Développement d'une nouvelle mesure d'équilibre pour l'aide à la sélection des variables dans un modèle de score de propension

Emmanuel Caruana

Université Sorbonne
Paris Cité

Thèse de doctorat de l'Université Sorbonne Paris Cité

Préparée à l'Université Paris Diderot

ECOLE DOCTORALE PIERRE LOUIS DE SANTE PUBLIQUE A PARIS

EPIDEMIOLOGIE ET SCIENCES DE L'INFORMATION BIOMEDICALE

Spécialité: Biostatistique et/ou Biomathématiques

Service de Biostatistique et Information Médicale INSERM 1153, équipe ECSTRA

Développement d'une nouvelle mesure d'équilibre pour l'aide à la sélection des variables à inclure dans un modèle de score de propension

Par Emmanuel Caruana

Thèse de doctorat de Santé Publique

Dirigée par le Pr Romain Pirracchio

Présentée et soutenue publiquement à Paris le 1 Mars 2017

Président de jury : Gilles Potel, professeur des universités, Nantes
Rapporteur: Yohann Foucher, maître de conférence universitaire, Nantes
Rapporteur: Denis Frasca, maître de conférence universitaire, Poitiers
Examinateur: Enrique Casalino, professeur des universités, Paris Diderot
Examinateur: Didier Journois, professeur des universités, Paris Descartes
Co-directeur de thèse: Sylvie Chevret, professeur des universités, Paris Diderot

Cette thèse s'appuie sur les articles suivants:

- **Caruana E,** Chevret S, Resche-Rigon M, et al. A new weighted balance measure helped to select the variables to be included in a propensity score model. J Clin Epidemiol 2015;68:1415–1422.e2.

- Caruana E, Chevret S, Pirracchio R. Effect of cricoid pressure on laryngeal view during prehospital tracheal intubation: a propensity-based analysis. *Emerg Med J* 2016:emermed-2016-205715 *(en attente d'impression)*

Titre: Développement d'une nouvelle mesure d'équilibre pour l'aide à la sélection des variables à inclure dans un modèle de score de propension

Résumé: Le score de propension s'est progressivement imposé comme l'une des méthodes de référence dans l'analyse des données observationnelles afin de prendre en compte le biais potentiel lié à l'existence de facteurs de confusion dans l'estimation de l'effet du traitement sur le critère de jugement. Parmi les recommandations de bonnes pratiques d'utilisation, le processus de sélection des variables à inclure dans le score final utilisé est essentiel, ainsi que l'évaluation de l'équilibre obtenu sur les covariables après appariement ou pondération sur ce score. Dans l'objectif de prioriser l'inclusion et l'équilibre des variables ayant une relation avec le critère de jugement une nouvelle mesure d'équilibre est proposée dans ce travail de thèse. Une première partie de ce travail a eu pour objectif de développer une mesure globale pondérée permettant d'évaluer l'équilibre global des covariables obtenu après appariement et ainsi d'aider à la sélection d'un modèle de propension le plus parcimonieux possible, en éliminant notamment les variables instrumentales. En effet ces variables ne doivent pas être introduites dans le modèle de score de propension au risque de majorer le biais final d'estimation. Lors des étapes d'évaluation de l'équilibre final obtenu, les différentes mesures d'équilibres disponibles ne prennent le plus souvent pas en compte cette information et concluent souvent à l'intérêt d'inclure une telle variable afin de réduire au maximum le déséquilibre entre les groupes. L'évaluation des performances de cette mesure a dans un premier temps fait appel à des simulations de type Monte Carlo. Dans une seconde partie, une mise en application sur des données réelles issues de la médecine d'urgence a permis de préciser la pratique d'utilisation d'une telle mesure.

Mots clefs: inférence causale, score de propension, appariement sur score de propension, équilibre, différences standardisées, variable instrumentale.

Title: Development of a new weighted balance measure to help to select the variables to be included in a propensity score model.

Abstract: Propensity score (PS) methods have become increasingly used to analyze observational data and take into account confusion bias in final estimate of treatment effects. The goal of the PS is to balance the distribution of potential confounders across treatment groups. The performance of the PS strongly relies on variable selection in PS construction and balance assessment in PS analysis. Specifically, the choice of the variables to be included in the PS model is of paramount importance. In order to priorize inclusion and balance of variables related to the outcome, a new balance measure was proposed in this thesis.

First, a new weighted balance measure was studied to help in construction of PS model and to obtain the most parsimonious model, by excluding instrumental variables known to be related with increasing bias in final treatment estimate. Several balances measures are proposed to assess final balance, but none of them help researchers to not include instrumental variables. We propose a new weighted balance measure that takes into account, for each covariate, its strength of association with the outcome. This measure was evaluated using a simulation study to assess whether minimization of the measure coincided with minimally biased estimates. Secondly, we propose to apply this measure to a real data set from an observational cohort study.

Keywords: propensity score, propensity score matching, balance, standardized mean difference, instrumental variable, causal inference.

Remerciements

Aux professeurs Romain Pirracchio et Sylvie Chevret : votre confiance, vos enseignements et votre patience m'ont plus que guidé tout au long de ce parcours. Partager avec vous tout ce chemin représente beaucoup pour moi et constitue ma plus belle réussite professionnelle.

Aux docteurs Yohann Foucher, Denis Frasca : je tiens à vous remercier sincèrement d'avoir accepté d'être les rapporteurs de ce travail.

Aux professeurs Enrique Casalino et Didier Journois : je vous remercie de m'avoir fait l'honneur d'accepter d'être les examinateurs de ce travail en intégrant mon jury.

Au professeur Gilles Potel : bien que notre rencontre soit récente, votre accueil, votre disponibilité ainsi que nos échanges m'ont profondément inspiré. Je vous remercie d'avoir accepté de juger ce travail.

A mes parents, mes grands-parents, ma soeur et mes frères: merci pour votre soutien indéfectible.

A Maniso, mon épouse, et Nola, ma fille: mes plus belles réussites, je vous dédie ce travail.

Laboratoire d'accueil

Service de Biostatistique et Information Médicale

INSERM 1153, équipe ECSTRA

Epidémiologie **C**linique, **St**atistique, pour la **R**echerche en S**a**nté

Université Paris 7 Diderot

Hôpital Saint-Louis, Assistance Publique - Hôpitaux de Paris

1 rue Claude Vellefaux

Paris 75010

France

TABLE DES MATIERES

Table des illustrations

1 INTRODUCTION et RATIONNEL

La recherche médicale s'appuie de plus en plus souvent sur des études observationnelles. En effet, lorsqu'on réalise une requête sur Pubmed en utilisant le mot clé "Observational Study", 7,002 publications en 2013 sont sélectionnées contre 13,050 en 2015. Bien que largement majoritaires, les publications d'études contrôlées randomisées ("randomized controlled trial") diminuent en nombre de 6% sur la même période (de 32 423 à 30 439).

En situation observationnelle, l'investigateur n'attribue pas l'intervention étudiée aux sujets participants mais celle-ci est déterminée par de multiples facteurs tels que les recommandations d'experts, les préférences individuelles, les modèles de pratiques, ou encore des choix politiques [1]. Pour évaluer l'efficacité d'un traitement médical dans cette situation, les méthodes statistiques sont basées sur une comparaison entre des individus "exposés" au traitement et d'autres non exposés afin de déterminer l'effet du traitement sur un critère de jugement au sein de la population d'étude. Mais cette simple comparaison ne permet pas d'accéder directement à l'effet causal de l'exposition. L'estimation d'un effet causal nécessite de plus la définition et l'acceptation d'un certain nombre d'hypothèses concernant la structure des données.

L'effet causal d'un traitement est défini par une différence entre deux évolutions : la différence d'évolution observée chez un groupe de sujets exposés à un régime de traitement et celle potentiellement observée si ce même groupe de sujets avait été exposé à l'autre régime de traitement, qu'il n'a, en pratique, pas reçu. En réalité, une seule évolution peut-être observée pour un groupe de sujets (*évolution dite "factuelle"*). C'est pourquoi certains auteurs, dont Judea Pearl, ont proposé d'estimer l'effet causal d'une exposition ou d'un traitement en introduisant la notion théorique de "contrefactuel" (c'est-à-dire d'*évolution potentielle*) [2],[3],[4]. Supposons que N individus soient inclus dans une expérience qui attribue à chaque individu i l'exposition au traitement $A \in \{0,1\}$ et que l'on observe le critère de jugement

binaire Y ∈ {0,1}. Y_a est le devenir tel qu'il aurait été observé sous le régime de traitement A=a. Pour les sujets traités, Y_1 est ainsi observé alors que Y_0 ne l'est pas, et à l'inverse pour les sujets non traités. Lorsque le devenir n'est pas observé, on parle donc de devenir *"contrefactuel"*.

Cela permet de définir l'effet causal ψ de A sur Y telle que:

$$\psi = E(Y_1 - Y_0) \qquad (1)$$

Ce paramètre est dénommé l'effet moyen du traitement ou "*Average Treatment Effect"* (ATE) [5]. Autrement dit, l'effet causal ψ que nous souhaiterions estimer est la différence de devenirs potentiels ou *contrefactuels* qui pourraient être observés sous les différents régimes de traitement pour l'ensemble des individus de la population [6]. Cette estimation est rendue complexe car, dans la réalité, le devenir contrefactuel n'est, par définition, pas observé. Aussi est-il nécessaire de disposer de méthodes statistiques ainsi que d'hypothèses spécifiques permettant d'identifier l'effet causal à partir des données observées.

1.1 Randomisation

Parmi les méthodes proposées afin d'estimer l'effet causal à partir des données observées, la randomisation est celle de référence [7]. Idéalement, les conditions suivantes doivent être réunies : tout patient randomisé dans un groupe de traitement doit être analysé en intention de traiter, la procédure d'allocation du traitement doit se faire en aveugle, et une fois le régime de traitement alloué, l'adhésion ne doit pas changer tout au long de l'étude [8]. De plus, l'échantillon de patients randomisés doit être de taille suffisante pour bénéficier des propriétés asymptotiques du tirage au sort. Ainsi, si la taille de l'échantillon tend vers l'infini, les caractéristiques des deux groupes de sujets tirés au sort seront en tout point comparables, et la

probabilité de survenue de l'événement Y dans le groupe des sujets traités est égale à celle qui aurait pu être observée dans le groupe des sujets non traités s'ils avaient été exposés au même régime de traitement. Une randomisation idéale garantie donc l'*échangeabilité* des groupes en assurant leur comparabilité initiale sur le vecteur V des facteurs de confusion connus et inconnus. L'indépendance entre Y et A (conditionnellement au résultat de la randomisation) permet ainsi d'estimer l'effet causal de A sur Y en comparant directement Y chez les sujets A=1 et A=0 [9],[10],[11],[12],[13].

Ainsi, d'après (1) et en acceptant l'hypothèse d'indépendance entre A et Y ($Y \perp A$):

$$\psi = E(Y_1) - E(Y_0) = E(Y_1|A=1) - E(Y_0|A=0) = E(Y|A=1) - E(Y|A=0) \qquad (2)$$

L'effet marginal est ainsi égal à l'effet observé, ce qui permet de considérer le lien associatif mesuré comme un lien de causalité [6].

En pratique, les conditions d'une randomisation sont difficiles à réunir, ne serait-ce parce que la taille de l'échantillon d'étude ne tend jamais vers l'infini. L'hypothèse d'indépendance stricte entre A et Y est donc difficile à garantir. Une condition moins restrictive consiste à considérer qu'il existe un sous-ensemble W de V, de covariables observables conditionnellement pour lesquelles l'indépendance entre Y et A peut être vérifiée. W étant observable, il est possible de vérifier si la randomisation a permis d'équilibrer sa distribution entre les deux groupes [14]. Si un déséquilibre important sur W est constaté, la similarité des deux groupes du point de vue des facteurs pronostiques est discutable, et l'estimation de l'effet de A sur Y potentiellement biaisée. Il en est de même en l'absence de randomisation.

1.2 Absence de Randomisation

En situation observationnelle, l'existence de différences systématiques de caractéristiques entre les groupes de traitements est responsable d'un biais dans l'estimation de l'effet causal du traitement sur le devenir [15]. De nombreuses méthodes statistiques ont été développées pour tenter d'estimer de manière non biaisée un effet causal en l'absence de randomisation. Ces méthodes reposent toutes sur un certain nombre d'hypothèses concernant la structure des données et ont en général pour but d'équilibrer les caractéristiques (W) entre les groupes de traitement afin de créer les conditions d'une indépendance conditionnelle entre Y et A : $Y_a \perp A|W$. Pour chaque niveau de W, l'échangeabilité conditionnelle peut alors s'écrire:

$$P[Y_a = 1|A = 1, W = w] = P[Y_a = 1|A = 0, W = w] \qquad (3)$$

L'objectif des méthodes d'inférence causale que nous détaillerons dans ce document est globalement de créer les conditions d'indépendance entre Y et A conditionnellement à une variable g(W) de dimension réduite pour permettre d'estimer de manière non biaisée l'effet de A sur Y [16],[17],[18]. Le score de propension est défini comme la probabilité individuelle de recevoir un traitement donné A conditionnellement à un vecteur de covariables observées W :

$$g(W) = P(A = 1|W) \qquad (4)$$

Rosenbaum et Rubin ont ainsi démontré que pour des classes de patients avec des valeurs similaires de $g(W)$, les distributions des facteurs confondants W entre les groupes de traitement était identiques [19]. Ainsi la situation d'indépendance de Y à l'accès au traitement A conditionnellement aux variables W, est retrouvée conditionnellement au score de propension $g(W)$.

Pour que cela soit vérifié, il suffit de prouver que

$$P(A = 1|W) = P(A = 1|g) \qquad (5)$$

En effet, l'indépendance entre A et W conditionnellement à $g(W)$ peut s'écrire

$$P(W, A|g) = P(W|g).P(A|g) \qquad (6)$$

Plus généralement,

$$P(W, A|g) = P(W|g).P(A|W, g) \quad (7)$$

En sachant que g est une fonction de W l'équation (7) peut s'écrire

$$P(W, A|g) = P(W|g).P(A|W) \qquad (8)$$

Par définition, $P(A = 1|W) = g(W)$.

De plus, $P(A = 1|g) = E(A|g) = E\{E(A|W)|g\} = E(g|g) = g = P(A = 1|W)$

L'équation (5) ainsi vérifiée permet d'admettre l'indépendance conditionnelle entre A et W conditionnellement au score de propension $g(W)$. L'interprétation causale de l'estimation à partir des données observées repose ensuite sur un certain nombre d'hypothèses concernant la structure des données, certaines vérifiables, d'autres non.

1.2.1 Identifiabilité

En inférence causale, il s'agit avant tout d'envisager des hypothèses permettant de considérer si le paramètre causal d'intérêt est identifiable à partir des données observées. Ces conditions d'identifiabilité regroupent *l'échangeabilité conditionnelle*, la *positivité*, et la *consistance* [11],[20].

L'échangeabilité conditionnelle décrite ci-dessus permet l'utilisation des sujets non exposés à l'intervention pour construire un groupe contrefactuel non biaisé.

La *positivité* peut formellement s'écrire

$$0 < P(A = 1|W) < 1 \qquad (9)$$

Cette hypothèse permet de s'assurer que les individus avec un même ensemble de covariables peuvent-être à la fois traités ou non traités, autrement dit les individus se ressemblent suffisamment pour que la comparaison ait un sens et que le paramètre d'intérêt soit identifiable pour toutes les strates de W [16].

La *consistance* traduit le fait que le devenir potentiel, pour un sujet exposé à un régime hypothétique de traitement, lorsqu'il se matérialise est précisément le résultat connu par cette personne. L'observation d'une association se retrouve dans différentes populations et dans différentes circonstances, formellement pour tout sujet avec A=a, Y_a=Y. Cette hypothèse exclut des modifications d'effets du traitement sur le critère de jugement, l'effet individuel du traitement est identique pour tous les individus [21].

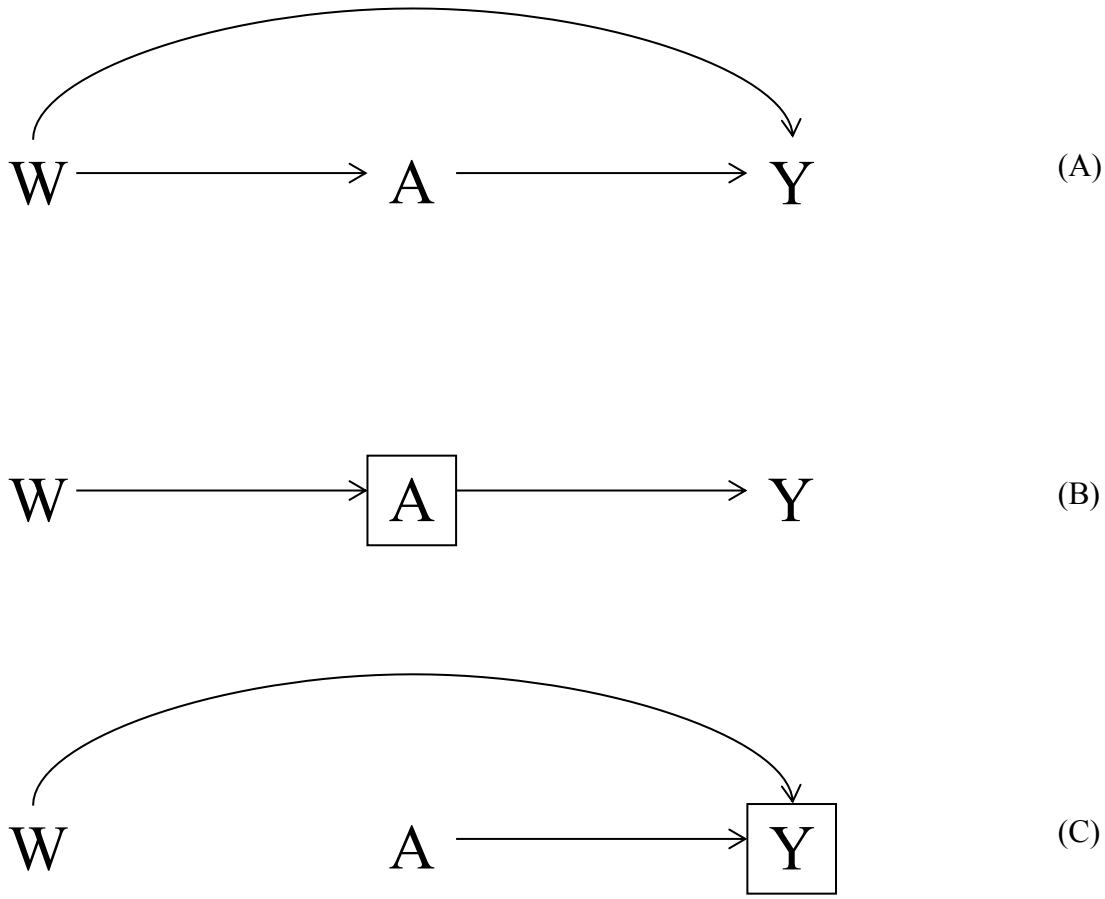
1.2.2 Diagrammes de causalité

Certains outils ont été proposés pour évaluer ces hypothèses lors de l'analyse de données réelles. Pearl [22] a ainsi proposé une méthode de modélisation structurale des données permettant de décrire les relations entre les variables observées et non observées, et d'identifier les conditions nécessaires à l'estimation du lien causal d'intérêt. Pour cela, il est nécessaire d'ordonner temporellement les variables et de représenter de manière non paramétrique (i.e., sans a priori sur la forme fonctionnelle de la relation) les relations entre les variables endogènes (i.e., variables dont les déterminants sont situés dans le modèle) et les variables exogènes (i.e., variables non mesurées dont les déterminants sont situés en dehors

des limites du modèle) [2]. Ce type de modèle structural causal peut être illustré à l'aide d'outils graphiques, les *Directed Acyclic Graphs*, DAG [20] (Figure 1).

Figure 1: Exemples de diagrammes de causalité.

Soit A le traitement à l'étude, Y le critère de jugement, et W une covariable d'étude. Les variables exogènes (U) ne sont pas représentées. Lorsqu'une variable est encadrée le chemin est bloqué. (A) A n'est marginalement pas indépendante de la variable observée W. (B) après conditionnement sur A, Y est ***conditionnellement indépendante*** de W. (C) A et W sont ***marginalement indépendante*** car le seul chemin les séparant est bloqué par le ***collider*** Y. Conditionner sur le ***collider*** Y annule l'indépendance.



Le DAG représente les variables endogènes sous forme de noeuds, et le lien entre elles par des arcs unidirectionnels représentant les relations directes mais non réciproques entre les nœuds. La présence d'une flèche suppose un effet direct d'une variable sur une autre. Il devient alors possible de déterminer graphiquement si deux variables sont indépendantes (*marginalement ou conditionnellement* à d'autres variables). Le chemin désigne un arc directionnel entre deux variables du graphique. En fonction de règles graphiques pré-établies, ce chemin est considéré comme "bloqué" ou "ouvert". Deux variables seront considérées comme directionnellement séparées si tous les chemins entre elles sont bloqués. Un chemin est dit "bloqué" si et seulement si deux arcs directionnels convergent vers une variable sur le même chemin. Lorsque le chemin séparant deux covariables est bloqué sans conditionnement nécessaire sur d'autres variables, elles sont considérées comme *marginalement indépendantes*. Par exemple sur la figure (A), le chemin $\boldsymbol{A \rightarrow Y \leftarrow W}$ est bloqué car deux arcs convergent sur le chemin vers Y: Y est alors appelé une convergence (*collider*) sur le chemin $\boldsymbol{A \rightarrow Y \leftarrow W}$. Le chemin $\boldsymbol{W \rightarrow A \rightarrow Y}$ est par contre ouvert. Par contre, pour bloquer un chemin passant par une variable *non collider* le conditionnement est nécessaire. Pour le second exemple de la figure (B), le chemin entre W et Y est ainsi bloqué en conditionnant sur A ; W et Y sont alors *conditionnellement indépendantes*. Lorsque le conditionnement est fait sur un *collider* le chemin reste ouvert. Sur le troisième exemple (figure (C)), conditionner sur Y, permet de laisser ouvert le chemin entre W et A. L'intérêt de ce type d'analyse graphique est de permettre de déterminer l'indépendance *marginale ou conditionnelle* entre deux variables [23]. Sur la figure (A), A et W ne sont pas marginalement indépendantes car le chemin $W \rightarrow A$ est ouvert. Sur la figure (C), A et W sont marginalement indépendantes car le chemin est bloqué par le *collider* Y; en revanche en conditionnant sur Y, W n'est pas indépendante de A car le chemin reste ainsi ouvert. Sur la figure (B), Y est indépendante de W conditionnellement à A.

