consistency, non-interference, positivity and conditional exchangeability.

(i) The consistency assumption is often stated such that an individual's observed outcome under their observed treatment is the same as their potential outcome under the hypothetical treatment (Robins et al., 2000; Hernan, 2004). This assumes that observing is the same as intervening and that there is only one version of the treatment. A precise definition of the treatment (components, timing, dose, etc.) is required to increase the plausibility of the consistency assumption. The consistency is defined

in terms of the potential outcomes for an individual i as $Y_i(1) = Y_i|z = 1$ Cole and Frangakis (2009).

(ii) The assumption of no-interference states that the treatment received by an individual is independent of other individuals potential outcomes: for two individuals i and j with $i \neq j$, $Z_i \perp Y_j(0), Y_j(1)$. An example of a violation of this assumption can be observed in vaccine studies where vaccines may have an indirect protective effect on other individuals potential outcomes.

(iii) Conditional exchangeability refers to the assumption of no unmeasured confounding, that is, there is no variable associated both with the treatment and the outcome that are not available for the analysis. This hypothesis can be formally defined as $Y(0), Y(1) \perp Z|C]$ with C a set of confounders. This assumption cannot be tested empirically and its plausibility can only be discussed using context-specific knowledge. (Cole and Hernan, 2008).

(iv) Positivity states that given their own characteristics, every individual has a non-zero probability of receiving any level of treatment (Cole and Hernan, 2008): for all i $P(Z_i = z) > 0$. For example, the existence of formal contraindications to one of the treatments evaluated in the target population is a violation of the this assumption.

In randomized controlled trials (RCTs), the assumptions of exchangeability – a stronger assumption than conditional exchangeability – and positivity are ensured by design, which is not the case in observational studies. In these studies, assuming conditional exchangeability, statistical methods have been proposed to incorporate confounders in the analysis to obtain unbiased estimates of the treatment effect. Two categories of methods can be considered: i) methods based on the outcome such as regression adjustment or g-computation and ii) methods based on the treatment assignment mechanism, such as propensity score methods. In this work we focus on propensity score-based methods, that model the treatment assignment separately from its effect on the outcome, which brings it closer to RCTs, as confounding is addressed before analysing outcome data. In this section, we first introduce propensity score methods for binary treatments, and then their extension to multilevel treatments.

2.2 Propensity scores and IPTW

The propensity score is the probability for each patient $i, i = 1, ..., n$ of belonging to their own treatment group, knowing their individual characteristics $C_1, ..., C_k$, with k number of chacharacteristics. $PS = P(Z = z|C_1 = c_1, ..., C_k = c_k)$. Propensity scores are balancing scores, as defined by (ROSENBAUM and RUBIN, 1983). A balancing score, b(c), is a function of the observed covariates c such that the conditional distribution of c given b(c) is the same for treated (Z = 1) and control (Z = 0) unit. This means that at each value of the propensity score, individuals have, on average, similar distributions of their characteristics.

Therefore, propensity scores are used to balance covariates between groups for the estimation of the ATE. This requires, in addition to the four identification assumptions, the propensity score model to be correctly specified, i.e. the model for the treatment assignment includes all the confounders, in their correct functional form, and includes potential interactions. Current guidelines recommend the inclusion of all risk factors (confounders or not) and avoid including predictors of treatment (instrumental variables), in the propensity score model (Lefebvre et al., 2008). The validity of propensity score estimators for the estimation of the treatment effect relies on their ability to balance the covariates between treatment groups. This can be assessed using metrics such as the standardised mean difference (SMD). The SMD, for a covariate C and a binary treatment, is defined as follows: $SMD = \frac{(\bar{C}_1 - \bar{C}_0)}{\sqrt{\sigma_1^2 + \sigma_0^2}}$ with $\bar{C}_0$ and $\bar{C}_1$ the sample mean of the variables C and σ_0^2 and σ_1^2 the sample variance of variable C in treatment group 0 and 1, respectively. SMDs can be calculated for each variable before weighting, and after weighting, by using the weighted means and variances instead. Typically, SMDs ¡10% are considered acceptable.

Once estimated and achieving a good covariate balance, the propensity score can be used in different ways for the estimation of the treatment effect. Most common approaches are matching, stratification and inverse-probability-of-treatment weighting (IPTW). Matching, which involves the creation of pairs of treated and untreated individuals with similar propensity score values, allows us to estimate the ATT. However, matching typically leads to a loss in sample size, affecting both the power and generalisability of the study. This phenomenon is amplified when the number of treatment levels increases. Stratification creates sub-classes of patients with similar propensity scores in which treatment effects are estimated and pooled across strata. Stratification often leads to residual imbalances, especially if the number of strata is too small. Thus, in this study, we will focus on IPTW, which has very good balancing properties. In IPTW, individuals are re-weighted by the inverse of their propensity score to create a pseudo-population in which treatment groups are comparable on measured covariates. Individuals with a low probability of receiving the treatment they actually received will therefore be upweighted. The presence of extreme weights indicates violations or near violations of the positivity assumption. These inverse weights allow the estimation of the ATE, but other weighting schemes can be used to estimate the ATT or other estimands such as the average treatment effect in the overlapping population (ATO) Crump et al. (2006). The causal treatment effect can then be estimated using either a difference in weighted mean outcomes, or using a weighted regression model. However, standard errors of the treatment effect estimates must account for (i) the correlation in the data induced by the weighting procedure and (ii) the uncertainty in propensity score estimation. This can be done using the Delta method, or via non-parametric bootstrap Austin (2016).

2.3 Multilevel treatment

IPTW theory was originally developed for binary treatments and its implementation for multilevel treatments has received little attention, which may explain its limited use in practice Yoshida et al. (2019). Some examples of the application of propensity scores for multilevel treatments can however be found in Ali et al. (2015) who investigated the association between long-acting beta-agonists, oral corticosteroids and both for asthma exacerbation and Ahn et al. (2021) who investigated the association between multiple level of exercise habits and stroke, heart failure, and mortality. With multiple or multilevel treatments, each individual has multiple propensity scores. For a binary treatment, each patient will have two propensity scores, but only one is directly estimated, since the other is its complement. For multilevel treatments, there are as many scores as treatment groups. These can be obtained from the predictions of a multinomial regression model, but alternative methods are presented in the next subsections. When estimating the ATT, multilevel treatments offer a wide variety of choices of causal estimands. The choice of the comparator should be guided by the objective of the study and underlying clinical question. For the ATE, there are as many contrasts as the number of two-by-two comparisons, although they may not all be relevant for the research question of interest.

2.4 Machine learning for propensity score estimation

2.4.1 Standard prediction

Propensity scores are individual predictions and therefore classification or prediction methods based on machine learning could be thought to be useful, especially when the parametric form of the propensity score model is complex. For instance, random forests are robust non-parametric prediction algorithms. The initial goal of machine learning algorithms is to minimize the prediction error. However, for causal inference using propensity scores, the goal is not to predict treatment allocation as best as possible, but to achieve balance between treatment groups. Furthermore, including in the propensity score variables strongly associated to the treatment (but not the outcome) increases the bias in the estimate of the causal treatment effect. Therefore, some machine learning algorithms, such as Generalized Boosted Models (GBMs) have been proposed to overcome this problem, by changing the targeted metrics for optimisation.

2.4.2 Generalized Boosted Models

GBMs are non-parametric methods based on boosting and decision trees, and present the advantage of not making any *a priori* hypothesis on the distribution of the data. GBMs are commonly used for estimating propensity scores (Shi et al., 2020; Wang et al., 2019) especially in the multilevel treatment setting. Their introduction for this purpose was proposed by McCaffrey, and GBMs for propensity scores are implemented in the R package *twang* Ridgeway et al. (2022).

Boosting creates a classifier from a set of weak classifiers (Hastie et al., 2009b). For propensity score estimation with GBMs, the weak classifiers are a set of regression trees, that capture complex relationships between treatment assignment and confounders, while avoiding over-fitting (McCaffrey et al., 2004; Ridgeway, 1999). This set of predictions are combined to obtain a strong classifier. The shrinkage parameter constrains the maximum contribution of each classifier to the strong classifier. Individuals misclassified by the weak classifiers have a higher probability of being incorporated into the next weak classifier. To improve performance (Friedman, 2001) each weak classifier is trained on a subset of the individuals controlled by the sub-sampling rate. When using GBMs for the estimation of propensity scores, the optimal number of trees is chosen to minimize a stopping rule criterion. This criterion focuses on differences in the weighted distributions of the confounders between treatment groups, often summarised with the standardised mean difference or the Kolmogorov-Smirnov distance.

GBMs showed interesting results for propensity score estimation (McCaffrey et al., 2004). They allow a more stable estimation of weights than parametric models and a less biased estimation of the treatment effect in non-linear situations (Harder et al., 2010; Lee et al., 2009). Moreover GBMs incorporate the variable selection step (McCaffrey et al., 2004). This property is particularly useful when dealing with a large number of confounding factors.

2.5 Alternative parametric methods for IPTW

Fully non-parametric methods typically require large sample sizes, and alternative parametric methods have been proposed to improve the performances of the standard IPTW estimator. Covariate Balancing Propensity Score (CBPS) aims to improve propensity-score based estimation by optimizing confounder balance, whereas augmented inverse-probability weighting incorporate an additional step of confounder adjustment to correct for remaining residual imbalances.

2.5.1 Covariate balancing propensity score

CBPS is a method for estimating PS, a major difference with other methods is that it combines in a single model the mechanism of treatment assignment and the balance between covariates.We estimate the CBPS by using the moment conditions based on the covariate balancing property under the GMM or EL framework (Hayashi, 2000; Owen, 2001) . The CBPS possesses various appealing attributes. First, CBPS estimation reduces the impact of potential errors in specifying a parametric propensity score model by selecting parameter values that optimize the resulting covariate balance. Second this method is easily generalizable to treatments with more than 2 levels or continous treatments (Imbens, 1999; Imai and van Dyk, 2004). In this article CBPS are estimated using generalized method of moments as proposed by Imai, Kosuke et al. (Imai and Ratkovic, 2013).

2.5.2 Augmented inverse probability of treatment weighted estimator

IPTW estimators often have inflated imprecision compared to multivariable regression adjustment. By combining both approaches, it is possible to gain efficiency. AIPTW is a doubly robust method that works by augmenting the IPTW estimator with a mean-zero term based on regression on the outcome. The estimation of the weights is done via regression modelling, as in the standard IPTW estimation. The AIPTW estimator is a doubly robust estimator, which means that the estimator remains consistent even if one of the propensity score or outcome model is misspecified (Tsiatis et al., 2008).

$$\hat{ATE}_{AIPW} = \frac{1}{n}\sum_{i=1}^{n}\left\{\left[\frac{X_i Y_i}{\hat{\pi}(Z_i)} - \frac{(1-X_i)Y_i}{1-\hat{\pi}(Z_i)}\right] - \frac{(X_i-\hat{\pi}(Z_i))}{\hat{\pi}(Z_i)(1-\hat{\pi}(Z_i))}\right.$$
$$\left.[(1-\hat{\pi}(Z_i))\hat{\mathbb{E}}(Y_i \| X_i = 1, Z_i) + \hat{\pi}(Z_i)\hat{\mathbb{E}}(Y_i \| X_i = 0, Z_i)]\right\}$$

Note that while GBM and CBPS are alternative methods to multinomial (or logistic) regression for the estimation the propensity score (first step), AIPTW modifies the treatment effect estimation model (second step).

3 Simulation study

3.1 Aims

We evaluate the performance of GBM and multinomial logistic regression for PS estimation under several scenarios including model misspecification and varying sample size. The aim is to evaluate the balance properties of the weights obtained with different specifications of GBM and multinomial logistic regression models in each scenario, and to investigate both the bias and precision in the estimation of the ATE.

3.2 Simulation structure

We place ourselves in the context of an observational study with a continuous outcome and a multilevel treatment with 3 modalities, as well as 26 variables, 14 continuous and 12 binary. 7 continuous and 8 binary variables are independently associated with the treatment and the outcome, 3 continuous and 2 binary variables are independently associated with the outcome only, 2 continuous and 1 binary variables are independently associated with the treatment only and 2 continuous and 1 binary variables are not associated with either the treatment or the outcome (noise variables).

We use these variables in 3 treatment assignment scenarios : (A) a model with simple confounder effect,(B) a model with confounder and interaction terms and a (C) a model with confounder, noise and instrumental variables.

In order to evaluate the impact of the sample size on the performance of the methods, we consider 3 hypothetical cohorts with different sample sizes, $n = 1000$, $n = 5000$ and $n = 10000$.

1000 datasets are generated for each of the 3 simulation scenarios.

The variables involved in the simulations are summarized in the diagrams and table below.

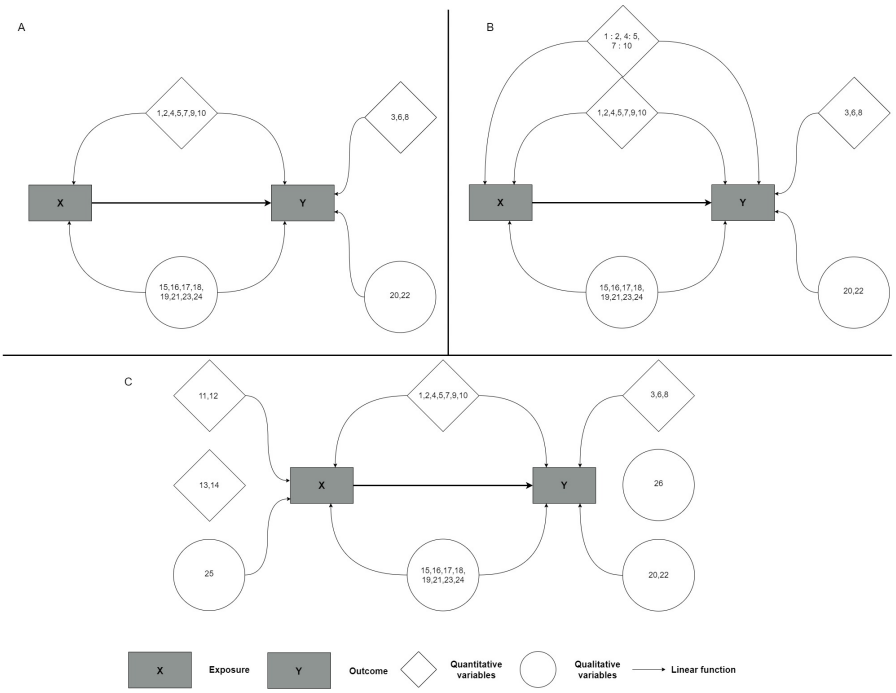

Figure 1: Variable Relationships in simulation for the three simulation scenarios Diagram presenting the three simulated scenarios, (A) additive scenario (B) scenario with interactions terms (C) scenario with instrumentals variables and noise Variables 13,14 and 26 are noise

3.3 Data-generating mechanisms

We generate datasets with 26 variables, 14 continuous variables following a standard normal distribution $\mathcal{N}(0, 1)$ and 12 binary variables following binomial distributions with probabilities : 0.05,0.2 and 0.5.

3.4 Treatment assignment mechanism

The variables use to calculate the probability of the multinomial distribution vary according to the scenario, with parameters defined as:

$Z_A = softmax(\beta_1 continuous_1 + \beta_2 continuous_2 + \beta_3 continuous_4 + \beta_4 continuous_5 + \beta_5\, continuous_7 + \beta_6\, continuous_9 + \beta_7\, continuous_{10} + \beta_8\, binary_{15} + \beta_9\, binary_{16} + \beta_{10}\ binary_{17} + \beta_{11}\ binary_{18} + \beta_{12}\ binary_{19} + \beta_{13}\ binary_{21} + \beta_{14}\ binary_{23} + \beta_{15}\ binary_{24})$

$Z_B = softmax(Z_A + \beta_{18}\ continuous_1 \times continuous_2 + \beta_{19}\ continuous_4 \times continuous_5 + \beta_{20}\ continuous_7 \times continuous_{10})$

$Z_C = softmax(Z_A + \beta_{21}\ continuous_{11} + \beta_{22}\ continuous_{12} + \beta_{23}\ binary_{25})$

All coefficient used in the treatment assignment mechanism are presented in supplementary table S1, S2 and S3.

3.5 Outcome simulation

The outcome Y is a continuous variables simulated in all scenarios using 20 of the 26 covariates and a noise ϵ variable that is generate using a $\mathcal{N}(0, 0.1)$. The treatment prevalence is similar in the 3 scenarios for each group is $Z = 0 = 30\%$, $Z = 1 = 37\%$ and $Z = 2 = 33\%$ .The true contrast between $Y(0)$ and $Y(1)$ and $Y(2)$ are 0.75 and 1.5.

The generation of the outcome in Scenario (A) and (C) use the same formula:
$Y = -1.206x_1 + 1.23x_2 - 0.3x_3 + 0.506x_4 + 0.146x_5 + 1.543x_6 - 0.656x_7 + 0.6x_8 + 0.035x_9 - 0.78x_{10} + 1.648x_{11} + 0.092x_{12} + 0.818x_{13} + 0.92x_{14} + 0.193x_{15} + 0.549x_{16} + 0.967x_{17} + 0.612x_{18} - 0.766x_{19} - 1.383x_{20} + 0.75\mathbb{1}_{Treatment=1} + 1.5\mathbb{1}_{Treatment=2} + \epsilon$
Scenario (B) :
$Y_B = Y_{A,C} + continuous_1 \times continuous_2 + continuous_4 \times continuous_5 + continuous_7 \times continuous_{10}$

variables	type	description	outcome	Treatment_additive	Treatment_interaction	Treatment_noise
x1	continous	treatment/outcome	X	X	X	X
x2	continous	treatment/outcome	X	X	X	X
x3	continous	outcome	X			
x4	continous	treatment/outcome	X	X	X	X
x5	continous	treatment/outcome	X	X	X	X
x6	continous	outcome	X			
x7	continous	treatment/outcome	X	X	X	X
x8	continous	outcome	X			
x9	continous	treatment/outcome	X	X	X	X
x10	continous	treatment/outcome	X	X	X	X
x11	continous	instrumental				X
x12	continous	instrumental				X
x13	continous	noise				
x14	continous	noise				
x15	boolean	treatment/outcome	X	X	X	X
x16	boolean	treatment/outcome	X	X	X	X
x17	boolean	treatment/outcome	X	X	X	X
x18	boolean	treatment/outcome	X	X	X	X
x19	boolean	treatment/outcome	X	X	X	X
x20	boolean	outcome	X			
x21	boolean	treatment/outcome	X	X	X	X
x22	boolean	outcome	X			
x23	boolean	treatment/outcome	X	X	X	X
x24	boolean	treatment/outcome	X	X	X	X
x25	boolean	instrumental				X
x26	boolean	noise				
x1:x2	continous	treatment/outcome	X*		X	
x4:x5	continous	treatment/outcome	X*		X	
x7:x10	continous	treatment/outcome	X*		X	

3.6 Methods compared

The different methods compared for the estimation of propensity scores are (i) a purely additive multinomial logistic regression model using all risk factors (confounders or not), (ii) a GBM, (iii) a multinomial logistic regression, (iv) an AIPTW and (v) CBPS using all risk factors and taking into account the interaction.The GBMs minimises the mean standardized difference between the variables using all risk factors. In the third scenario we compare seven models (i) a GBM, (ii) a multinomial logistic regression, (iii) a CBPS and (iv) an AIPTW with same variables as GBM and (v) a multinomial logistic regression, (vi) a CBPS and (vii) an AIPTW fit without instrumental variables and noise variables. The GBM minimises the mean standardized difference between the variables using all risk factors, the instrumental variables and noise variables. In the second and third scenario the linear model which does not take into account the right variables is therefore misspecified.

The ATE is estimated with a weighted linear regression estimating the effect of the treatment on the outcome. No covariates are included in the outcome model. Non-parametric bootstrap with 500 replications was used to estimate the standard error of the ATE and normal-based 95% confidence intervals were constructed after checking the distribution of the bootstrap replicates.

3.7 Measure of interest

We compare for each propensity score estimation method the distribution of weights and the presence of extreme weights.

We also evaluate in each scenarios the mean (95 % CI) ATE and the bias in the estimation of the ATE and the coverage rate of the true value of the ATE. The empirical variance of the ATE among the 1000 simulation, the bootstrap variances of the ATE and the model based variance for the regression. We also compute the Monte-Carlo errors among the 1000 simulations.

Finaly, we compare the SMD of the variables for each treatment levels obtain with the different methods.
The statistical analysis has been done using `R version 4.0.3`, the implementation of the GBMs is the one of the `R packages GBM version 2.18`.

4 Results

The results are presented for a sample size of 10,000. The full results for the the sample sizes are presented in the Supplementary Tables.

4.1 Distribution and balancing ability of the weights

We compared the distribution of the weights obtained using the different approaches in the three scenarios.

Figure 2 presents the distribution of the mean weights across the 1000 simulations for the three scenarios. In the three scenarios GBMs led to lower weights and a lower dispersion of weights. The maximum weight for correctly specified models is 178. This weight is observed for multinomial regressions in scenario 2. The maximum weight observed among the GBMs is 66.8 in the case of scenario 3. CBPS leads to a maximum weight of 125 in scenario 2. The weights obtained with the different models for the 3 scenarios are summarised in figure 4.1 and table 1.

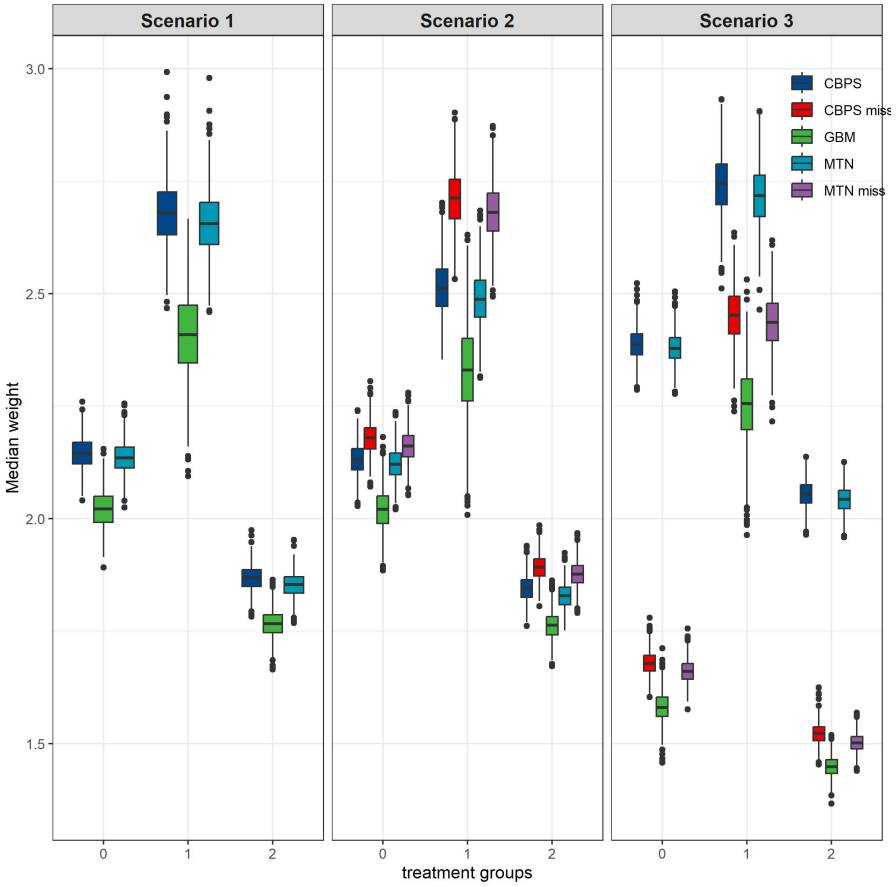

Figure 2: Boxplot of median weight over 1000 replication (n = 10000)

Table 1: Table of weights according to the model scenarios (n = 10000)

Treatment group	IPTW	GBM	CBPS	IPTW misspecified	CBPS misspecified
Simple confounders					
0	2.1 (1.6; 3)	2 (1.6; 2.8)	2.1 (1.7; 3)		
1	2.7 (1.8; 4.5)	2.4 (1.7; 4)	2.7 (1.8; 4.5)		
2	1.9 (1.4; 2.7)	1.8 (1.4; 2.5)	1.9 (1.4; 2.7)		
Simple confounders and interaction terms					
0	2.1 (1.6; 3)	2 (1.6; 2.8)	2.1 (1.6; 3)	2.2 (1.7; 3.1)	2.2 (1.7; 3.1)
1	2.5 (1.7; 4.3)	2.3 (1.6; 3.8)	2.5 (1.7; 4.3)	2.7 (1.8; 4.5)	2.7 (1.9; 4.5)
2	1.8 (1.4; 2.8)	1.8 (1.4; 2.5)	1.8 (1.4; 2.8)	1.9 (1.4; 2.8)	1.9 (1.5; 2.8)
Simple confounders and noise					
0	2.4 (1.9; 3.2)	1.6 (1.2; 2.4)	2.4 (1.9; 3.2)	1.7 (1.3; 2.6)	1.7 (1.3; 2.6)
1	2.7 (1.8; 4.5)	2.3 (1.6; 3.7)	2.7 (1.9; 4.5)	2.4 (1.6; 4.2)	2.5 (1.7; 4.2)
2	2 (1.6; 2.8)	1.4 (1.2; 2.1)	2.1 (1.6; 2.8)	1.5 (1.2; 2.3)	1.5 (1.2; 2.3)

[a] Results are presented in the form: median(1st quartile; 3rd quartile)

4.2 Covariates balance

It is interesting when comparing causal inference models to compare whether one of the models leads to a better balance between the variables for each treatment group. In figure 4.2 we present the mean SMD obtained from the 1000 simulations for each of the variables in the 3 scenarios.