Le DAG permet ainsi de représenter une structure causale avec des relations d'indépendance et de dépendance marginales et conditionnelles entre les variables. Il inclut les variables endogènes et suppose l'inclusion des variables exogènes (dénotées U), habituellement omises des représentations graphiques. Ainsi toutes les causes communes de deux variables sont explicitées sur un graphique. Ils sont très utiles pour analyser les relations de dépendance causale entre variables et permettent d'évaluer la possibilité d'identifier le lien causal entre une exposition et un devenir en s'appuyant sur l'hypothèse de positivité. Les DAG permettent ainsi d'envisager des hypothèses quant à la structure des données, hypothèses dites d'identifiabilité. Lorsqu'elles sont garanties, le paramètre causal d'intérêt est alors identifiable à partir des données observées [23],[24].

Ainsi, la plupart des méthodes d'inférence causale reposent-elles sur des hypothèses structurelles concernant les données, et ont en général pour but d'équilibrer, entre les groupes de traitement, la distribution de W* défini à partir du DAG comme le sous-ensemble minimal de W permettant d'estimer sans biais le paramètre d'intérêt [17],[18]. Parmi les différentes méthodes existantes, ce travail de thèse s'est intéressé particulièrement au score de propension (SP).

1.3 Méthodologie du score de propension

Le score de propension (SP), $g(W^*) = P(A = 1|W^*)$ est défini comme la probabilité pour un patient de recevoir un traitement conditionnellement à certaines de ses caractéristiques telles que définies par le DAG (W^*) [25]. Comme abordé dans la première partie de ce manuscrit, si g(W^*), fonction de W variables permettant d'en réduire la dimension, était connue, alors l'indépendance conditionnelle $Y \perp A|g(W^*)$ permet d'estimer directement et de manière non biaisée l'effet de A sur Y. En pratique, g(W^*) n'est pas connu et doit donc être estimé à partir des données observées et disponibles [26].

En recherche médicale, le traitement étudié est le plus souvent en « tout ou rien », il peut-être modélisé par une variable binaire (traitement versus contrôle ou traitement A versus traitement B). De ce fait, l'estimation du SP est souvent spécifiée au moyen d'un modèle de régression logistique [27]. Pour *j* variables W^* :

$$logit\{g(W^*)\} = \beta_0 + \beta_1.W^*{}_1 + \cdots + \beta_j.W^*{}_j \qquad (10)$$

Des méthodes alternatives d'apprentissage automatisé et non paramétrique peuvent également être utilisées pour cette procédure d'estimation. Les réseaux neuronaux, les arbres de classification/régression ont ainsi montré des performances supérieures aux méthodes paramétriques en termes de réduction de biais et de précision de l'estimation, particulièrement dans les situations de non-additivité et de non-linéarité de la relation entre A et W^* [28],[29],[30]. Les méthodes hybrides telles que le Super Learning améliorent l'équilibre final et réduisent le biais d'estimation de manière significative dans les situations où le modèle de SP est mal spécifié [31].

En plus du choix de la méthode de modélisation de la relation entre A et W^*, l'utilisateur doit également choisir quel sous-ensemble de W définit W^*. Comme nous l'avons vu ci-dessus, il

est possible et souhaitable d'utiliser les DAGs pour initier cette démarche. Néanmoins, dans la pratique, il existe souvent des incertitudes quant à la position des variables dans le DAG ou quant à leur interconnexion. Il nous est donc apparu utile de chercher à développer des outils statistiques permettant d'aider l'utilisateur dans son choix des variables à inclure dans le modèle de propension. Ce développement ainsi que la mise en application pratique au moyen de données réelles sont l'objet principal de ce travail de thèse.

1.3.1 Sélection des variables pour le modèle de score de propension

L'estimation du SP varie en fonction des variables incluses dans le modèle. Le choix des variables à inclure dans le modèle est crucial pour minimiser le biais et optimiser la précision de l'estimation de l'effet traitement [32],[33],[34],[35],[36]. Les recommandations actuelles suggèrent que le choix des variables explicatives, des termes d'interaction et/ou des termes d'ordre supérieur est avant tout basé sur les connaissances cliniques relatives au sujet d'étude [35],[37]. Les variables dites *vrais confondeurs*, c'est-à-dire associées à la fois à Y et A doivent impérativement être incluses. L'inclusion des variables pronostiques, liées uniquement à Y améliore la précision de l'estimation de l'effet du traitement sans augmenter le biais [32],[37]. En revanche, les variables instrumentales, liées uniquement à A, ne doivent pas être incluses dans le modèle de propension. En effet, plusieurs études ont démontré que l'inclusion de variables instrumentales dans le modèle de propension expose à un risque d'augmentation du biais et de diminution de la précision de l'estimation particulièrement dans les modèles non linéaires [37],[38]. Dans la suite du manuscrit, nous utiliserons W* comme notation pour définir le sous-ensemble de W comprenant les variables devant être incluses dans le modèle de SP.

1.3.2 Différents estimateurs basés score de propension

Il existe plusieurs méthodes utilisant le score de propension afin d'estimer l'effet causal de l'intervention. Les plus utilisés actuellement sont l'appariement, la stratification, l'ajustement, et la pondération inverse [39]. En pratique et sur des échantillons de taille finie, divers travaux ont rapporté la supériorité de l'appariement et de la pondération sur les deux autres estimateurs [40],[41],[42],[43],[44]. Dans la suite de ce document, nous nous focaliserons donc sur ces deux estimateurs. Il est à noter que les paramètres estimés avec la méthode d'appariement et celle de pondération sont différents [45]. En effet, tandis que la pondération est habituellement utilisée pour estimer l'effet moyen du traitement (*average treatment effect* ou ATE), l'appariement est un estimateur de l'effet moyen du traitement chez les traités (*average treatment effect among the treated* ou ATT) [46],[47].

L'appariement sur le score de propension consiste à apparier un sujet traité avec un sujet non traité ayant un score de propension *similaire* [48],[49]. Il apparaît dès lors évident qu'une des étapes essentielles de cette méthode sera de définir un degré de similitude. Après appariement, l'analyse de l'effet du traitement entre les groupes de sujets exposés et non-exposés se rapproche de celle obtenue lors d'un essai randomisé [50].

La seconde méthode est la pondération inverse. Un poids correspondant à l'inverse de la probabilité de recevoir le traitement effectivement reçu est attribué à chaque sujet de l'étude.

$$\omega = \frac{1}{P(A=a|W^*)} \qquad (11)$$

Ainsi, si A ∈ {0,1}, $\omega = \frac{1}{SP}$ quand A = 1, et $\omega = \frac{1}{1-SP}$ quand A=0.

Cette pondération permet de créer une nouvelle population pondérée à partir de laquelle une estimation de l'ATE pourra être obtenue.

L'ATE se définit de la manière suivante :

$$\psi(A, W^*) = E\left[\left\{\frac{A}{g(W^*)} - \frac{1-A}{1-g(W^*)}\right\}.Y\right] \qquad (12)$$

Si g(W*) était connu, la distribution W* dans la population pondérée serait indépendante de l'exposition au traitement [41], permettant ainsi une estimation directe et non biaisée de l'effet de A sur Y en appliquant des méthodes de régression adaptées aux données pondérées [51]. Mais ici encore, dans la pratique, g(W*) n'est pas connu. Il est donc nécessaire de vérifier la qualité de l'équilibration de la distribution de W* dans la population pondérée.

1.3.3 Evaluation de l'équilibre des facteurs de confusion

Comme nous l'avons expliqué auparavant, la plupart des estimateurs basés sur le score de propension ont pour but de réduire le déséquilibre de distributions des covariables entre les sujets traités et non traités [52]. L'approche par appariement et la pondération ont des performances supérieures comparativement aux méthodes d'ajustement et de stratification [40],[41]. Ainsi, la quantification de l'équilibre obtenu par appariement ou pondération est considérée comme un indicateur de la performance de l'estimateur, plus cette situation d'équilibre sera atteinte, meilleure sera la performance de l'estimateur [53]. Différentes mesures existent et permettent d'évaluer, avant et après appariement ou pondération, le déséquilibre, soit variable par variable, soit de manière globale.

1.3.3.1 Evaluation de l'équilibre de distribution variable par variable

Les tests de significativité

Ils sont fréquemment utilisés pour comparer la distribution des variables mesurées entre les groupes de traitement [42],[48]. Ils aident à faire un choix entre deux hypothèses statistiques concernant les caractéristiques d'une population issue d'une population plus large hypothétique [53]. En outre, ces tests sont influencés par la taille de l'échantillon (via le nombre d'observations restantes, le ratio de sujets traités et non traités restants, et la variance des groupes traités et non traités [54]).

Bien qu'étant faciles à utiliser et à interpréter, leur utilisation est déconseillée dans le contexte de comparaison des équilibres de distribution des variables sur l'échantillon apparié ou pondéré sur le PS [55]. Ainsi, notamment, en cas de taille d'échantillon insuffisante (ce qui peut singulièrement être le cas après appariement), l'absence de significativité du test ne

conduit pas à pouvoir accepter l'hypothèse nulle, même si la réduction du nombre d'observations améliore l'équilibre entre les groupes.

La différence standardisée (DS)

Il s'agit de la mesure de référence pour l'évaluation de l'équilibre de variables entre groupes de traitement [42]. Elle est définie comme la différence de moyenne exprimée en unités d'écart-type [56],[57]. Pour les variables continues :

$$DS = \frac{(\bar{W}^*_{A=1} - \bar{W}^*_{A=0})}{\sqrt{\frac{s^2_{A=1} + s^2_{A=0}}{2}}} \qquad (13)$$

avec $\bar{W}^*_{A=1}$et $\bar{W}^*_{A=0}$, les moyennes de la variable, et $s^2_{A=1}$ et $s^2_{A=0}$ les variances estimées respectivement dans le groupe des sujets traités (A=1) et non traités (A=0). Ainsi, les DS peuvent être comparées pour des variables continues mesurées sur différentes échelles. Cohen a également démontré pour les variables continues [58], que les DS étaient directement reliées au coefficient de corrélation de Pearson (r) entre la variable continue et la variable binaire d'affectation au groupe de traitement:

$$r = \frac{DS}{\sqrt{DS^2 + 1/pq}} \qquad (14)$$

avec p la proportion de sujets dans un groupe comparée au nombre total de sujets dans l'échantillon (et q=1-p). En outre, sous l'hypothèse de variables normalement distribuées et de variances égales, Cohen a déterminé les relations entre les DS et le degré de chevauchement des distributions des covariables entre les groupes de traitements [59].

Pour les variables binaires, la DS est obtenue à partir de la formule

$$DS = \frac{(\hat{p}_{A=1} - \hat{p}_{A=0})}{\sqrt{\frac{\hat{p}_{A=1}(1-\hat{p}_{A=1}) - \hat{p}_{A=0}(1-\hat{p}_{A=0})}{2}}} \qquad (15)$$

avec $\hat{p}_{A=1}$et $\hat{p}_{A=0}$ les proportions parmi les traités (A=1) et les non traités (A=0). Dans le cadre de variables binaires, la corrélation entre une variable binaire et la variable d'affectation au groupe de traitement s'évalue différemment via l'utilisation d'un coefficient (φ) basé sur la statistique du khi-deux non corrigé (χ_u^2)

$$\varphi = \sqrt{\chi_u^2 / N} \qquad (16)$$

avec N le nombre total de sujets. Equivalent au coefficient de corrélation de Pearson (r) [60], Austin a pu démontrer qu'une différence standardisée de 10% pour une variable continue correspondait approximativement à un coefficient phi de 5% pour une variable binaire (indiquant une corrélation négligeable) [61], et ainsi permettre d'envisager une comparaison des DS entre variables de différents types.

La DS a pour avantage un calcul simple, d'être indépendante de l'unité de mesure de la variable (car il s'agit d'une quantité sans unité), et non-influencée par la taille de l'échantillon [62],[63]. Cependant, il n'y a pas de consensus définissant clairement l'importance d'un déséquilibre résiduel sur l'échelle de DS (théoriquement entre $-\infty$ et $+\infty$). Le seuil habituellement choisi pour définir l'existence d'un déséquilibre résiduel est de 0,1 (10%) [59]. De plus, les différences standardisées n'analysent pas l'ensemble de la distribution ce qui peut masquer des différences résiduelles persistantes pour des zones particulières de distribution de la covariable malgré des valeurs de différences standardisées acceptables. Le dépistage d'un tel déséquilibre résiduel peut s'envisager par des diagrammes Quantile-Quantile du score de propension des sujets traités et contrôles [53].

La différence Z

Pour évaluer l'équilibre obtenu, la statistique utilisée doit idéalement être indépendante de la taille de l'échantillon. Kuss, lui, a proposé d'utiliser des mesures dérivées de la statistique de l'écart-réduit, Z [64]. En divisant la mesure d'équilibre propre à chaque type de variable par l'erreur-standard de celle-ci, la distribution normale centrée réduite de la différence Z obtenue permet une comparaison entre des variables de différents types. Sous l'hypothèse nulle d'équilibre des variables, Kuss propose une méthode graphique Quantile-Quantile (Q-Q plot) permettant une comparaison de l'ensemble des différences Z obtenues à une référence idéale définie par Rubin et Thomas [65][66]. Facile à utiliser et applicable à l'ensemble des types de variables, cette méthode basée sur des tests de comparaison peut cependant uniquement s'envisager en cas d'utilisation d'une procédure d'appariement.

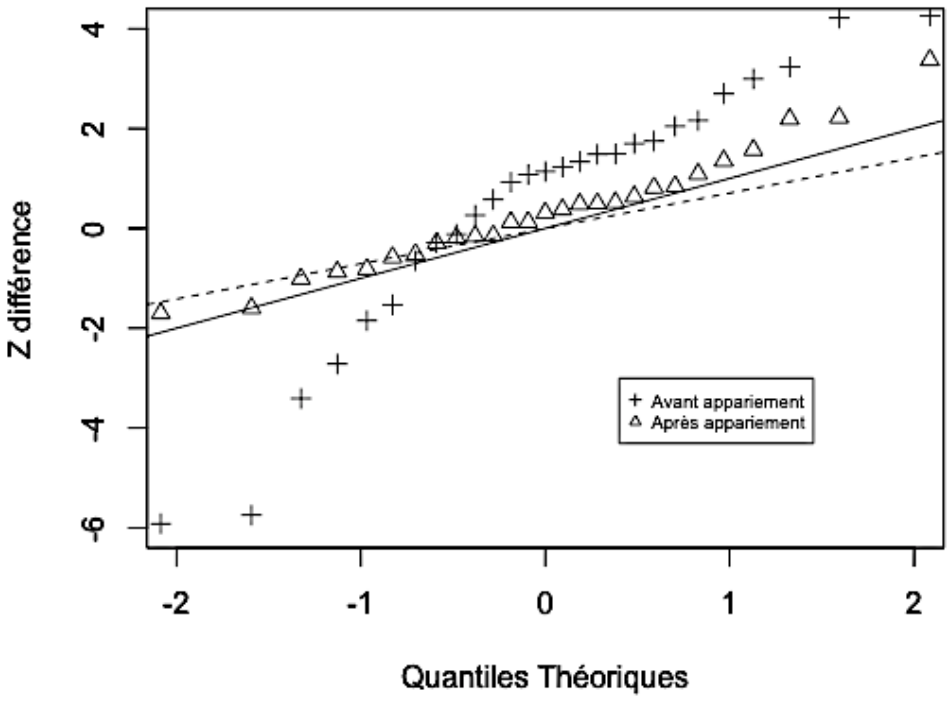


Figure 2: exemple de graphiques QQ de la différence Z avant et après une procédure d'appariement sur le score de propension à partir des données l'étude EFICA (Etude Française de l'Insuffisance Cardiaque Aiguë) La ligne en pointillé représente la distribution attendue des différences Z lors d'un essai randomisé. La ligne grise continue correspond à la distribution idéale des différences Z définie par Rubin/Thomas dans une analyse par appariement du score de propension

La distance de Malahanobis M_m

Elle est définie par

$$\widehat{M}_m = (\overline{W^*}^{A=1} - \overline{W^*}^{A=0})' \Sigma^{-1} (\overline{W^*}^{A=1} - \overline{W^*}^{A=0}) \quad (17)$$

avec $\overline{W^*}^{A=1}$ et $\overline{W^*}^{A=0}$ les vecteurs des moyennes de la variable du groupe traité et non traité et Σ la matrice de variance/covariance de la variable [67]. Une valeur basse de cette distance indique un déséquilibre faible [68].

La distance de Kolmogorov-Smirnov (D)

Elle est définie comme la distance verticale maximale entre deux fonctions de répartition [69]:

$$\widehat{D} = max|\hat{F}_{A=0}(W^*), \hat{F}_{A=1}(W^*)| \quad (18)$$

avec $\hat{F}_{A=1}(W^*)$, $\hat{F}_{A=0}(W^*)$ les fonctions de répartition de la variable étudiée W^* dans chacun des groupes de traitement. En se basant sur les fonctions de répartition, l'évaluation de la dispersion au sein des groupes est ainsi plus complète qu'à travers la comparaison de moyenne proposée via le calcul des différences standardisées. Cette mesure est aussi indépendante de l'unité de mesure, et ses seuils d'interprétation arbitraires varient entre 0 (l'équilibre optimal) et 1 (le déséquilibre complet). D'estimation plus complexe que la DS [70], elle a pour inconvénient principal d'être influencée par la taille de l'échantillon [62],[63].

Le coefficient de recouvrement (Overlapping coefficient, OVL)

Son obtention est plus complexe, et repose sur la densité de probabilité, permettant ainsi d'explorer l'ensemble de la distribution de la variable [71]. Il quantifie pour une variable binaire le chevauchement des densités de probabilité entre les deux groupes de sujets traités et non traités (figure 2). Pour une variable continue, le coefficient de recouvrement est estimé par la méthode non-paramétrique d'estimation par noyau de la densité de probabilité (*kernel density estimation*) [72]. Sa valeur indépendante de l'unité de mesure varie entre 0 (l'absence de chevauchement) et 1 (l'équilibre optimal).

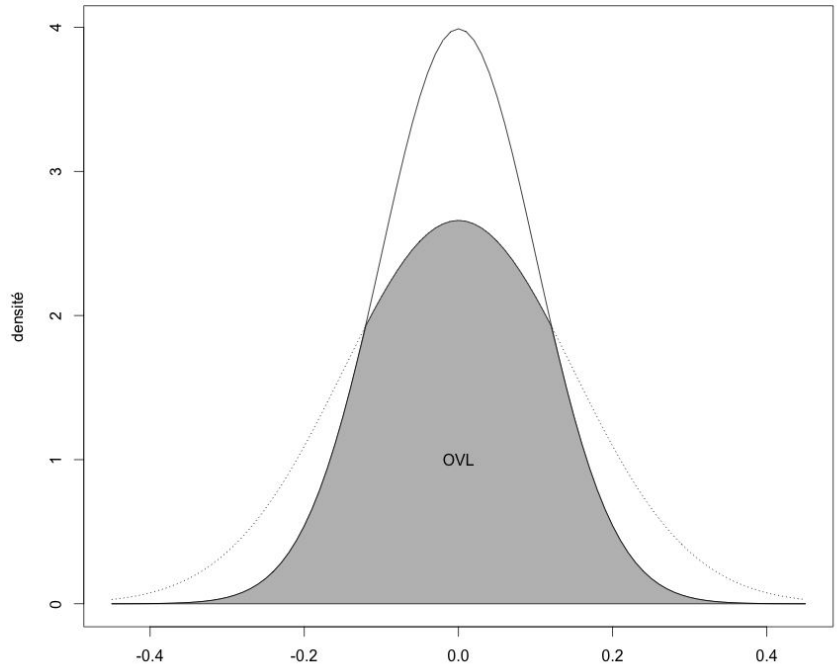


Figure 3: Coefficient de recouvrement (OVL) d'une covariable entre deux groupes de sujets

La distance de Lévy

Peu utilisée car d'estimation et d'interprétation encore plus complexes, cette mesure non-paramétrique explore l'ensemble de la distribution de la variable en prenant en compte la convergence en loi [73]. La distance verticale et horizontale entre deux fonctions de répartition d'une variable pour les sujets traités et non traités peut varier entre 0 (l'équilibre optimal) et 1 (le déséquilibre complet). Elle a également pour inconvénient de dépendre de l'unité de mesure et de la taille de l'échantillon [74].

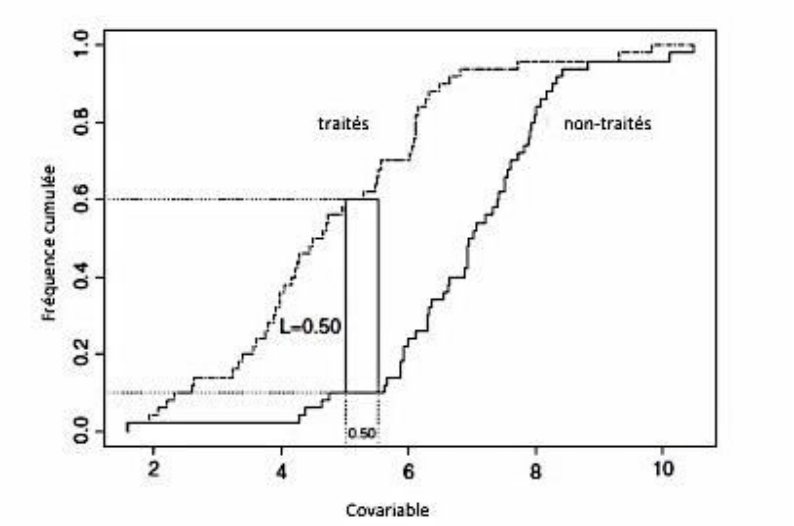


Figure 4: Distance de Lévy (L) d'une covariable entre deux groupes de sujet

1.3.4 Evaluation globale de l'équilibre de distribution des variables

Par global, nous entendons ici une mesure qui permet d'évaluer l'équilibre obtenu après appariement ou pondération sur l'ensemble des covariables d'intérêt W*.

La mesure $\mathbf{L_1}$

Elle a été décrite initialement par Iacus *et coll.* [75]. Elle mesure la proportion de recouvrement entre les histogrammes multidimensionnels des sujets traités et non traités. Une stratification spécifique pour chaque variable continue est établie (définie par chacune des classes pour les variables catégorielles), et il résulte du croisement des strates spécifiques à chaque variable un ensemble multidimensionnel de regroupement de variables *h* pour chacun des groupes de traitement. La mesure $\mathbf{L_1}$ est calculée comme la somme des déséquilibres entre chaque groupe :

$$L_1 = 0.5 \sum_{H \in h} |\hat{f}_1(H) - \hat{f}_0(H)| \qquad (23)$$

avec $\hat{f}_1(H)$, $\hat{f}_0(H)$ les proportions de sujets dans le groupe traité et non traité dans chacun des regroupements *h*. Cette mesure varie entre zéro (indiquant l'équilibre optimal) et un. Elle a ainsi pour avantage de prendre en compte également l'équilibre des termes d'interaction d'ordre élevé et de termes d'interaction [68]. Cependant, il n'y a pas de consensus sur le choix du niveau de stratification à utiliser.

La médiane $L_1(m_{L1})$ est une variante de la précédente mesure ayant pour but de diminuer l'influence du niveau de stratification choisi pour les différentes variables [75].

La statistique c après appariement

L'aire sous la courbe ROC (*Receiving Operating Charactistics*) mesure la capacité d'un modèle de SP à discriminer les sujets traités des sujets non traités [47]. Austin recommande de ne pas utiliser la c-statistique du modèle de SP calculée avant appariement pour évaluer la performance prédictive des modèles de score de propension et ses capacités à équilibrer les distributions des covariables entre les sujets traités et non traités [44],[76].

Cependant, certains auteurs au contraire ont proposé de calculer cette aire après appariement sur le score de propension pour évaluer l'équilibrer dans les échantillons appariés [68]. Sa valeur varie entre 0.5 (incapacité du modèle de SP à discriminer les sujets traités des sujets non traités après appariement) et 1 (équilibre optimal après appariement). En cas d'échantillon suffisamment grand, sa valeur est fortement corrélée au biais dans l'estimation de l'effet traitement [68].

Mesures d'équilibre pondérées

L'équilibre des "vrais facteurs confondants", c'est-à-dire des variables connues pour être liées à la fois au traitement A et au critère de jugement Y, ainsi que celui des variables pronostiques (variables liées uniquement à Y) est prioritaire pour minimiser le biais dans l'estimation de l'effet du traitement [63],[77]. Une mesure priorisant l'équilibre de telles covariables pourrait donc s'avérer utile dans ce contexte. Une des approches possibles consiste à pondérer le calcul de la mesure d'équilibre en attribuant à chaque variable un poids fonction de la force de son association avec le critère de jugement. Belitser et coll. ont proposé d'étudier les performances de plusieurs mesures d'équilibre (le coefficient de recouvrement, la distance de Levy, les différences standardisées, et la distance de Kolmogorov-Smirnov) et de pondérer ces différentes mesures d'équilibre avec un poids, reflétant la force d'association de chaque variable avec le critère de jugement [62]. L'utilisation de tels poids, par exemple sur la mesure du coefficient de recouvrement (OVL) permettait d'obtenir la mesure pondérée (OVL_p) :

$$\widehat{OVL_p} = \frac{1}{JI}\sum_{i=1}^{I} \sum_{j=1}^{J} \omega_i \widehat{OVL}_{ij} \qquad (19)$$

avec I le nombre de covariables W, J le nombre de strates du score de propension.

Trois poids différents ont été proposés et étudiés [62]. Chaque poids reflétait la force d'association au travers du calcul de l'odds ratio de la i^{ème} variable W avec le critère de jugement Y via un modèle de régression logistique, et la moyenne des forces d'association des k autres variables W restantes avec le critère de jugement Y.

Les différents poids ω_i ainsi dérivés sont décrits ci dessous:

poids 1 $$\omega_i = 1 + log(\widehat{OR}_{W_iY}) - \frac{1}{I}\sum_{k=1}^{I} log(\widehat{OR}_{W_kY}) \quad (20)$$

poids 2 $$\omega_i = 1 + \sqrt{\left|log(\widehat{OR}_{W_iY})\right|} - \frac{1}{I}\sum_{k=1}^{I} \sqrt{\left|log(\widehat{OR}_{W_kY})\right|} \quad (21)$$

poids 3 $$\omega_i = 1 + \left|log(\widehat{OR}_{W_iY})\right| - \frac{1}{I}\sum_{k=1}^{I} \left|log(\widehat{OR}_{W_kY})\right| \quad (22)$$

Parmi toutes les mesures comparées, celles basées sur les différences standardisées avaient les meilleures performances pour sélectionner le modèle de SP le moins biaisé, avec des performances similaires pour la mesure pondérée par le poids 3 basée elle aussi sur les différences standardisées [62]. Une des principales limites réside dans l'utilisation d'une méthode par stratification sur le score de propension, la méthode par pondération inverse ayant rapporté des performances moins bonnes. Cette approche peut cependant s'avérer intéressante pour aider le clinicien dans la construction et la sélection du score de propension le plus performant.

La mesure du score pronostique

Le score pronostique (h), $h(W) = P(Y_{a=0} = 1|W)$ est défini comme la probabilité de présenter un événement en situation de non exposition au traitement. Il vise à assurer que les groupes de sujets sont comparables sur leur pronostic, i.e., sur leur propension initiale à présenter l'événement d'intérêt [78]. Cette probabilité est estimée en modélisant le devenir dans le groupe non exposé au traitement. Ensuite, ce modèle est utilisé pour obtenir les prédictions du devenir en situation de non exposition au traitement (situation contrefactuelle pour une partie de l'échantillon) pour l'ensemble des individus. En cas d'indépendance

conditionnelle entre le devenir et l'exposition au traitement, la distribution du score pronostic doit être la même chez les traités et les non traités.

Stuart *et coll.* ont proposé une mesure d'équilibre basée sur le calcul des différences standardisées du score pronostique. En utilisant des simulations de Monte Carlo, Stuart *et coll.* ont rapporté une excellente corrélation entre le score pronostique et le biais dans l'estimation de l'effet du traitement [79], si le score pronostique incluait bien les vrais confondeurs. L'éventuelle non-inclusion de vrais confondeurs dans le modèle de score pronostique finalement retenu limite l'utilisation d'une telle mesure pour sélectionner le modèle final de SP le plus performant.

En résumé, comparativement à la mesure de référence qui repose sur le calcul des différences standardisées pour chaque covariable W, les mesures globales résumant l'équilibre sur l'ensemble des W d'intérêt (i.e. W*) semble intéressantes. La mesure L_1 est complexe d'interprétation et l'absence d'un consensus clairement établi sur ses modalités de calcul limite son utilisation. Le score pronostique s'avère plus simple d'utilisation et, contrairement à la mesure globale basée sur la C-statistique, permet, en plus de l'évaluation de l'équilibre, d'aider le clinicien dans le choix des variables à inclure. En raison de performances inégales [78] quant à sa capacité à capacité à identifier les variables reconnues comme essentielles à inclure dans le modèle de propension (i.e. les variables vrais confondeurs et pronostiques), le score pronostique reste peu utilisé en pratique et l'utilisation d'une mesure globale de l'équilibre permettant également de sélectionner le modèle de SP le plus performant dans l'estimation finale de l'effet traitement reste à élaborer.

2 OBJECTIF

Face à la diversité de méthodes décrites ci-dessus pour estimer l'équilibre obtenu après appariement ou pondération sur le score de propension en situation observationnelle, il n'existe aucune recommandation claire sur la (ou les) méthode(s) à privilégier en pratique. Le choix des variables à inclure dans le modèle de propension reste essentiellement guidé par la connaissance clinique, sans qu'aucune des métriques sus-citées ne permettent d'assister l'utilisateur dans ce choix pourtant crucial [37],[42]. L'impact des variables instrumentales est l'une des raisons limitant l'utilisation des premières méthodes décrites ci-dessus pour sélectionner les variables à inclure dans le modèle de propension. En effet, ces variables fortement associées à l'exposition au traitement mais non au devenir, sont par définition très déséquilibrées entre les groupes. Ainsi toute mesure d'équilibre ne tenant pas compte du caractère instrumental potentiel de la variable conclura à l'intérêt d'inclure une variable instrumentale dans le modèle de propension afin de réduire au maximum ce déséquilibre majeur entre les groupes. Or, il a été largement documenté dans la littérature que ce type de variable ne doit pas être introduite dans le modèle de propension [37],[38].

L'objectif de ce travail a donc été de développer une mesure globale pondérée permettant d'évaluer l'équilibre global des covariables obtenu après appariement et ainsi d'aider à la sélection d'un modèle de propension le plus parcimonieux possible, tout en éliminant les variables instrumentales. La première étape de ce travail de thèse a été d'élaborer une mesure satisfaisante, qui prenne en compte les forces d'association entre chaque variable et le critère de jugement, en pondérant ainsi plus fortement l'équilibre obtenu pour les "vrais confondeurs" par rapport aux variables instrumentales. Cette mesure a tout d'abord été utilisée dans une étude de simulation de Monte Carlo afin de démontrer qu'elle pouvait être utilisée pour sélectionner les variables à inclure dans le modèle de SP. Dans un second travail, cette mesure a été appliquée à des données réelles.

3 NOUVELLE MESURE D'ÉQUILIBRE AIDANT À SÉLECTIONNER LES VARIABLES À INCLURE DANS LE MODÈLE DE SCORE DE PROPENSION

3.1 Méthodes

3.1.1 Mesures d'équilibre

Comme nous l'avons rappelé dans la première partie de ce manuscrit, l'évaluation de l'équilibre de distribution des covariables d'intérêt entre les groupes de traitements est le plus souvent basé sur la valeur absolue des différences standardisées des moyennes de chaque covariable d'intérêt [59],[57], également appelées les tailles d'effet de Cohen [58]. Toute valeur absolue de différence de moyennes standardisée au-dessus du seuil de 10% (voire 20%) indique un déséquilibre résiduel entre les groupes de traitement concernant la distribution de la covariable correspondante [58] (Appendice A).

Nous avons proposé une version pondérée de la moyenne absolue des différences de moyennes standardisées qui prend en compte pour chaque covariable, sa force d'association avec le critère de jugement. Dans cette nouvelle mesure globale pondérée (Weighted Balance Measure, WBM), chaque différence de moyennes standardisée $\hat{d}_i$ est pondérée en fonction de sa valeur pronostique :

$$W\hat{B}M = \frac{1}{k}\sum_{i=1}^{k}\left|\widehat{\omega}_i \cdot \hat{d}_i\right| \qquad (24)$$

où $\widehat{\omega}_i$ est un poids prenant en compte l'association entre la covariable W_i et le critère de jugement.

Soit $\hat{\beta}_i$ le coefficient obtenu pour la covariable W_i issu du maximum de vraisemblance d'une régression multivariée de Y sur W sur l'ensemble de la population d'étude.

Les poids $\widehat{\omega}_i$ correspondent aux coefficients de régression standardisés :

$$\widehat{\omega}_i = \hat{\beta}_i \cdot SD_i \qquad (25)$$

avec SD_i l'écart-type d'échantillon pour la covariable W_i dans la cohorte originale [80].

Notre hypothèse est que le modèle de SP qui minimiserait une telle mesure pondérée est le modèle optimal (Appendice A).

3.1.2 Simulations de Monte Carlo

Pour évaluer la performance de cette nouvelle mesure résumée et la comparer avec la moyenne absolue des différences de moyennes standardisées et la distance de Mahalanobis, une première série de simulations de Monte Carlo a été conduite. Mille répétitions indépendantes d'un échantillon de 1000 observations ont été générées en utilisant un plan de simulation dérivé de précédents travaux publiés par Setoguchi *et al* [29].

3.1.2.1 Génération des données

Les détails sur le plan de simulation sont fournis dans l'appendice B. Onze variables normalement distribuées de moyenne 0 et d'écart-type 1 ont été générées :

- quatre "vrais confondeurs" ($W_{1,\dots,4}$) avec une force d'association au traitement A constante et une force d'association avec le critère de jugement Y décroissante, W_1 étant le plus fort des confondeurs et W_4 le plus faible ;

- trois prédicteurs ($W_{5,6,7}$) du traitement (A), à savoir, des variables plus ou moins corrélées avec les confondeurs : W_5 étant corrélée avec le confondeur le plus fort (W_1), W_6 avec un confondeur plus faible (W_2) et W_7 étant liée seulement avec le traitement ;

- trois variables pronostiques ($W_{8,9,10}$): W_8 corrélée avec un vrai confondeur (W_3), W_9 corrélée avec le confondeur le plus faible (W_4), et W_{10} uniquement associée au critère de jugement Y ;

- une variable associée ni avec l'exposition au traitement A ni avec le critère de jugement Y (W_{11}).

Différents niveaux de corrélation ont été introduits entre W_1 et W_5, W_2 et W_6, W_3 et W_8, et W_4 et W_9. Une dichotomisation basée sur la moyenne pour atténuer la force de la corrélation des coefficients utilisés a été réalisée: W_1, W_3, W_5, W_6, W_8 et W_9 devenant alors des variables binaires.

Une exposition au traitement binaire $A \in \{0,1\}$ a été générée avec une prévalence définie comme une fonction de $W_1 - W_7$ sur une échelle logit. La probabilité marginale d'exposition au traitement a été fixée à 0.4.

Un critère de jugement binaire $Y \in \{0,1\}$ a ensuite été généré, sa prévalence étant une fonction de A, W_{1-4}, W_{8-10} sur une échelle logit. La probabilité marginale de Y a été fixée à 0.5. La valeur d'effet marginal du traitement sur le critère de jugement a été fixée à -0.4 (Odds ratio = 0.67). La valeur de l'effet moyen chez les traités (ATT = *Average Treatment Effect among the Treated*) était fixée pour être égale à -0.6.

Pour vérifier le comportement de la mesure WBM dans une situation où l'allocation au traitement serait randomisée, une série supplémentaire de simulations a été effectuée en fixant les forces d'association entre les covariables et l'allocation au traitement égale à zéro.

3.1.2.2 Modèles de SP et estimation de l'effet

Le SP a été estimé à partir des données simulées en utilisant une régression logistique multivariée. En fonction des différentes variables incluses dans le modèle, neuf scores de propension ont été définis. Le premier modèle (modèle de référence ou SP 1) contenait trois

vrais confondeurs (W_1, W_2 et W_3). Pour chacun des huit autres modèles de SP (noté PS + W_i), l'une des huit autres variables ($W_{4,\ldots,11}$) était ajoutée séparément dans le modèle :

- SP 2 : W_1, W_2, W_3 et W_4 (quatre vrais confondeurs)
- SP 3 : W_1, W_2, W_3 et W_5 (trois vrais confondeurs et une variable corrélée à une IV)
- SP 4 : W_1, W_2, W_3 et W_6 (trois vrais confondeurs et une variable corrélée à une IV)
- SP 5 : W_1, W_2, W_3 et W_7 (trois vrais confondeurs et une variable non liée à une IV)
- SP 6 : W_1, W_2, W_3 et W_8 (trois vrais confondeurs et une variable corrélée à un facteur pronostique)
- SP 7 : W_1, W_2, W_3 et W_9 (trois vrais confondeurs et une variable corrélée à un facteur pronostique)
- SP 8 : W_1, W_2, W_3 et W_{10} (trois vrais confondeurs et une variable non liée à un facteur pronostique)
- SP 9: W_1, W_2, W_3 et W_{11} (trois vrais confondeurs et une variable indépendante ni liée à A et ni à Y)

Un appariement un à un sur le SP a été réalisé en utilisant un algorithme d'appariement au voisin le plus proche et une procédure sans remise. Chaque sujet traité était sélectionné au hasard et apparié avec le sujet le plus proche en fixant une différence maximale entre deux sujets (ou *caliper*) à 0.2 fois l'écart-type du logit du SP.

L'effet moyen du traitement chez les sujets traités était estimé [5]. Un estimateur de la variance d'Abadie-Imbens [81] (package Matching dans R [82]) prenant en compte l'incertitude liée à la procédure d'appariement a été utilisé.

3.1.2.3 Mesures de performance

Pour l'étude de simulation, N=1000 échantillons indépendants de données avec n=1000 individus ont été générés. Dans chaque échantillon simulé, ont été estimés :

- la performance de l'estimation de l'effet du traitement en termes de biais et d'erreur quadratique moyenne (EQM). Les biais absolu et relatif ont été reportés ;
- l'équilibre des covariables entre les groupes de traitement, dans la cohorte originale et celle appariée, au travers de la mesure WBM, la moyenne absolue des différences de moyennes standardisées, et de la distance de Mahalanobis. Chaque mesure était moyennée sur les N répétitions ;
- La moyenne des N variances des trois mesures d'équilibre, basées sur un ré-échantillonnage non-paramétrique de 1000 répétitions au sein de chaque échantillon de données.

Le but était de comparer, en termes de performance et d'équilibre, le modèle de référence (SP 1) à chaque nouveau modèle obtenu en ajoutant une variable. En théorie, le modèle attendu avec les meilleures performances en termes de biais et d'efficacité est le modèle de SP 2 incluant tous les vrais confondeurs et excluant les variables instrumentales.

Toutes les analyses ont été réalisées en utilisant le logiciel statistique R 2.15.1 sur une plate-forme Mac OsX (Fondation R pour le calcul statistique, Vienne, Autriche).

3.2 Résultats des simulations de Monte Carlo

Les résultats obtenus avec les neuf modèles de SP sont présentés dans le tableau 2.

3.2.1 Performance des estimations

L'estimation naïve de l'effet du traitement était nettement biaisée (biais=0.169). L'estimation via l'appariement sur le score de propension basée sur le modèle de SP 1 permettait d'éliminer le biais, avec une erreur quadratique moyenne proche de zéro (EQM=0.0015). L'ajout de W_4 (vrai confondeur avec une association faible avec le critère de jugement mais une forte association avec l'exposition au traitement) diminuait encore le biais et l'EQM. A l'inverse, l'inclusion d'une variable instrumentale (W_7) dans le modèle de SP (SP 5) était associée à une augmentation du biais et une diminution de la précision comparativement à SP 1. En incluant les covariables $W_{5,6}$ (variables instrumentales n'ayant pas d'association directe avec le critère de jugement mais corrélées avec les vrais confondeurs) (SP 3 et SP 4), l'estimation de l'effet traitement n'était pas substantiellement altérée. Similairement, l'inclusion des variables pronostiques $W_{8,9,10}$ (SP 6-8) ne modifiait pas substantiellement les estimations. Enfin, l'inclusion dans le modèle de SP d'une variable indépendante (W_{11}, SP 9) n'avait pas d'impact sur l'estimation.

Tableau 1 : Résultats des simulations. Biais, erreur quadratique moyenne (EQM) et mesure d'équilibre en fonction du modèle de SP. (Biais exprimé en valeur absolue (biais relatif en %); IC: intervalle de confiance. * signale une différence significative dans la minimisation de la mesure comparativement à la mesure d'équilibre obtenue pour le SP 1 avec un test t.

	Echantillon original	SP 1	SP 2 (+W4)	SP 3 (+W5)	SP 4 (+W6)	SP 5 (+W7)	SP 6 (+W8)	SP 7 (+W9)	SP 8 (+W10)	SP 9 (+W11)
Moyenne absolue des différences de moyennes standardisées	0.2767	0.2541	0.2147*	0.2326*	0.2357*	0.2245*	0.2519	0.2504	0.2513	0.2519
IC 95%	0.2752-0.2768	0.2539-0.2544	0.2144-0.2150	0.2324-0.2329	0.2355-0.2360	0.2242-0.2249	0.2517-0.2522	0.2502-0.2507	0.2511-0.2516	0.2517-0.2522
Distance Mahalanobis	1.5087	1.5169	1.3627*	1.5793	1.5707	0.9802	1.5195	1.5188	1.5161	1.5136
IC 95%	1.5022-1.5152	1.5153-1.5186	1.3607-0.3648	1.5776-1.5812	1.5691-1.5723	0.9783-0.9822	1.5179-1.5211	1.5173-1.5204	1.5146-1.5177	1.5121-1.5152
Weighted Balance Measure	0.0531	0.0417	0.0319*	0.0414	0.0416	0.0564	0.0399	0.0392	0.0379	0.0417
IC 95%	0.0522-0.0539	0.0416-0.0418	0.0317-0.0321	0.0413-0.0416	0.0414-0.0417	0.0562-0.0566	0.0398-0.0401	0.0391-0.0394	0.0378-0.0380	0.0416-0.0419
Taille d'échantillon	1000	795	668	758	761	804	793	792	793	793
Biais (%)	0.169 (265%)	0.018 (28%)	0.001 (3%)	0.018 (28%)	0.017 (27%)	0.022 (35%)	0.017 (28%)	0.023 (36%)	0.017 (27%)	0.017 (27%)
Erreur standard estimée	0.131	0.034	0.037	0.034	0.035	0.055	0.032	0.033	0.031	0.034
EQM	0.0459	0.0015	0.0013	0.0015	0.0015	0.0035	0.0014	0.0017	0.0013	0.0014

3.2.2 Mesures d'équilibre

La distribution de la moyenne absolue des différences des moyennes standardisées, la distance de Mahalanobis, et la WBM sont illustrées dans la Figure 5. Alors que la distribution de la moyenne des différences de moyenne semble normale, la distance de Mahalanobis a une distribution déviée à droite, et la WBM a une distribution proche de la normale, avec une discrète déviation à gauche. Comme attendu, l'inclusion dans le modèle de SP, d'une des variables associées avec l'exposition au traitement permettait de diminuer la moyenne absolue des différences de moyenne comparativement à la valeur obtenue avec le SP 1. Ensuite, bien que cela altère la performance d'estimation, l'inclusion d'une variable instrumentale résultait dans une valeur significativement plus petite de la moyenne absolue des différences des moyennes standardisées comparativement à la valeur obtenue avec le modèle de référence SP 1. Des résultats similaires ont été observés avec la distance de Mahalanobis.