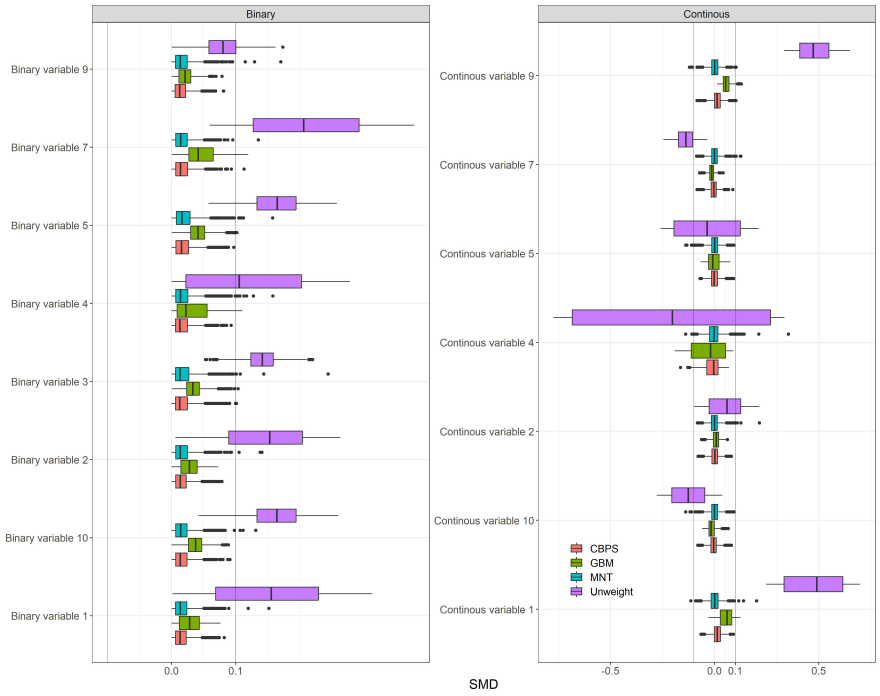

Figure 3: boxplot of SMD of covariates for simple confounders scenario

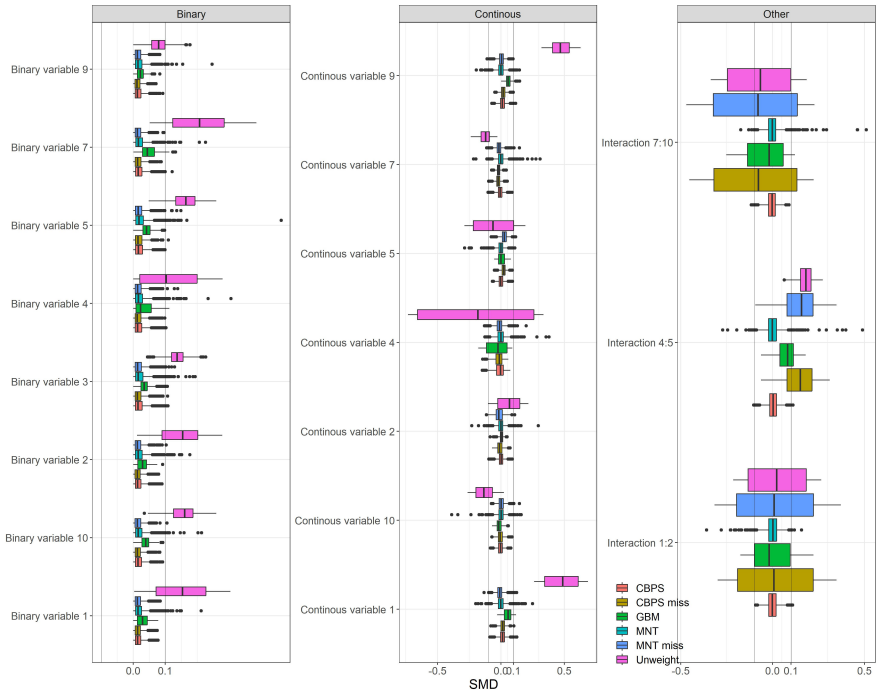

Figure 4: boxplot of SMD of covariates for interaction terms scenario

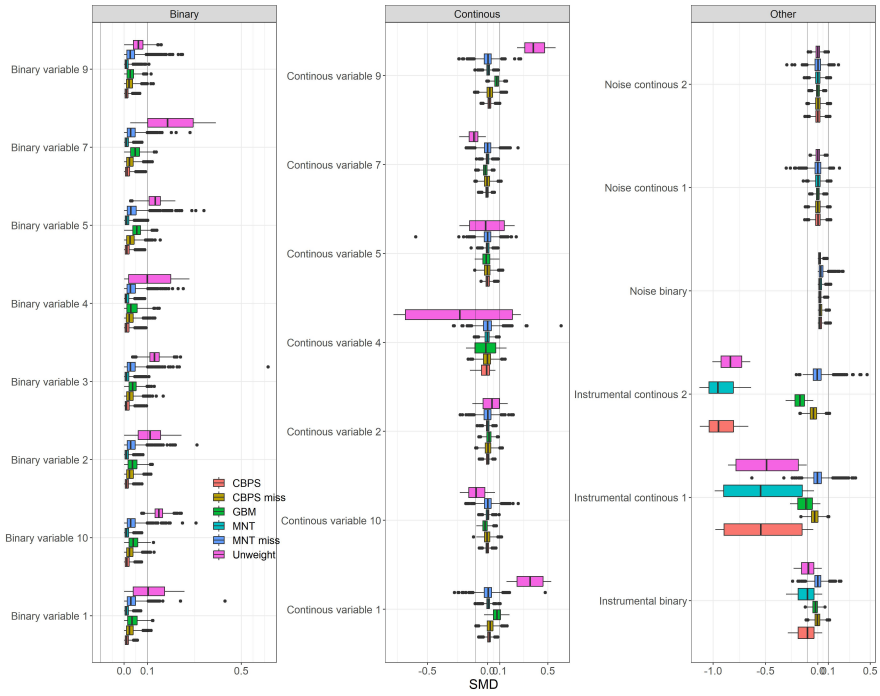

Figure 5: boxplot of SMD of covariates for noise and instrumental variables scenario

The GBMs lead to a slightly poorer balance than the logistic regression in the simple confounders scenario.

The GBMs lead to a slightly poorer balance than the correctly specified logistic regression models but better than the misspecified models in the scenario with interaction terms. This scenario is the only one in which one of the models leads to an imbalance with a SMD greater than 0.1 for the regression model without interaction terms.

In the scenario with instrumental variables and noise the GBMs lead to a lower quality balance in even misspecified regressions.

4.3 Average Treatment Effect

The result of this part presents the average ATE over the 1000 simulations and its bootstrap confidence interval.

The GBMs lead in all three scenarios and for all sample sizes to a slightly biased estimate of the ATE. In the scenario with interaction terms the ATE estimate of the GBMs remains less biased than that of the misspecified regres-

sion. In the case of the scenario with simple confounders effect the GBM led to a mean ATE of for contrast Y(1) - Y(0) 0.79 [0.59; 0.99], for contrast Y(2) - Y(0) 1.5 [1.34; 1.65]. The multinomial logistic regression led to a mean ATE of for contrast Y(1) - Y(0) 0.75 [0.51; 0.99], for contrast Y(2) - Y(0) 1.5 [1.33; 1.67]. The CBPS led to a mean ATE of for contrast Y(1) - Y(0) 0.77 [0.54; 0.99], for contrast Y(2) - Y(0) 1.5 [1.33; 1.67]. The AIPTW led to a mean ATE of for contrast Y(1) - Y(0) 0.75 [0.74; 0.76], for contrast Y(2) - Y(0) 1.5 [1.49; 1.51]. The adjustment led to a mean ATE of for contrast Y(1) - Y(0) 0.75 [0.74; 0.76], for contrast Y(2) - Y(0) 1.5 [1.5; 1.5]. This difference between regression and GBMs seems particularly pronounced when the sample size is small.

The GBMs seem to lead to a higher ATE estimate with narrower 95% confidence intervals than the regressions. This phenomenon is particularly noticeable when noise is introduced into the regression models which leads to a significant widening of the 95% confidence interval. In the case of the scenario with confounders, noise and instrumental variables the GBM led to a mean ATE of for contrast Y(1) - Y(0) 0.79 [0.56; 1.02], for contrast Y(2) - Y(0) 1.49 [1.28; 1.7]. The multinomial logistic regression with noise and instrumental variables led to a mean ATE of for contrast Y(1) - Y(0) 0.76 [0.44; 1.09], for contrast Y(2) - Y(0) 1.51 [1.21; 1.8]. The multinomial logistic regression correctly specify led to a mean ATE of for contrast Y(1) - Y(0) 0.76 [0.53; 0.99], for contrast Y(2) - Y(0) 1.5 [1.34; 1.66]. The CBPS with noise and instrumental variables led to a mean ATE of for contrast Y(1) - Y(0) 0.77 [0.47; 1.06], for contrast Y(2) - Y(0) 1.51 [1.24; 1.77]. The CBPS correctly specify led to a mean ATE of for contrast Y(1) - Y(0) 0.77 [0.55; 0.99], for contrast Y(2) - Y(0) 1.5 [1.34; 1.66]. The AIPTW with noise and instrumental variables led to a mean ATE of for contrast Y(1) - Y(0) 0.75 [0.68; 0.82], for contrast Y(2) - Y(0) 1.5 [1.44; 1.57]. The AIPTW correctly specify led to a mean ATE of for contrast Y(1) - Y(0) 0.75 [0.69; 0.81], for contrast Y(2) - Y(0) 1.5 [1.45; 1.56]. The adjustment with noise and instrumental variables led to a mean ATE of for contrast Y(1) - Y(0) 0.75 [0.69; 0.81], for contrast Y(2) - Y(0) 1.5 [1.45; 1.56]. The adjustment correctly specify led to a mean ATE of for contrast Y(1) - Y(0) 0.75 [0.69; 0.81], for contrast Y(2) - Y(0) 1.5 [1.45; 1.55].

The different estimates of the ATE and the 95% confidence interval for the 3 scenarios are presented in the figure 4.3.

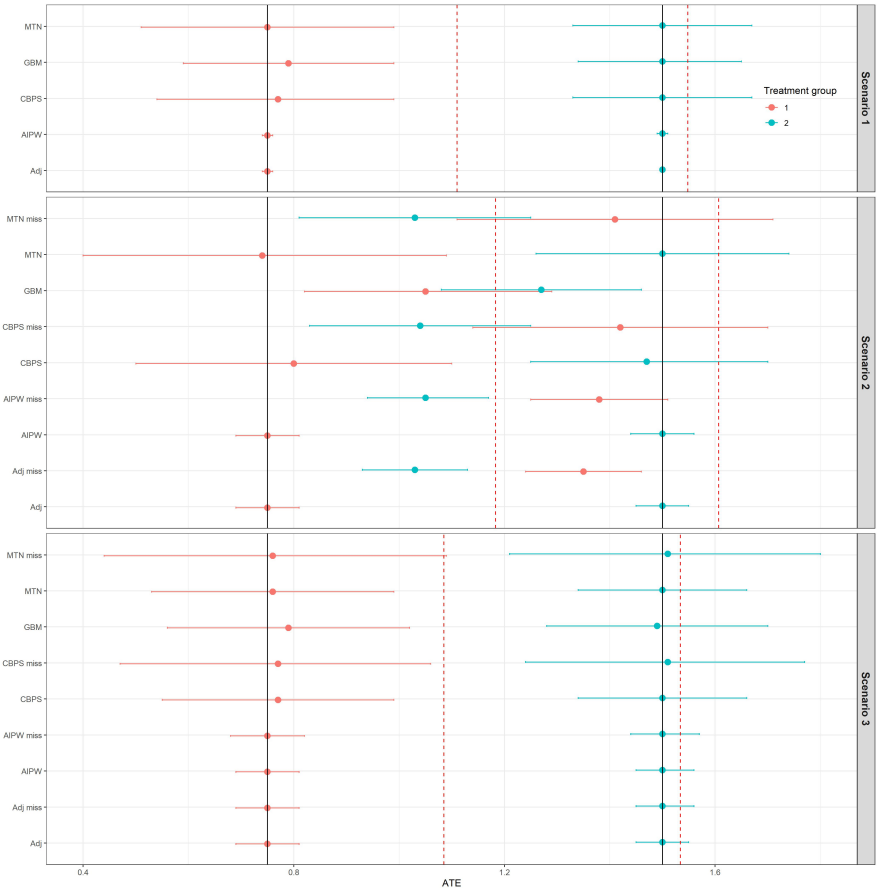

Figure 6: Mean ATE

The results for all sample sizes and the coverage rate of the actual value are presented in the Table 2.

Table 2: Table of bias according to the model scenarios (n = 10000)

Treatment group	Scenario 1		Scenario 2		Scenario 3	
	1 - 0	2 - 0	1 - 0	2 - 0	1 - 0	2 - 0
AIPW	0 [0; 0]94 %	0 [0; 0]97 %	0 [-0.1; 0.1]95 %	0 [-0.1; 0.1]95 %	0 [-0.1; 0.1]95 %	0 [-0.1; 0.1]96 %
AIPW miss			0.6 [0.5; 0.8]0 %	-0.4 [-0.6; -0.3]0 %	0 [-0.1; 0.1]96 %	0 [-0.1; 0.1]94 %
CBPS	0 [-0.2; 0.2]100 %	0 [-0.2; 0.2]100 %	0.1 [-0.2; 0.4]98 %	0 [-0.2; 0.2]100 %	0 [-0.2; 0.2]100 %	0 [-0.2; 0.2]100 %
CBPS miss			0.7 [0.4; 0.9]0 %	-0.5 [-0.7; -0.2]0 %	0 [-0.3; 0.3]99 %	0 [-0.3; 0.3]99 %
GBM	0 [-0.2; 0.2]100 %	0 [-0.2; 0.1]100 %	0.3 [0.1; 0.5]23 %	-0.2 [-0.4; 0]27 %	0 [-0.2; 0.3]99 %	0 [-0.2; 0.2]100 %
MTN	0 [-0.2; 0.2]100 %	0 [-0.2; 0.2]100 %	0 [-0.3; 0.3]99 %	0 [-0.2; 0.2]100 %	0 [-0.2; 0.2]100 %	0 [-0.2; 0.2]100 %
MTN miss			0.7 [0.4; 1]1 %	-0.5 [-0.7; -0.2]0 %	0 [-0.3; 0.3]98 %	0 [-0.3; 0.3]99 %
Adj	0 [0; 0]94 %	0 [0; 0]96 %	0 [-0.1; 0.1]94 %	0 [-0.1; 0.1]96 %	0 [-0.1; 0.1]96 %	0 [-0.1; 0.1]96 %
Adju miss			0.6 [0.5; 0.7]0 %	-0.5 [-0.6; -0.4]0 %	0 [-0.1; 0.1]96 %	0 [-0.1; 0.1]95 %

4.4 Standard error

In this section we present the standard error obtained with the different methods in the 3 scenarios. GBMs lead to a standard error that is generally lower than that obtained with IPTW or CBPS, but higher than that obtained with AIPTW or adjustment.

In the case of the scenario with simple confounders effect the GBM led to a montecarlo error of for Z=1 0.05 and 0.03 for Z=2. A bootstrap standard error of 0.1 for Z=1 and 0.08 for Z=2. And a model based error of 0.07 for Z=1 and 0.07 for Z=2. The multinomial logistic regression led to a montecarlo error of for Z=1 0.09 and 0.05 for Z=2. A bootstrap standard error of 0.12 for Z=1 and 0.09 for Z=2. And a model based error of 0.07 for Z=1 and 0.07 for Z=2. The CBPS led to a montecarlo error of for Z=1 0.08 and 0.04 for Z=2. A bootstrap standard error of 0.12 for Z=1 and 0.09 for Z=2. And a model based error of 0.07 for Z=1 and 0.07 for Z=2. The AIPTW led to a montecarlo error of for Z=1 0 and 0 for Z=2. A bootstrap standard error of 0 for Z=1 and 0 for Z=2. And a model based error of 0 for Z=1 and 0 for Z=2. The adjustment led to a montecarlo error of for Z=1 0 and 0 for Z=2. A bootstrap standard error of 0 for Z=1 and 0 for Z=2. And a model based error of 0 for Z=1 and 0 for Z=2.

In figure 4.4 we present the standard error estimator obtained from the 1000 simulations in the 3 scenarios. Full standard error obtained in each scenarios are summarized in table 3 and the ratio of model standard error on empirical standard error is presented in table 4.

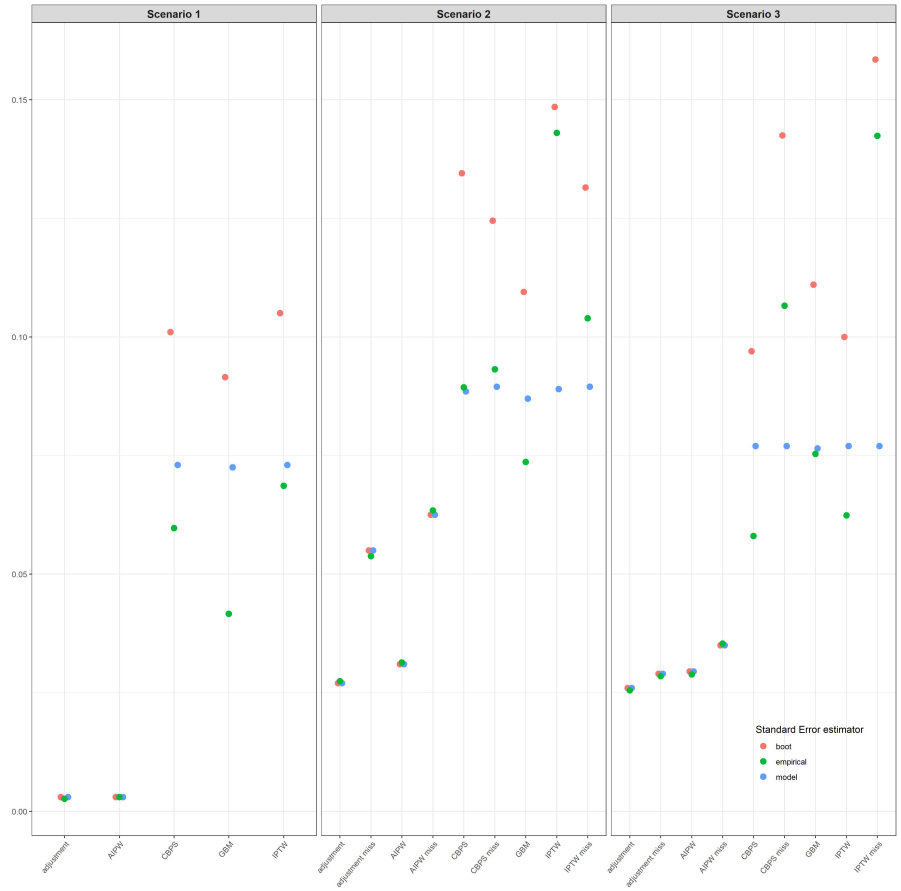

Figure 7: Standard Error estimator

Table 3: Table of standard error according to the model scenarios (n = 10000)

	Scenario 1		Scenario 2		Scenario 3	
Treatment group	1	2	1	2	1	2
Bootstrap standard error						
AIPW	0	0	0.03	0.03	0.03	0.03
AIPW miss			0.07	0.06	0.04	0.03
CBPS	0.12	0.09	0.15	0.12	0.11	0.08
CBPS miss			0.14	0.11	0.15	0.14
GBM	0.1	0.08	0.12	0.1	0.12	0.1
MTN	0.12	0.09	0.17	0.12	0.12	0.08
MTN miss			0.15	0.11	0.17	0.15
Adj	0	0	0.03	0.03	0.03	0.02
Adju miss			0.06	0.05	0.03	0.03
Montecarlo error						
AIPW	0	0	0.03	0.03	0.03	0.03
AIPW miss			0.07	0.06	0.04	0.03
CBPS	0.08	0.04	0.11	0.07	0.07	0.04
CBPS miss			0.1	0.08	0.11	0.1
GBM	0.05	0.03	0.09	0.06	0.08	0.07
MTN	0.09	0.05	0.2	0.09	0.08	0.04
MTN miss			0.12	0.09	0.15	0.13
Adj	0	0	0.03	0.03	0.03	0.02
Adju miss			0.06	0.05	0.03	0.03
Simple confounders and noise						
AIPW	0	0	0.03	0.03	0.03	0.03
AIPW miss			0.07	0.06	0.04	0.03
CBPS	0.07	0.07	0.09	0.09	0.08	0.08
CBPS miss			0.09	0.09	0.08	0.08
GBM	0.07	0.07	0.09	0.09	0.08	0.08
MTN	0.07	0.07	0.09	0.09	0.08	0.08
MTN miss			0.09	0.09	0.08	0.08
Adj	0	0	0.03	0.03	0.03	0.02
Adju miss			0.06	0.05	0.03	0.03

Table 4: Table of ratio of standard error according to the model scenarios (n = 1000)

	Scenario 1		Scenario 2		Scenario 3	
Contrast	0-1	0-2	0-1	0-2	0-1	0-2
AIPW	0.93	1.09	1.02	1.03	0.98	1
AIPW miss			1	0.98	0.99	0.99
CBPS	0.95	1.7	1.03	1.86	0.82	1.26
CBPS miss			0.67	0.78	0.86	1.09
GBM	1.39	2.34	0.99	1.04	0.97	1.51
MTN	0.81	1.56	0.93	1.84	0.45	1
MTN miss			0.51	0.58	0.73	1.04
Adj	1.03	1.27	1.04	1	0.98	0.99
Adju miss			1.05	0.99	1	1.05

5 Illustrative example

To illustrate the practical implementation of multinomial regression, GBM, CBPS and AIPTW, we applied them to a real data example. The data come from a French prospective national cohort study that uses the research database of the Observatoire Sommeil de la Fédération de Pneumologie (OSFP). The OSFP is a web-based registry containing de-identified data collected on individuals with sleep disorders. We focused on patients diagnosed with obstructive sleep apnea (OSA) and treated with continuous positive airway pressure (CPAP).

Adherence to CPAP is a major issue in the relief of OSA symptoms. A commonly accepted minimal adherence threshold for efficacy is 4 hours per night. However, it may be useful to refine our understanding of the causal effect of CPAP adherence on sleep outcomes, by increasing the number of adherence categories. Therefore, we divided adherence to CPAP into four levels (0-4h, 4-6h, 6-7h, 7-10h).

A major problem for OSA patients is daytime sleepiness. We therefore investigated the causal effect of CPAP adherence on a continuous score, the Epworth Sleepiness Scale, which is a daytime sleepiness rating scale. We included all patients with a follow-up visit between 6 and 18 months after initiation of CPAP to assess its impact on sleepiness. We included all patients with an Epworth score completed for the diagnostic and follow-up visits.

Age, sex, body mass index (kg/m^2), smoking status, score on Pichot de-

pression scale, apnea hypopnea index, presence of morning headache, presence of hypertension, presence of diabetes, presence of hypercholesterolemia all on the initiation of treatment were considered as potential confounders in this study.Figure 8 is a DAG representing the relationships between confounding factors, treatment and outcome.

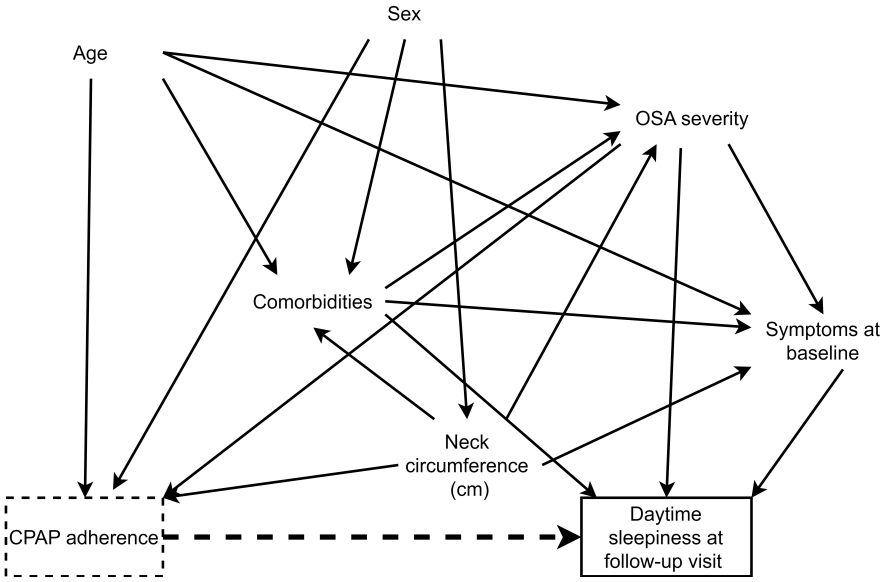

Figure 8: Causal directed acyclic graph
Causal directed acyclic graph for the relation between multilevel CPAP adherence and residual daytime sleepiness under CPAP. Dotted straight arrow indicates causal relation under investigation; solid arrows indicate known relations. CPAP: Continuous Positive Airway Pressure; OSA: Obstructive Sleep Apnea
Dotted frame: exposure; Solid frame: outcome
Comorbidities : depression measured by Pichot's depression scale presence of morning headache, presence of hypertension, presence of diabetes and presence of hypercholesterolemia.

We targeted the ATE quantified using mean differences, parametric models were specified without interaction between potential confounders. In our example, it seems reasonable to consider that the hypotheses are respected. For exchangeability, we have taken into account the major confounding factors in the context of our clinical problem. There does not seem to be any interference or violation of positivity and we can assume that consistency is respected.

Of the 2565 patients included, 591 (23%) belong to the 0-4h adherence group, 977 (38%) to the 4-6h group, 553 (21.6%) to the 6-7h group and 444 (17.3%) to the 7-10h group. The mean absolute reduction in Epworth score over the course of treatment was 3.4(5) for patients in the 0-4h adherence group, 4.7(5.2) for

the 4-6h group, 5.0(5.1) for the 6-7h group and 5.2(5.2) for the 7-10h group. The characteristics of the population according to the adherence groups are summarised in Table 5.