La valeur de WBM diminuait avec l'inclusion d'une variable associée avec le critère de jugement. Cependant, elle augmentait significativement comparativement à la valeur obtenue avec SP 1, avec l'inclusion d'une variable instrumentale dans le modèle de SP. La plus petite valeur de WBM était observée pour le modèle SP 2, qui correspondait également au modèle de SP associée avec la meilleure performance d'estimation. Finalement, en re-créant des conditions de simulation proches de celles d'un essai avec une allocation de traitement randomisée, la moyenne de WBM était de 0.021 avec une distribution normale (Figure 6).

Figure 5: Distributions de la moyenne absolue des différences de moyenne standardisées (haut), distance de Mahalanobis (milieu), et WBM (bas)

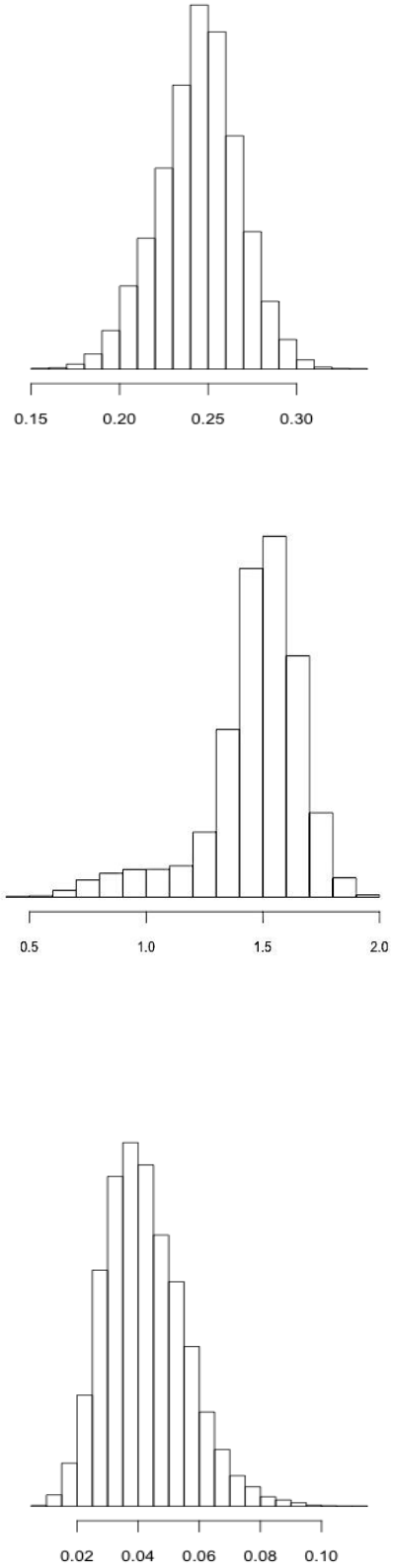
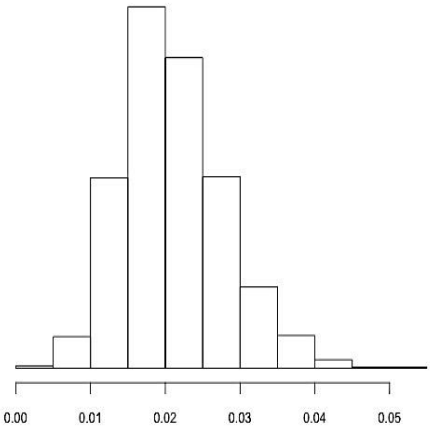


Figure 6: Distribution de WBM dans une situation d'essai randomisé



3.3 Discussion de l'étude de simulation

Nous avons proposé une mesure globale d'équilibre donnant plus de poids aux vrais confondeurs qu'aux variables instrumentales. Cette mesure résumée et pondérée semble être utile pour sélectionner un ensemble minimal de covariables devant être incluses dans le modèle de SP, en évitant l'inclusion de variables instrumentales. Il est souvent difficile de distinguer les vrais confondeurs, les facteurs pronostiques et les variables instrumentales en se basant sur les seules connaissances du sujet dans la littérature. La mesure WBM pourrait ainsi aider à identifier le modèle de SP fournissant l'estimation d'effet du traitement la moins biaisée. La sélection devrait être basée sur les connaissances déjà existantes sur le sujet mais aussi sur des mesures d'équilibre appropriées. Une mesure pondérée dérivée de la moyenne absolue des différences de moyennes standardisées pourrait s'avérer utile pour identifier le modèle de SP le plus parcimonieux et efficace en termes d'estimation. En se basant sur ces résultats, l'utilisation des connaissances déjà existantes sur le sujet d'étude pour sélectionner les variables devant être incluses dans le modèle de SP de référence, et ensuite en incluant une par une les autres variables en vérifiant si cela est associé avec une diminution de la WBM peut être une stratégie de sélection adaptée. Les variables qui minimisent la mesure devraient être incluses dans le modèle final, alors que celles qui ne l'influencent pas ou l'augmentent devraient être probablement écartées.

L'utilisation pratique de cette nouvelle mesure globale a fait l'objet d'un second travail détaillé dans la suite de ce manuscrit.

3.4 Article

ELSEVIER

CrossMark

Journal of Clinical Epidemiology 68 (2015) 1415–1422

Journal of Clinical Epidemiology

A new weighted balance measure helped to select the variables to be included in a propensity score model

Emmanuel Caruana[a], Sylvie Chevret[a], Matthieu Resche-Rigon[a], Romain Pirracchio[a,b,*]

[a]*Service de Biostatistique et Information Médicale, Hôpital Saint-Louis, AP-HP, Inserm 1153 ECSTRA Team, Université Paris 7 Diderot, 1 rue Claude Vellefaux, Paris 75010, France*

[b]*Service d'Anesthésie-Réanimation, Hôpital Européen Georges Pompidou, Université Paris V Descartes—Sorbonne Paris Cité, 20 rue Leblanc, Paris 75015, France*

Accepted 24 April 2015; Published online 1 May 2015

Abstract

Objectives: The propensity score (PS) is a balancing score. Following PS matching, balance checking usually relies on estimating separately the standardized absolute mean difference for each baseline characteristic. The average standardized absolute mean difference and the Mahalanobis distances have been proposed to summarize the information across the covariates. However, they might be minimized when nondesirable variables such as instrumental variables (IV) are included in the PS model. We propose a new weighted summary balance measure that takes into account, for each covariate, its strength of association with the outcome.

Study Design and Setting: This new measure was evaluated using a simulation study to assess whether minimization of the measure coincided with minimally biased estimates. All measures were then applied to a real data set from an observational cohort study.

Results: Contrarily to the other measures, our proposal was minimized when including the confounders, which coincided with minimal bias and mean squared error, but increased when including an IV in the PS model. Similar findings were observed in the real data set.

Conclusion: A balance measure taking into account the strength of association between the covariates and the outcome may be helpful to identify the most parsimonious PS model.

Keywords: Propensity score; Propensity score matching; Balance; Standardized mean difference; Instrumental variables; Causal inference

1. Introduction

Propensity Score (PS) estimators have become very popular for causal inference in observational medical research. The PS is the individual probability of receiving the treatment conditional on patients' characteristics [1]. It is often referred to as a "balancing" score [2,3], that is, treated and untreated subjects have the same distribution of observed baseline covariates conditionally on the true PS. Indeed, whatever the estimator, the goal of the PS is to balance the distribution of potential confounders. Because the true PS is unknown, it has to be estimated from the observed data. The performance of PS estimators strongly relies on this estimation [4–6]. Specifically, the choice of the variables to be included in the PS model is of paramount importance. For instance, omitting a confounder (i.e., a variable associated both with the treatment and the outcome) in the PS model may produce substantially biased estimates [7,8]. To avoid this risk of bias, some authors have recommended the use of nonparsimonious exposure models [7]. However, such an approach carries several drawbacks. First, when the sample size and/or the treatment prevalence are limited, nonparsimonious models may be overparameterized [9]. Second, including instrumental variables (IV, i.e., strong predictors of treatment allocation but not of the outcome) in the PS model may compromise the efficiency of the estimator [8,10]. This emphasizes the need for methods that would provide some guidance to select the variables to be included in the PS model. Goodness-of-fit and discrimination measures do not provide any reliable information in this context [11]. Balance measures, for example, metrics that aim at assessing whether adequate balance is achieved, appear more suitable. Several balancing measures have been proposed, the most widely used in practice being the standardized absolute mean difference, computed separately for each important baseline characteristic [3,12,13]. More recently, some authors have

Conflict of interest: None.
Funding: None.
* Corresponding author. Tel.: +1-56-09-25-13; fax: +1-56-09-25-51.
E-mail address: romainpirracchio@yahoo.fr (R. Pirracchio).

http://dx.doi.org/10.1016/j.jclinepi.2015.04.009
0895-4356/

What is new?

- Balance measures are crucial in propensity score (PS) analysis. Global balance measures may be easier to use.
- Including an instrumental variable in the PS model has been shown to impair the efficiency of the PS estimator, although it improves balance.
- A global balance measure that takes into account for each covariate its estimated strength of association with the outcome could prove useful to select the appropriate minimal set of covariates to be included in the PS model.
- Applied to simulation study, the new balance measure identified the PS model associated with minimal bias and mean squared error.
- This global balance measure may help to identify the most parsimonious PS model thereby avoiding the inclusion of useless and even potentially nondesirable variables.

advocated the use of a single global balance measure such as the average standardized absolute mean difference [14,15]. Several other candidate metrics have been tested [2,14,16–18]. Although appealing, most of these approaches may be flawed because including an IV in the PS model may improve the overall balance although harming the final estimate. Hence, there is a need for a "clever" summary measure of the balance that would take into account the strength of the association between each covariate and the outcome thereby putting more weight on the balance reached for a confounder than that for an IV.

The objective of this study was to propose a procedure to select the minimum set of variables to be included in the PS model based on such weighted balance measure (WBM). In the first part, we used Monte Carlo simulations to assess the performance of the new measure; then, we report the results of an application to a real data set from an observational cohort study conducted to evaluate the benefit of continuous positive airway pressure (CPAP) on short-term mortality based on 575 patients admitted to an intensive care unit (ICU) for a severe acute decompensated heart failure.

2. Methods

2.1. Balance measures

Usually, the measure of imbalance between the treatment groups is based on the standardized absolute mean differences [3,19], also referred to as the Cohen's effect sizes [20]. Any value of the standardized absolute mean difference above the threshold of 10% indicates a meaningful imbalance between the treatment groups with regard to the distribution of the corresponding covariate [20] (Appendix A at www.jclinepi.com).

Several global balance measures have been proposed, the most widely used being the average standardized absolute mean difference [14,15] and the Mahalanobis distance [21] (Appendix A at www.jclinepi.com). For both measures, the lower the value, the lower the imbalance between treatment groups. The best PS model is the one that minimizes the measure. However, such an approach may be flawed.

IVs are strongly associated with treatment allocation but not with the outcome. Therefore, IV distribution is, by definition, strongly imbalanced between treatment groups. Thus, including an IV in the PS model is likely to substantially minimize the average standardized absolute mean difference or the Mahalanobis distance, although such variables have been shown to be undesirable in PS models [7,8]. We hypothesized that a weighted version of average standardized absolute mean difference, taking into account, for each covariate, its strength of association with the outcome, could prove useful to select the appropriate minimal set of covariates to be included in the PS model and to avoid the inclusion of IVs. We proposed a WBM where each standardized mean difference $\widehat{d}_i$ is weighted according to its prognostic value, as follows:

$$\widehat{\mathrm{WBM}} = \frac{1}{k}\sum_{i=1}^{k}\left|\widehat{\omega}_i \cdot \widehat{d}_i\right| \quad (1)$$

where ω_i is a weight accounting for the association between the covariate W_i and the outcome. Specifically, $\widehat{\beta}_i$ is the coefficient obtained for the covariate W_i from a multivariate maximum likelihood regression of Y on the W:

$$\widehat{\omega}_i = \widehat{\beta}_i \cdot \mathrm{SD}_i \quad (6)$$

where SD_i is the sample standard deviation for the covariate W_i in the original cohort [22]. The PS model that would minimize such a weighted measure would be the one associated with the most efficient estimator (Appendix A at www.jclinepi.com).

2.2. Monte Carlo simulations

To evaluate the performance of this new summary measure and compare it with the average standardized absolute mean difference and the Mahalanobis distance, we first conducted a series of Monte Carlo simulations. A thousand replicates of sample size 1,000 were simulated using a data generation plan derived from the one previously published by Setoguchi et al. [23].

2.2.1. Data generation

Details about the simulation plan are provided in Appendix B at www.jclinepi.com. Eleven normally distributed covariates were generated:

- Four true confounders ($W_{1,\ldots,4}$) with a decrease strength of association with the outcome, W_1 being the strongest confounder and W_4 the weaker;
- Three predictors of exposure, that is, IV with different strengths of association with the confounders: W_5 was correlated with the strongest confounder (W_1), W_6 with the weaker confounder (W_2), and W_7 was only related to exposure;
- Three prognostic covariates: W_8 correlated with a true confounder (W_3), W_9 with the weaker (W_4), and W_{10} related to outcome only;
- One associated neither with the exposure nor with the outcome (W_{11}).

Various amounts of correlation were introduced between W_1 and W_5, W_2 and W_6, W_3 and W_8, and W_4 and W_9. A dichotomization according to the mean to attenuate the magnitude of correlation coefficients used was realized so that (W_1, W_3, W_5, W_6, W_8, and W_9) were considered binary in scale. We generated a binary exposure $A \in \{0,1\}$, which prevalence was a function of W_1-W_7 on a logit scale. The overall exposure prevalence was set to be 0.4. We finally generated a binary outcome $Y \in \{0,1\}$, its prevalence being a function of A, W_{1-4}, W_{8-10} on a logit scale. The overall outcome prevalence was fixed at 0.5. The true effect of the exposure on the outcome was set to be −0.4 (odds ratio = 0.67).

To verify the behavior of the WBM under randomized treatment allocation, we performed an additional simulation series where the coefficients defining the association between the covariates and the treatment allocation were set at zero.

2.2.2. PS models and effect estimation

The PS was estimated from the data using multivariate logistic regressions. According to the variables included in the model, we defined nine different PS models. The first PS model (reference model or PS 1) contained three of the major confounders (W_1, W_2, and W_3). For each of the eight other PS models (denoted PS + W_i), one of the eight other variables ($W_{4,\ldots,11}$) was added separately in the model:

- PS 2: $\mathbf{W}_1$, $\mathbf{W}_2$, $\mathbf{W}_3$, and $\mathbf{W}_4$ (four confounders)
- PS 3: $\mathbf{W}_1$, $\mathbf{W}_2$, $\mathbf{W}_3$, and $\mathbf{W}_5$ (three main confounders and a correlated IV)
- PS 4: $\mathbf{W}_1$, $\mathbf{W}_2$, $\mathbf{W}_3$, and $\mathbf{W}_6$ (three main confounders and a correlated IV)
- PS 5: $\mathbf{W}_1$, $\mathbf{W}_2$, $\mathbf{W}_3$, and $\mathbf{W}_7$ (three main confounders and an uncorrelated IV)
- PS 6: $\mathbf{W}_1$, $\mathbf{W}_2$, $\mathbf{W}_3$, and $\mathbf{W}_8$ (three main confounders and a correlated prognostic covariate)
- PS 7: $\mathbf{W}_1$, $\mathbf{W}_2$, $\mathbf{W}_3$, and $\mathbf{W}_9$ (three main confounders and a correlated prognostic covariate)
- PS 8: $\mathbf{W}_1$, $\mathbf{W}_2$, $\mathbf{W}_3$, and $\mathbf{W}_{10}$ (three main confounders and an uncorrelated prognostic covariate)
- PS 9: $\mathbf{W}_1$, $\mathbf{W}_2$, $\mathbf{W}_3$, and $\mathbf{W}_{11}$ (three main confounders and an independent covariate)

PS matching was then performed using a 1:1 nearest neighbor without replacement algorithm. Each treated subject was randomly selected and matched once to the nearest untreated subject with a caliper width set at 0.2 of the standard deviation of the logit of the PS. We estimated the average treatment effect on the treated subject (set to be equal to −0.06 in our simulation) [24]. The Abadie–Imbens estimator [25] (package Matching for R [26]) was used for variance estimation, as it takes into account the uncertainty related to the matching procedure.

2.2.3. Performance metrics

For the simulation study, $N = 1{,}000$ independent data sets with $n = 1{,}000$ individuals were generated. In each simulated data set, we estimated

- the performance of the PS matching estimator in terms of bias and mean squared error (MSE). We reported absolute and relative bias;
- the balance in the covariates between treatment groups, in the original and the matched cohorts, as evaluated by our WBM, the average standardized absolute mean difference and the Mahalanobis distance. Each measure was averaged over the N replications;
- the variance of the three balance measures, based on 1,000 nonparametric bootstrap replicates within each simulated data set.

The goal was to compare, in terms of performance and balance, the reference model (PS 1) to each new model obtained when adding one more variable. According to the

Table 1. Results of univariate analyses for death and use of CPAP in clinical illustration

	Univariate analyses for death		Univariate analysis for use of CPAP	
Variable	**OR (95% CI)**	***P*-value**	**OR (95% CI)**	***P*-value**
Age	2.63 (1.52, 4.57)	0.001		
Obesity	0.48 (0.30, 0.77)	0.003		
Shock	2.34 (1.43, 3.82)	0.001		
Acute coronary syndrome	1.78 (1.20, 2.62)	0.004		
NYHA	2.27 (1.52, 3.39)	<0.001	0.79 (0.45, 1.38)	0.031
First systolic BP	0.99 (0.98, 1.01)	0.002	1.006 (1.004, 1.013)	0.035
Teaching hospital	0.92 (0.63, 1.35)	0.11	2.15 (1.65, 3.82)	<0.001

Abbreviations: CPAP, continuous positive airway pressure; OR, odds ratio; CI, confidence interval; NYHA, New York Heart Association; BP, blood pressure.

Table 2. Simulation results

Metrics and performances	Original sample	PS 1	PS 2 (+ W_4)	PS 3 (+ W_5)	PS 4 (+ W_6)
Average standardized absolute mean difference	0.2767	0.2541	0.2147[a]	0.2326[a]	0.2357[a]
CI 95%	0.2752, 0.2768	0.2539, 0.2544	0.2144, 0.2150	0.2324, 0.2329	0.2355, 0.2360
Mahalanobis distance	1.5087	1.5169	1.3627[a]	1.5793	1.5707
CI 95%	1.5022, 1.5152	1.5153, 1.5186	1.3607, 1.3648	1.5776, 1.5812	1.5691, 1.5723
Weighted balance measure	0.0531	0.0417	0.0319[a]	0.0414	0.0416
CI 95%	0.0522, 0.0539	0.0416, 0.0418	0.0317, 0.0321	0.0413, 0.0416	0.0414, 0.0417
Sample size	1,000	795	668	758	761
Bias (%)	0.169 (265)	0.018 (28)	0.001 (3)	0.018 (28)	0.017 (27)
SD Estimates	0.131	0.034	0.037	0.034	0.035
MSE	0.0459	0.0015	0.0013	0.0015	0.0015

Abbreviations: PS, propensity score; CI, confidence interval; SE, standard error; MSE, mean squared error.
Bias, mean square error, and balance diagnosis according to the PS model. [Bias is expressed as absolute value (relative bias in %).]
[a] Stands for a significant difference in minimization as compared with the balance measure obtained for PS 1 with *t*-test.

theoretical background, we expected to find the best properties in terms of consistency and efficiency with the PS model including all the confounders but not the IVs (PS 2).

All analyses were performed using R 2.15.1 statistical software running on a Mac OsX platform (The R Foundation for Statistical Computing, Vienna, Austria). Details of the simulations with basic R codes for computing the proposed summary measure are provided on request.

2.3. Clinical illustration

2.3.1. Patients and data

We used the data from EFICA (*Etude Française de l'Insuffisance Cardiaque Aiguë*), an observational cohort study of severe acute decompensated heart failure patients admitted to an ICU in 60 French hospitals from April to October 2001. A total of 575 patients with the diagnostic criteria for acute cardiac failure were included and 139 (24%) received CPAP. The following data were collected: age (classified into two classes 0: for ages $\leq$70 and 1: for ages $>$70); gender; obesity (0: absence, 1: presence), shock (0: absence, 1: presence); New York Heart Association classification (NYHA 0: for classes I/II, 1: for classes II/III); hospitalization site (0: for coronary care unit, 1: ICU); atrial fibrillation; acute coronary syndrome (ACS); first systolic blood pressure (mmHg, presented as median), and first heart rate (beats per minute, presented as median). The goal of the study was to assess the benefit on short-term mortality of applying CPAP within the first 24 hours in this population of patients.

2.3.2. Statistical analysis

The variables were classified as confounders, IVs, or prognostic factors according to subject-matter knowledge [27–31] and/or statistical analysis (Table 1). We identified true confounders (NHYA, first systolic BP), IV (teaching hospital), prognostic factors [obesity (strongest one), age, shock, ACS], and unclassified factors (gender, cardiogenic pulmonary edema, first heart rate, atrial fibrillation). Based on this classification, PS 1 (reference model) included true confounders and prognostic factors except obesity. Eight other PS models were constructed by adding a variable to PS 1:

- PS 2: added obesity (strongest risk factor)
- PS 3: added teaching hospital (instrumental variable)
- PS 4: added gender (unclassified variable)
- PS 5: added cardiogenic pulmonary edema (unclassified variable)
- PS 6: added first heart rate (unclassified variable)
- PS 7: added atrial fibrillation (unclassified variable)
- PS 8: included all the study variables

The estimated PSs were then used to match patients who received and did not receive CPAP, using a 1:1 nearest neighbor matching algorithm with calipers fixed at 0.2 standard deviation of the logit of the PS. The three global measures of balance described previously were estimated from the original and each matched data set.

A final PS model (PS 9) included the true confounders, the prognostic factors, and the variables that significantly minimized the WBM. For each PS model, 1,000 nonparametric bootstrap replicates were used to estimate the standard deviation of these measures. The average treatment effect on the treated [expressed as risk difference together with their 95% confidence intervals (CIs)] was estimated, and variance estimation was based on the Abadie–Imbens variance estimator [32].

3. Results

3.1. Monte Carlo simulations

The results obtained with the nine PS models are provided in Table 2.