The different methods lead to median weights of between 2.9 and 3.7. There are no extreme weights, the maximum weight being 17.3, obtained with the GBMs. As shown in Figure 9, all the methods achieve a satisfactory balance between the various confounding factors. GBMs are the method leading to the greatest imbalance. The results are shown in figure 10. Overall, adjusted methods led to the estimation of a smaller effect of adherence on the reduction of the Epworth score compared to the unadjusted value, suggesting the presence of positive confounding. The GBMs and regressions lead in this case to extremely similar results, with slightly narrower confidence intervals for the regression. But the calculation time needed for the estimation of the ATE by the GBM is more than a hundred times longer than that of the regressions. The fact that the two methods lead to extremely close results can be explained by the not very strong confounding as presented in figure 9.

Table 5: Patient characteristics according to the adherence group

	All groups 2,565	0-4 h (1) 591 (23.0%)	4-6 h (2) 977 (38.1%)	6-7 h (3) 553 (21.6%)	7-10 h (4) 444 (17.3%)
Variables at diagnosis					
Gender (male)	1,764 (68.8%)	381 (64.5%)	677 (69.3%)	396 (71.6%)	310 (69.8%)
Age (years)	56.6 $(12.8)_{1,4}$	54.6 $(13.1)_{3,4}$	55.9 $(12.1)_{3,4}$	58.1 $(12.5)_{1,2}$	58.7 $(13.5)_{1,2}$
Body mass index (kg/m2)	31.8 (6.8)	31.8 (6.8)	31.4 $(6.2)_4$	31.9 (6.8)	32.8 $(8.1)_2$
Depression scale	4.1 (3.8)	4.3 (3.9)	4.0 (3.8)	4.1 (3.8)	3.9 (3.6)
ESS score	10.7 (5.1)	11.0 $(5.2)_4$	10.9 $(5.1)_4$	10.4 (5.0)	10.1 $(5.2)_{1,2}$
Apnea hypopnea index (event/h)	39.2 $(20.6)_{1,3}$	36.5 $(20.3)_{3,4}$	38.0 $(19.7)_{3,4}$	42.0 $(21.1)_{1,2}$	41.7 $(21.7)_{1,2}$
Morning headaches	1,040 (40.5%)	253 (42.8%)	405 (41.5%)	220 (39.8%)	162 (36.5%)
Hypertension	1,193 (46.5%)	258 (43.7%)	451 (46.2%)	256 (46.3%)	228 (51.4%)
Diabetes	689 (26.9%)	180 (30.5%)	247 (25.3%)	138 (25%)	124 (27.9%)
Hypercholesterolemia	755 (29.4%)	164 (27.7%)	279 (28.6%)	179 (32.4%)	133 (30%)
Smoker	403 (15.7%)	115 (19.5%)$_4$	165 (16.9%)	73 (13.2%)	50 $(11.3\%)_1$
Former smoker	763 $(29.7\%)_1$	140 $(23.7\%)_{3,4}$	286 (29.3%)	186 $(33.6\%)_1$	151 $(34\%)_1$
Variables at follow-up					
ESS score	6.3 $(4.2)_{1,3,4}$	7.6 $(4.6)_{2,3,4}$	6.4 $(4.1)_{1,3,4}$	5.4 $(3.8)_{1,2}$	5.5 $(4.1)_{1,2}$
Duration since diagnosis (year)	0.8 (0.3)	0.8 (0.3)	0.8 (0.3)	0.8 (0.3)	0.8 (0.3)

ESS : Epworth Sleepiness Scale

Quantitative variables are presented as mean (standard deviation). Qualitative variables are expressed in number of individuals (% of individuals)

t-test was performed for the quantitative variables and a Pearson's Chi squared test for the categorical variables after application of a Bonferroni correction for multiple testing.

1,2,3,4 numbers in subscript refers to columns statistically different at the 5% threshold. e.g. 1 means that there is a statistically significant difference between the adherence group of that column and the 0-4 adherence group (1) for the variable in question.

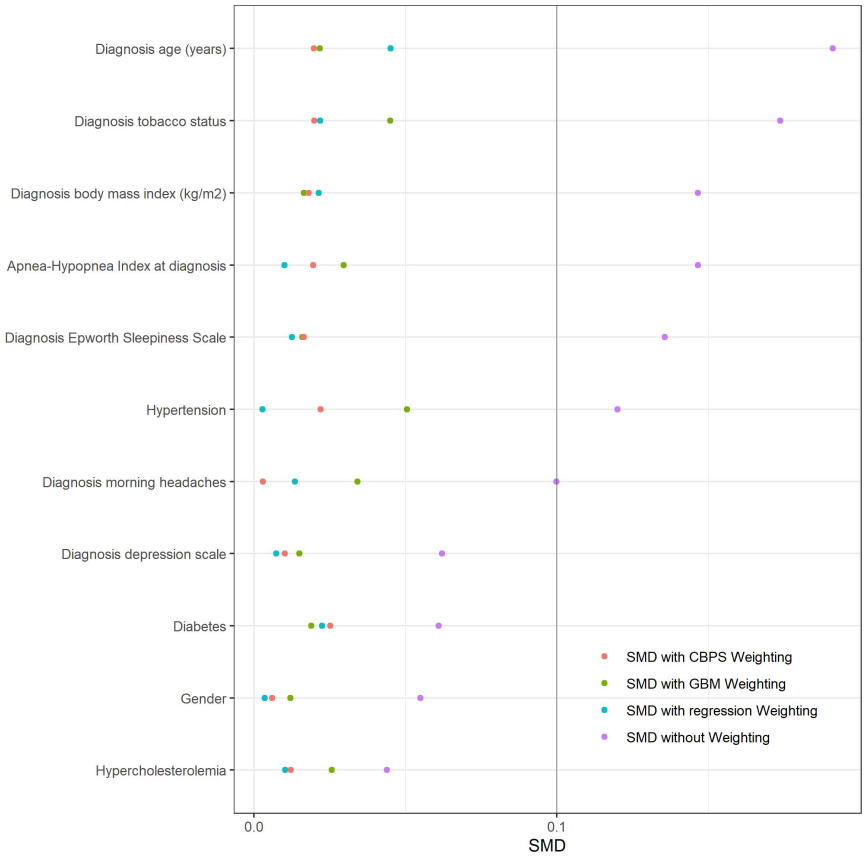

Figure 9: Standardized mean difference before and after weighting
SMD: Standardized Mean Difference; CPAP : Continuous Positive Airway Pressure; ADR: adverse drug reaction; SaO2: arterial oxygen saturation

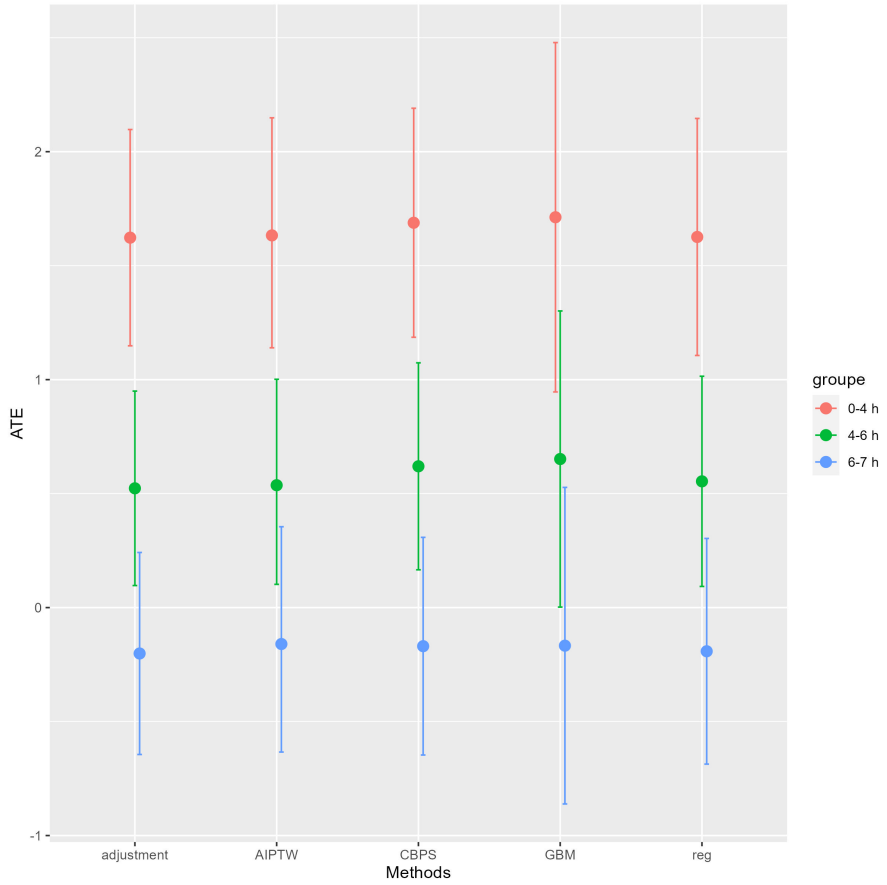

Figure 10: Difference mean in Epworth score between each adherence group and the reference group using different methods
Each point represents the difference mean in Epworth score between each adherence group and the reference group (7-10h). The vertical bars represent the 95% confidence intervals of these estimates.

6 Discussion

In order to limit the biases of observational studies, a wide range of methods has been developed for estimating propensity scores which are widely applied in medical research and epidemiology Granger et al. (2020) an received an increasing attention. GBMs are not an exception, they are the second most common method used for propensity score estimation in medical research in the case of multi-level treatments Cook et al. (2019); Chu et al. (2019); Chou et al. (2020).

The present study aimed to evaluate the performances of GBM with multinomial regression and CBPS for the estimation of the propensity score for multilevel treatments when estimating the ATE with IPTW, and compare them to the performance of multivariable adjustment and AIPTW for the estimation of causal effects.

Using simulations, we showed that in all the studied scenarios, GBMs lead to smaller propensity score weights, and smaller variance estimates of the ATE than multinomial regression, which is consistent with the other study on the topic McCaffrey et al. (2013). As expected, misspecified models lead to biased ATE estimates with larger variance and larger weight in scenario two. The bias is only present in the second scenario, the GBMs have a bias of 0.3, while the misspecified models have a bias of 0.6. Misspecified models lead to a bootstrap standard error twice as large as their correctly specified equivalent. The bootstrap standard error of GBMs is comparable to that of misspecified models.

The estimates of the ATE after estimating propensity scores using GBMs were slightly biased, unlike those obtained with other methods when the underlying models are correctly specified, although the bias was typically very small. This is expected as the models corresponding to the true data generation mechanism will always exhibit better performances. For weighted estimators, the unbiasedness of the ATE estimates relies on the ability of the different methods to balance covariates after weighting. Weights obtained using GBMs led to a worst balance than those obtained with a correctly specified multinomial regression, especially in scenarios with interaction terms and noise and instrumental variables. These results may be explained by limitations in the ability of GBMs to select variables, with in particular the inclusion of instrumental variables, which is known to introduce bias. Therefore, as with multinomial regression, a preliminary step of variable selection, based on expert knowledge is required. Residual imbalances on covariates and their interactions could also be explained by the fact that during the iterative process of constructing the GBM model, the weak classifiers focus on the most unbalanced individuals Hastie et al. (2009a), giving them increasing importance. This can be mitigated by defining the bag fraction, which is the fraction of data used in each weak learner.

Regarding precision, the bootstrap variance estimates after estimating propensity scores using GBMs were much smaller than those of the multinomial regression (with therefore narrower confidence intervals) but larger than those obtained with a doubly robust method such as AIPTWs. This is coherent with the literature, as AIPTW often improves precision when the treatment and outcome model are correctly specified Robins et al. (1994). In addition, these

differences between GBMs and other correctly specified methods are exacerbated when the sample size is small. This can easily be explained by the fact that non-parametric models require large sample sizes.

Although GBMs showed good performances in situations were the parametric form of the propensity score model is hard to define, the implementation of GBMs can be challenging since they are relatively expensive in terms of computational resources, making them time-consuming to use for large data-sets, especially in combination with non parametric bootstrap to calculate confidence intervals. Computation time can be improved by reducing the interaction depth in the algorithm, but this may have an impact on the resulting covariate balance.

To investigate the implementation of those methods,we applied them to analyse a study assessing the effect of adherence to continuous positive airway pressure on daytime sleepiness in obstructive sleep apnea. In this clinical application, all the methods lead to an acceptable SMD (< 0.1) for the different confounding factors and none of the methods led to extreme weights. All the methods resulted in similar estimates of the ATE, but in this particular case the GBMs lead to the widest confidence intervals, which was expected given the moderate sample size.

The main strength of this study is the comparison of a wide range of methods, under various scenarios on several key parameters (balance, bias and precision). Although GBMs and multinomial regressions are the most commonly used methods with multilevel treatments, our results suggest that AIPW should be considered in these settings. However, this study also has some limitations. First, we focused on a few simple scenarios, but we can assume that if a method fails in this context, the problem would also be encountered in more complex situations. Second, like many machine learning methods, GBM remains sensitive to the tuning of its hyper-parameters Parast et al. (2016). In this study we did not focus on tuning the hyper-parameters, but it is possible that a method that adapts the hyper-parameters to each data-set using grid search would give better results. However, GBMs are relatively computationally expensive, and therefore such tuning may be more challenging to implement. In this study we have chosen to focus on hyper-parameters close to the default twang hyper-parameters which are likely to be used by researchers. For this reason, we focussed on the use of the mean SMD as the criterion for stopping GBMs, although the Kolmogorov-Smirnov statistic may be preferred in certain circumstances Parast et al. (2016).

In conclusion, PS for multi-level treatment can be estimated by different methods and our results showed that GBMs lead to a higher SMD than the other models compared, and the ATE estimate obtained using the weights estimated with GBMs may be biased compared to other correctly specified models. Although GBMs are an useful tool in clinical research thanks to their easy-to-use implementation and variable selection capability, researchers should carefully select the appropriate method in their specific setting and consider the use of doubly robust estimators.

Conflict of Interest

The authors have declared no conflict of interest.

References

H.-J. Ahn, S.-R. Lee, E.-K. Choi, K.-D. Han, J.-H. Jung, J.-H. Lim, J.-P. Yun, S. Kwon, S. Oh, and G. Y. H. Lip. Association between exercise habits and stroke, heart failure, and mortality in Korean patients with incident atrial fibrillation: A nationwide population-based cohort study. *PLoS Med*, 18(6): e1003659, June 2021. ISSN 1549-1676. doi: 10.1371/journal.pmed.1003659.

A. K. Ali, A. G. Hartzema, A. G. Winterstein, R. Segal, X. Lu, and L. Hendeles. Application of multicategory exposure marginal structural models to investigate the association between long-acting beta-agonists and prescribing of oral corticosteroids for asthma exacerbations in the Clinical Practice Research Datalink. *Value Health*, 18(2):260–270, Mar. 2015. ISSN 1524-4733. doi: 10.1016/j.jval.2014.11.007.

P. C. Austin. Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis. *Statistics in Medicine*, 35(30):5642–5655, aug 2016. doi: 10.1002/sim.7084.

F. Bettega, C. Leyrat, R. Tamisier, M. Mendelson, Y. Grillet, M. Sapène, M. R. Bonsignore, J. L. Pépin, M. W. Kattan, and S. Bailly. Application of inverse-probability-of-treatment weighting to estimate the effect of daytime sleepiness in patients with obstructive sleep apnea. *Annals of the American Thoracic Society*, 19(9):1570–1580, sep 2022. doi: 10.1513/annalsats.202109-1036oc.

S. Bozorgmehri, H. Aboud, G. Chamarthi, I.-C. Liu, O.-B. Tezcan, A. M. Shukla, A. Kazory, R. Rupam, M. S. Segal, A. Bihorac, and R. Mohandas. Association of early initiation of dialysis with all-cause and cardiovascular mortality: A propensity score weighted analysis of the united states renal data system. *Hemodialysis International*, 25(2):188–197, feb 2021. doi: 10.1111/hdi.12912.

Y.-H. Chou, J.-Y. Huang, E. Kornelius, J.-Y. Chiou, and C.-N. Huang. Major adverse cardiovascular events after treatment in early-stage breast cancer patients receiving hormone therapy. *Scientific Reports*, 10(1), jan 2020. doi: 10.1038/s41598-020-57726-z.

H. H. Chu, J. H. Kim, H.-K. Yoon, H.-K. Ko, D. I. Gwon, P. N. Kim, K.-B. Sung, G.-Y. Ko, S. Y. Kim, and S. H. Park. Chemoembolization combined with radiofrequency ablation for medium-sized hepatocellular carcinoma: A propensity-score analysis. *Journal of Vascular and Interventional Radiology*, 30(10):1533–1543, oct 2019. doi: 10.1016/j.jvir.2019.06.006.

S. R. Cole and C. E. Frangakis. The consistency statement in causal inference. *Epidemiology*, 20(1):3–5, jan 2009. doi: 10.1097/ede.0b013e31818ef366.

S. R. Cole and M. A. Hernan. Constructing Inverse Probability Weights for Marginal Structural Models. *American Journal of Epidemiology*, 168(6):656–664, July 2008. ISSN 0002-9262, 1476-6256. doi: 10.1093/aje/kwn164.

R. R. Cook, J. A. Fulcher, N. H. Tobin, F. Li, D. Lee, M. Javanbakht, R. Brookmeyer, S. Shoptaw, R. Bolan, G. M. Aldrovandi, and P. M. Gorbach. Effects of HIV viremia on the gastrointestinal microbiome of young MSM. *AIDS*, 33 (5):793–804, apr 2019. doi: 10.1097/qad.0000000000002132.

R. Crump, V. J. Hotz, G. Imbens, and O. Mitnik. Moving the goalposts: Addressing limited overlap in the estimation of average treatment effects by changing the estimand. Technical report, oct 2006.

J. H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, Oct. 2001. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1013203451.

E. Granger, T. Watkins, J. C. Sergeant, and M. Lunt. A review of the use of propensity score diagnostics in papers published in high-ranking medical journals. *BMC Medical Research Methodology*, 20(1), may 2020. doi: 10.1186/s12874-020-00994-0.

V. S. Harder, E. A. Stuart, and J. C. Anthony. Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods*, 15(3):234–249, 2010. doi: 10.1037/a0019623.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer-Verlag GmbH, Aug. 2009a. ISBN 9780387848587.

T. Hastie, R. Tibshirani, and J. Friedman. *Boosting and Additive Trees*, pages 337–387. Springer Series in Statistics. Springer New York, New York, NY, 2009b. ISBN 978-0-387-84857-0 978-0-387-84858-7. doi: 10.1007/978-0-387-84858-7_10.

F. Hayashi. *Econometrics*. Princeton University Press, 2000. ISBN 9780691010182.

M. A. Hernan. A definition of causal effect for epidemiological research. *Journal of Epidemiology & Community Health*, 58(4):265–271, Apr. 2004. ISSN 0143-005X. doi: 10.1136/jech.2002.006361. URL `https://jech.bmj.com/lookup/doi/10.1136/jech.2002.006361`.

K. Imai and M. Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):243–263, jul 2013. doi: 10.1111/rssb.12027.

K. Imai and D. A. van Dyk. Causal inference with general treatment regimes. *Journal of the American Statistical Association*, 99(467):854–866, sep 2004. doi: 10.1198/016214504000001187.

G. Imbens. The role of the propensity score in estimating dose-response functions. Technical report, apr 1999.

B. K. Lee, J. Lessler, and E. A. Stuart. Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29(3):337–346, dec 2009. doi: 10.1002/sim.3782.

G. Lefebvre, J. A. C. Delaney, and R. W. Platt. Impact of mis-specification of the treatment model on estimates from a marginal structural model. *Statist. Med.*, 27(18):3629–3642, Aug. 2008. ISSN 02776715, 10970258. doi: 10.1002/sim.3200.

D. F. McCaffrey, G. Ridgeway, and A. R. Morral. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychol Methods*, 9(4):403–425, Dec. 2004. ISSN 1082-989X. doi: 10.1037/1082-989X.9.4.403.

D. F. McCaffrey, B. A. Griffin, D. Almirall, M. E. Slaughter, R. Ramchand, and L. F. Burgette. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statist. Med.*, 32(19):3388–3414, Aug. 2013. ISSN 02776715. doi: 10.1002/sim.5753.

A. B. Owen. *Empirical Likelihood*. Taylor & Francis Group, 2001. ISBN 9781420036152.

L. Parast, D. F. McCaffrey, L. F. Burgette, F. H. de la Guardia, D. Golinelli, J. N. V. Miles, and B. A. Griffin. Optimizing variance-bias trade-off in the TWANG package for estimation of propensity scores. *Health Services and Outcomes Research Methodology*, 17(3-4):175–197, dec 2016. doi: 10.1007/s10742-016-0168-2.

C. Pawlowski, P. Lenehan, A. Puranik, V. Agarwal, A. Venkatakrishnan, M. J. Niesen, J. C. O'Horo, A. Virk, M. D. Swift, A. D. Badley, J. Halamka, and V. Soundararajan. FDA-authorized mRNA COVID-19 vaccines are effective per real-world evidence synthesized across a multi-state health system. *Med*, 2(8):979–992.e8, aug 2021. doi: 10.1016/j.medj.2021.06.007.

G. Ridgeway. The State of Boosting. 1999.

G. Ridgeway, D. F. McCaffrey, A. R. Morral, M. Cefalu, L. F. Burgette, J. D. Pane, and B. A. Griffin. *Toolkit for Weighting and Analysis of Nonequivalent Groups: A Tutorial for the R TWANG Package*. RAND Corporation, 2022. doi: 10.7249/tl-a570-5.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, sep 1994. doi: 10.1080/01621459.1994.10476818.

J. M. Robins, M. Á. Hernán, and B. Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, sep 2000. doi: 10.1097/00001648-200009000-00011.

C. L. Rodríguez-Bernal, Y. Santa-Ana-Téllez, A. García-Sempere, I. Hurtado, S. Peiró, and G. Sanfélix-Gimeno. Clinical outcomes of nonvitamin k oral anticoagulants and acenocoumarol for stroke prevention in contemporary practice: A population-based propensity-weighted cohort study. *British Journal of Clinical Pharmacology*, 87(2):632–643, jul 2020. doi: 10.1111/bcp.14430.

P. R. ROSENBAUM and D. B. RUBIN. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983. doi: 10.1093/biomet/70.1.41.

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.

D. B. Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469):322–331, mar 2005. doi: 10.1198/016214504000001880.

F. Shi, C. Wang, Y. Kong, L. Yang, J. Li, G. Zhu, J. Guo, Q. Zheng, B. Zhang, and S. Wang. Assessing the Survival Benefit of Surgery and Various Types of Radiation Therapy for Treatment of Hepatocellular Carcinoma: Evidence from the Surveillance, Epidemiology, and End Results Registries. *Journal of Hepatocellular Carcinoma*, 7:201–218, 2020. ISSN 2253-5969. doi: 10.2147/JHC.S272813.

A. A. Tsiatis, M. Davidian, M. Zhang, and X. Lu. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Statistics in Medicine*, 27(23):4658–4677, oct 2008. doi: 10.1002/sim.3113.

Z.-X. Wang, L.-P. Yang, H.-X. Wu, D.-D. Yang, P.-R. Ding, D. Xie, G. Chen, Y.-H. Li, F. Wang, and R.-H. Xu. Appraisal of Prognostic Interaction between Sidedness and Mucinous Histology in Colon Cancer: A Population-Based Study Using Inverse Probability Propensity Score Weighting. *Journal of Cancer*, 10(2):388–396, 2019. ISSN 1837-9664. doi: 10.7150/jca.28014.

K. Yoshida, D. H. Solomon, S. Haneuse, S. C. Kim, E. Patorno, S. K. Tedeschi, H. Lyu, J. M. Franklin, T. Stürmer, S. Hernández-Díaz, and R. J. Glynn. Multinomial Extension of Propensity Score Trimming Methods: A Simulation Study. *Am J Epidemiol*, 188(3):609–616, Mar. 2019. ISSN 0002-9262. doi: 10.1093/aje/kwy263. URL `https://academic.oup.com/aje/article/188/3/609/5231606`.