3.1.1. Performance of the estimators

The naive estimator of treatment effect was severely biased (bias = 0.169). The PS matching estimator based on PS 1 was able to remove most of the bias, as reflected by a MSE close to zero (MSE = 0.0015). Adding W_4

PS 5 (+ W_7)	PS 6 (+ W_8)	PS 7 (+ W_9)	PS 8 (+ W_{10})	PS 9 (+ W_{11})
0.2245[a]	0.2519	0.2504	0.2513	0.2519
0.2242, 0.2249	0.2517, 0.2522	0.2502, 0.2507	0.2511, 0.2516	0.2517, 0.2522
0.9802	1.5195	1.5188	1.5161	1.5136
0.9783, 0.9822	1.5179, 1.5211	1.5173, 1.5204	1.5146, 1.5177	1.5121, 1.5152
0.0564	0.0399	0.0392	0.0379	0.0417
0.0562, 0.0566	0.0398, 0.0401	0.0391, 0.0394	0.0378, 0.0380	0.0416, 0.0419
804	793	792	793	793
0.022 (35)	0.017 (28)	0.023 (36)	0.017 (27)	0.017 (27)
0.055	0.032	0.033	0.031	0.034
0.0035	0.0014	0.0017	0.0013	0.0014

(confounder with a weak association with the outcome but strong association with the exposure) further decreased the bias. Hence, as expected, the PS 2 was associated with the smallest bias and MSE. On the contrary, including an IV (W_7) in the PS model (PS 5) was associated with an increase in bias and a decrease in precision as compared with PS 1. Including the covariates $W_{5,6}$ (IVs that had no direct association with the outcome but were correlated with the confounders) (PS 3 and PS 4) did not substantially alter the estimator performance. Similarly, including the prognostic variables $W_{8,9,10}$ (PS 6–8) did not substantially modify the performance of the estimator. Finally, including in the PS model, an independent covariate ($W_{11,}$ PS 9) had no impact on estimator performance.

3.1.2. Balance measures

The distribution of the average standardized absolute mean difference, the Mahalanobis distance, and the WBM is illustrated in Fig. 1. Although the distribution of the average mean difference seems normal, the Mahalanobis distance has a skewed right wide distribution and the WBM has a distribution close to the normal one, with a trend toward skewed left. As expected, including in the PS model, any variable associated with the exposure further decreased the average standardized absolute mean difference as compared with PS 1. Hence, although it impaired the performance of the estimator, including an IV resulted in a significantly smaller value of the standardized absolute mean difference as compared with the reference model PS 1. Similar results were observed with the Mahalanobis distance.

The value of the WBM decreased when including variables associated with the outcome. However, it significantly increased as compared with PS 1, when including an IV in the PS model. The smallest measure value of the WBM was observed for PS 2, which was also the PS model associated with the best performance. Eventually, when altering the simulation procedure to mimic a randomized treatment allocation, the mean of WBM was 0.021 with a normal distribution (Fig. 2).

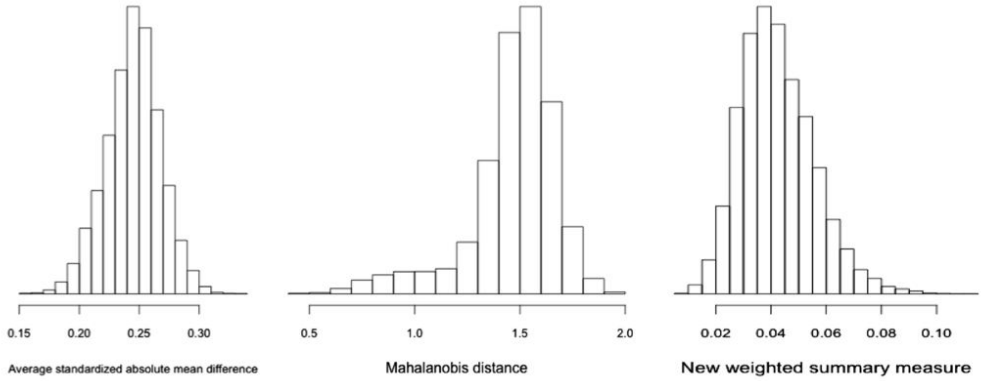


Fig. 1. Distributions of the average standardized absolute mean difference, the Mahalanobis distance, and the new weighted summary measure.

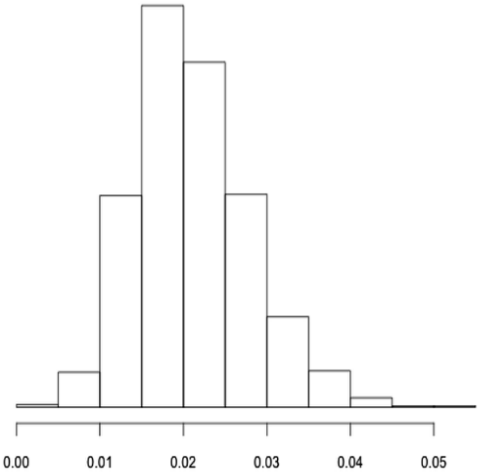


Fig. 2. Distribution of the new weighted summary measure in a randomized trial situation.

3.2. Clinical illustration

As expected from the nonrandomized design, most of the baseline covariates were severely imbalanced between treatment groups (Table 3). Nine PS models were tested. They resulted in different profiles in terms of covariate balance as reflected by the standardized differences in the matched samples. Whatever the PS model, none of the estimates of treatment effect reached statistical significance.

As expected, all balance measures were maximal in the original cohort. The WBM was minimal for the PS 8 (nonparsimonious) and the PS 9 (final PS model) and significantly different from the value obtained with PS1. These two PS models resulted in treatment effect estimates nonsignificantly different from zero, although the absolute value of the point estimate for the risk different was larger for the nonparsimonious model [risk difference: −0.102; 95% CI: −0.28, 0.15 for PS 8 and risk difference: −0.065, 95% CI: −0.32, 0.11 for PS 9]. The nonweighted average measure was also minimal for PS 8 and PS 9. However, the Mahalanobis distance was minimal for the PS 1 (reference model). This latter model lead to a treatment effect estimate very close to the one obtained with the PS9 (risk difference: −0.05, 95% CI: −0.26, 0.16).

4. Discussion

Different PS models, that is, including different sets of covariates, may differ in their ability to achieve conditional exchangeability and thus may yield substantially different estimates [33]. Hence, identification of the best PS model is a crucial endeavor. Because the PS is defined as a balancing score, the best PS model is often considered as the one that achieves the best balance in the distribution of baseline characteristics [34]. However, including an IV in the PS model can in turn decrease the estimator's efficiency [8,10]. A single balance measure that would put more weight on confounders than on IVs could prove useful to select a minimal set of covariates to be included in the PS model, thereby avoiding the inclusion of strong IVs. It is often difficult to distinguish between confounders, prognostic factors,

Table 3. Standardized differences, summary measures, and treatment effect

Variable	Original sample (n = 575)	Reference model PS 1 (n = 278)	+ Instrumental variable PS 2 (n = 272)	+ Prognostic factor PS 3 (n = 262)
Age	−0.064	0.019	0.128	−0.101
NYHA	0.131	0.046	0.078	−0.238
Gender	0.197	0.193	0.287	0.299
Teaching hospital	0.424	0.454	0.449	0.139
Obesity	−0.028	−0.103	0.075	−0.177
Shock	−0.315	0.098	0.099	0.02
Atrial fibrillation	0.107	0.115	0.135	0.07
CPE	0.291	0.172	0.174	0
ACS	0.12	0.029	0.134	0.061
First heart rate	−0.057	0.022	−0.065	−0.157
First SBP	0.29	0.032	−0.012	−0.078
Average standardized absolute mean difference (SD)	0.184 (0.02)	0.116 (9.6e-3)	0.149 (1.0e-3)	0.122 (9.6e-3)
Weighted balance measure (SD)	0.0435 (0.01)	0.0213 (1.7e-3)	0.0281 (2.1e-3)	0.0258 (1.9e-3)
Mahalanobis distance (SD)	182 (129)	2.43 (1.88)	3.50 (5.52)	28.5 (10.0)
Risk difference [95% CI]	−0.093 [−0.34, 0.15]	−0.05 [−0.26, 0.16]	−0.059 [−0.27, 0.15]	−0.031 [−0.24, 0.18]

Abbreviations: PS, propensity score; NYHA, New York Heart Association; CPE: cardiogenic pulmonary edema, ACS acute coronary syndrome, SBP systolic blood pressure (mmHg); SD, standard deviation; CI, confidence interval.
Treatment effect is expressed as risk differences (95% confidence intervals).
[a] Stands for a significant difference in minimization as compared with the balance measure obtained for PS 1.

and IVs based on expert-matter knowledge. We claim that our proposed WBM could help to identify the PS model that will provide the least bias estimate of treatment effect.

Using Monte Carlo simulation, we were able to confirm that the PS model minimizing this WBM was the one that included all confounders but not the IVs. Consistently with previous studies [8], this model was also the one associated with the smallest MSE. The decrease in the matched sample size was probably compensated by an increase in performance of PS 2 prediction accuracy due to the inclusion of a variable correlated with a confounder. Including an IV in the PS model resulted in a significantly larger value of our WBM. This result is reinforced by the fact that we used a complex simulation scenario with correlation among variables. On the contrary, as compared with the reference model, both the nonweighted average standardized absolute mean difference and the Mahalanobis distance were minimized when including an IV in the PS model. Thus, variable selection based on these metrics would lead to including IVs in the PS model. However, as previously reported, such a PS model was associated with a larger MSE [33,35,36]. Whatever the metrics, including a prognostic variable did not substantially alter overall balance estimation and had very limited impact on treatment effect estimate.

Accordingly, in our clinical illustration, we proposed a practical strategy to identify the best PS model. The set of variable included in the best PS model differed according to the balance measure. The final PS model, that is, the one including the true confounders, the prognostic factors, and the variables that significantly minimized WBM, and the nonparsimonious model lead to similar WBM and average standardized absolute mean difference. However, the Mahalanobis distance lead to identify the reference PS model as best model.

This study carries some limitations. First, valid variance estimation and subsequent statistical inference would be of great interest for our measure, to compare several candidate PS models. Given variance estimator is still lacking for our measure, we used nonparametric bootstrap to estimate its variance. Further research is needed to derive its asymptotic properties. Second, our measure has been tested in the context of PS matching. Further work would be needed to assess whether it could be extended to PS weighting estimators. Third, it should be noted that our measure cannot be strictly described as a weighted average, as the sum is divided by the number of covariates k, rather than the sum of the weights. The version chosen here is simpler, and its performance should not be altered as the sum of the weights remains constant whatever the variable included in the PS model. Fourth, the simulation framework used here is different from the one widely used proposed by Austin et al. and thus does not allow direct comparison of the results between studies [33].

In conclusion, selection of the variables to be included in the PS model is of paramount importance. It should be based on subject-matter knowledge but also on appropriate balance metrics. We derived a weighted version of the average standardized absolute mean difference that has proved useful to identify the most parsimonious efficient PS model. Based on these results, we suggest using subject-matter knowledge to select the variables that have to be included in the reference PS model, and thereafter, to include one-by-one the other variables and verify if it is associated with a decrease in our WBM. The variables that minimize this measure should be included in the final

	+ Unclassified variables			Nonparsimonious model	Final model
PS 4 (n = 278)	PS 5 (n = 278)	PS 6 (n = 278)	PS 7 (n = 272)	PS 8 (n = 254)	PS 9 (n = 276)
0.019	−0.118	0	0.038	−0.041	0.055
0.03	−0.015	0.061	0.016	−0.05	0.062
−0.031	0.208	0.252	0.122	−0.084	0.12
0.67	0.488	0.454	0.381	0	0.391
−0.103	0.036	−0.152	−0.141	−0.077	0
0.038	0.077	0	0.12	−0.096	0.038
0.115	0.132	0.202	0	−0.069	0
0.284	−0.103	0.132	0.113	0.117	0.153
0.058	0.146	0.014	0.044	0.016	0.087
−0.027	−0.051	0.1	−0.176	−0.054	0.018
−0.004	−0.058	−0.04	0.07	−0.069	−0.058
0.125 (6.5e-3)	0.13 (9.6e-3)	0.128 (0.01)	0.111 (8.5e-3)[a]	0.061 (6.2e-3)[a]	0.089 (7.9e-3)[a]
0.0224 (1.3e-3)	0.0249 (1.9e-3)	0.0204 (2.0e-3)[a]	0.0211 (1.3e-3)[a]	0.0143 (1.4e-3)[a]	0.0173 (1.6e-3)[a]
0.699 (2.1)	8.50 (4.3)	11.3 (4.6)	34.4 (7.1)	11.3 (2.1)	7.0 (10.3)
−0.079 [−0.29, 0.13]	−0.036 [−0.25, 0.18]	−0.043 [−0.26, 0.17]	−0.044 [−0.26, 0.17]	−0.102 [−0.28, 0.15]	−0.065 [−0.32, 0.11]

PS model, although those which do not alter or increase it should probably be discarded.

Acknowledgment

The authors thank Prof Alexandre Mebazaa and all the EFICA investigators who provided the data set.

References

[1] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70:41.

[2] Ali MS, Groenwold RHH, Pestman WR, Belitser SV, Roes KCB, Hoes AW, et al. Propensity score balance measures in pharmacoepidemiology: a simulation study: propensity score balance measures. Pharmacoepidemiol Drug Saf 2014;23:802–11.

[3] Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat Med 2009;28:3083–107.

[4] Drake C. Effects of misspecification of the propensity score on estimators of treatment effect. Biometrics 1993;49:1231–6.

[5] Kang JDY, Schafer JL. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. Stat Sci 2007;22:523–39.

[6] Smith JA, Todd PE. Does matching overcome LaLonde's critique of nonexperimental estimators? J Econom 2005;125:305–53.

[7] Rubin DB, Thomas N. Matching using estimated propensity scores: relating theory to practice. Biometrics 1996;52:249.

[8] Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. Am J Epidemiol 2006;163:1149–56.

[9] Robins JM, Mark SD, Newey WK. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. Biometrics 1992;48:479–95.

[10] Myers JA, Rassen JA, Gagne JJ, Huybrechts KF, Schneeweiss S, Rothman KJ, et al. Effects of adjusting for instrumental variables on bias and precision of effect estimates. Am J Epidemiol 2011; 174:1213–22.

[11] Westreich D, Cole SR, Funk MJ, Brookhart MA, Stürmer T. The role of the c-statistic in variable selection for propensity score models. Pharmacoepidemiol Drug Saf 2011;20:317–20.

[12] Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. Stat Med 2008;27: 2037–49.

[13] Austin PC. Assessing balance in measured baseline covariates when using many-to-one matching on the propensity-score. Pharmacoepidemiol Drug Saf 2008;17:1218–25.

[14] Belitser SV, Martens EP, Pestman WR, Groenwold RHH, Boer A, Klungel OH. Measuring balance and model selection in propensity score methods: balance measure for propensity scores methods. Pharmacoepidemiol Drug Saf 2011;20:1115–29.

[15] Groenwold RHH, Vries F, Boer A, Pestman WR, Rutten FH, Hoes AW, et al. Balance measures for propensity score methods: a clinical example on beta-agonist use and the risk of myocardial infarction: balance measures for propensity score methods. Pharmacoepidemiol Drug Saf 2011;20:1130–7.

[16] Franklin JM, Rassen JA, Ackermann D, Bartels DB, Schneeweiss S. Metrics for covariate balance in cohort studies of causal effects. Stat Med 2014;33:1685–99.

[17] Stuart EA, Lee BK, Leacy FP. Prognostic score–based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. J Clin Epidemiol 2013;66: S84–90.e1.

[18] Kuss O. The z-difference can be used to measure covariate balance in matched propensity score analyses. J Clin Epidemiol 2013;66: 1302–7.

[19] Tritchler D. Interpreting the standardized difference. Biometrics 1995;51:351–3.

[20] Cohen J. Statistical Power Analysis for the Behavioral Sciences (2nd edn). Hillsdale, NJ: Lawrence Erlbaum Associates Publishers; 1988.

[21] Rubin DB. Bias reduction using Mahalanobis-metric matching. Biometrics 1980;36:293.

[22] Bring J. How to standardize regression coefficients. Am Stat 1994;48: 209–13.

[23] Setoguchi S, Schneeweiss S, Brookhart MA, Glynn RJ, Cook EF. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. Pharmacoepidemiol Drug Saf 2008;17:546–55.

[24] Imbens GW. Nonparametric estimation of average treatment effects under exogeneity: a review. Rev Econ Stat 2004;86:4–29.

[25] Abadie A, Imbens GW. Bias-corrected matching estimators for average treatment effects. J Bus Econ Stat 2011;29:1–11.

[26] Sekhon JS. Multivariate and propensity score matching software with automated balance optimization: the matching package for R. J Stat Software 2008;42(7).

[27] Zannad F, Mebazaa A, Juillière Y, Cohen-Solal A, Guize L, Alla F, et al. Clinical profile, contemporary management and one-year mortality in patients with severe acute heart failure syndromes: the EFICA study. Eur J Heart Fail 2006;8:697–705.

[28] Mebazaa A, Parissis J, Porcher R, Gayat E, Nikolaou M, Boas FV, et al. Short-term survival by treatment among patients hospitalized with acute heart failure: the global ALARM-HF registry using propensity scoring methods. Intensive Care Med 2011;37:290–301.

[29] Tremblay A, Bandi V. Impact of body mass index on outcomes following critical care. Chest 2003;123:1202–7.

[30] El-Solh A, Sikka P, Bozkanat E, Jaafar W, Davies J. Morbid obesity in the medical ICU. Chest 2001;120:1989–97.

[31] Goulenok C, Cariou A. Obésité en réanimation, définition, épidémiologie, pronostic. Réanimation 2006;15:421–6.

[32] Abadie A, Imbens GW. Large sample properties of matching estimators for average treatment effects. Econometrica 2006;74:235–67.

[33] Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. Stat Med 2007;26:734–53.

[34] Imai K, Ratkovic M. Covariate balancing propensity score. J Roy Stat Soc B 2014;76:243–63.

[35] Wooldridge J. Should instrumental variables be used as matching variables. MI: Michigan State University; 2009.

[36] Pearl J. Invited commentary: understanding bias amplification. Am J Epidemiol 2011;174:1223–7.

Appendixes

Appendix A

Let consider a binary exposure A, representing two treatment groups, one treated group ($A = 1$) and one control group ($A = 0$), from one experiment; a binary outcome measure, Y and a vector W of k distinct covariates.

The standardized absolute mean difference, also referred to as the Cohen's d effect size, is the difference in population means divided by the average population standard deviation for each of the k continuous covariates.

$$d_i = \frac{\left(\mu_i^{A=1} - \mu_i^{A=0}\right)}{\sigma_i}; i = 1, \ldots, k \quad (1)$$

They are estimated as the ratio of the difference in sample means for each covariate between the treatment and the control groups over the pooled standard deviation as estimated on the whole sample by:

$$d_i = \frac{\left(p_i^{A=1} - p_i^{A=0}\right)}{\sqrt{\frac{\left(p_i^{A=1}\left(1-p_i^{A=1}\right)+p_i^{A=0}\left(1-p_i^{A=0}\right)\right)}{2}}} \quad (2)$$

For the k covariates binary in scale, the standardized difference was defined as:

$$d_i = \frac{\left(p_i^{A=1} - p_i^{A=0}\right)}{\sqrt{\frac{\left(p_i^{A=1}\left(1-p_i^{A=1}\right)+p_i^{A=0}\left(1-p_i^{A=0}\right)\right)}{2}}}, i = 1, \ldots, k$$

with $p_i^{A=1}$ and $p_i^{A=0}$ denote the prevalence of dichotomous variable in treated and untreated subjects, respectively.

The first proposed summary measure of imbalance is the average standardized absolute mean difference estimated as:

$$\widehat{M}_{\text{ASAM}} = \frac{1}{k}\sum_{i=1}^{k}\left|\widehat{d}_i\right| \quad (3)$$

The second proposed measure is the Mahalanobis distance estimated as:

$$\widehat{M}_m = \left(\overline{W}^{A=1} - \overline{W}^{A=0}\right)' \Sigma^{-1}\left(\overline{W}^{A=1} - \overline{W}^{A=0}\right) \quad (4)$$

where $\overline{W}^{A=1}$ and $\overline{W}^{A=0}$ are the vectors of covariate sample means in treatment and control groups, and Σ is the sample variance–covariance matrix of the covariates.

We proposed a weighted balance measure where each standardized mean difference is weighted according to its prognostic value, as follows:

$$W\widehat{B}M = \frac{1}{k}\sum_{i=1}^{k}\left|\widehat{\omega}_i \times \widehat{d}_i\right| \quad (5)$$

where ω_i is a weight accounting for the association between the covariate W_i and the outcome. Specifically is the coefficient obtained for the covariate W_i from a multivariate maximum likelihood regression of Y on the W:

$$\widehat{\omega}_i = \widehat{\beta}_i \times \text{SD}_i \quad (6)$$

where SD_i is the sample standard deviation for the covariate W_i in the original cohort [22]. When the outcome is binary, the proposed weights are directly derived from the estimated standardized coefficients obtained from a multivariate logistic regression of the outcome on the covariates.

Appendix B

Let Y be the binary outcome, A be the binary exposure and W a vector of 11 covariates (four confounders, $W_{1,\ldots,4}$ associated with both exposure and outcome, three exposure predictors $W_{5,6,7}$, three outcome predictors $W_{8,9,10}$, and one without association with exposure and outcome W_{11}).

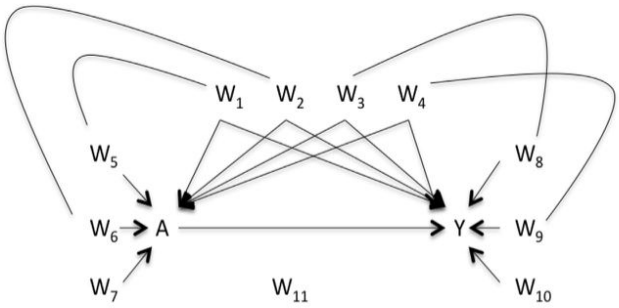
For each simulated data set, the 11 covariates W_i ($i = 1,\ldots, 11$) were generated from a normal distribution with zero mean and unit variance. First, eight covariates (V_i, $i = 1\ldots6, 8, 9$) were generated as independent standard normal random variables. Second, another final eight covariates (W_i, $i = 1\ldots6, 8, 9$) were generated as a linear combination of V_i, $i = 1\ldots6, 8, 9$. In addition, three covariates (W_7, W_{10}, and W_{11}) were generated as independent standard normal random variables. In the second step, correlations between some of the variables were introduced, with correlation coefficients varying from 0.1 to 0.2. To attenuate the magnitude of correlation coefficients, a dichotomization was realized (according to the mean of each variable). Eventually, these variables (W_1, W_3, W_5, W_6, W_8, and W_9) were thus considered binary in scale.

The coefficients used for data generation are:

Coefficient	True PS model (β)	Outcome model (α)
Intercept	0	0.8
Coefficient 1	0.2	1.2
Coefficient 2	0.4	0.8
Coefficient 3	0.8	0.4
Coefficient 4	−1.2	0.2
Coefficient 5	−1.2	−1.2
Coefficient 6	−1.2	−1.2
Coefficient 7	−1.8	−1.2

The correlation matrix for data generation models and the corresponding directed acyclic graph is displayed below.

	Confounders				Exposure predictors			Outcome predictors		
Properties of covariates	W_1	W_2	W_3	W_4	W_5	W_6	W_7	W_8	W_9	W_{10}
Confounders										
W_1	1									
W_2	0	1								
W_3	0	0	1							
W_4	0	0	0	1						
Exposure predictors										
W_5	0.1	0	0	0	1					
W_6	0	0.2	0	0	0	1				
W_7	0	0	0	0	0	0	1			
Outcome predictor										
W_8	0	0	0.1	0	0	0	0	1		
W_9	0	0	0	0.1	0	0	0	0	1	
W_{10}	0	0	0	0	0	0	0	0	0	0



Treatment status: A
Confounders: W_1,W_2,W_3,W_4
Outcome predictos: W_8,W_9,W_{10}

Outcome: Y
Treatment predictors: W_5,W_6,W_7
No relation with A/Y: W_{11}

The probability of exposure, that is, the true PS, was generated as a function of the covariates W_i:

$$\Pr(A = 1|W_i) = f(W_i, \beta)$$

with f corresponding to a linear and additive function on log odds scale, guarantying the true PS to be bounded away from 0 to 1 and the average exposure probability to be of approximately 0.4. The vector A was then drawn from a Bernoulli distribution with probability set by $f(W_i, \beta)$.

Similarly, the binary outcome Y was generated from a linear combination of A and W_i:

$$\Pr(Y = 1|A, Wi) = f(A, W, \gamma, \alpha)$$

with f a logistic regression model, guarantying the true outcome probability to be bounded away from 0 to 1 and the average outcome probability to be approximately of 0.5. The effect of exposure γ was fixed at -0.4. The vector Y was drawn from a Bernoulli distribution with probability set by $f(A, W, \gamma, \alpha)$.

4 ANALYSE DE PROPENSION ÉTUDIANT L'EFFET DE LA PRESSION CRICOÏDIENNE SUR L'EXPOSITION LARYNGÉE LORS DE L'INTUBATION OROTRACHÉALE EN PRÉHOSPITALIER

4.1 Introduction

La prise en charge des voies aériennes supérieures en pré-hospitalier peut-être particulièrement difficile étant données les mauvaises conditions d'intubation trachéale (IT). La pression cricoidïenne (PC), également appelée la manoeuvre de Sellick, a été décrite en 1961 pour prévenir la régurgitation du contenu gastrique durant l'induction anesthésique [83]. Cette manœuvre a pour but d'occlure transitoirement le segment supérieur de l'œsophage par l'application d'une pression postérieure sur le cartilage cricoïdien contre le rachis cervical. Une telle pression cricoïdienne permet la compression de l'hypopharynx rétrocricoïdien et permettrait de prévenir les régurgitations gastriques dans le pharynx. Les recommandations pour la réanimation cardiopulmonaire et les soins d'urgence cardiovasculaire recommandent, en dehors de l'arrêt cardiaque [84], l'utilisation de la PC pour l'intubation dans toutes les situations, notamment pour les patients traumatisés. Cependant, le bénéfice escompté et donc l'utilisation de la PC sont toujours débattus dans la littérature [85]. De plus, plusieurs études ont rapporté un impact négatif de cette pression sur la perméabilité des voies aériennes et ainsi sur la ventilation [86]. Les résultats concernant l'impact de la PC sur l'exposition laryngée (conditionnant la difficulté d'intubation) sont contradictoires et proviennent soit d'études sur cadavres, ou d'études de petite taille incluant des patients intubés dans des blocs opératoires pour une chirurgie programmée [87]. Ainsi, il manque des données robustes quant à l'impact de la PC sur l'exposition laryngée durant l'intubation orotrachéale en urgence. Dans ce contexte, la preuve d'une association entre l'utilisation de la PC et une augmentation du taux de laryngoscopies difficiles pourrait être un argument fort contre son utilisation en

pratique courante. L'objectif principal de cette étude était d'utiliser une méthode type score de propension pour estimer l'impact de la PC sur le taux de laryngoscopies difficiles en se basant sur une base de données observationnelle de patients intubés en dehors de l'hôpital.

4.2 Méthodes

4.2.1 Schéma de l'étude

Tous les patients intubés entre Avril 2008 et Novembre 2012 par l'équipe médicale du service mobile d'urgence et de réanimation d'un hôpital universitaire ont été inclus prospectivement dans cette cohorte. Les critères d'exclusion étaient : un âge inférieur à 18 ans, une contre-indication à la succinylcholine (i.e une allergie connue, l'hyperthermie maligne, une myopathie, la tétraplégie, la pré-éclampsie, l'hyperkaliémie), à la kétamine ou l'étomidate. Cette étude a été déclarée à la commission nationale de l'informatique et des libertés (numéro de déclaration 167412v0).

4.2.2 Protocole

L'équipe médicale d'urgence était composée d'au moins un ambulancier, d'un infirmier anesthésiste et d'un sénior urgentiste. Les internes en dernière année de formation pouvaient aussi participer à l'équipe. Tous les médecins étaient titulaires du diplôme de médecine d'urgence et avaient au moins deux années d'expérience en médecine d'urgence en plus de leurs trois années d'internat. Ils recevaient tous une formation théorique et pratique sur la gestion de l'intubation difficile. Tous les internes devaient accomplir une formation initiale d'entrainement à l'intubation en bloc opératoire. Pour être complet cet entrainement devait comporter au moins dix intubations réussies successives. L'équipe suivait une procédure d'intubation standardisée : la position du patient était optimisée pour l'intubation trachéale (IT) en plaçant le patient sur un brancard à hauteur maximale avant le geste. Après 3 minutes

de pré-oxygénation, tous les patients recevaient une induction à séquence rapide (ISR) avec soit de l'étomidate (0.3 mg/kg), de la kétamine (2 mg/kg) ou du thiopental (3-5 mg/kg) en association avec la succinylcholine (1-1.5 mg/kg). Cette procédure était basée sur les recommandations de la société française de médecine d'urgence. Une PC pouvait être réalisée sur décision médicale. Les patients en arrêt cardiaque étaient intubés sans pré-oxygénation ni induction anesthésique, leur position était optimisée sur un brancard à hauteur maximale si les conditions le permettaient. Une lame de laryngoscope était utilisée (Macinstosh taille 4). L'algorithme d'intubation était basé sur les recommandations de la société française d'anesthésie et de réanimation [88] qui comportent l'utilisation de manœuvres laryngées (la pression à droite vers le haut et vers l'arrière) et l'utilisation d'une technique alternative comme le mandrin d'Eschmann et/ou le masque laryngé dans les situations d'intubation difficile ou d'échec. La courbe de capnographie était utilisée chez tous les patients.

4.2.3 Recueil des données

Le médecin responsable de l'intervention recueillait les données. Les variables collectées comportaient : l'âge, le sexe, l'obésité (0:absence, 1:présence) définie par un indice de masse corporelle supérieur à 30, les indications de l'intubation orotrachéale (arrêt cardiaque, statut neurologique altéré, i.e., coma/agitation/intoxication/traumatisme, détresse respiratoire aiguë et état de choc), la position des patients durant la procédure d'intubation (au sol ou sur un brancard à hauteur maximale), le statut de l'opérateur (sénior ou interne), la présence d'une immobilisation cervicale. La PC était réalisée d'après la méthode décrite par Sellick [83] par un infirmier anesthésiste avec une expérience en médecine préhospitalière et en bloc opératoire. Cette procédure imposait de bien reconnaître le cartilage cricoïde avec le pouce et le majeur, et comportait l'application d'une pression continue maintenue par l'index. Une force

d'application de 30 Newtons était souhaitée [89] mais ne pouvait être mesurée en dehors de l'hôpital.

4.2.4 Critère de jugement

Le critère de jugement principal était le taux de laryngoscopie difficile définie sur l'échelle de Cormack et Lehane [90] (CL). Le score de CL classe les vues laryngoscopiques en 4 grades (I : vue de toute la glotte, II : vue partielle de la glotte, III : vue isolée de l'épiglotte et non de la glotte, IV : épiglotte et glotte non vues). Le médecin qui réalisait l'IT reportait le grade de CL. Une laryngoscopie difficile était définie par un grade de CL III ou IV [91].

Les critères secondaires étaient le taux d'intubation difficile (ID) définie par au moins deux tentatives de laryngoscopies ou la nécessité d'utiliser une technique alternative pour l'IT (comme définie par la société française d'anesthésie et de réanimation [88]) ; le nombre de complications liées à l' IT jusqu'à 5 minutes après l'intubation : la désaturation (SpO_2<90%), l'inhalation (régurgitation visible durant la laryngoscopie), le vomissement, le bronchospasme et/ou laryngospasme, l'intubation sélective, l'intubation oesophagienne (diagnostiquée en utilisant la courbe de capnographie/cpanométrie sur 6 cycles respiratoires combinée à l'auscultation pulmonaire) et les traumatismes des voies aériennes supérieures.

4.2.5 Analyses statistiques

Une approche par score de propension (SP) a été utilisée pour prendre en compte les possibles biais de sélection dans la réalisation de la PC [16]. Le SP était estimé en utilisant un modèle de régression logistique avec la PC comme variable dépendante à expliquer.

Plusieurs modèles de SP ont été évalués, chacun caractérisé par un ensemble de variables explicatives différent :

- le premier modèle de SP (SP 1) incluait toutes les variables

- le second modèle de SP (SP 2) incluait les « vrais» confondeurs et les facteurs pronostiques, mais pas les variables dites instrumentales [32]. En se basant sur les données de la littérature [92],[93],[94] et sur les associations statistiques multivariées déterminées empiriquement à partir de notre échantillon, il a été identifié quatre vrais confondeurs (l'arrêt cardiaque, le statut neurologique altéré, l'état de choc, la détresse respiratoire), trois facteurs pronostiques (le sexe, l'obésité, la position du patient), et aucune variable instrumentale. Les sept variables ont été incluses dans le modèle de SP 2.

- le troisième modèle de SP (SP 3) incluait seulement les variables pronostiques statistiquement significatives après utilisation d'un modèle de régression logistique multivarié utilisant une procédure de sélection des variables basée sur le critère d'information d'Akaike. Trois variables ont été sélectionnées et incluses dans le modèle SP 3 : le sexe, la position du patient et l'obésité.

Les trois différents modèles de SP ont été ensuite utilisés pour apparier les patients qui bénéficiaient d'une PC et ceux qui n'en bénéficiaient pas, en utilisant un algorithme d'appariement 1:1 avec remise (autorisant les ex-aequo).

L'équilibre de distribution des variables initiales était vérifié en utilisant la différence des moyennes standardisées (DMS), (une DMS supérieure à 0.1 (10%) traduisait un déséquilibre résiduel entre les covariables [59]), la valeur absolue de la moyenne des différences standardisées (Average Standardized Absolute Mean differences, ASAM) et la mesure d'équilibre pondérée (WBM) développée dans notre précédent travail. Le meilleur modèle de

SP était celui qui permettait d'obtenir le meilleur équilibre des covariables et donc minimisait la WBM.

L'effet moyen du traitement chez les traités exprimé sous la forme d'une différence de risque avec son intervalle de confiance à 95% était le paramètre d'intérêt. L'estimation de la variance était basée sur l'estimateur robuste de la variance d'Abadie-Imbens [81]. Cinquante-six patients (4%) ayant des données manquantes ont été exclus de l'analyse. Les données manquantes concernaient la variable obésité, non-disponible pour 43 patients (3%, 27 non exposés à l'intervention et 16 patients exposés).

Toutes les tests statistiques étaient bilatéraux, et une valeur p <0.05 était considérée comme statistiquement significative. Une correction de Bonferroni était utilisée pour prendre en compte l'inflation d'erreur de type I liée aux comparaisons multiples des analyses univariées. Toutes les analyses statistiques ont été réalisées avec le logiciel statistique R.3.2.0 en utilisant une plateform Mac OsX.

4.3 Résultats

4.3.1 Population d'étude

Pendant la période d'étude 1 291 patients ont été intubés. Trente-deux patients âgés de moins de 18 ans, huit adultes ayant une contre-indication à la succynylcholine (utilisation de rocuronium à la place), et 56 patients avec des données manquantes ont été exclus de l'analyse finale qui a inclus un total de 1 195 patients (figure 7). Les caractéristiques initiales sont présentées dans le tableau 4. Sur l'échantillon original, les principales différences concernaient les variables suivantes : le statut neurologique altéré (59% dans le groupe avec PC contre 8% dans le groupe non exposé, DMS=1.05), la position des patients lors de l'IT (56% dans le groupe avec PC contre 12%, DMS=0.88), et l'âge (moyenne=63 (ET=18) dans le groupe avec PC contre moyenne=57 (ET=18) dans le groupe non-exposé, DMS=-0.33). Comme attendu, les patients intubés pour un arrêt cardiaque étaient plus nombreux dans le groupe non exposé à la PC (83% contre 18%, DMS=-1.75).

4.3.2 Appariement sur le score de propension et mesures d'équilibre

La différence des moyennes standardisées (DMS) sur l'échantillon original et l'échantillon apparié sont représentées dans le tableau 2. Le meilleur équilibre global a été obtenu pour le modèle de SP 2 qui incluait les vrais confondeurs et les facteurs pronostiques (figure 8). La mesure résumée WBM était minimale pour ce modèle SP 2 (WBM=0.41), ainsi que la valeur absolue de la moyenne des différences standardisées (ASAM=0.030). Ce dernier a donc été sélectionné comme le meilleur modèle de SP pour la suite des analyses. L'âge était la seule variable qui restait légèrement déséquilibrée après appariement (DMS=-0.15 pour le modèle SP 2).

Figure 7: Diagramme de l'étude

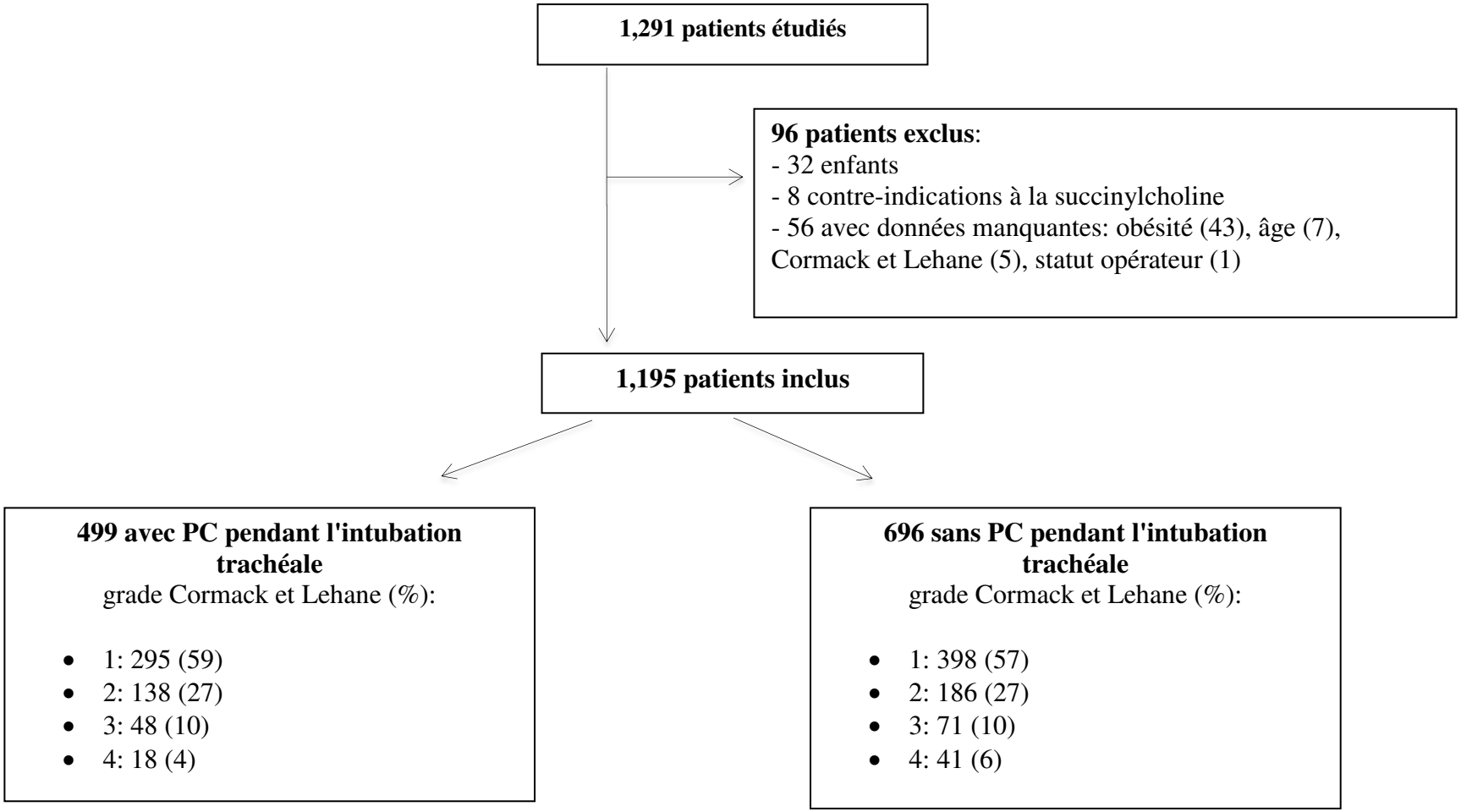


Figure 8: Différences standardisées avant et après appariement sur le score de propension

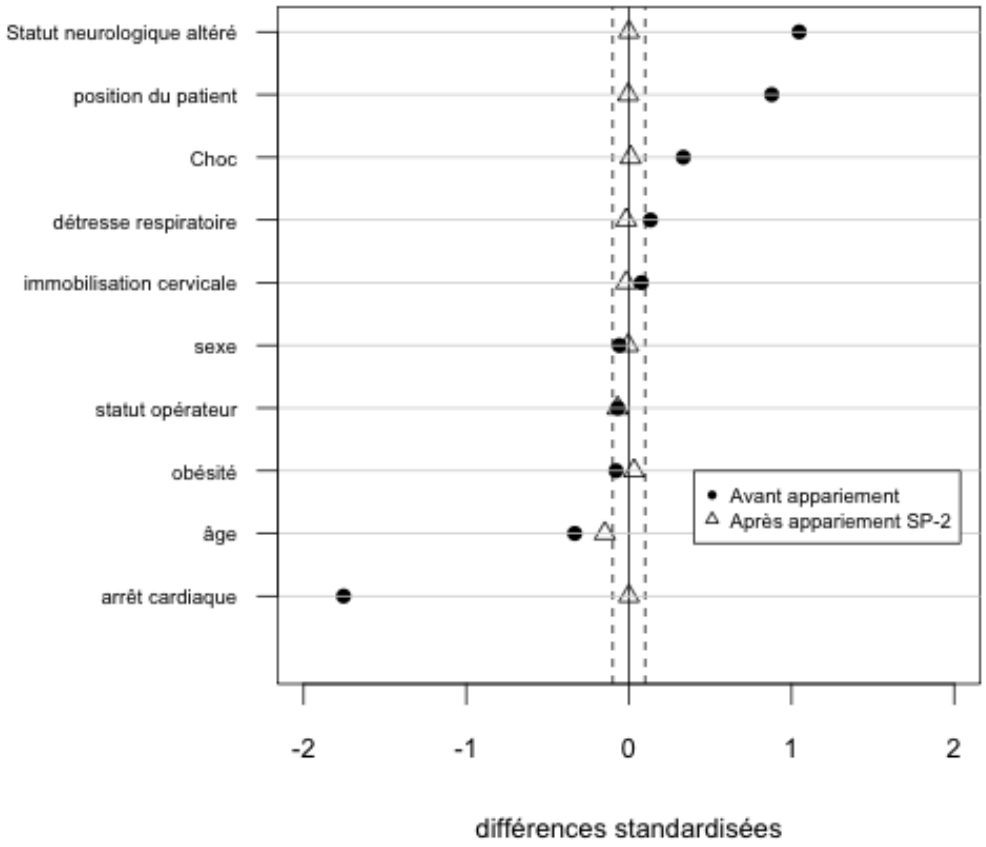


Tableau 2 : Caractristiques initiales d'équilibre dans l'échantillon original et dans chaque échantillon apparié.