Comparison of multinomial logistic regression and generalized boosted models for propensity score estimation for multilevel treatments

Supplementary Material

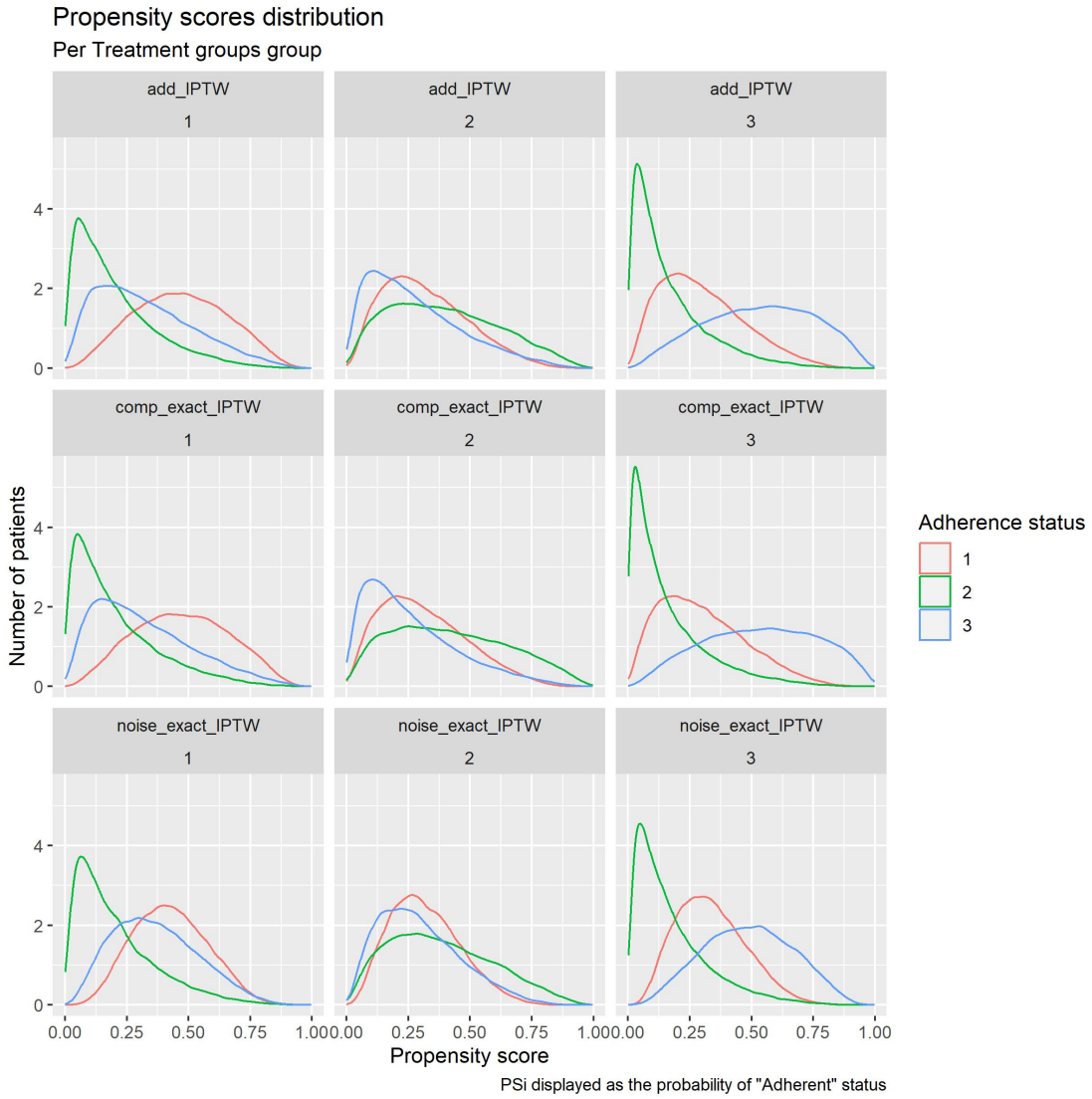

Figure 1: Mean ATE

1 Treatment assignment coefficient

In the first scenario the additive scenario, we use 15 of the 26 covariates in probability of treatment level assignment . The covariates include and coefficients associate with each levels are summarize in table 1 below.

S1 Table

	0	1	2
intercept	0.000	0.000	0.000
continuous 1	1.290	0.932	0.562
continuous 2	-0.838	-0.798	-0.998
continuous 4	0.778	1.588	0.419
continuous 5	0.641	0.478	0.893
continuous 7	0.294	0.500	0.417
continuous 9	1.172	0.538	0.702
continuous 10	-0.639	-0.594	-0.385
binary 15	1.435	1.295	0.871
binary 16	0.767	0.578	0.266
binary 17	-0.161	-0.624	0.318
binary 18	1.432	0.772	1.524
binary 19	-0.444	0.330	0.710
binary 21	0.514	-0.215	0.869
binary 23	0.652	0.483	0.358
binary 24	0.462	0.106	0.957

In the second scenario we add interaction terms, we use 15 of the 26 covariates in probability of treatment level assignment. The covariates include associate with each levels are summarize in table 2 below.

S2 Table

	0	1	2
intercept	0.000	0.000	0.000
continuous 1	1.290	0.932	0.562
continuous 2	-0.838	-0.798	-0.998
continuous 4	0.778	1.588	0.419
continuous 5	0.641	0.478	0.893
continuous 7	0.294	0.500	0.417
continuous 9	1.172	0.538	0.702
continuous 10	-0.639	-0.594	-0.385
binary 15	1.435	1.295	0.871
binary 16	0.767	0.578	0.266
binary 17	-0.161	-0.624	0.318
binary 18	1.432	0.772	1.524
binary 19	-0.444	0.330	0.710
binary 21	0.514	-0.215	0.869
binary 23	0.652	0.483	0.358
binary 24	0.462	0.106	0.957
continuous 1 : continuous 2	-0.413	-0.649	-0.212
continuous 4 : continuous 5	0.471	0.397	0.249
continuous 7 : continuous 10	0.315	0.654	0.173

In the third scenario we use 18 of the 26 covariates in probability of treatment level assignment, $continuous_{11}$, $continuous_{12}$ and $binary_{25}$ are instrumentals variables only associate with treatment. The covariates include and coefficients associate with each levels are summarize in table 3 below.

S3 Table

	0	1	2
intercept	0.000	0.000	0.000
continuous 1	1.290	0.932	0.562
continuous 2	-0.838	-0.798	-0.998
continuous 4	0.778	1.588	0.419
continuous 5	0.641	0.478	0.893
continuous 7	0.294	0.500	0.417
continuous 9	1.172	0.538	0.702
continuous 10	-0.639	-0.594	-0.385
binary 15	1.435	1.295	0.871
binary 16	0.767	0.578	0.266
binary 17	-0.161	-0.624	0.318
binary 18	1.432	0.772	1.524
binary 19	-0.444	0.330	0.710
binary 21	0.514	-0.215	0.869
binary 23	0.652	0.483	0.358
binary 24	0.462	0.106	0.957
continuous 11*	0.173	0.476	1.371
continuous 12*	-1.144	-0.189	0.217
binary 25*	0.170	0.602	0.284

[a] Variables with * are instrumental variables

2 Distribution and balancing ability of the weights

S4 Table: Table of weights according to the model scenarios (n = 1000)

Treatment group	IPTW	GBM	CBPS	IPTW misspecified	CBPS misspecified
Simple confounders					
0	2.1 (1.6; 3)	1.8 (1.4; 2.4)	2.1 (1.7; 3)		
1	2.6 (1.7; 4.4)	2.1 (1.5; 3.1)	2.7 (1.9; 4.3)		
2	1.8 (1.4; 2.7)	1.6 (1.3; 2.1)	1.9 (1.5; 2.7)		
Simple confounders and interaction terms					
0	2.1 (1.6; 3)	1.8 (1.4; 2.4)	2.1 (1.6; 3)	2.1 (1.6; 3)	2.2 (1.7; 3)
1	2.4 (1.6; 4.2)	2.1 (1.5; 3.1)	2.5 (1.7; 4.1)	2.6 (1.8; 4.4)	2.7 (1.9; 4.3)
2	1.8 (1.3; 2.7)	1.6 (1.3; 2.2)	1.8 (1.4; 2.7)	1.8 (1.4; 2.8)	1.9 (1.5; 2.7)
Simple confounders and noise					
0	2.3 (1.8; 3.2)	1.4 (1.2; 2)	2.4 (1.9; 3.1)	1.6 (1.2; 2.6)	1.7 (1.3; 2.6)
1	2.6 (1.8; 4.4)	1.9 (1.5; 2.8)	2.7 (1.9; 4.3)	2.3 (1.6; 4.1)	2.4 (1.7; 4)
2	2 (1.6; 2.8)	1.4 (1.2; 1.8)	2 (1.6; 2.8)	1.5 (1.2; 2.3)	1.5 (1.2; 2.3)

[a] Results are presented in the form: median(1st quartile; 3rd quartile)

S5 Table: Table of weights according to the model scenarios (n = 5000)

Treatment group	IPTW	GBM	CBPS	IPTW misspecified	CBPS misspecified
Simple confounders					
0	2.1 (1.6; 3)	2 (1.5; 2.7)	2.1 (1.7; 3)		
1	2.6 (1.8; 4.5)	2.3 (1.6; 3.8)	2.7 (1.8; 4.5)		
2	1.9 (1.4; 2.7)	1.7 (1.4; 2.4)	1.9 (1.4; 2.7)		
Simple confounders and interaction terms					
0	2.1 (1.6; 3)	2 (1.5; 2.7)	2.1 (1.6; 3)	2.2 (1.7; 3.1)	2.2 (1.7; 3)
1	2.5 (1.7; 4.3)	2.3 (1.6; 3.7)	2.5 (1.7; 4.2)	2.7 (1.8; 4.5)	2.7 (1.9; 4.4)
2	1.8 (1.4; 2.8)	1.7 (1.4; 2.5)	1.8 (1.4; 2.7)	1.9 (1.4; 2.8)	1.9 (1.5; 2.8)
Simple confounders and noise					
0	2.4 (1.9; 3.2)	1.5 (1.2; 2.3)	2.4 (1.9; 3.1)	1.7 (1.3; 2.6)	1.7 (1.3; 2.6)
1	2.7 (1.8; 4.5)	2.2 (1.6; 3.5)	2.8 (1.9; 4.5)	2.4 (1.6; 4.2)	2.5 (1.7; 4.2)
2	2 (1.6; 2.8)	1.4 (1.2; 2)	2.1 (1.6; 2.8)	1.5 (1.2; 2.3)	1.5 (1.2; 2.3)

[a] Results are presented in the form: median(1st quartile; 3rd quartile)

2.1 Average Treatment Effect

S6 Table: Table of bias according to the model scenarios (n = 1000)

Treatment group	Additive		Interaction		Noise	
	1 - 0	2 - 0	1 - 0	2 - 0	1 - 0	2 - 0
AIPW	0 [0; 0]94 %	0 [0; 0]95 %	0 [-0.2; 0.2]94 %	0 [-0.2; 0.2]94 %	0 [-0.2; 0.2]96 %	0 [-0.2; 0.2]94 %
AIPW miss			0.6 [0.2; 1]13 %	-0.4 [-0.8; -0.1]36 %	0 [-0.2; 0.2]94 %	0 [-0.2; 0.2]95 %
CBPS	0.1 [-0.6; 0.7]100 %	0 [-0.5; 0.5]100 %	0.2 [-0.6; 1]99 %	-0.1 [-0.7; 0.5]100 %	0.1 [-0.6; 0.7]100 %	0 [-0.5; 0.5]100 %
CBPS miss			0.7 [-0.1; 1.5]64 %	-0.4 [-1; 0.2]80 %	0.1 [-0.7; 0.8]100 %	0 [-0.7; 0.7]100 %
GBM	0.1 [-0.5; 0.7]100 %	0 [-0.5; 0.4]100 %	0.5 [-0.2; 1.2]73 %	-0.4 [-0.9; 0.2]87 %	0.1 [-0.5; 0.7]99 %	0 [-0.6; 0.5]99 %
MTN	0 [-0.8; 0.7]99 %	0 [-0.6; 0.5]100 %	0 [-1; 1]99 %	0 [-0.8; 0.7]100 %	0 [-0.7; 0.8]100 %	0 [-0.5; 0.5]100 %
MTN miss			0.6 [-0.3; 1.6]74 %	-0.5 [-1.2; 0.2]78 %	0 [-0.9; 1]98 %	0 [-0.8; 0.9]99 %
Adj	0 [0; 0]96 %	0 [0; 0]94 %	0 [-0.2; 0.2]95 %	0 [-0.2; 0.2]95 %	0 [-0.2; 0.2]96 %	0 [-0.1; 0.1]95 %
Adju miss			0.6 [0.2; 1]10 %	-0.5 [-0.8; -0.1]19 %	0 [-0.2; 0.2]96 %	0 [-0.2; 0.2]95 %

S7 Table: Table of bias according to the model scenarios (n = 5000)

Treatment group	Additive		Interaction		Noise	
	1 - 0	2 - 0	1 - 0	2 - 0	1 - 0	2 - 0
AIPW	0 [0; 0]96 %	0 [0; 0]94 %	0 [-0.1; 0.1]95 %	0 [-0.1; 0.1]94 %	0 [-0.1; 0.1]95 %	0 [-0.1; 0.1]94 %
AIPW miss			0.6 [0.4; 0.8]0 %	-0.4 [-0.6; -0.3]0 %	0 [-0.1; 0.1]94 %	0 [-0.1; 0.1]94 %
CBPS	0 [-0.3; 0.3]100 %	0 [-0.2; 0.2]100 %	0.1 [-0.3; 0.5]98 %	0 [-0.4; 0.3]100 %	0 [-0.3; 0.3]99 %	0 [-0.2; 0.2]100 %
CBPS miss			0.7 [0.3; 1.1]2 %	-0.4 [-0.7; -0.2]5 %	0 [-0.4; 0.4]99 %	0 [-0.3; 0.4]99 %
GBM	0.1 [-0.2; 0.3]100 %	0 [-0.2; 0.2]100 %	0.4 [0.1; 0.7]29 %	-0.3 [-0.5; 0]45 %	0.1 [-0.2; 0.4]99 %	0 [-0.3; 0.3]100 %
MTN	0 [-0.3; 0.3]99 %	0 [-0.2; 0.2]100 %	0 [-0.4; 0.5]99 %	0 [-0.3; 0.3]100 %	0 [-0.3; 0.3]99 %	0 [-0.2; 0.2]100 %
MTN miss			0.7 [0.2; 1.1]8 %	-0.5 [-0.8; -0.2]7 %	0 [-0.5; 0.4]98 %	0 [-0.4; 0.4]98 %
Adj	0 [0; 0]96 %	0 [0; 0]95 %	0 [-0.1; 0.1]95 %	0 [-0.1; 0.1]95 %	0 [-0.1; 0.1]96 %	0 [-0.1; 0.1]95 %
Adju miss			0.6 [0.4; 0.8]0 %	-0.5 [-0.6; -0.3]0 %	0 [-0.1; 0.1]95 %	0 [-0.1; 0.1]95 %

2.2 Standard error

S8 Table: Table of standard error according to the model scenarios (n = 1000)

	Scenario 1		Scenario 2		Scenario 3	
Treatment group	1	2	1	2	1	2
Bootstrap standard error						
AIPW	0.01	0.01	0.1	0.09	0.1	0.09
AIPW miss			0.2	0.19	0.11	0.11
CBPS	0.32	0.26	0.4	0.32	0.32	0.25
CBPS miss			0.4	0.31	0.38	0.33
GBM	0.29	0.24	0.35	0.28	0.31	0.26
MTN	0.38	0.28	0.5	0.38	0.37	0.27
MTN miss			0.47	0.35	0.48	0.42
Adj	0.01	0.01	0.09	0.08	0.09	0.08
Adju miss			0.18	0.16	0.1	0.09
Montecarlo error						
AIPW	0.01	0.01	0.11	0.1	0.1	0.09
AIPW miss			0.21	0.19	0.12	0.11
CBPS	0.17	0.11	0.24	0.17	0.18	0.12
CBPS miss			0.28	0.22	0.25	0.21
GBM	0.16	0.11	0.23	0.18	0.2	0.17
MTN	0.31	0.16	0.48	0.3	0.28	0.14
MTN miss			0.43	0.29	0.49	0.4
Adj	0.01	0.01	0.09	0.08	0.09	0.08
Adju miss			0.18	0.16	0.09	0.09
Simple confounders and noise						
AIPW	0.01	0.01	0.1	0.09	0.1	0.09
AIPW miss			0.2	0.19	0.11	0.11
CBPS	0.23	0.23	0.28	0.28	0.24	0.24
CBPS miss			0.28	0.28	0.24	0.24
GBM	0.23	0.22	0.28	0.27	0.25	0.24
MTN	0.23	0.23	0.28	0.28	0.24	0.24
MTN miss			0.28	0.28	0.24	0.24
Adj	0.01	0.01	0.09	0.08	0.09	0.08
Adju miss			0.18	0.16	0.1	0.09

S9 Table: Table of standard error according to the model scenarios (n = 5000)

	Scenario 1		Scenario 2		Scenario 3	
Treatment group	1	2	1	2	1	2
Bootstrap standard error						
AIPW	0	0	0.05	0.04	0.04	0.04
AIPW miss			0.09	0.08	0.05	0.05
CBPS	0.16	0.12	0.2	0.16	0.16	0.12
CBPS miss			0.2	0.15	0.2	0.18
GBM	0.14	0.11	0.17	0.14	0.16	0.14
MTN	0.17	0.12	0.23	0.17	0.17	0.12
MTN miss			0.21	0.16	0.23	0.21
Adj	0	0	0.04	0.04	0.04	0.03
Adju miss			0.08	0.07	0.04	0.04
Montecarlo error						
AIPW	0	0	0.05	0.04	0.04	0.04
AIPW miss			0.09	0.08	0.05	0.05
CBPS	0.1	0.06	0.14	0.09	0.1	0.06
CBPS miss			0.14	0.1	0.15	0.13
GBM	0.07	0.05	0.11	0.07	0.1	0.09
MTN	0.13	0.07	0.2	0.12	0.12	0.06
MTN miss			0.17	0.12	0.23	0.2
Adj	0	0	0.04	0.04	0.04	0.03
Adju miss			0.08	0.07	0.04	0.04
Simple confounders and noise						
AIPW	0	0	0.05	0.04	0.04	0.04
AIPW miss			0.09	0.08	0.05	0.05
CBPS	0.1	0.1	0.13	0.12	0.11	0.11
CBPS miss			0.13	0.13	0.11	0.11
GBM	0.1	0.1	0.12	0.12	0.11	0.11
MTN	0.1	0.1	0.13	0.13	0.11	0.11
MTN miss			0.13	0.13	0.11	0.11
Adj	0	0	0.04	0.04	0.04	0.03
Adju miss			0.08	0.07	0.04	0.04

VI. Conclusion et perspectives

En recherche médicale, les analyses sur les données observationnelles se développent. De plus en plus d'études et de résultats sont basés sur les approches d'inférence causale. Nous avons vu que les approches binaires sont les plus utilisées mais ne sont pas adaptées dans le cas de traitement multi-niveaux. C'est le cas lorsque l'on étudie le traitement par PPC dans le cadre du SAOS où le fait de dichotomiser une variable continue entraîne une perte d'informations. Pour autant, j'ai montré, dans le cadre de la revue systématique de la littérature, que les approches multi-niveaux sont peu utilisées et de nombreuses lacunes persistent dans le report de la méthode et des résultats. Aussi, cette thèse avait pour objectif d'approfondir la connaissance sur les approches multi-niveaux avec un exemple d'application concret pour permettre une meilleure utilisation de ces méthodes et leur diffusion dans la communauté scientifique médicale.

L'observance à la PPC est associée à de nombreux facteurs et présente un cas d'usage intéressant pour explorer les méthodes d'inférence causale. Dans le cadre d'une revue de la littérature nous avons pu lister les différentes problématiques associées à l'observance des patients sous PPC.

Dans le second chapitre, j'ai présenté les fondements théoriques sur lesquels se basent les IPTWs, notamment le "potential outcome framework", les hypothèses associées, différents estimateurs pour les scores de propension, différents estimands et les problématiques spécifiques aux traitements multi-niveaux. Je me suis aussi intéressé aux méthodes permettant d'évaluer l'équilibre des facteurs de confusion post pondération et à l'estimation de la variance.

Dans le troisième chapitre j'ai effectué une revue systématique pour évaluer l'utilisation des IPTWs multi-niveaux dans la littérature médicale. J'ai ainsi pu déterminer la fréquence d'utilisation de ces méthodes qui reçoivent une attention grandissante ces dernières années. J'ai étudié la manière dont ces méthodes étaient implémentées, mettant en évidence que les GBMs étaient la seconde méthode la plus utilisée pour l'estimation des SP dans la recherche médicale. J'ai également détaillé les différents éléments rapportés dans les publications et la qualité de leur reporting. J'ai décrit de multiples limites dans la manière dont les résultats issus des IPTWs multi-niveaux étaient rapportés par rapport à ce qui est recommandé [8, 87]. J'ai produit une représentation graphique simple sur les éléments clefs nécessaires à la présentation des résultats de ces méthodes. Ce travail est actuellement en révision dans la revue Journal of Clinical Epidemiology.

Dans le quatrième chapitre, j'ai mis en oeuvre un IPTW avec traitement multi-niveaux afin de comparer l'effet causal de plusieurs niveaux d'observance à la PPC sur la somnolence diurne mesurée par le score d'Epworth. L'observance à la PPC est souvent simplifiée en variables binaires (inférieure ou supérieure à 4H), mais on peut supposer qu'il existe des effets de seuil dans l'effet causal de l'observance à la PPC. Cette étude avait aussi pour but de présenter une application des méthodes IPTW à traitements multi-niveaux de manière pédagogique afin de faire connaître et de rendre ces méthodes plus accessibles aux cliniciens. J'ai pu montrer que les méthodes statistiques conventionnelles tendent à surévaluer l'effet de l'observance à la PPC sur la somnolence diurne. Cette application a aussi soulevé pour moi des questions sur l'efficacité et les limites des méthodes autres que la régression multinomiale pour l'estimation des scores de propension. Ce travail a

conduit à une publication dans la revue Annals of the American Thoracic Society.

Dans le cinquième chapitre j'ai évalué, par une étude de simulation, les performances des GBMs, seconde méthode la plus utilisée dans la littérature médicale. L'objectif était l'estimation des scores de propension quand le traitement était multi-niveaux. Nous avons comparé les performances des GBMs, des CBPSs, de la régression multinomiale, des AIPTW et de l'ajustement entre 3 scénarios se voulant proches de conditions réelles. Ce travail m'a permis d'en apprendre beaucoup sur la conduite d'une étude de simulation. Les GBMs, malgré leur apparente simplicité d'utilisation, se sont révélés complexes principalement à cause de la nécessité d'optimiser des hyperparamètres. Au final dans les 3 scénarios, les GBMs n'aboutissaient pas à une estimation moins biaisée de l'ATE que celle des autres méthodes. Une hypothèse pour expliquer ce résultat serait que les GBMs n'accordent pas un poids assez élevé aux individus extrêmes conduisant à un équilibre imparfait entre les groupes de traitements.

Mon travail de thèse comporte certaines limitations. En matière de traitement des valeurs manquantes, j'ai choisi de me cantonner à l'usage de l'imputation multiple pour leur gestion. Il aurait été intéressant d'explorer d'autres méthodes [46] notamment "Inverse-probability-of-missingness weighting" [67], ainsi que leur utilisation dans le cadre des traitements multi-niveaux. Dans le cadre de mon travail, je me suis concentré sur l'estimation de la variance en utilisant le bootstrap, sans explorer plus avant les différentes méthodes possibles pour estimer la variance [5]. Pour les analyses causales concernant le SAOS, une importante limitation est qu'il m'a été impossible de prendre en compte la durée de sommeil dans nos analyses car la variable contenait trop de valeurs aberrantes pour être utilisable. Ces valeurs aberrantes conduisaient à des viols de l'hypothèse de positivité, les patients avec des durées de sommeil excessivement faibles ne pouvant appartenir au groupe des patients très observants.

Le premier prolongement de mon travail de thèse serait de poursuivre l'utilisation des bases de données observationnelles de patients souffrant de SAOS. En effet ces bases de données sont massives et contiennent des données de qualité grâce à la télé-transmission. Les méthodes d'inférence causale semblent prometteuses pour estimer les effets du traitement par PPC sur les conséquences du SAOS en prenant en compte les différents facteurs de confusion et de restaurer l'échangeabilité conditionnelle. J'ai eu l'occasion d'encadrer un stage et d'amorcer une publication sur les effet de l'observance à la PPC sur le score de dépression. En effet, plusieurs items pour évaluer la dépression reposent sur la somnolence. Il nous paraissait intéressant d'évaluer l'effet causal de la PPC sur le score de dépression et de vérifier que cet effet était exclusivement dû à l'effet sur la somnolence.

Une autre piste que les grandes bases de données observationnelles comme les données du Système National des Données de Santé permettrait d'évaluer est l'effet de l'observance à la PPC au cours du temps en utilisant un modèle marginal structurel. Il serait en effet intéressant d'évaluer l'effet de variations des niveaux d'observance sur l'effet du traitement en prenant en compte les facteurs de confusion temps-dépendant. Cette approche permettrait d'évaluer l'effet de l'observance à la PPC sur une longue durée. Dans le cas des facteurs de confusion temps-dépendant la boucle traitement-facteurs de confusion rend impossible l'ajustement par les régressions [30, 15]. Robins [67] a proposé un nouveau type de modèle : les modèles marginaux structurels permettant d'évaluer l'effet causal dans ce cas. De plus, les approches longitudinales avec les traitements multi-niveaux représentent un défi important car avec de multiples points de temps on multiplie d'autant plus le nombre de scores de propension et donc le risque de poids extrêmes. Cela n'a pas pu être abordé dans le cadre de ma thèse.

Comme nous l'avons vu dans le chapitre 2, binariser un traitement multi-niveaux conduit à une perte d'informations. Le même raisonnement est vrai pour le cas des traitements continus transformés en traitement multi-niveaux. Un prolongement

naturel de mon travail de thèse serait donc de travailler sur les scores de propension généralisés et leur application à l'évaluation de l'effet causal de l'observance à la PPC, l'observance étant une variable continue. Les scores de propension généralisés ne sont pas fondamentalement différents du cas binaire : au lieu d'un score de propension par individu on manipule des densités [31]. Certains modèles comme les CBPSs ou les Generalized Additive Models, permettent nativement l'estimation des SP pour les traitements continus. Mais l'emploi de ces méthodes représente un challenge méthodologique important notamment le contrôle de la balance après pondération n'est pas trivial [6]. L'évaluation des performances de ces modèles semble variable selon le type de modèle de l'estimation de l'estimand et la puissance de l'effet des facteurs de confusions sur les variations du traitement [7]. L'usage et le développement de ces méthodes semblent donc un challenge important que je n'ai pas eu le temps d'explorer dans ma thèse.

Le dernier prolongement de mon travail de thèse est l'émulation d'essais ciblés (emulate target trials) à partir de larges bases de données observationnelles. En 2016, Hernán et Robins ont proposé le "target trial framework" [29]. C'est un guide pour concevoir et analyser les études observationnelles en évitant les biais les plus courants. Dans ce guide, ils recommandent de : (1) définir clairement une question causale sur une intervention, (2) spécifier le protocole de l'essai hypothétique et (3) expliquer comment les données d'observation seront utilisées pour l'émuler. Le but des études émulées est de concevoir des essais pragmatiques, c'est-à-dire où les différentes stratégies de traitement sont comparées dans les conditions usuelles d'utilisation du traitement. Les patients de la base de données répondants aux critères d'inclusion sont inclus dans l'essai émulé. Il est recommandé de n'inclure, dans les patients traités, que les nouveaux utilisateurs afin de ne pas passer à coté d'effets précoces du traitement. Ce type d'études ne peut presque jamais être réalisé en aveugle car les professionnels de santé en charge des patients sont au courant des traitements que ces derniers reçoivent. Une fois les patients inclus, il est nécessaire d'émuler une stratégie d'assignation aléatoire entre les différents groupes de traitements. Pour cela il faut ajuster pour tous les facteurs de confusion à baseline afin d'assurer l'échangeabilité conditionnelle entres les groupes. Il est possible d'utiliser différentes méthodes comme l'appariement, la stratification, l'ajustement, la g-computation ou encore les IPTWs. Je pense que l'utilisation des IPTWs pour le traitement multi-niveaux combinées à l'émulation d'essais pourrait être un outil prometteur pour l'analyse des bases de données liées au SAOS.

VII. Annexes

7.1 Manuscrit publié (Expert Review of Respiratory Medicine)

Taylor & Francis
Taylor & Francis Group

Expert Review of Respiratory Medicine

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ierx20

The individual and societal prices of non-adherence to continuous positive airway pressure, contributors, and strategies for improvement

Monique Mendelson, Jeremy Duval, François Bettega, Renaud Tamisier, Sébastien Baillieul, Sébastien Bailly & Jean-Louis Pépin

To cite this article: Monique Mendelson, Jeremy Duval, François Bettega, Renaud Tamisier, Sébastien Baillieul, Sébastien Bailly & Jean-Louis Pépin (2023): The individual and societal prices of non-adherence to continuous positive airway pressure, contributors, and strategies for improvement, Expert Review of Respiratory Medicine, DOI: 10.1080/17476348.2023.2202853

To link to this article: https://doi.org/10.1080/17476348.2023.2202853

Published online: 24 Apr 2023.

EXPERT REVIEW OF RESPIRATORY MEDICINE
https://doi.org/10.1080/17476348.2023.2202853

Taylor & Francis
Taylor & Francis Group

REVIEW

The individual and societal prices of non-adherence to continuous positive airway pressure, contributors, and strategies for improvement

Monique Mendelson [a*], Jeremy Duval [a,b*], François Bettega [a], Renaud Tamisier [a], Sébastien Baillieul [a], Sébastien Bailly [a*] and Jean-Louis Pépin [a*]

[a]HP2 Laboratory, Inserm U1300, Grenoble Alps University, Grenoble, France; [b]LVL Médical, 44 Quai Charles de Gaulle, Lyon, France

ABSTRACT

Introduction: Continuous positive airway pressure (CPAP) is the first-line therapy for obstructive sleep apnea (OSA). CPAP is highly effective for improving symptoms and quality of life, but the major issue is adherence, with up to 50% of OSA discontinuing CPAP in the first 3 years after CPAP initiation.

Areas covered: We present the individual and societal costs of non-adherence to CPAP, factors associated with non-adherence to CPAP, as well as current strategies for improving adherence including telehealth, couple-based interventions, and behavioral interventions. We also report on challenges and pitfalls for the visualization and analysis of CPAP remote monitoring platforms.

Expert opinion: CPAP termination rates and adherence to therapy remain major issues despite technical improvements in CPAP devices. The individual and societal price of non-adherence to CPAP for OSA patients goes beyond excessive sleepiness and includes cardiovascular events, all-cause mortality, and increased health costs. Strategies for improving CPAP adherence should be individually tailored and aim to also improve lifestyle habits including physical activity and diet. Access to these strategies should be supported by refining visualization dashboards of CPAP remote monitoring platforms, and by disseminating telehealth and innovative analytics, including artificial intelligence.

ARTICLE HISTORY
Received 5 January 2023
Accepted 11 April 2023

KEYWORDS
Adherence; continuous positive airway pressure; couple; data visualization; obstructive sleep apnea; socio-economic; telemedicine

1. Introduction

Obstructive sleep apnea (OSA) is a major health concern, affecting nearly 1 billion individuals worldwide with multiorgan consequences that result in considerable economic and social burdens [1,2]. OSA is independently associated with cardiovascular comorbidities, alteration of quality of life, alteration of neurocognitive function, and depression [3,4].

Continuous positive airway pressure (CPAP) is the first-line treatment for moderate–severe OSA. However, poor compliance brings a great challenge to its effectiveness. CPAP has been shown to be effective for alleviating symptoms, restoring neurocognitive function, and improving quality of life [5,6]. Treatment adherence, according to the World Health Organization, is 'the extent to which a person's behavior – taking medication, following a diet, and/or executing lifestyle changes – corresponds with the agreed recommendations from a healthcare provider'. It is admitted that CPAP adherence is a key factor of treatment efficiency [7] and as an illustration, it has been established that a minimum of 4 hours of CPAP is required in order to decrease blood pressure in minimally symptomatic patients [8].

Poor adherence to continuous positive airway pressure (CPAP) treatment is associated with persistent symptoms (i.e. residual sleepiness [9–11], altered quality of life, and work productivity [12]), substantial health-care costs, and excess in morbidity and mortality. Thus, low levels of adherence are a significant obstacle in the effective management of OSA. Early identification and management of poor adherence to CPAP treatment is of major clinical importance to optimize treatment outcomes in patients with OSA.

The novelty of this review is to summarize the individual and societal consequences of non-adherence to CPAP, as well as challenges and pitfalls for the visualization and analysis of CPAP adherence data. Factors associated with non-adherence to CPAP such as comorbidities, psychological factors, couples' profile, socio-economic status, and access to care are presented. We also expose tailored strategies for improving adherence that target these factors including telehealth, couple-based interventions, and behavioral interventions. Expanding awareness of factors associated with non-adherence to CPAP can be used to design effective interventions that will improve patients' integrated care.

2. CPAP adherence data

2.1. Definitions of CPAP adherence

Continuous positive airway pressure (CPAP) is the first-line treatment for OSA. There are relatively arbitrary definitions to define *optimal CPAP adherence*. The most widespread threshold used is the one established by the United

CONTACT Jean-Louis Pépin Jpepin@chu-grenoble.fr SHU Pneumologie – Physiologie, Centre Hospitalier Universitaire Grenoble Alpes, Grenoble, Cedex 9 38043, France

*These authors contributed equally.

Article highlights

- CPAP termination rates remain very high, and adherence to therapy is a major issue. In the last 20 years, no significant improvement in CPAP adherence has been observed despite obvious improvements in technical aspects of devices.
- OSA severity, as assessed by apnea–hypopnea index, and technical aspects relative to CPAP devices seem to be minor contributors in explaining CPAP adherence.
- Other factors (comorbidities, psychological, couples profile, socio-economic status, access to care, and cultural diversity) should be better acknowledged and included in tailored interventions.
- Strategies for improving CPAP adherence should be individually personalized and aim to improve not only CPAP adherence but also global lifestyle habits including physical activity and diet.
- Access to these strategies should be supported by improving visualization dashboards of CPAP remote monitoring platforms and by disseminating telehealth and innovative analytics including artificial intelligence.

States Centers for Medicare and Medicaid Services (CMS), which requires device usage for 4 h/night on ≥4 nights/week during a 30-day period in the first 90 days of therapy. If these CMS adherence criteria are not achieved, in the United States, OSA patients have therapy withdrawn at the end of the 90 days following CPAP initiation. The CMS thresholds and, in particular, the 4 h/night device usage, have been adopted as targets in many countries around the world. In France, these metrics of adequate adherence and patients' acceptance of CPAP remote monitoring are associated with different levels of reimbursement.

The minimal clinically important difference (MCID) for improvement in adherence used in clinical studies and for evaluating the impact of interventions is 30 min [13].

2.2. Big data analyses of short-term CPAP adherence

Hundreds of millions of OSA patients are treated by CPAP worldwide and remote monitoring of CPAP adherence is becoming available in developed countries. Nightly objective measurements of CPAP adherence are available at large scale, and short-term CPAP adherence based on big data has been reported in unselected populations [14,15]. This collection of nightly objective adherence from millions of patients is a unique characteristic of the sleep apnea field. From over 2.6 million US OSA patients, CMS adherence in the first 90 days has been reported as high as 75% with an overall mean daily usage (all days) of 5.54 h/night [14]. The proportion of days with non-zero usage was 93%. In another big data study [15], across 789,260 patients initiated on CPAP, adherence with CMS criteria was 72.6% but varied dramatically by age and gender ranging from 51.3% in young women to 80.6% in 71–80-year-old men. Over the past decade, CPAP adherence improved over time, but association between socio-economic status and health inequities in CPAP adherence remained highly significant as strong determinants for poor adherence [16].

2.3. Long-term adherence and CPAP termination rates

Long-term randomized controlled trials have been essentially conducted on specific OSA phenotypes (i.e. non-sleepy, minimally symptomatic patients with cardiovascular disease). These trials have shown that fewer than 50% of patients are using CPAP for >4 h/night after several years of follow-up [17,18], and the mean adherence on the long term was close to 3 hours per night. In clinical cohorts including more diverse OSA phenotypes, including sleepy individuals, adherence data are higher. In 5,138 OSA patients with a median follow-up of 6.6 years, 1,311 patients (25.5%) were considered as CPAP non-adherent (mean daily CPAP use 0–4 h) [19]. In half a million OSA patients from a national exhaustive database in France [20], CPAP termination rates were investigated in new CPAP users initiating the therapy in 2015 and 2016. Overall CPAP termination rates after 1, 2, and 3 years were 23.1%, 37.1%, and 47.7%, respectively. In a multivariable analysis, age categories, female sex (1.09 (1.08–1.10)), COPD (1.12 (1.10–1.13)), and diabetes (1.18 (1.16–1.19)) were significantly associated with higher CPAP termination risk. Therapy termination rates were highest in younger or older patients with ≥ 1 comorbidity.

All together, these data consistently demonstrate that long-term CPAP continuation and adherence are significant concerns that require the development of specific personalized management pathways and patients' engagement tools in populations at risk for non-adherence.

3. The individual and societal price of non-adherence to CPAP

3.1. Optimal amount of CPAP usage is still debated and may vary depending on outcomes

A substantial body of evidence suggests that the presence of OSA contributes to a number of poor health outcomes, including neurocognitive impairment and sleepiness, hypertension, and cardiovascular diseases leading to early mortality [2,21,22].

Reference thresholds of CPAP adherence targeted by caregivers and patients are relatively arbitrary. The quality of evidence is poor to establish an optimal duration of CPAP usage. It is unclear whether improvements related to CPAP therapy are driven by a threshold effect (e.g. occurring when the usage is >4 h/night) or a dose–response relationship (i.e. 'more is better'). The situation is even more complex, as the optimal duration of CPAP adherence may vary depending on the outcome of interest. For example, one additional hour of CPAP usage was associated with an additional reduction in systolic blood pressure of 1.5 mmHg and an additional reduction in diastolic blood pressure of 0.9 mmHg [23]. In clinical cohorts, a nightly usage of 6 hours seems to be required for reducing incident cardiovascular events [19]. For improvement in clinical symptoms, a linear dose–response relationship was found between increased use of CPAP and achieving normal levels of both objective and subjective daytime sleepiness. Up to 7-hour use seems necessary for full benefit in terms of quality of life assessed by the Functional Outcomes associated with Sleepiness Questionnaire (FOSQ) [24] (Figure 1). A clear dose–response relationship between CPAP usage and health-care resource utilization was recently shown in

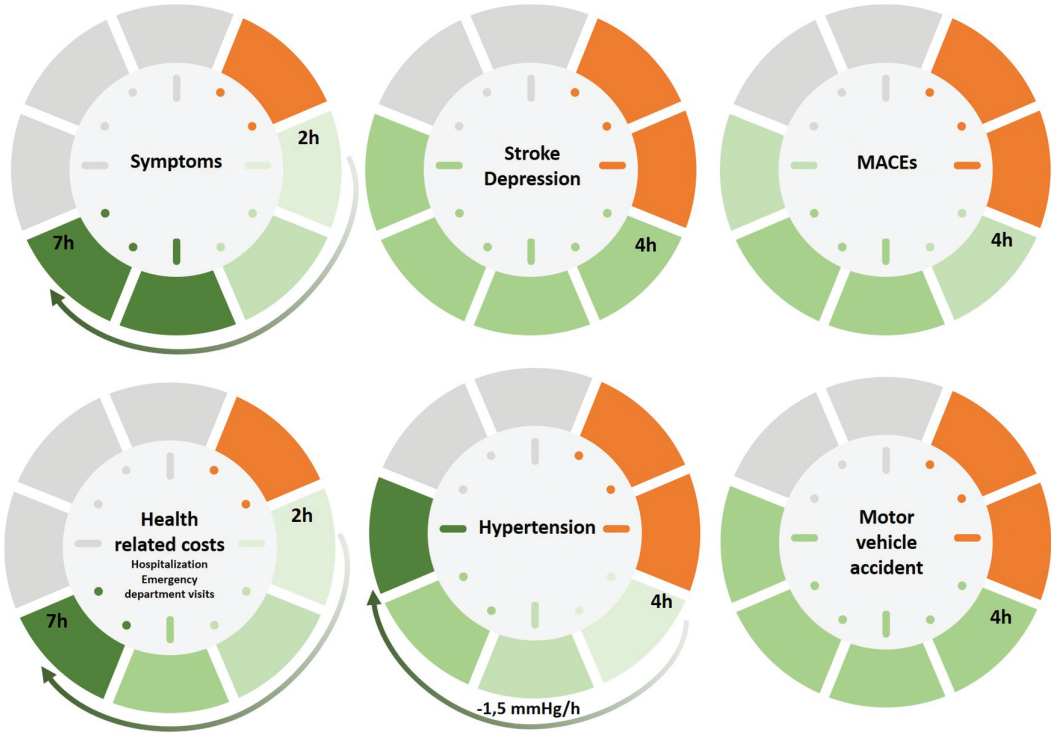

Figure 1. Relationship between CPAP adherence and effectiveness.

a study using a linked data set, with benefits seen even when usage was as low as 1–2 hours per night [25].

Interestingly, despite being widely adopted, the 4-hour adherence rule has not undergone empirical testing to demonstrate superiority nor is this threshold validated as significant to patient outcomes. Moreover, the 4-hour rule has the unintended consequence of worsening access to CPAP in patients with the fewest resources and who are often the most vulnerable to the effects of untreated OSA [26]. As mentioned previously, the CMS policy requires a minimum of 4 hours of adherence to qualify for long-term coverage of CPAP treatment. Unfortunately, this policy exacerbates disparities by unequally affecting patients with greater barriers and challenges to meeting the CMS adherence requirement because of socio-economic and structural factors. Specifically, people with lower SES tend to have lower CPAP use than those with a higher SES. For example, in an observational study of veterans, CPAP adherence was associated with neighborhood SES, with only 34% of participants using CPAP 4 hours or more among those residing in the lowest SES block compared with 62% among those in the highest SES blocks [27]. In this study, the difference could not be attributed to financial burden because all CPAP costs were covered under the Veterans Administration benefits. Thus, a recently published Official American Thoracic Society Policy Statement advocates for a revision of the adherence requirements for CPAP coverage in order to promote health equity and align with patient-centered outcomes, including improved sleepiness and sleep quality [26].

3.2. The price of non-adherence to CPAP for OSA patients: residual sleepiness, burden of cardiovascular events, all-cause mortality, and COVID-19

3.2.1. Cardiovascular events and all-cause mortality

There have been three major randomized controlled trials (RCTs) conducted to assess the effects of CPAP on secondary prevention of cardiovascular events: SAVE [17], RICCADSA [28], and ISAACC [18]. These three studies provided neutral results, as the use of CPAP did not reduce the primary composite endpoint of major adverse cardiac events (MACEs). However, the results of pre-specified sensitivity analysis in the SAVE trial showed that the subgroup with CPAP adherence >4 h per night demonstrated a reduction in the risk of stroke and total cerebrovascular events [17]. Accordingly, in the RICCADSA trial [28], the subgroup using CPAP >4 h per night demonstrated a lower rate of cardiovascular events.

There is also real-world evidence demonstrating a link between CPAP adherence and cardiovascular and risk of premature mortality . In a French nationwide database analysis including half a million CPAP-treated patients using propensity score matching [29], the authors demonstrated a relationship between CPAP termination and incident cardiovascular events (e.g. incident heart failure and incident hypertension) and all-cause mortality. Data from the *Pays de la Loire* Cohort linked to health administrative data aimed at identifying incident MACEs demonstrated that the reduction in cardiovascular events occurred only for nightly CPAP usage above 6 hours per night [19]. The association with CPAP adherence was stronger in male patients, OSA without

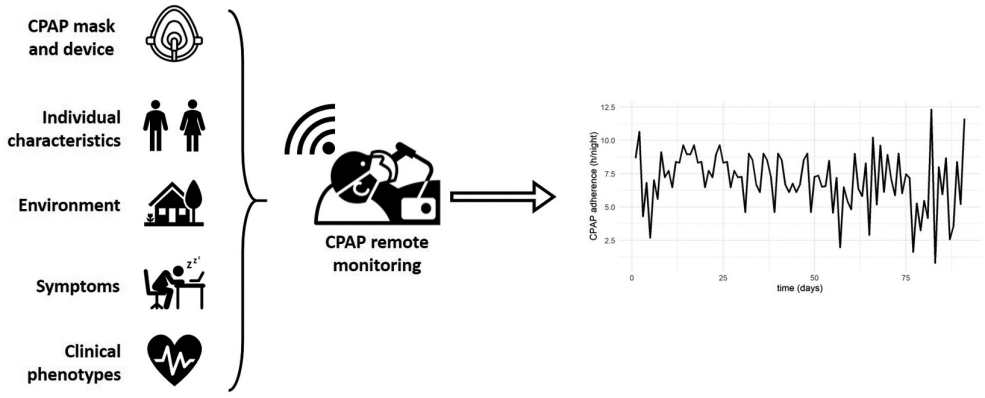

Figure 2. CPAP adherence remote monitoring.

overt CV disease at diagnosis, and the subgroup reporting excessive daytime sleepiness at initiation of therapy.

3.2.2. Residual sleepiness

Excessive daytime sleepiness (EDS) affects approximately half of patients with obstructive sleep apnea (OSA) and can persist despite primary OSA therapy using CPAP [9,11]. However, in the largest study to date available in the field [10], residual excessive sleepiness prevalence in CPAP-treated obstructive sleep apnea patients was 13% (18% for those with an initial Epworth Sleepiness Scale score >11), and significantly decreased with CPAP use (9% in ≥6 h night^{-1} CPAP users).

3.2.3. The societal price of non-adherence to CPAP: health-care utilization

The economic costs associated with OSA are substantial for both the individual and society [30]. As an example, the economic burden of OSA-related motor vehicle accidents was estimated at 810,000 collisions and 1,400 fatalities in the United States in 2000 [30]. Untreated OSA leads to aggregation and progression of comorbidities that can potentially increase health-care utilization. There are now data providing compelling evidence for a dose–response relationship between CPAP usage and health-care utilization.

In the Medicare fee-for-service database, the total episode-of-care costs of CPAP-adherent patients ($6825) were lower than those of non-adherent ($11,312; $P < .05$) and control ($8102) participants [31]. In a national analysis of older adult Medicare beneficiaries with OSA, CPAP usage was associated with a 2% reduction in risk of stroke for each month of CPAP adherence [32]. The co-occurrence of OSA and chronic obstructive pulmonary disease, termed overlap syndrome, is associated with poor prognosis and a high burden of care consumption. CPAP usage by patients with overlap syndrome was associated with reduced all-cause hospitalizations and emergency room visits, severe acute exacerbations, and health-care costs [33].

There are more data to come to demonstrate the crucial role of CPAP adherence for reducing the health-related costs in the comorbid associations of sleep apnea with diabetes, heart failure, and atrial fibrillation.

3.2.4. COVID-19

The results of a study examining the risk of severe COVID-19 in non-adherent OSA patients showed that the patients who were adherent to CPAP were less likely to experience a severe course of COVID-19 or death than OSA patients non-compliant with therapy. This was observed despite the fact that the former group had more severe OSA. This result underlines the importance of adherence to CPAP therapy in OSA in the context of a global pandemic [34].

4. CPAP adherence: challenges and pitfalls for data analyses, new metrics, and visualization tools to depict diversity and complexity of CPAP adherence patterns

4.1. Size of the problem and data science challenges

The emergence and validation of communicating CPAP devices have enabled the materialization of remote monitoring platforms for the collection and visualization of data generated nightly by hundreds of millions of patients worldwide. CPAP remote monitoring platforms generate an avalanche of data collected bedside daily and processed by CPAP manufacturers and caregivers. Algorithms have been developed to compute and aggregate datasets. These algorithms include information regarding adherence (hours of usage/night), and efficacy indexes of residual apnea–hypopnea index (rAHI) events and leaks (Figure 2).

There is an urgent need to not only improve the data cleaning and processing of these datasets but also solve major concerns for data science applications [35]. The main concerns that need to be addressed are the following: (1) additional validations of rAHI reliability, especially for central events, (2) lack of standardization of the summarized data provided by the different PAP brands, (3) handling of missing values and 4) correct appreciation of treatment interruptions.

4.2. Innovative analytic methods for identification of trajectories of adherence and CPAP efficacy: guidance and alerts to clinicians, caregivers, and patients

Longitudinal data from CPAP remote telemonitoring since the time of CPAP initiation provide interesting insights and can be analyzed as time series, including trends, cyclic components, or can be used to identify specific events responsible for changes in the pattern of the series. For example, events such as change in CPAP device can generate significant changes in the observed rAHI in a patient, due to different reports provided by different CPAP manufacturers [36]. Furthermore, a patient who interrupts CPAP use due to traveling or because of the onset of an incident medical condition will generate missing or null values that should be managed with early information to the caregivers.

CPAP adherence data that have been managed correctly can provide important information on CPAP adherence trajectories and identification of acute events, as it has been shown for COVID-19 infection [37] or acute cardiovascular events [38]. Various approaches to consider CPAP adherence trajectories have been proposed. Babbin et al. performed unsupervised clustering in CPAP adherent patients after exclusion of CPAP terminations [39]. Their study aimed at describing individual patterns of CPAP use over time, quantitatively creating subgroups of individuals with similar patterns, and identifying variables related to the subgroups. Ultimately, their aim was to assess the utility of combining time-series analysis to identify patterns over time. In this study, a four-cluster solution was found, and participants were distributed among the following groups: great users, good users, low users, and slow decliners. The authors of this study concluded that combining time-series analysis and dynamic cluster analysis was a useful way to evaluate longitudinal patterns of CPAP adherence at both the individual and subgroup level. However, in order to improve generalization, future studies should include groups of patients who terminated CPAP. Bottaz-Bosson et al. [40] explored clustering approaches based on dynamic time warping to consider the time course patterns of CPAP adherence and to identify different clusters of CPAP adherence trajectories, including patients who had stopped CPAP. Such analyses are useful to identify low, intermediate, and high levels of CPAP adherence groups, as well as to provide a risk stratification for CPAP termination.

Another clinically relevant topic is to assess the relationship between CPAP treatment adherence and patient outcomes by using causal inference approaches because the binary threshold of 4 h of CPAP adherence probably limits the understanding of this relation. Bettega et al. [41] addressed this question by using different methods to approach CPAP adherence. In their study, they illustrated how two causal inference methods (i.e. inverse-probability-of-treatment-weighting and inverse-probability-of-treatment-weighting with regression adjustment) can be applied on observational data for the estimation of the effect of different ranges of CPAP adherence on daytime sleepiness. Thus, multiple patterns of CPAP adherence should be considered to truly investigate a dose–response effect.

4.3. Dashboards and visualization tools

This type of approach using trajectories can be visually reported on dashboards that can contribute to the implementation of personalized medicine in OSA (Figure 2). Also, the identification of patient trajectories of CPAP use versus the variability of treatment efficiency based on rAHI measures is of major interest [42]. A recent study proposed automated analyses of the impact of changes in CPAP masks as a tool to be included in remote monitoring platforms for raising alerts after harmful interventions (i.e. when a change from a nasal to facial mask increases rAHI) [43].

Raw data on patterns of CPAP adherence and not aggregated data are informative of some of the OSA phenotypes. In patients with comorbid insomnia and sleep apnea (COMISA) [44], one of the most frequent OSA phenotypes, several distinct periods of CPAP use occur during the night and during daytime naps, reflecting the severity of insomnia and sleep deprivation. Such information can now be easily available for caregivers in routine practice.

5. Factors associated with CPAP adherence

The factors influencing CPAP adherence (or non-adherence) are multidimensional and involve characteristics and behaviors by the physician, patient and health-care system (Figure 3).

The severity or symptoms of OSA have been shown to affect CPAP adherence. AHI and oxygen desaturation index (ODI) have been positively associated with CPAP adherence in patients who present excessive daytime sleepiness [45,46] and who do not [47].

Despite the increase in CPAP treatment options (i.e. bilevel positive airway pressure, and auto-adjusting CPAP) and numerous advances in machine dynamics including quieter pumps, softer masks, and improved portability, treatment acceptance and adherence remain low and stable over the last 20 years [48]. No clinically significant improvement in CPAP adherence has been observed in recent years despite efforts toward behavioral intervention and patient coaching as well as advances in machine dynamics. Technical advances are probably not key targets, and dissemination at scale of behavioral therapies is lacking.