[a] La différence des moyennes standardisées (DMS) la différence de moyenne entre les groupes de traitement divisée par une estimation de l'écart-type intra-groupe

[b] Brancard en position haute

* signale les variables incluses dans chaque modèle de SP

Variables	échantillon original (n=1195)			échantillon apparié SP 1	échantillon apparié SP 2	échantillon apparié SP 3
	sans pression cricoïdienne (n=696)	avec pression cricoïdienne (n=499)	DMS[a]	DMS [a]	DMS [a]	DMS [a]
Age (moyenne, ET)	63 (18)	57 (18)	-0.33	-0.02*	-0.15	-0.22
Sexe (femme) (n,%)	245 (35)	190 (38)	-0.06	0.09*	0	0*
Obésité (n,%)	97 (14)	57 (11)	-0.08	-0.03*	0.03*	0*
Statut opérateur (senior) (n,%)	318 (46)	211 (42)	-0.07	0.04*	-0.07*	-0.09
Position patient [b] (n,%)	87 (12)	280 (56)	0.88	0*	0*	0*
Immobilisation cervicale (n,%)	55 (8)	51 (10)	0.08	-0.04*	-0.02	-0.10
Indication d'intubation (n,%)						
arrêt cardiaque	576 (83)	89 (18)	-1.75	0*	0*	-0.90
statut neurologique altéré	53 (8)	295 (59)	1.05	-0.05*	0*	0.73
détresse respiratoire aiguë	22 (3)	32 (6)	0.13	-0.02*	-0.02*	-0.15
choc	29 (4)	83 (17)	0.33	0.09*	0.01*	0.05

	Weighted Balance Measure			
	échantillon original (n=1195)	**échantillon apparié SP 1**	**échantillon apparié SP 2**	**échantillon apparié SP 3**
	29.70	1.48	0.41	15.58

	Valeur absolue de la moyenne des différences standardisées			
	échantillon original	**SP 1**	**SP 2**	**SP 3**
	0.476	0.038	0.030	0.22

4.3.3 Critères de jugement

Les estimations d'effet de la PC dans l'échantillon original et l'échantillon apparié sur le SP 2 sont présentées dans le tableau 3. Dans l'échantillon original (différence de risques de -0.03, 95% IC [-0.17 ; 0.18], p=0.37) comme dans l'échantillon apparié (différence de risques de 0.001, 95% IC [-0.07 ; 0.08], p=0.50), le taux de laryngoscopie difficile ne différait pas entre les groupes de traitement. En ajustant sur l'âge pour tenir compte du déséquilibre résiduel, les résultats ne furent pas modifiés (différence de risques de 0.01, 95% IC [-0.06 ; 0.08], p=0.39). Le risque d'intubation difficile était similaire dans le groupe exposé à la PC et dans le groupe non-exposé dans l'échantillon original (différence de risques = -0.02, 95% IC [-0.22 ; 0.17], p=0.43), de même que dans l'échantillon apparié (différence de risques =0.06, 95% IC [-0.13,0.25], p=0.28). Un total de 195 complications liées à l'intubation ont eu lieu chez 153 patients (13%). Dans l'échantillon original (avant appariement), les complications étaient plus fréquentes dans le groupe exposé à la PC (n=84 (17%) contre 69 (10%), p<0.001; différence de risques = 0.07, 95 % IC [-0.05, 0.19], p=0.13). L'intubation œsophagienne (n=29, 6%) et la désaturation (n=40, 8%) étaient les complications les plus fréquentes. Après appariement il n'y avait pas de différence en termes de complications entre les patients exposés et non-exposés à la PC excepté pour les lésions traumatiques des voies aériennes qui étaient plus fréquentes dans le groupe avec PC avec 3.6% comparé à 5% dans l'autre groupe (différence de risques = 0.03, 95% IC [0.002,0.05], p = 0.01).

Tableau 3: Laryngoscopie difficile, complications et intubation difficile en fonction de la pression cricoidienne pendant l'intubation trachéale.

	Echantillon global			Echantillon apparié SP 2
Critère de jugement	**Sans pression cricoïdienne (n=696)**	**Avec pression cricoïdienne (n=499)**	**p value***	**Différence de risque [95% IC]**
Laryngoscopie difficile (n,%)	112 (17)	66 (13)		
Différence de risques [95% IC]	-0.03 [-0.07,0.01]			0.001 [-0.07,0.08]
Complication (n,%)	69 (10)	84 (17)	< 0.001	0.04 [-0.06,0.15]
Intubation oesophagienne	34 (5)	29 (6)	0.47	0.05 [-0.03,0.13]
Désaturation	14 (2)	40 (8)	< 0.001	-0.005 [-0.05,0.04]
Intubation sélective	16 (2)	18 (4)	0.17	0.016 [-0.02,0.05]
Traumatisme	4 (1)	11 (2)	0.01	0.03 [0.002,0.05]
Vomissement, régurgitation	5 (1)	14 (3)	0.004	0.02 [-0.006,0.05]
Bronchospasme, Laryngospame	6 (1)	1 (0)	0.27	-0.005 [-0.03,0.02]
Intubation difficile (n,%)	283 (41)	190 (38)	0.36	0.06 [-0.13,0.25]

* correction de Bonferroni pour les tests multiples à appliquer (i.e. multiplier les p values par le nombre de comparaisons, à savoir 8 au total).
L'effet traitement est exprimé en différence de risques (Intervalle de confiance à 95 %). La laryngoscopie difficile était définie par un grade de Cormack III ou IV obtenu par laryngoscopie directe et basé sur les structures visibles. L'intubation difficile était définie par au moins deux tentatives échouées, ou l'utilisation d'une méthode alternative.

4.4 Discussion de l'étude clinique

En dépit d'un manque de preuve, la pression cricoidienne (PC) est largement utilisée lors de l'induction anesthésique pour réduire l'incidence des inhalations de contenu gastrique. En médecine d'urgence, l'efficacité de la PC pour prévenir les inhalations reste à ce jour à documenter [95]. De plus, l'utilisation de la PC pendant l'intubation d'urgence pourrait être associée à une mauvaise exposition laryngée [96]. L'objectif principal de cette étude était d'évaluer l'impact de la PC sur l'exposition laryngée en se basant sur une grande base de données observationnelle. En utilisant l'appariement basé sur le score de propension, aucune relation entre l'utilisation de la PC et le taux d'exposition laryngée difficile, définie par un grade de Cormack et Lehane III ou IV [90], n'a été trouvée. Ce résultat est en accord avec un travail précédent d'une étude randomisée contrôlée réalisée au bloc opératoire qui n'objectivait aucune relation entre le grade de Cormack et Lehane et l'utilisation de la PC [97]. Dans l'échantillon original (avant l'appariement), les complications liées à l'intubation trachéale étaient plus fréquentes dans le groupe de patients non-exposés à la PC. Ceci pouvait s'expliquer par la prévalence plus importante de patients en arrêt cardiaque reconnus pour être associés avec des meilleures conditions d'intubation. Cependant, l'intubation œsophagienne était plus fréquente dans le groupe des patients en arrêt cardiaque non-exposés à la PC. Une position non optimale d'intubation pouvait en partie expliquer ces résultats (brancard en position haute pour 3% de ces patients contre 83 % pour les patients exposés à la PC sans arrêt cardiaque, $p<0.001$). En effet le positionnement optimal du patient avec un brancard en position haute est associé à de meilleures conditions d'intubation. Dans l'échantillon apparié, les lésions traumatiques des voies aériennes étaient plus fréquentes dans le groupe des patients avec PC. Aucune différence de risques significative n'était observée au niveau des inhalations pulmonaires. Cependant, étant donné sa faible incidence, cette étude n'avait sans

doute pas la puissance suffisante pour conclure sur l'effet de la PC sur la prévention des inhalations de contenu gastrique.

4.5 Article

EMJ Online First, published on September 30, 2016 as 10.1136/emermed-2016-205715
Original article

Effect of cricoid pressure on laryngeal view during prehospital tracheal intubation: a propensity-based analysis

Emmanuel Caruana,[1,2] Sylvie Chevret,[1] Romain Pirracchio[2,3,4]

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/emermed-2016-205715).

[1]Service Mobile d'Urgences et Réanimation, Hôpital Beaujon, Clichy, France
[2]Service de Biostatistiques et Informatique Médicale, Hôpital Saint-Louis, AP-HP; INSERM 1153 ECSTRA Team; Université Paris 7 Diderot, Paris, France
[3]Service de Biostatistique et Information Médicale, Hôpital Européen Georges Pompidou, Université Paris V Descartes—Sorbonne Paris Cité, Paris, France
[4]Department of Anesthesia and Perioperative Care, San Francisco General Hospital & Trauma Center, University of California San Francisco, San Francisco, California, USA

Correspondence to
Dr Emmanuel Caruana, A&E Department, Beaujon hospital university AP-HP, 100, boulevard du Général Leclerc, Clichy 92110, France; manucaru@hotmail.com

Received 15 January 2016
Revised 11 September 2016
Accepted 12 September 2016

ABSTRACT

Background The benefit of cricoid pressure during tracheal intubation is still debated and, due to its potential negative impact on laryngeal views, its routine use is questioned. The goal of this study was to estimate its impact on laryngeal view.

Methods All patients intubated in the prehospital setting were included. Three different propensity score (PS) models were used and compared in terms of the balance achieved between those patients who received cricoid pressure and those who did not. The PS model that optimised the balance was retained in order to estimate the relationship between cricoid pressure and the following outcomes: difficult laryngoscopy, intubation-related complications and difficult intubation.

Results Among the 1195 patients included, 499 (41.7%) received cricoid pressure. The optimal PS included seven variables (cardiac arrest, altered neurological status, shock, respiratory distress, gender, obesity, patient's position). After PS matching, no significant risk difference (RD) in the rate of difficult laryngoscopy was found between the patients who received cricoid pressure and those who did not (RD=0.001, 95% CI −0.07 to 0.08, p=0.50). No significant difference was found in terms of difficult intubation (RD=0.06, 95% CI −0.13 to 0.25, p=0.28) and in terms of prevalence of complications, except for airway trauma that were more frequent in cricoid pressure group (RD=0.03, 95% CI 0.002 to 0.05, p=0.01).

Conclusions No significant relationship was found between the use of cricoid pressure for prehospital intubation and difficult laryngoscopy. Cricoid pressure was found to be associated with more airway trauma. This finding could question its routine use.

Key messages

What is already known on this subject?
- Cricoid pressure is widely used to reduce the incidence of gastric content aspiration during anaesthesia induction, but its potential adverse impact on laryngeal views challenged its routine use.

What this study adds?
- Cricoid pressure is widely used to reduce the incidence of gastric content aspiration during intubation, but previous literature has suggested it may have a negative impact on laryngeal views.
- In this study of 1195 patients intubated in the prehospital setting, we used a propensity score to account for potential selection bias in the use of cricoid pressure. Cricoid pressure had no impact on laryngeal views and did not increase the rate of difficult intubations. Cricoid pressure was associated with more airway trauma.

INTRODUCTION

Out-of-hospital airway management can be particularly cumbersome, with poor tracheal intubation conditions. Cricoid pressure, also referred to as the Sellick manoeuvre, was described in 1961 to prevent regurgitation of gastric content during anaesthesia induction.[1] This manoeuvre aims at transiently occluding the upper end of the oesophagus through the application of a backward pressure on cricoid cartilage against the cervical spine. Such cricoid pressure results in the compression of the postcricoid hypopharynx and should prevent gastric regurgitation into the pharynx. Guidelines for cardiopulmonary resuscitation and emergency cardiovascular care recommend the use of cricoid pressure for intubation in all patients, except those in cardiac arrest,[2] and especially in trauma patients. However, the expected benefit and thus the need to use cricoid pressure are still debated in the literature. First, there is clinical evidence that cricoid pressure might fail to prevent aspiration during intubation.[3] In addition, several studies have reported a negative impact on airway patency and therefore on ventilation.[4]

The results concerning the impact of cricoid pressure on laryngeal view are conflicting and come either from cadaver studies or from small studies including patients intubated in the operating room for elective surgery. Hence, there is a lack of robust information on the impact of cricoid pressure on laryngoscopic conditions during urgent intubation. In this context, any credible proof of a link between the use of cricoid pressure and an increase rate of difficult laryngoscopy would be a strong argument against its routine use.

Causal inference methods have been suggested in order to estimate causal relationships between an intervention and an outcome, when the study is potentially confounded by selection bias due to the absence of randomisation.[5] Among other things, propensity score (PS) matching has been shown to well balance measured baseline covariates across treatment groups and thus to better estimate the causal relationship between the exposure and the outcome.

To cite: Caruana E, Chevret S, Pirracchio R. *Emerg Med J* Published Online First: [*please include* Day Month Year] doi:10.1136/emermed-2016-205715

The primary goal of this study was to use a propensity-based approach to estimate the impact of cricoid pressure on the rate of difficult laryngoscopy based on a prospective observational study of patients intubated outside the hospital.

METHODS

Study design

All consecutive patients intubated in the prehospital setting between April 2008 and November 2012 by a French physician-staffed emergency medical services (EMS) unit from a teaching hospital were prospectively included in a registry. The exclusion criteria were anyone aged under 18, contraindications to succinylcholine (ie, known allergy, malignant hyperthermia, myopathy, tetraplegia, preeclampsia and hyperkalemia), ketamine or etomidate. The study was based on this registry, which was declared to the French data protection authority (Commission Nationale de l'Informatique et des Libertés, number of declaration 1674124v0).

Setting

Our EMS team is composed of at least one ambulance driver, an anaesthetic nurse and a senior emergency physician. Residents in their last year of training may also be present. All emergency physicians are board-certified in emergency medicine, that is, they have at least a 2-year experience in emergency medicine in addition to their 3-year residency. They all received theoretical and practical training in emergency airway management. All residents have to attend an initial training for intubation in the operating theatre. To be completed, this training requires at least 10 successive successful intubations.

The team followed a standardised intubation procedure: the patient's position was optimised for tracheal intubation (TI) by placing the patients on a stretcher at full height before intubation. After a 3-min preoxygenation period, all patients received a rapid sequence intubation with either etomidate (0.3 mg/kg), ketamine (2 mg/kg) or thiopentone (3–5 mg/kg) in association with succinylcholine (1–1.5 mg/kg), based on the guidelines from the French Society of Emergency Medicine. The physician in charge decided whether cricoid pressure should be applied or not. Cricoid pressure was performed by a nurse anaesthetist with substantial experience in both EMS and operating theatres, and using the method described by Sellick.[1] This procedure involved careful identification of cricoid cartilage with the thumb and index finger, followed by the application of a steady pressure mainly by the index finger. A force of 30 N was considered desirable[6] but could not be measured. Patients in cardiac arrest were intubated without preoxygenation or induction, their position was optimised on a stretcher at full height whenever possible. Either single-use or reusable laryngoscope blades (Macintosh size 4) were used. The intubation algorithm was based on the guidelines endorsed by the French Society of Anesthesiology and Intensive Care Medicine.[7] After a first attempt, or in case of a suspected difficult airway, the physicians were encouraged to follow a two-step predefined airway management algorithm: first, direct laryngoscopy with an external laryngeal manipulation and/or the use of alternative techniques such as Eschmann bougie; second, in case of failure, tracheal intubation through a laryngeal mask. Waveform capnography was used in all patients.

Data collection

The physicians collected clinical data immediately after the intervention. One of the physicians from the EMS unit was responsible for quality assessment throughout the study. The baseline variables were age, gender, obesity (0: absence, 1: presence) defined by a body mass index over 30 (estimated from the height and weight measures reported by the patient or his/her next of kin when available, or estimated by the physician when unavailable), reasons for intubation (cardiac arrest, altered neurological status, ie, coma/agitation/drug intoxication/trauma, acute respiratory distress and shock), patient's position during the procedure (on the floor or on a stretcher at full height), operator's status (senior or resident) and presence of cervical immobilisation.

Outcomes

The primary outcome was the rate of difficult laryngoscopy as defined by the Cormack and Lehane[8] (CL). The CL score classifies laryngoscopic views in four grades (I, full view of glottis; II, partial view of glottis; III, only epiglottis seen, glottis not seen; IV, neither glottis nor epiglottis seen). The physician who performed the intubation reported the CL grade. A difficult laryngoscopy was defined as a CL grade III or IV.[9]

The secondary outcomes were the rate of difficult intubation as defined by two or more failed laryngoscopic attempts or the need for any alternative methods for intubation (as defined by the French Society of Anesthesiology and Intensive Care Medicine)[7]; and the number of intubation-related complications occurring within 5 min after the procedure: oxygen desaturation (SpO_2<90%), aspiration (regurgitation visualised during the laryngoscopy), vomiting, bronchospasm and/or laryngospasm, mainstem intubation, recognised or unrecognised oesophageal intubation (diagnosed using a waveform capnography/capnometry at sixth breath combined with pulmonary auscultation) and airway trauma.

Statistical analysis

Continuous variables are expressed as mean with SD or median with IQR when appropriate. Counts and percentages are provided for categorical variables.

We used a PS approach to account for selection bias.[5] The PS was defined as the individual probability of receiving cricoid pressure given baseline observed covariates. This probability was estimated using a logistic regression model with cricoid pressure as dependent variable. Several PS models were tested, each characterised by a different set of explanatory variables.

- The *first* PS model (PS 1) included all collected baseline variables.
- The *second* PS model (PS 2) included (i) confounders: variables suspected to be related to both the treatment and the outcome and (ii) prognostic factors: strong predictors of the outcome but not of treatment allocation. A third type of variables, called instrumental variables, are variables that would act as strong predictors of treatment allocation but not of the outcome; however, instrumental variables in the PS model may jeopardise the efficiency of the estimator and thus were not included.[10] Based on subject-matter knowledge and significant statistical association in multivariate analysis, we identified four confounders (cardiac arrest, altered neurological status, shock and respiratory distress) and three prognostic factors (gender, obesity and patient's position). All seven variables were included in the PS 2.
- For the *third* PS model (PS 3), we only included statistically significant prognostic variables, as identified using a multivariate logistic regression model with a stepwise selection procedure based on the Akaike Information Criterion. Three variables were selected and included into the PS model 3: gender, patient's position and obesity.

The three different PS models were then used to match the patients who received cricoid pressure to those who did not receive it. The nearest neighbour matching algorithm allowed for replacement and for ties, so that subjects not exposed to cricoid pressure could be matched to several subjects with cricoid pressure if their PS were equal or close.

The balance in the distribution of baseline covariates was checked using the standardised mean difference (SMD). A SMD greater than 0.1 (10%) denotes meaningful imbalance in baseline covariates.[10] Caruana *et al*[11] recently proposed a weighted balance measure (WBM) that takes into account the strength of association between each covariate and the outcome. We defined the best PS model as the one optimising covariate balance and thereby minimising the WBM.

The average treatment effect on the treated (expressed as a risk difference together with its 95% CI) was defined as the parameter of interest. Variance estimation was based on the Abadie-Imbens robust variance estimator. Fifty-six (4%) patients with missing data were excluded from the analysis. Missing values mainly concerned the variable 'obesity', which was not available in 43 patients (3%, 27 non-exposed patients and 16 exposed patients).

All statistical tests were two-tailed, and $p<0.05$ considered statistically significant. Bonferroni correction (ie, p values multiplied by the number of comparisons, ie, eight) was used to handle the risk of type I error inflation due to multiple comparisons in the univariate analyses. Due to the use of matching with replacement, secondary outcomes were compared in PS matched sample with the risk difference and its 95% CI.

All analyses were performed using R.3.2.0 statistical software on a Mac OsX platform.

RESULTS

Study population

During the study period, 1291 patients were intubated. Thirty-two patients aged more than 18 years, 8 adults with a contraindication to succinylcholine (rocuronium used instead), and 56 patients with missing data were excluded from the final analysis that included a total of 1195 patients (figure 1). Baseline characteristics are reported in table 1. In the original sample, the main observed differences between groups concerned the following variables: altered neurological status (59% in the cricoid pressure group vs 8% in the non-cricoid pressure group, SMD=1.05), patient's position (56% in the cricoid pressure group vs 12%, SMD=0.88) and age (mean=63 (SD=18) in the cricoid pressure group vs mean=57 (SD=18), SMD=−0.33). As expected, patients intubated for cardiac arrest were more prevalent in the non-cricoid pressure group (83% vs 18%, SMD=−1.75).

Propensity score matching and balance measures

The SMDs for the original and the matched datasets are reported in table 1. The best overall balance was obtained for PS 2 that included the confounders and the prognostic factors (figure 2). Accordingly, the WBM was minimised by the PS 2 (WBM=0.41). Therefore, the second PS model (PS 2) was selected as the best PS model for further analyses. Age was the only variable that remained slightly imbalanced after matching (SMD=−0.15 for the PS 2).

Outcomes

Estimated effects of cricoid pressure in the original cohort and in PS 2 matched cohort are reported in table 2. In the original cohort (risk difference=−0.03, 95% CI −0.17 to 0.18, p=0.37) as well as in the matched cohort (risk difference =0.001, 95% CI −0.07 to 0.08, p=0.50), the rate of difficult laryngoscopy did not differ between treatment groups. Further adjustment on age to account for residual imbalance did not markedly alter the results (risk difference=0.01, 95% CI −0.06 to 0.08, p=0.39).

The risk of difficult intubation was similar in the cricoid pressure and the non-cricoid pressure group in the original sample (risk difference=−0.02, 95% CI −0.22 to 0.17, p=0.43), as

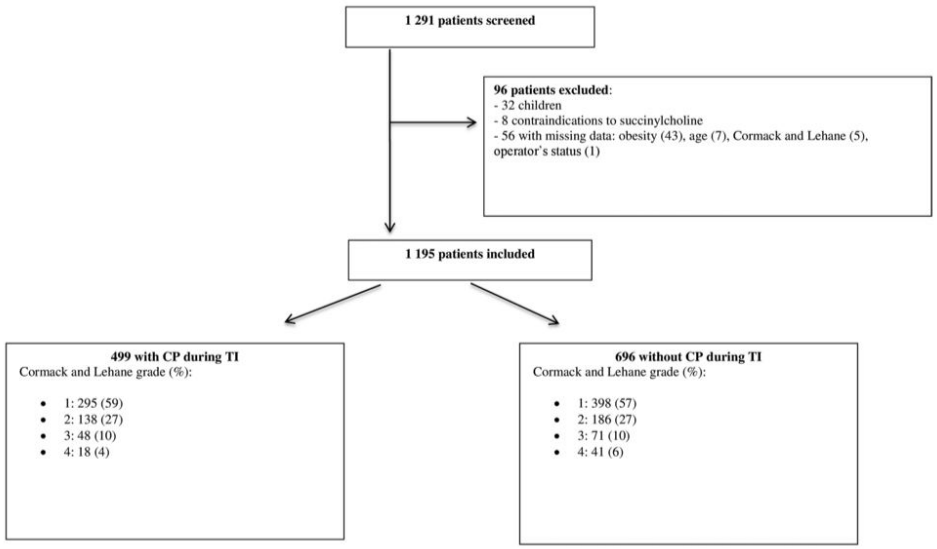


Figure 1 Study flow chart

Table 1 Baseline characteristics balance in the original ant matched cohort for each PS models

Characteristics	Overall cohort (n=1195)			PS 1 matched cohort	PS 2 matched cohort	PS 3 matched cohort
	No cricoid pressure (n=696)	Cricoid pressure (n=499)	SMD*	SMD*	SMD*	SMD*
Age (mean, SD)	63 (18)	57 (18)	−0.33	−0.02†	−0.15	−0.22
Gender (female) (n, %)	245 (35)	190 (38)	−0.06	0.09†	0	0†
Obesity (n, %)	97 (14)	57 (11)	−0.08	−0.03†	0.03†	0†
Operator's status (senior) (n, %)	318 (46)	211 (42)	−0.07	0.04†	−0.07†	−0.09
Patient's position‡(n, %)	87 (12)	280 (56)	0.88	0†	0†	0†
Cervical immobilisation (n, %)	55 (8)	51 (10)	0.08	−0.04†	−0.02	−0.10
Reasons for intubation (n, %)						
Cardiac arrest	576 (83)	89 (18)	−1.75	0†	0†	−0.90
Altered neurological status	53 (8)	295 (59)	1.05	−0.05†	0†	0.73
Acute respiratory distress	22 (3)	32 (6)	0.13	−0.02†	−0.02†	−0.15
Shock	29 (4)	83 (17)	0.33	0.09†	0.01†	0.05

Weighted balance measure			
Overall cohort (n=1195)	PS 1 matched cohort	PS 2 matched cohort	PS 3 matched cohort
29.70	1.48	0.41	15.58

*SMD is the mean difference divided by the pooled SD.
†Variables included in each PS model.
‡Stretcher at full height.
PS, propensity score; SMD, standardised mean difference.