Furthermore, to date, no single factor has been consistently identified as predictive of CPAP acceptance and adherence [49,50]. Therefore, understanding obstacles and critical elements associated with a patient's decision to use CPAP is crucial in promoting treatment acceptance and designing interventions [51–54]. It is now increasingly clear that factors influencing adherence to CPAP go beyond disease severity and machine options and include patient characteristics, and psychological and social factors.

5.1. CPAP-related side effects

Side effects to CPAP treatment are relatively common and include mask leakage, mask pressure, dry mouth, nasal congestion, claustrophobia, and difficulties exhaling [55]. However, a number of investigations have failed to demonstrate a correlation between a reduction in side effects offered

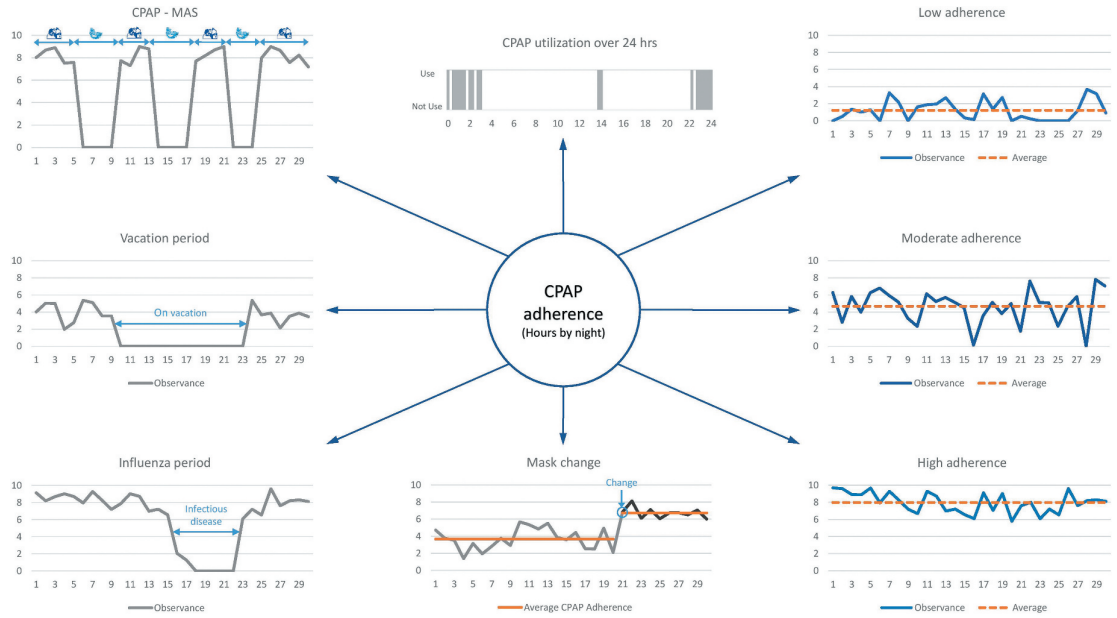

Figure 3. Visualization of CPAP adherence.

by technical solutions (i.e. air humidifiers and different types of devices) and increased adherence to CPAP [54,56–58].

5.2. Socio-economic status (SES)

Higher income [59,60], educational level [61], socio-economic status in neighborhood [27,59,61], number of household members, and civil status [27,62] have been associated with increased CPAP adherence in some, but not all studies [63–66].

Furthermore, socio-economic status is positively associated with health literacy and access to health-care services [60]. Previous studies have shown that low SES predicts poor adherence to CPAP, and patients with higher SES are more likely to commence treatment [67]. In a study that evaluated SES based on monthly income level, for each increase in income-level category, the odds for CPAP acceptance increased by 140% [60]. Furthermore, low-income level was found to be an independent determinant for not accepting CPAP. Low SES has been recognized as a potential barrier to chronic treatment due to multiple and complex interactions [60]. These complex interactions generally encompass not only income and education level but also a wide range of associated factors that may affect the quality of health-care patients receive, including insurance status, access to care, patients' health beliefs, and many facets of the doctor–patient relationship, such as trust and communication [68–70]. Patients with lower SES have a greater likelihood of exposure to SES risk behaviors including smoking, excessive alcohol consumption, physical inactivity, poor diet, nonattendance to medical visits, and poor adherence with treatment for chronic conditions [71].

A more recent large population-based study with extended follow-up showed that civil status, educational level, household income, and foreign background predict CPAP adherence in a clinically significant manner [72].

5.3. Treatment efficacy and interfaces

It has been consistently reported that a high rAHI (i.e. above 5–10 per hour) is associated with poor adherence and an increased rate of CPAP termination [73]. This is particularly true for emergent central sleep apnea (CSA) under CPAP. CSA emerging or persisting under CPAP is associated with male sex, age, sedentary lifestyle, OSA severity, cardiovascular comorbidities (heart failure and arrhythmia), and type of interface (orofacial mask versus nasal mask). Early recognition of such clinical situations facilitates early intervention to reduce the risk of therapy discontinuation and shift to more efficient ventilator modalities allowing to restore efficacy and good adherence [74,75].

The type of interface and supply also play a significant role in driving CPAP adherence. It has been demonstrated that the regular replacement of mask interface components is associated with better long-term adherence to CPAP therapy versus no resupply [76]. Case studies, small randomized controlled trials, and a meta-analysis demonstrated deteriorations in rAHI [43] and lower adherence to CPAP in OSA patients when they switch from nasal to facial masks [77–80].

5.4. Auto-CPAP versus fixed pressure CPAP

Non-behavioral interventions aimed at improving CPAP adherence, such as mechanical interventions which involve changing the way that positive pressure is delivered, have also been investigated. A systematic review and meta-analysis that compared automatic-CPAP (auto-CPAP) to fixed pressure CPAP found no statistical difference in machine use or in adherence. However, patients preferred auto-CPAP over fixed pressure CPAP [81]. Similarly, a more recent systematic review and meta-analysis found a mild improvement in CPAP adherence, patient preference, and sleep architecture in favor of

auto-CPAP compared to fixed-CPAP [82]. However, the clinical relevance of these findings required further study.

5.5. Social support–bed partners

Spouses can influence patients' health behaviors and be an integral component to successful CPAP adherence [72]. In a retrospective cohort of 330 OSA patients, being married was associated with a higher nocturnal CPAP use of >4 hours after 1 week [27]. In another study with 80 patients and a 1-month follow-up, patients living with a partner showed higher CPAP use [62]. Gentina et al. found that the quality of the marriage and engagement of the partner affected nightly CPAP use [83].

5.6. Adherence during the COVID-19 pandemic

During the COVID-19 lockdown period, a large prospective cohort conducted in France showed that there was a significant improvement in CPAP adherence [84]. Interestingly, there was an inverse relationship between pre-COVID-19 adherence and increase in adherence during lockdown. Furthermore, a particularly significant increase was observed among individuals with poor adherence during the pre-COVID-19 period. Notably, the increase in CPAP adherence observed exceeds the threshold of adherence improvement considered clinically significant (i.e. 30 minutes) [13]. The authors proposed a number of speculations to explain their observations. First, it is possibly that a proportion of adult men under 65 years old who were still working might have been suffering from chronic sleep deprivation with reduced sleep time and thus reduced adherence. Lockdown curtailed their activities and allowed these individuals to recover from sleep loss and increase their duration of nightly CPAP usage and thus adherence. Second, CPAP-treated patients who started therapy in the spring before the pandemic (the reference period used in this study) might have increased their adherence over 1 year independently of lockdown effects. Finally, OSA has been consistently reported to be associated with severe COVID-19 [85]. This forewarning had been widely disseminated to patients by physicians and home care providers [86]. Thus, better adherence might have been linked to a fear of hospitalization in patients with COVID-19. This anxiety may have triggered a change in behavior regarding CPAP use.

A smaller cross-sectional study of severe OSA patients showed that lockdown-related CPAP adherence improved in severe OSA patients, with a shift in almost half of poor pre-lockdown adherers toward good lockdown CPAP adherence [87]. Another prospective study examining the impact of the COVID-19 pandemic on CPAP adherence over the entire 2020 year failed to show a significant change in adherence [88]. However, this study compared adherence from 2019 to 2020, and the differences observed are not only related to the COVID-19 pandemic but also the consequence of circumstantial events like the 2019 heat wave in Europe.

Whether the COVID-19 pandemic impacted CPAP adherence over the long term is unknown and only preliminary short-term data have been reported to date. Nevertheless, these observations suggest that behavioral interventions, based on the patients' perception of both disease-related risk and CPAP-related benefits could improve adherence.

In addition to already established determinants of CPAP adherence, a number of socio-economic factors such as economic, educational, and marital status affect CPAP acceptance, as mentioned above. When treating patients with CPAP, a greater awareness of the impact of these different socio-economic factors on adherence, and, when necessary, individually tailored follow-up, may improve treatment adherence and may contribute to health equity.

6. Strategies for improving CPAP adherence

Several strategies are employed to promote adherence and enhance the efficacy of CPAP, i.e. multiple educational interventions, behavioral therapies, CPAP device modifications, and telehealth.

6.1. Couples-based interventions

Due to the dyadic (i.e. pairing of two individuals) nature of sleep for adults living with spouses or bed partners, the effects of OSA and its treatment expand beyond the individual patient. Results from studies examining facilitators and barriers to CPAP use perceived by patients have suggested that spouses play an important role in adherence to CPAP treatment [89]. Thus, there is a strong scientific premise for designing couple-based interventions aimed at improving CPAP adherence [90]. First, it has been shown that couples' sleep is highly interdependent, meaning that one partner's sleep affects and is affected by the other partner's sleep [91]. In a clinical trial, OSA treatment was associated with a 50% reduction in the other partner's nocturnal arousals [92]. Second, an important motivator for patients to seek OSA diagnosis is sleep disruption reported by a bedpartner. Third, there is strong evidence supporting the importance of couple-based interventions in other chronic diseases for improving treatment adherence, symptom management, and patient and partner health outcomes [93]. Finally, partner support is pivotal to encourage adherence to CPAP while as conflict in the relationship can reduce adherence [94]. A pilot investigation comparing a couples-oriented intervention with a patient-oriented intervention showed improvements in patients' CPAP adherence and in sleep quality and sleepiness in the couples-oriented intervention group [95]. A recent scoping review showed that the presence of a partner promotes adherence to CPAP therapy in patients with OSA, resulting in ameliorating their overall quality of life [96].

Taken together, these findings provide empirical support for designing interventions aimed at increasing CPAP adherence that integrate the partner. In this line, the protocol of a randomized pilot/feasibility trial 'We-Pap' including an intervention that uses a couple-based treatment model to target CPAP adherence, as well as the broader sleep health issues that affect both the patient and partner, was recently published [97].

6.2. Telemonitoring/Patient engagement tools

A novel aspect of CPAP devices is the machine- or web-based tracking systems that generate information both to the health-care provider and to the patient.

Over the last 20 years, telehealth technology has been applied to the field of CPAP, easing the follow-up process and allowing health-care providers to deliver more consistent care. Telehealth is defined as the application of telecommunications and digital communication technologies to deliver and facilitate health services [98]. Telehealth was developed to provide health-care services at a distance and can be used for assessment, diagnosis, treatment, obtaining/retrieving data, and evaluation.

The 2015 guidelines for telemedicine utilization published by the American Academy of Sleep Medicine (AASM) promoted telehealth technology development, and the COVID-19 pandemic has expedited internet-based home telemedicine for the diagnosis and treatment of OSA [99]. Recent updates from AASM advocated for the delivery of high-quality sleep care through telehealth interventions and suggested a significant role of telehealth in maintaining the continuity of sleep health [100].

Telemonitoring (TM) is a subset of telehealth. TM of patients treated with CPAP devices provides information on CPAP adherence, efficacy, and leaks. TM can also include the use of electronic messages [101] and self-management platforms. Health-care providers and physicians can use these mobile platforms to identify potential sources of poor CPAP adherence, including mask leaks, CPAP side effects, unfounded patient beliefs, and/or inappropriate behaviors regarding their therapy [99,102,103].

A limited number of studies have investigated the impact of telemonitoring on CPAP adherence and have produced conflicting results. Some studies have suggested higher CPAP usage with TM compared with usual care [101,103–105]. However, most studies have failed to show better CPAP adherence [101,106–110]. These discrepancies might be partly explained by the fact that the impact of TM on CPAP adherence is dependent on OSA phenotypes and that adherence is a multifactorial issue that cannot be solved only by TM. Recently, there have been a number of studies that indicate phenotypes or clusters of OSA symptoms. For example, Gagnadoux and coworkers identified five clusters based on gender, presence of insomnia and comorbidities, depressive symptoms, and daytime sleepiness, as well as other typical nocturnal and diurnal OSA symptoms [110]. In this study, CPAP use >4 hours/night and symptom improvement differed between the clusters.

A significant interest in TM is the capability for early identification of patients at risk of poor adherence. This would allow re-allocation of resources to low adherence subgroups while reducing the number of visits for those with good adherence for whom regular nightly TM is sufficient. The cost-effectiveness validation of such a strategy is essential because a number of studies have highlighted the greater burden of caregivers' interventions for both technical (masks, humidifiers, etc.) and medical (residual sleepiness and persistent high residual AHI) reasons.

TM has also shown promise for enhancing self-perceived efficacy of CPAP treatment and/or improving biological parameters in a multimodal approach. It is often assumed that support of any kind during CPAP treatment can improve patient's engagement for CPAP adherence and adopting a healthy lifestyle. Recently, the American Heart Association updated their construct of cardiovascular health by including healthy sleep [111]. Thus, well-designed multimodal telehealth interventions can encourage positive health promotion and preservation alongside CPAP adherence.

It must be recognized that access to the required technology, both from the provider and patient standpoint, is a limitation to widespread use. This drawback will likely diminish over time as the availability of technology increases and cost decreases. The success of telehealth/monitoring may also be dependent on knowing who best responds (age, education, and geographic setting) to the different telehealth approaches.

6.3. Supportive, educational, and behavioral interventions

A recent Cochrane review examined the effects of supportive, educational, and behavioral interventions on CPAP adherence [112]. Educational interventions were defined as interventions imparting information about CPAP treatment or OSA more generally, delivered through in-person sessions, group educational sessions, written materials, video format, or any combination of these. Supportive interventions referred to interventions in which participants were provided with additional clinical follow-up or with telemonitoring equipment that facilitated either self-monitoring of CPAP usage or monitoring by clinical staff to prompt 'as needed' clinical follow-up. Behavioral interventions employed psychotherapeutic techniques derived from behavioral, cognitive, or related models of health behavior change. By definition, behavioral interventions under any of the models used involve at least a minimal degree of direct participant engagement or interaction (as opposed to purely educational, in which information is merely imparted to participants). Behavioral interventions target a modifiable and measured construct known or hypothesized to influence health beliefs about OSA and CPAP therapy [112]. The authors found that all types of interventions increase CPAP usage with varying levels of certainty. Behavioral therapy increases machine usage by 79 min per night, and ongoing supportive interventions probably increase machine use by about 42 min per night. This is significantly greater than the MCID of 30 min [13]. Educational and mixed interventions may potentially improve machine usage; however, the certainty of this evidence is very low.

7. Expert opinion

Continuous positive airway pressure (CPAP) is the first-line treatment for OSA and has been shown to be effective for alleviating symptoms, restoring neurocognitive function, and

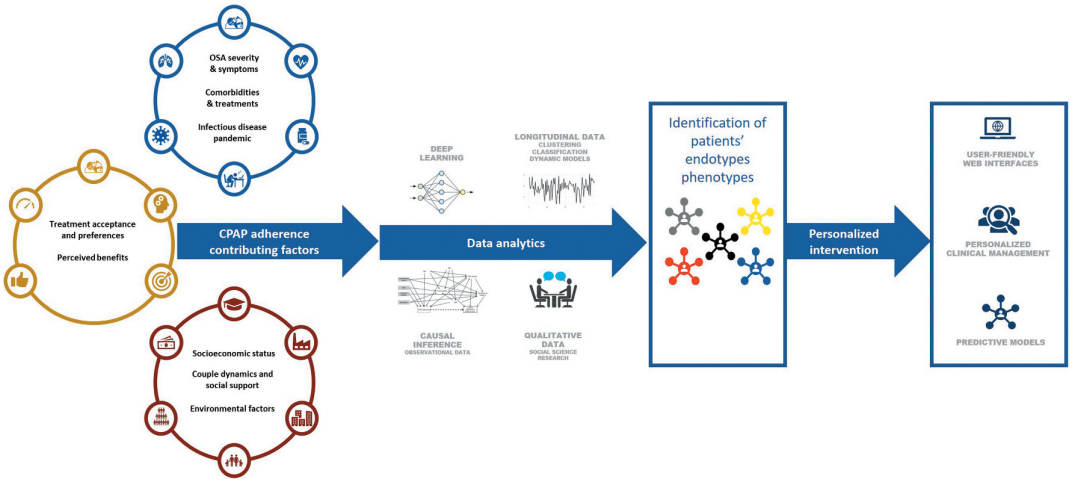

Figure 4. Factors influencing CPAP adherence, data analytic tools/methods, identification of patient phenotypes and development of personalized interventions.

improving quality of life. Adherence to treatment has been identified as a key factor for treatment efficiency and poor adherence is associated with persistent symptoms including residual sleepiness, altered quality of life, considerable health costs, and increased morbidity and mortality. Unfortunately, over the past 20 years, despite improvements in CPAP treatment options, machine dynamics, treatment acceptance, and adherence remain low among patients with OSA [48].

Due to the multidimensional nature of CPAP adherence, one type of intervention is unlikely to meet the need of all patients. A multi-dimensional evaluation of patient characteristics and their eco-system should allow the design and implementation of individualized interventions aimed at improving CPAP adherence.

More attention should be paid to health iniquities, health economic status, and health literacy. After appropriate baseline evaluation and risk stratification of non-adherence, the type of intervention might be prioritized and might evolve over the time course of patient follow-up. Interestingly, the emergence and validation of communicating PAP devices has enabled the materialization of remote monitoring platforms for the collection and visualization of data generated nightly by hundreds of millions of patients worldwide. Accumulated PAP data provide valuable and objective information regarding patient treatment adherence and efficiency. Improving the quality and standardizing data handling could facilitate data sharing among specialists worldwide and enable artificial intelligence strategies to be applied in the field of sleep apnea.

Remote CPAP monitoring platforms are already in place for informing caregivers. This is a unique opportunity to implement 'at scale' telehealth interventions to support adherence (Figure 4). Thus, future studies designed to evaluate individualized interventions are needed in order to implement these types of interventions at a large scale.

Funding

JL Pépin, F Bettega, S Bailly, S Baillieul, and R Tamisier are supported by the French National Research Agency in the framework of the 'Investissements d'aveni' program (ANR-15-IDEX-02) and Grenoble Alpes University Foundation (ANR-19-P3IA-0003) 'Chair of excellence' 'e-health and integrated care and trajectories medicine and MIAI artificial intelligence'. J Duval is supported by LVL Medical and The Grenoble Multidisciplinary Institute for Artificial Intelligence (MIAI) in the framework of a 'Convention Industrielle de Formation par la Recherche' (CIFRE) PhD.

Declaration of interest

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

Reviewer disclosures

Peer reviewers on this manuscript have no relevant financial or other relationships to disclose.

ORCID

Monique Mendelson http://orcid.org/0000-0001-8774-8510
Jeremy Duval http://orcid.org/0000-0001-6037-5486
François Bettega http://orcid.org/0000-0002-9736-5289
Renaud Tamisier http://orcid.org/0000-0003-1128-6529
Sébastien Baillieul http://orcid.org/0000-0002-2348-6918
Sébastien Bailly http://orcid.org/0000-0002-2179-4650
Jean-Louis Pépin http://orcid.org/0000-0003-3832-2358

References

Papers of special note have been highlighted as either of interest (•) or of considerable interest (••) to readers.

1. Benjafield AV, Ayas NT, Eastwood PR, et al. Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. Lancet Respir Med. 2019 Aug;7(8):687–698.
2. Levy P, Kohler M, McNicholas WT, et al. Obstructive sleep apnoea syndrome. Nat Rev Dis Primers. 2015 Jun 25;1:15015. DOI:10.1038/nrdp.2015.15

3. Javaheri S, Barbe F, Campos-Rodriguez F, et al. Sleep apnea: types, mechanisms, and clinical cardiovascular consequences. J Am Coll Cardiol. 2017 Feb 21;69(7):841–858. DOI:10.1016/j.jacc.2016.11.069
4. Marin JM, Carrizo SJ, Vicente E, et al. Long-term cardiovascular outcomes in men with obstructive sleep apnoea-hypopnoea with or without treatment with continuous positive airway pressure: an observational study. Lancet. 2005 Mar 19-25;365(9464):1046–1053. DOI:10.1016/S0140-6736(05)71141-7
5. Yu J, Zhou Z, McEvoy RD, et al. Association of positive airway pressure with cardiovascular events and death in adults with sleep apnea: a systematic review and meta-analysis. JAMA. 2017 Jul 11;318(2):156–166. DOI:10.1001/jama.2017.7967
6. Li Z, Cai S, Wang J, et al. Predictors of the efficacy for daytime sleepiness in patients with obstructive sleep apnea with continual positive airway pressure therapy: a meta-analysis of randomized controlled trials. Front Neurol. 2022;13:911996.
7. Patil SP, Ayappa IA, Caples SM, et al. Treatment of adult obstructive sleep apnea with positive airway pressure: an American Academy of Sleep Medicine Clinical Practice guideline. J Clin Sleep Med. 2019 Feb 15;15(2):335–343. DOI:10.5664/jcsm.7640

•• **This meta-analysis, systematic review and GRADE assessment provides supporting evidence for the clinical practice guideline for the treatment of obstructive sleep apnea (OSA) in adults using positive airway pressure (PAP).**

8. Bratton DJ, Stradling JR, Barbe F, et al. Effect of CPAP on blood pressure in patients with minimally symptomatic obstructive sleep apnoea: a meta-analysis using individual patient data from four randomised controlled trials. Thorax. 2014 Dec;69(12):1128–1135.
9. Craig S, Pepin JL, Randerath W, et al. Investigation and management of residual sleepiness in CPAP-treated patients with obstructive sleep apnoea: the European view. Eur Respir Rev. 2022 Jun 30;31(164):210230. DOI:10.1183/16000617.0230-2021
10. Gasa M, Tamisier R, Launois SH, et al. Residual sleepiness in sleep apnea patients treated by continuous positive airway pressure. J Sleep Res. 2013 Aug;22(4):389–397.
11. Rosenberg R, Schweitzer PK, Steier J, et al. Residual excessive daytime sleepiness in patients treated for obstructive sleep apnea: guidance for assessment, diagnosis, and management. Postgrad Med. 2021 Sep;133(7):772–783.
12. Weaver TE, Pepin JL, Schwab R, et al. Long-term effects of solriamfetol on quality of life and work productivity in participants with excessive daytime sleepiness associated with narcolepsy or obstructive sleep apnea. J Clin Sleep Med. 2021 Oct 1;17(10):1995–2007. DOI:10.5664/jcsm.9384
13. Patil SP, Ayappa IA, Caples SM, et al. Treatment of adult obstructive sleep apnea with positive airway pressure: an American Academy of Sleep Medicine systematic review, meta-analysis, and GRADE assessment. J Clin Sleep Med. 2019 Feb 15;15(2):301–334. DOI:10.5664/jcsm.7638
14. Cistulli PA, Armitstead J, Pepin JL, et al. Short-term CPAP adherence in obstructive sleep apnea: a big data analysis using real world data. Sleep Med. 2019 Jul;59:114–116. DOI:10.1016/j.sleep.2019.01.004

•• **This study, which provides a big data anlysis using real world data, reports mean adherence in a large database (Center for Medicare and Medicaid Services) and highlights that real world CPAP adherence is acceptable and compares favorably to pharmacotherapy in other chronic diseases.**

15. Patel SR, Bakker JP, Stitt CJ, et al. Age and Sex Disparities in Adherence to CPAP. Chest. 2021 Jan;159(1):382–389. DOI:10.1016/j.chest.2020.07.017

•• **This study using telemonitoring data from a CPAP manufacturer database showed that CPAP adherence rates vary substantially by demographics.**

16. Pandey A, Mereddy S, Combs D, et al. Socioeconomic inequities in adherence to positive airway pressure therapy in population-level analysis. J Clin Med. 2020 Feb 6;9(2):442. DOI:10.3390/jcm9020442
17. McEvoy RD, Antic NA, Heeley E, et al. CPAP for prevention of cardiovascular events in obstructive sleep apnea. N Engl J Med. 2016 Sep 8;375(10):919–931. DOI:10.1056/NEJMoa1606599
18. Sanchez-de-la-Torre M, Sanchez-de-la-Torre A, Bertran S, et al. Effect of obstructive sleep apnoea and its treatment with continuous positive airway pressure on the prevalence of cardiovascular events in patients with acute coronary syndrome (ISAACC study): a randomised controlled trial. Lancet Respir Med. 2020 Apr;8(4):359–367.
19. Gerves-Pinquie C, Bailly S, Goupil F, et al. Positive airway pressure adherence, mortality, and cardiovascular events in patients with sleep apnea. Am J Respir Crit Care Med. 2022 Dec 1;206(11):1393–1404. DOI:10.1164/rccm.202202-0366OC

•• **This study based on data from a French cohort demonstrated a dose-response relationship between PAP adherence and incident major adverse cardiovascular events (MACEs; a composite outcome of mortality, stroke, and cardiac diseases) in OSA.**

20. Pepin JL, Bailly S, Rinder P, et al. CPAP therapy termination rates by OSA Phenotype: a French nationwide database analysis. J Clin Med. 2021 Mar 1;10(5):936. DOI:10.3390/jcm10050936

•• **This study examining real-world data from a comprehensive, unbiased database, highlights the potential for ongoing use of CPAP treatment to reduce all-cause mortality in patients with OSA.**

21. Yeghiazarians Y, Jneid H, Tietjens JR, et al. Obstructive sleep apnea and cardiovascular disease: a scientific statement from the American heart association. Circulation. 2021 Jul 20;144(3):e56–67. DOI:10.1161/CIR.0000000000000988
22. Cowie MR, Linz D, Redline S, et al. Sleep disordered breathing and cardiovascular disease: jACC state-of-the-art review. J Am Coll Cardiol. 2021 Aug 10;78(6):608–624. DOI:10.1016/j.jacc.2021.05.048
23. Bratton DJ, Gaisl T, Wons AM, et al. CPAP vs mandibular advancement devices and blood pressure in patients with obstructive sleep apnea: a systematic review and meta-analysis. JAMA. 2015 Dec 1;314(21):2280–2293. DOI:10.1001/jama.2015.16303
24. Weaver TE, Maislin G, Dinges DF, et al. Relationship between hours of CPAP use and achieving normal levels of sleepiness and daily functioning. Sleep. 2007 Jun;30(6):711–719. DOI:10.1093/sleep/30.6.711

•• **This multi-center, quasi-experimental study showed that a greater percentage of patients will achieve normal functioning with longer nightly CPAP durations, but what constitutes adequate use varies between different outcomes.**

25. Malhotra A, Sterling KL, Cistulli PA, et al. Dose-response relationship between obstructive sleep apnea therapy adherence and healthcare utilization. Ann Am Thorac Soc. 2023 Feb 3. DOI:10.1513/AnnalsATS.202208-738OC
26. May AM, Patel SR, Yamauchi M, et al. Moving toward equitable care for sleep apnea in the United States: positive airway pressure adherence thresholds: an official American Thoracic Society policy statement. Am J Respir Crit Care Med. 2023 Feb 1;207(3):244–254. DOI:10.1164/rccm.202210-1846ST
27. Platt AB, Field SH, Asch DA, et al. Neighborhood of residence is associated with daily adherence to CPAP therapy. Sleep. 2009 Jun;32(6):799–806.
28. Peker Y, Glantz H, Eulenburg C, et al. Effect of positive airway pressure on cardiovascular outcomes in coronary artery disease patients with nonsleepy obstructive sleep apnea. The RICCADSA randomized controlled trial. Am J Respir Crit Care Med. 2016 Sep 1;194(5):613–620. DOI:10.1164/rccm.201601-0088OC
29. Pepin JL, Bailly S, Rinder P, et al. Relationship between CPAP termination and all-cause mortality: a French nationwide database analysis. Chest. 2022 Jun;161(6):1657–1665.
30. Faria A, Allen AH, Fox N, et al. The public health burden of obstructive sleep apnea. Sleep Sci. 2021 Jul;14(3):257–265.
31. Bock JM, Needham KA, Gregory DA, et al. Continuous positive airway pressure adherence and treatment cost in patients with obstructive sleep apnea and cardiovascular disease. Mayo Clin Proc Innov Qual Outcomes. 2022 Apr;6(2):166–175.

32. Wickwire EM, Bailey MD, Somers VK, et al. CPAP adherence is associated with reduced risk for stroke among older adult Medicare beneficiaries with obstructive sleep apnea. J Clin Sleep Med. 2021 Jun 1;17(6):1249–1255. DOI:10.5664/jcsm.9176
33. Sterling KL, Pepin JL, Linde-Zwirble W, et al. Impact of positive airway pressure therapy adherence on outcomes in patients with obstructive sleep apnea and chronic obstructive pulmonary disease. Am J Respir Crit Care Med. 2022 Jul 15;206(2):197–205. DOI:10.1164/rccm.202109-2035OC
34. Genzor S, Prasko J, Mizera J, et al. Risk of severe COVID-19 in non-adherent OSA patients. Patient Prefer Adherence. 2022;16:3069–3079.
35. Bottaz-Bosson G, Midelet A, Mendelson M, et al. Remote monitoring of positive airway pressure data: challenges, pitfalls and strategies to consider for optimal data science applications. Chest. 2022 [2022 Dec 1];0(0). Doi:10.1016/j.chest.2022.11.034
36. Midelet A, Borel JC, Tamisier R, et al. Apnea-hypopnea index supplied by CPAP devices: time for standardization? Sleep Med. 2021 May;81:120–122.
37. Pepin JL, Bailly S, Borel JC, et al. Detecting COVID-19 and other respiratory infections in obstructive sleep apnoea patients through CPAP device telemonitoring. Digit Health. 2021 Jan;7:20552076211002957.
38. Prigent A, Pellen C, Texereau J, et al. CPAP telemonitoring can track Cheyne-Stokes respiration and detect serious cardiac events: the AlertApnee study. Respirology. 2022 Feb;27(2):161–169.
39. Babbin SF, Velicer WF, Aloia MS, et al. Identifying longitudinal patterns for individuals and subgroups: an example with adherence to treatment for obstructive sleep apnea. Multivariate Behav Res. 2015;50(1):91–108. DOI:10.1080/00273171.2014.958211
40. Bottaz-Bosson G, Hamon A, Pepin JL, et al. Continuous positive airway pressure adherence trajectories in sleep apnea: clustering with summed discrete Frechet and dynamic time warping dissimilarities. Stat Med. 2021 Oct 30;40(24):5373–5396. DOI:10.1002/sim.9130
41. Bettega F, Leyrat C, Tamisier R, et al. Application of inverse-probability-of-treatment weighting to estimate the effect of daytime sleepiness in patients with obstructive sleep apnea. Ann Am Thorac Soc. 2022 Sep;19(9):1570–1580.
42. Midelet A, Bailly S, Tamisier R, et al. Hidden Markov model segmentation to demarcate trajectories of residual apnoea-hypopnoea index in CPAP-treated sleep apnoea patients to personalize follow-up and prevent treatment failure. Epma J. 2021 Dec;12(4):535–544. DOI:10.1007/s13167-021-00264-z
•• **This study implemented a data science method (Hidden Markov Models) and showed that this type of approach might constitute the backbone for deployment of patient-centred CPAP management improving the personalized interpretation of telemonitoring data, identifying individuals for targeted therapy and preventing treatment failure or abandonment.**
43. Midelet A, Bailly S, Borel JC, et al. Bayesian structural time series with synthetic controls for evaluating the impact of mask changes in residual apnea-hypopnea index telemonitoring data. IEEE J Biomed Health Inform. 2022 Oct;26(10):5213–5222.
44. Sweetman A, Lack L, Bastien C. Co-morbid Insomnia and Sleep Apnea (COMISA): prevalence, consequences, methodological considerations, and recent randomized controlled trials. Brain Sci. 2019 Dec 12;9(12):371.
45. Jacobsen AR, Eriksen F, Hansen RW, et al. Determinants for adherence to continuous positive airway pressure therapy in obstructive sleep apnea. PLoS ONE. 2017;12(12):e0189614. DOI:10.1371/journal.pone.0189614
46. Kohler M, Smith D, Tippett V, et al. Predictors of long-term compliance with continuous positive airway pressure. Thorax. 2010 Sep;65(9):829–832.
47. Campos-Rodriguez F, Martinez-Alonso M, Sanchez-de-la-Torre M, et al. Long-term adherence to continuous positive airway pressure therapy in non-sleepy sleep apnea patients. Sleep Med. 2016 Jan;17:1–6.
48. Rotenberg BW, Murariu D, Pang KP. Trends in CPAP adherence over twenty years of data collection: a flattened curve. J Otolaryngol Head Neck Surg. 2016 Aug 19;45(1):43.
49. Budhiraja R, Parthasarathy S, Drake CL, et al. Early CPAP use identifies subsequent adherence to CPAP therapy. Sleep. 2007 Mar;30(3):320–324.
50. Gay P, Weaver T, Loube D, et al. Evaluation of positive airway pressure treatment for sleep related breathing disorders in adults. Sleep. 2006 Mar;29(3):381–401.
51. Hui DS, Chan JK, Choy DK, et al. Effects of augmented continuous positive airway pressure education and support on compliance and outcome in a Chinese population. Chest. 2000 May;117(5):1410–1416.
52. Olsen S, Smith S, Oei T, et al. Health belief model predicts adherence to CPAP before experience with CPAP. Eur Respir J. 2008 Sep;32(3):710–717.
53. Popescu G, Latham M, Allgar V, et al. Continuous positive airway pressure for sleep apnoea/hypopnoea syndrome: usefulness of a 2 week trial to identify factors associated with long term use. Thorax. 2001 Sep;56(9):727–733.
54. Weaver TE, Grunstein RR. Adherence to continuous positive airway pressure therapy: the challenge to effective treatment. Proc Am Thorac Soc. 2008 Feb 15;5(2):173–178.
55. Brostrom A, Arestedt KF, Nilsen P, et al. The side-effects to CPAP treatment inventory: the development and initial validation of a new tool for the measurement of side-effects to CPAP treatment. J Sleep Res. 2010 Dec;19(4):603–611.
56. Brostrom A, Stromberg A, Ulander M, et al. Perceived informational needs, side-effects and their consequences on adherence - a comparison between CPAP treated patients with OSAS and healthcare personnel. Patient Educ Couns. 2009 Feb;74(2):228–235.
57. Sawyer AM, Canamucio A, Moriarty H, et al. Do cognitive perceptions influence CPAP use? Patient Educ Couns. 2011 Oct;85(1):85–91.
58. Zozula R, Rosen R. Compliance with continuous positive airway pressure therapy: assessing and improving treatment outcomes. Curr Opin Pulm Med. 2001 Nov;7(6):391–398.
59. Brin YS, Reuveni H, Greenberg S, et al. Determinants affecting initiation of continuous positive airway pressure treatment. Isr Med Assoc J. 2005 Jan;7(1):13–18.
60. Simon-Tuval T, Reuveni H, Greenberg-Dotan S, et al. Low socioeconomic status is a risk factor for CPAP acceptance among adult OSAS patients requiring treatment. Sleep. 2009 Apr;32(4):545–552.
61. Bakker JP, O'Keeffe KM, Neill AM, et al. Ethnic disparities in CPAP adherence in New Zealand: effects of socioeconomic status, health literacy and self-efficacy. Sleep. 2011 Nov 1;34(11):1595–1603. DOI:10.5665/sleep.1404
62. Lewis KE, Seale L, Bartle IE, et al. Early predictors of CPAP use for the treatment of obstructive sleep apnea. Sleep. 2004 Feb 1;27(1):134–138. DOI:10.1093/sleep/27.1.134
63. Billings ME, Auckley D, Benca R, et al. Race and residential socioeconomics as predictors of CPAP adherence. Sleep. 2011 Dec 1;34(12):1653–1658. DOI:10.5665/sleep.1428
64. Campbell A, Neill A, Lory R. Ethnicity and socioeconomic status predict initial continuous positive airway pressure compliance in New Zealand adults with obstructive sleep apnoea. Intern Med J. 2012 Jun;42(6):e95–101.
65. Gulati A, Ali M, Davies M, et al. A prospective observational study to evaluate the effect of social and personality factors on continuous positive airway pressure (CPAP) compliance in obstructive sleep apnoea syndrome. BMC Pulm Med. 2017 Mar 22;17(1):56. DOI:10.1186/s12890-017-0393-7
66. Ye L, Pack AI, Maislin G, et al. Predictors of continuous positive airway pressure use during the first week of treatment. J Sleep Res. 2012 Aug;21(4):419–426.
67. Becker MH, Maiman LA. Sociobehavioral determinants of compliance with health and medical care recommendations. Med care. 1975 Jan;13(1):10–24.

68. Andrulis DP. Access to care is the centerpiece in the elimination of socioeconomic disparities in health. Ann Intern Med. 1998 Sep 1;129(5):412–416.
69. Potosky AL, Breen N, Graubard BI, et al. The association between health care coverage and the use of cancer screening tests. Results from the 1992 national health interview survey. Med Care. 1998 Mar;36(3):257–270.
70. Schillinger D, Grumbach K, Piette J, et al. Association of health literacy with diabetes outcomes. JAMA. 2002 Jul 24-31;288 (4):475–482. DOI:10.1001/jama.288.4.475
71. Williams LK, Joseph CL, Peterson EL, et al. Race-ethnicity, crime, and other factors associated with adherence to inhaled corticosteroids. J Allergy Clin Immunol. 2007 Jan;119(1):168–175.
72. Palm A, Grote L, Theorell-Haglow J, et al. Socioeconomic factors and adherence to CPAP: the population-based course of disease in patients reported to the Swedish CPAP oxygen and ventilator registry study. Chest. 2021 Oct;160(4):1481–1491.
73. Liu D, Armitstead J, Benjafield A, et al. Trajectories of emergent central sleep apnea during CPAP therapy. Chest. 2017 Oct;152 (4):751–760.
74. Bailly S, Daabek N, Jullian-Desayes I, et al. Partial failure of CPAP treatment for sleep apnoea: analysis of the French national sleep database. Respirology. 2020 Jan;25(1):104–111.
75. Pepin JL, Woehrle H, Liu D, et al. Adherence to positive airway therapy after switching from CPAP to ASV: a big data analysis. J Clin Sleep Med. 2018 Jan 15;14(1):57–63. DOI:10.5664/jcsm.6880
76. Benjafield AV, Oldstone LM, Willes LA, et al. Positive airway pressure therapy adherence with mask resupply: a propensity-matched analysis. J Clin Med. 2021 Feb 12;10(4):720. DOI:10.3390/jcm10040720
77. Andrade RGS, Viana FM, Nascimento JA, et al. Nasal vs oronasal CPAP for OSA treatment: a meta-analysis. Chest. 2018 Mar;153 (3):665–674.
78. Rowland S, Aiyappan V, Hennessy C, et al. Comparing the efficacy, mask leak, patient adherence, and patient preference of three different CPAP interfaces to treat moderate-severe obstructive sleep apnea. J Clin Sleep Med. 2018 Jan 15;14(1):101–108. DOI:10.5664/jcsm.6892
79. Ebben MR, Narizhnaya M, Segal AZ, et al. A randomised controlled trial on the effect of mask choice on residual respiratory events with continuous positive airway pressure treatment. Sleep Med. 2014 Jun;15(6):619–624.
80. Manis E, Cheng H, Shelgikar AV. Elevated residual apnea-hypopnea index on continuous positive airway pressure download after transition to full-face mask. Ann Am Thorac Soc. 2021 Mar;18 (3):524–526.
81. Smith I, Lasserson TJ. Pressure modification for improving usage of continuous positive airway pressure machines in adults with obstructive sleep apnoea. Cochrane Database Syst Rev. 2019 Dec 2;12:CD003531. DOI:10.1002/14651858
82. Xu T, Li T, Wei D, et al. Effect of automatic versus fixed continuous positive airway pressure for the treatment of obstructive sleep apnea: an up-to-date meta-analysis. Sleep Breath. 2012 Dec;16 (4):1017–1026.
83. Gentina T, Bailly S, Jounieaux F, et al. Marital quality, partner's engagement and continuous positive airway pressure adherence in obstructive sleep apnea. Sleep Med. 2019 Mar;55:56–61.
84. Pepin JL, Daabek N, Bailly S, et al. Adherence to continuous positive airway pressure hugely improved during COVID-19 lockdown in France. Am J Respir Crit Care Med. 2021 Nov 1;204(9):1103–1106. DOI:10.1164/rccm.202103-0803LE
85. Strausz S, Kiiskinen T, Broberg M, et al. Sleep apnoea is a risk factor for severe COVID-19. BMJ Open Respir Res. 2021 Jan;8(1):e000845.
86. Pepin JL, Sauvaget O, Borel JC, et al. Continuous positive airway pressure-treated patients' behaviours during the COVID-19 crisis. ERJ Open Res. 2020 Oct;6(4):00508–2020.
87. Demirovic S, Lusic Kalcina L, Pavlinac Dodig I, et al. The COVID-19 lockdown and CPAP adherence: the more vulnerable ones less likely to improve adherence? Nat Sci Sleep. 2021;13:1097–1108.
88. Bertelli F, Suehs CM, Mallet JP, et al. Did COVID-19 impact positive airway pressure adherence in 2020? A cross-sectional study of 8477 patients with sleep apnea. Respir Res. 2022 Mar 4;23(1):46. DOI:10.1186/s12931-022-01969-z
89. Brostrom A, Nilsen P, Johansson P, et al. Putative facilitators and barriers for adherence to CPAP treatment in patients with obstructive sleep apnea syndrome: a qualitative content analysis. Sleep Med. 2010 Feb;11(2):126–130.
90. Merle R, Pison C, Logerot S, et al. Peer-driven intervention to help patients resume CPAP therapy following discontinuation: a multicentre, randomised clinical trial with patient involvement. BMJ Open. 2021 Oct 14;11(10):e053996. DOI:10.1136/bmjopen-2021-053996
91. Troxel WM, Robles TF, Hall M, et al. Marital quality and the marital bed: examining the covariation between relationship quality and sleep. Sleep Med Rev. 2007 Oct;11(5):389–404.
92. Beninati W, Harris CD, Herold DL, et al. The effect of snoring and obstructive sleep apnea on the sleep quality of bed partners. Mayo Clin Proc. 1999 Oct;74(10):955–958.
93. Martire LM, Schulz R, Helgeson VS, et al. Review and meta-analysis of couple-oriented interventions for chronic illness. Ann Behav Med. 2010 Dec;40(3):325–342.
94. Baron KG, Smith TW, Czajkowski LA, et al. Relationship quality and CPAP adherence in patients with obstructive sleep apnea. Behav Sleep Med. 2009;7(1):22–36. DOI:10.1080/15402000802577751
95. Luyster FS, Aloia MS, Buysse DJ, et al. A couples-oriented intervention for positive airway pressure therapy adherence: a pilot study of obstructive sleep apnea patients and their partners. Behav Sleep Med. 2019 Sep;17(5):561–572.
96. Rosa D, Amigoni C, Rimoldi E, et al. Obstructive sleep apnea and adherence to Continuous Positive Airway Pressure (CPAP) treatment: let's talk about partners! Healthcare (Basel). 2022 May 19;10 (5). DOI:10.3390/healthcare10050943
97. Baron KG, Gilles A, Sundar KM, et al. Rationale and study protocol for We-PAP: a randomized pilot/feasibility trial of a couples-based intervention to promote PAP adherence and sleep health compared to an educational control. Pilot Feasibility Stud. 2022 Aug 6;8(1):171. DOI:10.1186/s40814-022-01089-x
98. Catalyst N. What is telehealth? [BRIEF ARTICLE]. NEJM Catal. 2018 2018 Feb 1.
99. Singh J, Badr MS, Diebert W, et al. American Academy of Sleep Medicine (AASM) position paper for the use of telemedicine for the diagnosis and treatment of sleep disorders. J Clin Sleep Med. 2015 Oct 15;11(10):1187–1198. DOI:10.5664/jcsm.5098
100. Shamim-Uzzaman QA, Bae CJ, Ehsan Z, et al. The use of telemedicine for the diagnosis and treatment of sleep disorders: an American Academy of Sleep Medicine update. J Clin Sleep Med. 2021 May 1;17(5):1103–1107. DOI:10.5664/jcsm.9194
101. Hwang D, Chang JW, Benjafield AV, et al. Effect of telemedicine education and telemonitoring on continuous positive airway pressure adherence. The Tele-OSA randomized trial. Am J Respir Crit Care Med. 2018 Jan 1;197(1):117–126. DOI:10.1164/rccm.201703-0582OC
102. Hwang D. Monitoring progress and adherence with positive airway pressure therapy for obstructive sleep apnea: the roles of telemedicine and mobile health applications. Sleep Med Clin. 2016 Jun;11 (2):161–171.
103. Pepin JL, Tamisier R, Hwang D, et al. Does remote monitoring change OSA management and CPAP adherence? Respirology. 2017 Nov;22(8):1508–1517.
104. Olsen S, Smith SS, Oei TP, et al. Motivational interviewing (MINT) improves continuous positive airway pressure (CPAP) acceptance and adherence: a randomized controlled trial. J Consult Clin Psychol. 2012 Feb;80(1):151–163.
105. Schwab RJ, Badr SM, Epstein LJ, et al. An official American thoracic society statement: continuous positive airway pressure adherence tracking systems. The optimal monitoring strategies and outcome measures in adults. Am J Respir Crit Care Med. 2013 Sep 1;188 (5):613–620. DOI:10.1164/rccm.201307-1282ST

106. Fox N, Hirsch-Allen AJ, Goodfellow E, et al. The impact of a telemedicine monitoring system on positive airway pressure adherence in patients with obstructive sleep apnea: a randomized controlled trial. Sleep. 2012 Apr 1;35(4):477–481. DOI:10.5665/sleep.1728
107. Frasnelli M, Baty F, Niedermann J, et al. Effect of telemetric monitoring in the first 30 days of continuous positive airway pressure adaptation for obstructive sleep apnoea syndrome - a controlled pilot study. J Telemed Telecare. 2016 Jun;22 (4):209–214.
108. Turino C, de Batlle J, Woehrle H, et al. Management of continuous positive airway pressure treatment compliance using telemonitoring in obstructive sleep apnoea. Eur Respir J. 2017 Feb;49 (2):1601128.
109. Munafo D, Hevener W, Crocker M, et al. A telehealth program for CPAP adherence reduces labor and yields similar adherence and efficacy when compared to standard of care. Sleep Breath. 2016 May;20(2):777–785.
110. Tamisier R, Treptow E, Joyeux-Faure M, et al. Impact of a multimodal telemonitoring intervention on CPAP adherence in symptomatic OSA and low cardiovascular risk: a randomized controlled trial. Chest. 2020 Nov;158(5):2136–2145.
111. Lloyd-Jones DM, Allen NB, Anderson CAM, et al. Life's essential 8: updating and enhancing the American heart association's construct of cardiovascular health: a presidential advisory from the American heart association. Circulation. 2022 Aug 2;146(5):e18–43. DOI:10.1161/CIR.0000000000001078
112. Askland K, Wright L, Wozniak DR, et al. Educational, supportive and behavioural interventions to improve usage of continuous positive airway pressure machines in adults with obstructive sleep apnoea. Cochrane Database Syst Rev. 2020 Apr 7;4CD007736(4). DOI:10.1002/14651858.CD007736.pub3

7.2 Travaux annexes

7.2.1 Résumé des collaborations

- **API Beauty is in the eye of the Clients :** j'ai collaboré à un article de génie logiciel avec une équipe de l'université KTH de Stockholm. Ma contribution a été de rédiger et de vérifier le script d'analyse des données. Cette contribution est un travail d'ouverture dans un domaine dans lequel l'inférence causale sur données observationnelles pourrait être un outil intéressant et à ma connaissance peu exploité. (publié dans la revue Journal of Systems and Software [25])

- **Quality of life of patients with solid malignancies at 3 months after unplanned admission in the intensive care unit: A prospective case-control study. :** j'ai contribué à la rédaction d'un article qui traite des modifications de qualité de vie des patients avec un cancer après un passage en réanimation en réalisant l'analyse statistique. (publié dans la revue Plos one [78])

- **LAM :** l'objectif de ce travail est d'explorer les caractéristiques prédictives des patients avant un diagnostic de leucémie aiguë myéloïde afin de proposer un algorithme prédictif. (présenté au CLARA)

- **Late relapse after hematopoietic stem cell transplantation for acute leukemia: a retrospective study by SFGM-TC. :** l'objectif est d'explorer les caractéristiques et les déterminants des rechutes tardives post greffe chez les patients leucémiques.(Publié dans la revue Transplantation and Cellular Therapy [37]) (présenté dans 2 conférences d'hématologie)

- **Checkpoints inhibitors and allo-SCT in Hodgkin lymphoma : a matter of time - A study by the SFGM-TC :** La greffe de cellules souches allogéniques pour les patients atteints d'un lymphome hodgkinien récidivant/réfractaire après une greffe autologue est désormais une option établie. De nombreux patients reçoivent des inhibiteurs de points de contrôle IPC comme traitement de rechute. Néanmoins, des inquiétudes ont été exprimées quant au risque accru de maladie aiguë du greffon contre l'hôte après l'exposition aux IPC, sans qu'aucun groupe de contrôle ne permette d'élucider les facteurs de risque.(Soumis en vue de publication dans la revue American Society of Hematology)

- **Impact of extended interval dosing of immune checkpoint inhibitors in lung cancer patients during the COVID-19 pandemic. :** durant la pandémie certains centres ont choisi de doubler les doses de certaines immunothérapies afin de limiter le nombres de venues à l'hôpital. L'objectif était d'évaluer si cette nouvelle pratique conduisait à un surplus d'effets indésirables.(Publié dans la revue Respiratory Medicine and Research [83])

- **Safety profile of immune checkpoint inhibitors according to cancer type. Bulletin du Cancer :** Nous avons utilisé l'inférence bayésienne pour évaluer la probabilité d'apparition d'un effet indésirable selon la localisation du cancer. (Publié dans le Bulletin du cancer [23])

- **La cytométrie en image, un outil facile à utiliser pour analyser le contenu en chitine de 12 espèces de levures :** l'objectif est d'évaluer une nouvelle méthode de dosage de la chitine puis d'évaluer les capacités de cette méthode sur la prédiction de résistances aux anti-fongiques et l'identification d'espèces.(Soumis en vue de publication dans le journal of medical mycology)

- **La planification 3D en chirurgie orthognathique pour évaluer la résection osseuse pour l'impaction maxillaire :** le but de cette étude est de mesurer en 3D le rapport entre la résection osseuse maxillaire et l'impaction maxillaire et d'identifier les paramètres modifiant ce rapport. (Soumis dans le journal of cranio-maxillo-facial surgery)

7.2.2 Résumé des communications

Présentations orales

- **Présentation orale à la conférence de l'International Society for Clinical Biostatistics (2021):** Présentation sur une application des GBMs à la problématique de l'effet causal de l'observance à la PPC sur la somnolence diurne.
- **Présentation orale à la conférence FUNcausal (2023):** Présentation sur l'application des IPTWs à la problématique de l'effet causal de l'observance à la PPC.
- **Présentation orale au comité d'investigation clinique du CHU de Grenoble (2022):** Présentation sur l'application des IPTWs à la problématique de l'effet causal de l'observance à la PPC sur la somnolence diurne.
- **Présentation orale à la journée des doctorants du laboratoire HP2 (2022) :** Présentation sur l'application des IPTWs à la problématique de l'effet causal de l'observance à la PPC.
- **Présentation orale journée des doctorants du laboratoire HP2 (2023) :** Présentation de la revue systématique de la littérature sur l'usage des IPTWs multi-niveaux dans la recherche médicale.

Posters

- **Poster à l'EuroSim :** Poster sur l'application la revue systématique de la littérature sur l'usage des IPTWs multi-niveaux dans la recherche médicale.
- **Poster à FUNcausal :** Poster sur l'application la revue systématique de la littérature sur l'usage des IPTWs multi-niveaux dans la recherche médicale.
- **Poster au congrès du Sommeil (2021):** Poster sur l'application des IPTWs à la problématique de l'effet causal de l'observance à la PPC sur la somnolence diurne.

Bibliography

[1] H.-J. Ahn, S.-R. Lee, E.-K. Choi, K.-D. Han, J.-H. Jung, J.-H. Lim, J.-P. Yun, S. Kwon, S. Oh, and G. Y. H. Lip. Association between exercise habits and stroke, heart failure, and mortality in Korean patients with incident atrial fibrillation: A nationwide population-based cohort study. *PLoS Med*, 18(6):e1003659, June 2021.

[2] A. K. Ali, A. G. Hartzema, A. G. Winterstein, R. Segal, X. Lu, and L. Hendeles. Application of multicategory exposure marginal structural models to investigate the association between long-acting beta-agonists and prescribing of oral corticosteroids for asthma exacerbations in the Clinical Practice Research Datalink. *Value Health*, 18(2):260–270, Mar. 2015.

[3] A. J. M. H. Allen, N. Bansback, and N. T. Ayas. The effect of OSA on work disability and work-related injuries. *Chest*, 147(5):1422–1428, may 2015.

[4] D. G. Altman and P. Royston. The cost of dichotomising continuous variables. *BMJ*, 332(7549):1080.1, may 2006.

[5] P. C. Austin. Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis. *Statistics in Medicine*, 35(30):5642–5655, aug 2016.

[6] P. C. Austin. Assessing covariate balance when using the generalized propensity score with quantitative or continuous exposures. *Statistical Methods in Medical Research*, 28(5):1365–1377, feb 2018.

[7] P. C. Austin. Assessing the performance of the generalized propensity score for estimating the effect of quantitative or continuous exposures on binary outcomes. *Statistics in Medicine*, 37(11):1874–1894, mar 2018.

[8] P. C. Austin and E. A. Stuart. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*, 34(28):3661–3679, aug 2015.

[9] A. V. Benjafield, N. T. Ayas, P. R. Eastwood, R. Heinzer, M. S. M. Ip, M. J. Morrell, C. M. Nunez, S. R. Patel, T. Penzel, J.-L. Pépin, P. E. Peppard, S. Sinha, S. Tufik, K. Valentine, and A. Malhotra. Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. *The Lancet Respiratory Medicine*, 7(8):687–698, aug 2019.

[10] K. Benson and A. J. Hartz. A comparison of observational studies and randomized, controlled trials. *New England Journal of Medicine*, 342(25):1878–1886, jun 2000.

[11] R. B. Berry, R. Budhiraja, D. J. Gottlieb, D. Gozal, C. Iber, V. K. Kapur, C. L. Marcus, R. Mehra, S. Parthasarathy, S. F. Quan, S. Redline, K. P. Strohl, S. L. D. Ward, and M. M. Tangredi. Rules for scoring respiratory events in sleep:

Update of the 2007 AASM manual for the scoring of sleep and associated events. *Journal of Clinical Sleep Medicine*, 08(05):597–619, oct 2012.

[12] F. Bettega, C. Leyrat, R. Tamisier, M. Mendelson, Y. Grillet, M. Sapène, M. R. Bonsignore, J. L. Pépin, M. W. Kattan, and S. Bailly. Application of inverse-probability-of-treatment weighting to estimate the effect of daytime sleepiness in patients with obstructive sleep apnea. *Annals of the American Thoracic Society*, 19(9):1570–1580, sep 2022.

[13] S. Bozorgmehri, H. Aboud, G. Chamarthi, I.-C. Liu, O.-B. Tezcan, A. M. Shukla, A. Kazory, R. Rupam, M. S. Segal, A. Bihorac, and R. Mohandas. Association of early initiation of dialysis with all-cause and cardiovascular mortality: A propensity score weighted analysis of the united states renal data system. *Hemodialysis International*, 25(2):188–197, feb 2021.

[14] S. R. Cole and M. A. Hernan. Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, 168(6):656–664, jul 2008.

[15] R. Daniel, S. Cousens, B. D. Stavola, M. G. Kenward, and J. A. C. Sterne. Methods for dealing with time-dependent confounding. *Statistics in Medicine*, 32(9):1584–1618, dec 2012.

[16] C. R. Davies and J. J. Harrington. Impact of obstructive sleep apnea on neurocognitive function and impact of continuous positive air pressure. *Sleep Medicine Clinics*, 11(3):287–298, sep 2016.

[17] M. S. de-la Torre, A. S. de-la Torre, S. Bertran, J. Abad, J. Duran-Cantolla, V. Cabriada, O. Mediano, M. J. Masdeu, M. L. Alonso, J. F. Masa, A. Barceló, M. de la Peña, M. Mayos, R. Coloma, J. M. Montserrat, E. Chiner, S. Perelló, G. Rubinós, O. Mínguez, L. Pascual, A. Cortijo, D. Martínez, A. Aldomà, M. Dalmases, R. D. McEvoy, F. Barbé, L. Abad, A. Muñoz, E. Zamora, I. Vicente, S. Inglés, C. Egea, J. Marcos, A. Fernández, J. Ullate, J. D. Carro, J. L. Rodríguez, M. J. Mendoza, R. Labeaga, D. Diez, B. Muria, C. Amibilia, A. Urrutia, S. Castro, L. Serrano, I. Salinas, R. Diez, A. Martínez, M. Florés, E. Galera, A. Mas, M. Martínez, M. Arbonés, S. Ortega, A. Martín, J. M. Román-Sánchez, M. I. Valiente-Diaz, M. E. Viejo-Ayuso, C. Rodríguez-García, N. Sánchez-Rodríguez, N. Mayoral, F. J. Rubio, Y. Anta-Mejias, S. Romera-Peralta, P. Resano, R. Arroyo-Espilguero, M. Bienvenido-Villalba, L. Vigil, E. Ramírez, M. Piñar, E. Martínez, C. Múñoz, E. Ordax, N. Surname, J. Corral, F. J. G. de Terreros Caro, E. García-Ledesma, R. Gallego, J. L. Cabrero, R. Pereira, P. Giménez, M. Carrera, J. Pierola, C. Villena, M. Campaner, A. M. Fortuna, P. Peñacoba, A. J. M. García, S. G. Castillo, L. Navas, O. Garmendia, M. Suárez, J. Sancho, N. Farre, G. Bonet, A. Bardaji, A. Villares, and M. J. Vázquez. Effect of obstructive sleep apnoea and its treatment with continuous positive airway pressure on the prevalence of cardiovascular events in patients with acute coronary syndrome (ISAACC study): a randomised controlled trial. *The Lancet Respiratory Medicine*, 8(4):359–367, apr 2020.

[18] Y. Dodge, D. Cox, and D. Commenges. *The Oxford Dictionary of Statistical Terms*. Oxford University Press, USA, 2006.

[19] P. Escourrou, N. Meslier, B. Raffestin, R. Clavel, J. Gomes, E. Hazouard, J. Paquereau, I. Simon, and E. O. Frija. Quelle approche clinique et quelle procédure diagnostique pour le SAHOS ? *Revue des Maladies Respiratoires*, 27:S115–S123, oct 2010.

[20] A. Faria, A. H. Allen, N. Fox, N. Ayas, and I. Laher. The public health burden of obstructive sleep apnea. *Sleep Science*, 14(3):257, 2021.

[21] J. Fleetham, N. Ayas, D. Bradley, K. Ferguson, M. Fitzpatrick, C. George, P. Hanly, F. Hill, J. Kimoff, M. Kryger, D. Morrison, F. Series, and W. Tsai. Directives de la société canadienne de thoracologie: Diagnostic et traitement des troubles respiratoires du sommeil de l'adulte. *Canadian Respiratory Journal*, 14(1):31–36, 2007.

[22] E. Granger, T. Watkins, J. C. Sergeant, and M. Lunt. A review of the use of propensity score diagnostics in papers published in high-ranking medical journals. *BMC Medical Research Methodology*, 20(1), may 2020.

[23] C. Guérin, M. Laramas, F. Bettega, A. Bocquet, E. Berton, M. Lugosi, L. Bouillet, and A.-C. Toffart. Safety profile of immune checkpoint inhibitors according to cancer type. *Bulletin du Cancer*, 110(7-8):825–835, jul 2023.

[24] V. S. Harder, E. A. Stuart, and J. C. Anthony. Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods*, 15(3):234–249, 2010.

[25] N. Harrand, A. Benelallam, C. Soto-Valero, F. Bettega, O. Barais, and B. Baudry. API beauty is in the eye of the clients: 2.2 million maven dependencies reveal the spectrum of client–API usages. *Journal of Systems and Software*, 184:111134, feb 2022.

[26] T. Hastie, R. Tibshirani, and J. Friedman. *Boosting and Additive Trees*, pages 337–387. Springer Series in Statistics. Springer New York, New York, NY, 2009.

[27] M. A. Hernan. A definition of causal effect for epidemiological research. *Journal of Epidemiology and Community Health*, 58(4):265–271, apr 2004.

[28] M. A. Hernán. Methods of public health research — strengthening causal inference from observational data. *New England Journal of Medicine*, 385(15):1345–1348, oct 2021.

[29] M. A. Hernán and J. M. Robins. Using big data to emulate a target trial when a randomized trial is not available: Table 1. *American Journal of Epidemiology*, 183(8):758–764, mar 2016.

[30] M. A. Hernan and J. M. Robins. *Causal Inference*. Taylor & Francis Group, 2019.

[31] K. Hirano and G. W. Imbens. The propensity score with continuous treatments, jul 2004.

[32] D. S. Hui, J. K. Chan, D. K. Choy, F. W. Ko, T. S. Li, R. C. Leung, and C. K. Lai. Effects of augmented continuous positive airway pressure education and support on compliance and outcome in a chinese population. *Chest*, 117(5):1410–1416, may 2000.

[33] I. H. Iftikhar, C. M. Hoyos, C. L. Phillips, and U. J. Magalang. Meta-analyses of the association of sleep apnea with insulin resistance, and the effects of CPAP on HOMA-IR, adiponectin, and visceral adipose fat. *Journal of Clinical Sleep Medicine*, 11(04):475–485, apr 2015.

[34] I. H. Iftikhar, C. W. Valentine, L. R. Bittencourt, D. L. Cohen, A. C. Fedson, T. Gíslason, T. Penzel, C. L. Phillips, L. Yu-sheng, A. I. Pack, and U. J. Magalang. Effects of continuous positive airway pressure on blood pressure in patients with resistant hypertension and obstructive sleep apnea. *Journal of Hypertension*, 32(12):2341–2350, dec 2014.

[35] S. Javaheri, F. Barbe, F. Campos-Rodriguez, J. A. Dempsey, R. Khayat, S. Javaheri, A. Malhotra, M. A. Martinez-Garcia, R. Mehra, A. I. Pack, V. Y. Polotsky, S. Redline, and V. K. Somers. Sleep apnea. *Journal of the American College of Cardiology*, 69(7):841–858, feb 2017.

[36] J. D. Y. Kang and J. L. Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4), nov 2007.

[37] E. Kaphan, F. Bettega, E. Forcade, H. Labussière-Wallet, N. Fegueux, M. Robin, R. P. D. Latour, A. Huynh, L. Lapierre, A. Berceanu, A. Marcais, P.-E. Debureaux, N. Vanlangendonck, C.-E. Bulabois, L. Magro, A. Daniel, J. Galtier, B. Lioure, P. Chevallier, C. Antier, M. Loschi, G. Guillerm, J.-B. Mear, S. Chantepie, J. Cornillon, G. Rey, X. Poire, A. Bazarbachi, M.-T. Rubio, N. Contentin, C. Orvain, R. Dulery, J. O. Bay, C. Croizier, Y. Beguin, A. Charbonnier, C. Skrzypczak, D. Desmier, A. Villate, M. Carré, and A. Thiebaut-Bertrand. Late relapse after hematopoietic stem cell transplantation for acute leukemia: a retrospective study by SFGM-TC. *Transplantation and Cellular Therapy*, 29(6):362.e1–362.e12, jun 2023.

[38] T. Kennedy-Martin, S. Curtis, D. Faries, S. Robinson, and J. Johnston. A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. *Trials*, 16(1), nov 2015.

[39] M. Kohler, A.-C. Stoewhas, L. Ayers, O. Senn, K. E. Bloch, E. W. Russi, and J. R. Stradling. Effects of continuous positive airway pressure therapy withdrawal in patients with obstructive sleep apnea. *American Journal of Respiratory and Critical Care Medicine*, 184(10):1192–1199, nov 2011.

[40] G. Labarca, J. Dreyse, L. Drake, J. Jorquera, and F. Barbe. Efficacy of continuous positive airway pressure (CPAP) in the prevention of cardiovascular events in patients with obstructive sleep apnea: Systematic review and meta-analysis. *Sleep Medicine Reviews*, 52:101312, aug 2020.

[41] B. K. Lee, J. Lessler, and E. A. Stuart. Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29(3):337–346, dec 2009.

[42] C. H. K. Lee, L. C. Leow, P. R. Song, H. Li, and T. H. Ong. Acceptance and adherence to continuous positive airway pressure therapy in patients with obstructive sleep apnea (OSA) in a southeast asian privately funded healthcare system. *Sleep Science*, 10(2):57–63, 2017.

[43] G. Lefebvre, J. A. C. Delaney, and R. W. Platt. Impact of mis-specification of the treatment model on estimates from a marginal structural model. *Statist. Med.*, 27(18):3629–3642, Aug. 2008.

[44] L. E. Levesque, J. A. Hanley, A. Kezouh, and S. Suissa. Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes. *BMJ*, 340(mar12 1):b5087–b5087, mar 2010.

[45] P. Lévy, M. Kohler, W. T. McNicholas, F. Barbé, R. D. McEvoy, V. K. Somers, L. Lavie, and J.-L. Pépin. Obstructive sleep apnoea syndrome. *Nature Reviews Disease Primers*, 1(1), jun 2015.

[46] C. Leyrat, J. R. Carpenter, S. Bailly, and E. J. Willamson. A review and evaluation of standard methods to handle missing data on time-varying confounders in marginal structural models, 2019.

[47] J. M. Marin, S. J. Carrizo, E. Vicente, and A. G. Agusti. Long-term cardiovascular outcomes in men with obstructive sleep apnoea-hypopnoea with or without treatment with continuous positive airway pressure: an observational study. *The Lancet*, 365(9464):1046–1053, mar 2005.

[48] C. Maringe, S. B. Majano, A. Exarchakou, M. Smith, B. Rachet, A. Belot, and C. Leyrat. Reflection on modern methods: trial emulation in the presence of immortal-time bias. assessing the benefit of major surgery for elderly lung cancer patients using observational data. *International Journal of Epidemiology*, 49(5):1719–1729, may 2020.

[49] D. F. McCaffrey, G. Ridgeway, and A. R. Morral. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychol Methods*, 9(4):403–425, Dec. 2004.

[50] R. D. McEvoy, N. A. Antic, E. Heeley, Y. Luo, Q. Ou, X. Zhang, O. Mediano, R. Chen, L. F. Drager, Z. Liu, G. Chen, B. Du, N. McArdle, S. Mukherjee, M. Tripathi, L. Billot, Q. Li, G. Lorenzi-Filho, F. Barbe, S. Redline, J. Wang, H. Arima, B. Neal, D. P. White, R. R. Grunstein, N. Zhong, and C. S. Anderson. CPAP for prevention of cardiovascular events in obstructive sleep apnea. *New England Journal of Medicine*, 375(10):919–931, sep 2016.

[51] R. D. McEvoy, M. S. de-la Torre, Y. Peker, C. S. Anderson, S. Redline, and F. Barbe. Randomized clinical trials of cardiovascular disease in obstructive sleep apnea: understanding and overcoming bias. *Sleep*, 44(4), mar 2021.

[52] W. T. McNicholas. Diagnosis of obstructive sleep apnea in adults. *Proceedings of the American Thoracic Society*, 5(2):154–160, feb 2008.

[53] M. Mendelson, J. Duval, F. Bettega, R. Tamisier, S. Baillieul, S. Bailly, and J.-L. Pépin. The individual and societal prices of non-adherence to continuous positive airway pressure, contributors, and strategies for improvement. *Expert Review of Respiratory Medicine*, 17(4):305–317, apr 2023.

[54] S. B. Montesi, B. A. Edwards, A. Malhotra, and J. P. Bakker. The effect of continuous positive airway pressure treatment on blood pressure: A systematic review and meta-analysis of randomized controlled trials. *Journal of Clinical Sleep Medicine*, 08(05):587–596, oct 2012.

[55] T. P. Morris, I. R. White, and M. J. Crowther. Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11):2074–2102, jan 2019.

[56] S. Olsen, S. Smith, T. Oei, and J. Douglas. Health belief model predicts adherence to CPAP before experience with CPAP. *European Respiratory Journal*, 32(3):710–717, sep 2008.

[57] A. I. Pack, U. J. Magalang, B. Singh, S. T. Kuna, B. T. Keenan, and G. Maislin. Randomized clinical trials of cardiovascular disease in obstructive sleep apnea: understanding and overcoming bias. *Sleep*, 44(2), nov 2020.

[58] A. I. Pack, U. J. Magalang, B. Singh, S. T. Kuna, B. T. Keenan, and G. Maislin. To RCT or not to RCT? depends on the question. a response to McEvoy et al. *Sleep*, 44(4), mar 2021.

[59] J. Pearl. An introduction to causal inference. *The International Journal of Biostatistics*, 6(2), jan 2010.

[60] Y. Peker, H. Glantz, C. Eulenburg, K. Wegscheider, J. Herlitz, and E. Thunström. Effect of positive airway pressure on cardiovascular outcomes in coronary artery disease patients with nonsleepy obstructive sleep apnea. the RICCADSA randomized controlled trial. *American Journal of Respiratory and Critical Care Medicine*, 194(5):613–620, sep 2016.

[61] J.-L. Pépin, S. Bailly, P. Rinder, D. Adler, A. V. Benjafield, F. Lavergne, A. Josseran, P. Sinel-Boucher, R. Tamisier, P. A. Cistulli, A. Malhotra, and P. Hornus. Relationship between CPAP termination and all-cause mortality. *Chest*, 161(6):1657–1665, jun 2022.

[62] J.-L. Pépin, S. Bailly, P. Rinder, D. Adler, D. Szeftel, A. Malhotra, P. Cistulli, A. Benjafield, F. Lavergne, A. Josseran, R. Tamisier, and P. H. and. CPAP therapy termination rates by OSA phenotype: A french nationwide database analysis. *Journal of Clinical Medicine*, 10(5):936, mar 2021.

[63] G. Popescu. Continuous positive airway pressure for sleep apnoea/hypopnoea syndrome: usefulness of a 2 week trial to identify factors associated with long term use. *Thorax*, 56(9):727–733, sep 2001.

[64] G. Ridgeway. The State of Boosting. 1999.

[65] G. Ridgeway, D. F. McCaffrey, A. R. Morral, M. Cefalu, L. F. Burgette, J. D. Pane, and B. A. Griffin. *Toolkit for Weighting and Analysis of Nonequivalent Groups: A Tutorial for the R TWANG Package*. RAND Corporation, 2022.

[66] J. Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512, 1986.

[67] J. M. Robins, M. Á. Hernán, and B. Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, sep 2000.

[68] P. R. ROSENBAUM and D. B. RUBIN. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[69] B. W. Rotenberg, D. Murariu, and K. P. Pang. Trends in CPAP adherence over twenty years of data collection: a flattened curve. *Journal of Otolaryngology - Neck Surgery*, 45(1), aug 2016.

[70] D. B. Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469):322–331, mar 2005.

[71] S. Ryan, C. Arnaud, S. F. Fitzpatrick, J. Gaucher, R. Tamisier, and J.-L. Pépin. Adipose tissue as a key player in obstructive sleep apnoea. *European Respiratory Review*, 28(152):190006, jun 2019.

[72] A. Sabil, M. Le Vaillant, C. Stitt, F. Goupil, T. Pigeanne, L. Leclair-Visonneau, P. Masson, A. Bizieux-Thaminy, M.-P. Humeau, N. Meslier, and F. Gagnadoux. A CPAP data-based algorithm for automatic early prediction of therapy adherence. *Sleep Breath*, 25(2):957–962, June 2021.

[73] S. Schiza, P. Lévy, M. A. Martinez-Garcia, J.-L. Pepin, A. Simonds, and W. Randerath. The search for realistic evidence on the outcomes of obstructive sleep apnoea. *European Respiratory Journal*, 58(4):2101963, oct 2021.

[74] F. Shi, C. Wang, Y. Kong, L. Yang, J. Li, G. Zhu, J. Guo, Q. Zheng, B. Zhang, and S. Wang. Assessing the Survival Benefit of Surgery and Various Types of Radiation Therapy for Treatment of Hepatocellular Carcinoma: Evidence from the Surveillance, Epidemiology, and End Results Registries. *Journal of Hepatocellular Carcinoma*, 7:201–218, 2020.

[75] C. J. Stepnowsky and J. E. Dimsdale. Dose-response relationship between CPAP compliance and measures of sleep apnea severity. *Sleep Med*, 3(4):329–334, July 2002.

[76] E. A. Stuart. Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1), feb 2010.

[77] E. A. Stuart, B. K. Lee, and F. P. Leacy. Prognostic score–based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. *Journal of Clinical Epidemiology*, 66(8):S84–S90.e1, aug 2013.

[78] A.-C. Toffart, W. M'Sallaoui, S. Jerusalem, A. Godon, F. Bettega, G. Roth, J. Pavillet, E. Girard, L. M. Galerneau, J. Piot, C. Schwebel, and J. F. Payen. Quality of life of patients with solid malignancies at 3 months after unplanned admission in the intensive care unit: A prospective case-control study. *PLOS ONE*, 18(1):e0280027, jan 2023.

[79] W. Trzepizur, M. Blanchard, T. Ganem, F. Balusson, M. Feuilloy, J.-M. Girault, N. Meslier, E. Oger, A. Paris, T. Pigeanne, J.-L. Racineux, A. Sabil, C. Gervès-Pinquié, and F. Gagnadoux. Sleep apnea–specific hypoxic burden, symptom subtypes, and risk of cardiovascular events and all-cause mortality. *American Journal of Respiratory and Critical Care Medicine*, 205(1):108–117, jan 2022.

[80] A. A. Tsiatis, M. Davidian, M. Zhang, and X. Lu. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Statistics in Medicine*, 27(23):4658–4677, oct 2008.

[81] M. Tsuyumu, T. Tsurumoto, J. Iimura, T. Nakajima, and H. Kojima. Ten-year adherence to continuous positive airway pressure treatment in patients with moderate-to-severe obstructive sleep apnea. *Sleep and Breathing*, 24(4):1565–1571, feb 2020.

[82] J. P. Vandenbroucke, A. Broadbent, and N. Pearce. Causality and causal inference in epidemiology: the need for a pluralistic approach. *International Journal of Epidemiology*, 45(6):1776–1786, jan 2016.

[83] M. Veron, T. Pierret, M. Pérol, F. Bettega, J. Benet, N. Denis, D. Moro-Sibilot, A. Swalduz, and A.-C. Toffart. Impact of extended interval dosing of immune checkpoint inhibitors in lung cancer patients during the COVID-19 pandemic. *Respiratory Medicine and Research*, 83:101004, jun 2023.

[84] Z.-X. Wang, L.-P. Yang, H.-X. Wu, D.-D. Yang, P.-R. Ding, D. Xie, G. Chen, Y.-H. Li, F. Wang, and R.-H. Xu. Appraisal of Prognostic Interaction between Sidedness and Mucinous Histology in Colon Cancer: A Population-Based Study Using Inverse Probability Propensity Score Weighting. *Journal of Cancer*, 10(2):388–396, 2019.

[85] J. H. Ware and M. B. Hamel. Pragmatic trials — guides to better patient care? *New England Journal of Medicine*, 364(18):1685–1687, may 2011.

[86] E. M. Wickwire, S. E. Tom, A. Vadlamani, M. Diaz-Abad, L. M. Cooper, A. M. Johnson, S. M. Scharf, and J. S. Albrecht. Older adult US medicare beneficiaries with untreated obstructive sleep apnea are heavier users of health care than matched control patients. *Journal of Clinical Sleep Medicine*, 16(1):81–89, jan 2020.

[87] E. Williamson, R. Morley, A. Lucas, and J. Carpenter. Propensity scores: From naïve enthusiasm to intuitive understanding. *Statistical Methods in Medical Research*, 21(3):273–293, jan 2011.

[88] K. Yoshida, D. H. Solomon, S. Haneuse, S. C. Kim, E. Patorno, S. K. Tedeschi, H. Lyu, J. M. Franklin, T. Stürmer, S. Hernández-Díaz, and R. J. Glynn. Multinomial extension of propensity score trimming methods: A simulation study. *American Journal of Epidemiology*, 188(3):609–616, dec 2018.

[89] A. Zinchuk and H. K. Yaggi. Phenotypic subtypes of OSA. *Chest*, 157(2):403–420, feb 2020.