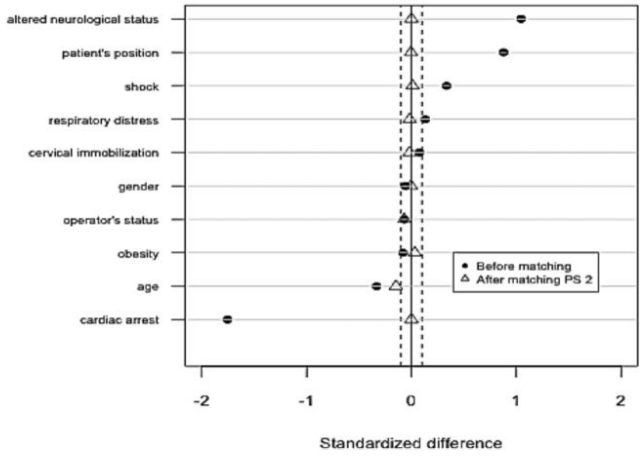


Figure 2 Standardised mean differences before and after matching. PS, propensity score.

well as in the matched sample (risk difference=0.06, 95% CI −0.13 to 0.25, p=0.28).

A total of 195 intubation-related complications occurred in 153 patients (13%). In the original cohort (before matching), more intubation-related complications were observed in cricoid pressure group (n=84 (17%) vs 69 (10%), p<0.001; risk difference=0.07, 95% CI −0.05 to 0.19, p=0.13). Oesophageal intubation (n=29, 6%) and desaturation (n=40, 8%) were the most frequent complications. When patients were exposed to cricoid pressure, oesophageal intubation occurred less frequently with a stretcher at full height (38% vs 62%, p=0.04, data not shown). After matching, there was no difference in terms of complications between patients exposed to cricoid pressure and those unexposed to it, except for airway trauma that were more frequent in cricoid pressure group with 3.6% as compared with 0.5% in non-cricoid pressure group (risk difference=0.03, 95% CI 0.002 to 0.05, p=0.01).

DISCUSSION

Despite the lack of scientific evidence, cricoid pressure is widely used during anaesthesia induction to reduce the incidence of gastric content aspiration. In emergency medicine, the efficiency of cricoid pressure to prevent aspiration still has to be documented.[12] In addition, the use of cricoid pressure during emergent intubation might be associated with impaired laryngeal views.[13] The primary objective of this study was to assess the impact of cricoid pressure on laryngeal view based on a large prehospital observational dataset. Using a PS matching approach, no

Table 2 Difficult laryngoscopy, complication and difficult intubation according to cricoid pressure during tracheal intubation

Outcome	Overall cohort			PS 2 matched cohort
	No cricoid pressure (n=696)	Cricoid pressure (n=499)	p Value*	Risk difference (95% CI)
Difficult laryngoscopy (n, %)	112 (17)	66 (13)		
Risk difference (95% CI)	−0.03 (−0.07 to 0.01)			0.001 (−0.07 to 0.08)
Complication (n, %)	69 (10)	84 (17)	<0.001	0.04 (−0.06 to 0.15)
Oesophageal intubation	34 (5)	29 (6)	0.47	0.05 (−0.03 to 0.13)
Desaturation	14 (2)	40 (8)	<0.001	−0.005 (−0.05, 0.04)
Mainstem intubation	16 (2)	18 (4)	0.17	0.016 (−0.02 to 0.05)
Trauma	4 (1)	11 (2)	0.01	0.03 (0.002 to 0.05)
Vomiting, aspiration during intubation	5 (1)	14 (3)	0.004	0.02 (−0.006 to 0.05)
Bronchospasm, laryngospasm	6 (1)	1 (0)	0.27	−0.005 (−0.03 to 0.02)
Difficult intubation (n, %)	283 (41)	190 (38)	0.36	0.06 (−0.13 to 0.25)

*Bonferroni correction for multiple tests (ie, p values multiplied by the number of comparisons, ie, eight).
Treatment effect is expressed as *risk differences* (95% CIs). *Difficult laryngoscopy* defined according to the Cormack and Lehane grade III or IV obtained by direct laryngoscopy and based on the structures seen. *Difficult intubation* defined as two or more failed laryngoscopic attempts, or the need for any alternative methods.
PS, propensity score.

significant relationship was found between the use of cricoid pressure and the rate of difficult laryngoscopy as defined by a CL[8] grade III or IV. This result is consistent with a previous randomised controlled study[14] reporting no association between the CL grade and the use of cricoid pressure in the operating theatre.

In the original sample (before matching), intubation-related complications were significantly less frequent in the group of patients without cricoid pressure. As a potential explanation, this group encompassed more cardiac arrests, known to be associated with better intubation conditions. However, oesophageal intubation tended to be more frequent in cardiac arrest patients unexposed to cricoid pressure. Suboptimal patient's position may in turn explain this result. Indeed optimal patient positioning is known to be associated with better intubation conditions. In the matched sample however, airway trauma were more frequent in the cricoid pressure group. No significant risk difference was observed in terms of pulmonary aspirations. However, given its low incidence, this study was clearly not powerful enough to conclude on the effect of cricoid pressure on preventing aspiration episodes.

For the patients in cardiac arrest, the use of cricoid pressure is not clearly recommended in the guidelines and is therefore less commonly used. Accordingly, in our sample, most of the cardiac arrests were observed in the non-exposed group and it was often difficult to match them with subject who received cricoid pressure. In this setting where there are few control individuals comparable to the treated individuals, it is helpful to use matching with replacement because controls that look similar to many treated individuals can be used multiple times.[15] In this situation, nearest neighbour matching may result in bad matches. However, this risk was limited by allowing replacement, which in turn increases the overall balance and may in turn decrease the bias.[16] Furthermore, this situation constitutes a threat for the positivity assumption.[17] Positivity is one of the key assumptions in causal inference and basically requires both exposed and unexposed observations to be represented for each possible covariate combination. Here, because almost all patients in cardiac arrest did not receive cricoid pressure, this assumption can be considered as nearly violated. The most appropriate way to interpret this result is to consider that the conclusions only apply to the subjects they were actually matched to.[18] Hence our results should not be extrapolated to patients intubated for a cardiac arrest.

The goal of PS matching is to create a counterfactual group where observed relevant baseline patient characteristics are balanced across groups.[19] PS matching may sometimes miss its target, especially when baseline imbalance is limited. In this case, PS matching may result in discarding part of the population and thus may decrease the efficiency of the estimator. As highlighted by King and Nielsen in their working paper,[20] the first step is to clearly define the causal quantity of interest. The average treatment effect in the treated, which may be more straightforwardly targeted by PS matching estimators, was the quantity of interest in our study.[21] A second point is to verify that the distribution of baseline covariates in the treated and the controls is indeed imbalance. This was indeed the case in the present sample. Finally, PS model specification is of paramount importance both in terms of the variables included in the model[11] and the functional form of the relationship between treatment allocation and the explanatory variables.[22] After matching, the best PS model is then the one that offers the best balance across groups. Multiple balance metrics have been proposed in this context, the most widely used in practice being the average SMD, that is, the empirical mean of the standardised difference obtained for each single covariate.[23] The major drawback of this metric is that it attributes the same weight to each patient characteristic, while, in practice, it may be more important to balance one variable than another. For instance, in the present case, it seems more important to balance the distribution of patients with obesity than the distribution of gender. An alternative metric, referred to as the WBM,[11] has been recently developed to take into account the strength of association between the covariates and the outcome. The goal is to emphasise the need to adequately balance the distribution of the covariates strongly associated with the outcome. Explanatory variables were classified into confounders, prognostic factors and instrumental variables based on expert-matter knowledge and on a quantification of their association with the outcome using multivariable logistic models. Subsequently, three PS models were tested with the goal to only include in the PS model the minimal set of covariates to minimise the WBM. Although PS matching succeeded in reducing the WBM and balancing most patient characteristics at baseline, patient's age was not evenly distributed after matching. Thus, the effect of cricoid pressure on the outcome was also estimated, in the matched dataset, using a regression model adjusting for age. Adjusting on

age did not modify the estimated association between cricoid pressure and the outcome. Patients in cardiac arrest were older and usually unexposed to cricoid pressure; however, adjusting on age did not change this final result.

This study carries some limitations. First, no measure of the actual pressure produced by cricoid pressure was possible in prehospital conditions. However, highly trained personnel only performed cricoid pressure and tracheal intubation. Second, the CL grade relies on a subjective assessment of the vocal chords visualisation. Such a subjective rating might have introduced a certain degree of evaluation bias. This classification is indeed known to carry a limited interobserver reliability and a poor intraobserver reproducibility, both questioning the validity of the CL classification to document laryngeal view during direct laryngoscopy.[24] Moreover, no subjective assessment of the laryngeal view was collected before its application to detect if cricoid pressure could enhance the laryngeal view for patients with CL grade IV. Further works are needed to detail this point. Third, the impact of backward upward rightward pressure could not be specifically addressed in this study as the CL grade was collected before any external manoeuvre. Finally, as previously stated, our conclusions only apply to a population where it was indeed possible to compare exposure to unexposure to cricoid pressure. Therefore, further studies would be needed to answer the question of the benefit of cricoid pressure in patients intubated for cardiac arrest.

In summary, cricoid pressure used for prehospital emergent tracheal intubation is not associated with difficult laryngoscopy as defined by the CL grade III or IV. However, in our cohort, its use was associated with more airway trauma. In this context, the real benefit of cricoid pressure to prevent gastric content aspiration still has to be demonstrated.

Acknowledgements François Xavier Duchateau, who provided the dataset.

Contributors EC performed the analysis and wrote the manuscript. SC supervised the study and helped draft the manuscript. RP contributed to the statistical analysis and helped draft the manuscript. All authors read and approved the final manuscript.

Competing interests None declared.

Ethics approval Commission Nationale de l'Informatique et des Libertés.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Sellick BA. Cricoid pressure to control regurgitation of stomach contents during induction of anaesthesia. *Lancet* 1961;2:404–6.
2 Link MS, Berkow LC, Kudenchuk PJ, *et al*. Part 7: adult advanced cardiovascular life support 2015 American Heart Association guidelines update for cardiopulmonary resuscitation and emergency cardiovascular care. *Circulation* 2015;132:S444–64.
3 Schwartz DE, Matthay MA, Cohen NH. Death and other complications of emergency airway management in critically ill adults. A prospective investigation of 297 tracheal intubations. *Anesthesiology* 1995;82:367–76.
4 Hartsilver EL, Vanner RG. Airway obstruction with cricoid pressure. *Anaesthesia* 2000;55:208–11.
5 Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983;70:41–55.
6 Vanner RG, Asai T. Safe use of cricoid pressure. *Anaesthesia* 1999;54:1–3.
7 Société Française d'Anesthésie et de Réanimation. Intubation difficile. 2006.
8 Cormack RS, Lehane J. Difficult tracheal intubation in obstetrics. *Anaesthesia* 1984;39:1105–11.
9 Langeron O, Birenbaum A, Le Saché F, *et al*. Airway management in obese patient. *Minerva Anestesiol* 2014;80:382–92.
10 Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat Med* 2009;28:3083–107.
11 Caruana E, Chevret S, Resche-Rigon M, *et al*. A new weighted balance measure helped to select the variables to be included in a propensity score model. *J Clin Epidemiol* 2015;68:1415–22.e2.
12 Trethewy CE, Burrows JM, Clausen D, *et al*. Effectiveness of cricoid pressure in preventing gastric aspiration during rapid sequence intubation in the emergency department: study protocol for a randomised controlled trial. *Trials* 2012;13:17.
13 Ellis DY, Harris T, Zideman D. Cricoid pressure in emergency department rapid sequence tracheal intubations: a risk-benefit analysis. *Annal Emerg Med* 2007;50:653–65.
14 Turgeon AF, Nicole PC, Trépanier CA, *et al*. Cricoid pressure does not increase the rate of failed intubation by direct laryngoscopy in adults. *Anesthesiology* 2005;102:315–19.
15 Dehejia RH, Wahba S. Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. *J Am Stat Assoc* 1999;94:1053–62.
16 Caliendo M, Kopeinig S. Some practical guidance for the implementation of propensity score matching. *J Econ Surveys* 2008;22:31–72.
17 Westreich D, Cole SR. Invited commentary: positivity in practice. *Am J Epidemiol* 2010;171:674–7;. discussion 678-81.
18 Petersen ML, Porter KE, Gruber S, *et al*. Diagnosing and responding to violations in the positivity assumption. *Stat Methods Med Res* 2012;21:31–54.
19 Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat Sci* 2010;25:1–21.
20 King G, Nielsen R. Why propensity scores should not be used for matching. Copy at http://j.mp/1sexgVw Export BibTex Tagged XML Download Paper 2015;481.
21 Pirracchio R, Carone M, Rigon MR, *et al*. Propensity score estimators for the average treatment effect and the average treatment effect on the treated may yield very different estimates. *Stat Methods Med Res* 2013. Published Online First: 6 November 2013.
22 Pirracchio R, Petersen ML, van der Laan M. Improving propensity score estimators' robustness to model misspecification using super learner. *Am J Epidemiol* 2015;181:108–19.
23 Austin PC. Assessing balance in measured baseline covariates when using many-to-one matching on the propensity-score. *Pharmacoepidem Drug Safe* 2008;17:1218–25.
24 Krage R, van Rijn C, van Groeningen D, *et al*. Cormack-Lehane classification revisited. *Br J Anaesth* 2010;105:220–7.

5 DISCUSSION

5.1 Résultats principaux

Au moyen d'une étude basée sur des simulations de type Monte Carlo, nous avons montré que le modèle de SP qui minimisait la valeur de WBM était celui qui incluait tous les confondeurs et aucune des variables instrumentales. En accord avec de précédentes études [32], ce modèle était également celui associé avec l'erreur quadratique moyenne la plus faible.

Dans notre première étude de simulation, le modèle final de SP retenu incluait tous les vrais confondeurs, et minimisaient les valeurs de WBM et d'ASAM. En revanche, la distance de Mahalanobis identifiait le modèle de SP incluant tous les vrais confondeurs et une variable instrumentale, alors que ce modèle était effectivement associé à une augmentation du biais et une diminution de la précision. L'inclusion d'une variable instrumentale dans le modèle de SP était responsable d'une augmentation significative de la WBM et d'une altération des performances de l'estimateur, comme précédemment rapporté dans la littérature [34],[41],[98]. D'après les résultats de cette étude, il semble que l'utilisation de la WBM comme mesure d'équilibre de la distribution des covariables après appariement sur le SP puisse permettre de sélectionner le jeu minimal de variables explicatives à inclure dans le modèle de SP tout en évitant de sélectionner des variables instrumentales

Actuellement, les recommandations de bonne pratique d'utilisation des méthodes de score de propension et de présentations des résultats préconisent de clairement rapporter la méthode de sélection des variables utilisée, les connaissances d'experts sur le sujet ayant contribuées aux choix des variables, et le profil d'association des variables avec le traitement à l'étude et/ou le critère de jugement. La sélection des variables incluses dans le modèle de SP dépend de la mesure d'équilibre utilisée et pourrait alors se baser sur les connaissances déjà existantes sur le sujet mais aussi sur des mesures d'équilibre appropriées, telles que la WBM.

Ensuite pour l'évaluation de l'équilibre, il est recommandé de rapporter la mesure utilisée d'évaluation de l'équilibre des distributions obtenu avant et après application des méthodes de SP, et d'exposer l'éventuel déséquilibre persistant sur les covariables au sein du modèle final de SP retenu. Une revue récente de la littérature rapportait une évaluation de l'équilibre dans moins de 60% des articles, et parmi celles-ci une utilisation des différences standardisées dans moins de 25% [42]. Une façon d'y remédier serait d'utiliser une mesure facile à calculer, bien comprises des épidémiologistes, et prenant en compte la force d'association de covariables avec le critère de jugement.En se basant sur les résultats de l'étude de simulation présentée dans ce travail, l'utilisation des connaissances cliniques déjà existantes sur le sujet d'étude pourrait être couplé à une étape d'inclusion une à une des variables restantes en vérifiant si cela est associé à une diminution de la WBM. Cette stratégie permettrait d'aider à minimiser le nombre final de variables incluses en sélectionnant les variables minimisant la mesure WBM, alors que celles qui ne l'influencent pas ou l'augmentent devraient être probablement écartées.

Dans un second article, nous avons appliqué l'utilisation de la WBM à un cas concret. L'utilisation de cette mesure d'équilibre pondérée a permis de sélectionner un modèle final basé sur la minimisation de la WBM, afin d'inclure un ensemble de variables minimal. Les variables explicatives ont été classées en facteur de confusion, facteurs pronostiques et variables instrumentales en se basant sur des d'experts et sur la quantification de leur association avec le critère de jugement en utilisant des modèles de régression logistique multivariés. Trois modèles ont ainsi pu être testés, comme précédemment décrit, dans l'objectif de minimiser la WBM et d'inclure un ensemble de variables minimales. Aucune relation significative n'a été trouvée entre la pression cricoïdienne (PC) et la laryngoscopie difficile. Cependant ces résultats ne peuvent être extrapolés aux patients en arrêt cardiaque. En effet, pour ces patients l'utilisation de la PC n'est pas préconisée dans les

recommandations, et par ce fait elle est donc moins utilisée en pratique. Par conséquent, dans notre échantillon la plupart des patients en arrêt cardiaque étaient observés dans le groupe des patients non-exposés à la PC, ce qui pouvait rendre difficile l'appariement de ces patients avec ceux exposés à la PC. Dans cette situation où les sujets contrôles sont moins nombreux que les sujets exposés, l'appariement de proche en proche peut alors s'avérer de mauvaise qualité et l'utilisation d'une procédure d'appariement avec remise permet alors d'utiliser plusieurs fois des sujets contrôles similaires à plusieurs sujets exposés au traitement [99], ce qui a en plus permis d'améliorer l'équilibre global et diminuer le biais d'estimation de l'effet traitement [100]. Du fait du non-respect total de l'hypothèse de positivité, qui requiert autant de sujets exposés que non-exposés pour chaque combinaison de variables possible, étant donné que la plupart des patients en arrêt cardiaque ne recevaient pas la PC, la façon la plus appropriée pour interpréter les résultats obtenus est de considérer que les conclusions ne peuvent s'appliquer qu'aux patients ayant été appariés [101]. D'autres travaux seraient alors à envisager pour répondre à la question du bénéfice de la PC chez les patients intubés pour arrêt cardiaque.

5.2 Limites et perspectives

Nos travaux portant sur le développement de la WBM présentent plusieurs limites. Premièrement, une estimation valide de la variance et une statistique de test seraient d'un grand intérêt pour la mesure, afin de pouvoir tester la différence entre les WBM obtenues à partir plusieurs modèles de SP. Etant donné l'absence de formulation algébrique explicite de la variance pour la mesure, un ré-échantillonnage non paramétrique pour l'estimer a été utilisé. Des travaux ultérieurs sont nécessaires afin de dériver les propriétés asymptotiques de cette mesure. Deuxièmement, la mesure a été testée seulement avec une méthode d'appariement basée sur le score de propension. D'autres travaux seraient à envisager pour

évaluer si elle pourrait être utilisée avec d'autres estimateurs basés sur le SP tels que la pondération. Troisièmement, la WBM ne peut être décrite comme une somme pondérée, la sommation étant divisée par le nombre de covariables k et non par la somme des poids. La version choisie est plus simple, et ses performances ne devraient pas être altérées étant donné que la somme des poids reste constante quelles que soient les variables incluses dans le modèle de SP. Quatrièmement, le plan de simulation utilisé ici est différent de celui plus largement répandu utilisé par Austin et al. et donc la comparaison directe des résultats avec ces travaux ne peut avoir lieu [41]. Enfin, l'élaboration d'un package R intégrant le calcul de cette nouvelle mesure résumée serait également à envisager.

Concernant l'étude clinique, plusieurs limitations sont à rapporter. L'échelle d'évaluation de Cormack et Lehane utilisée est une échelle subjective de la visualisation des cordes vocales, ce qui pourrait introduire un certain degrés de biais dans l'évaluation. De plus cette échelle est connue pour avoir une fiabilité inter-observateur limitée et une pauvre reproductibilité intra-observateur, ce qui remet en question la validité de cette classification pour documenter la vision laryngée pendant la laryngoscopie directe [102]. De plus, aucune évaluation subjective de la vue laryngée n'a pu être recueillie avant l'application de la PC pour évaluer l'éventuel impact de celle-ci sur l'amélioration de la vision laryngée chez les patients avec un grade de Cormack et Lehane IV. D'autres travaux sont à envisager pour cette évaluation.

6 CONCLUSION

L'utilisation des méthodes d'inférence causale basées sur le score de propension s'est largement répandue pour tenter de diminuer le biais lié aux facteurs de confusion dans les études observationnelles. L'absence de randomisation aboutit à un déséquilibre de caractéristiques entre les groupes de traitement responsable d'un biais dans l'estimation de l'effet causal du traitement. Une revue de la littérature des différentes études médicales utilisant le score de propension a récemment mis en évidence une faible qualité méthodologique quant à la sélection des variables à inclure dans le modèle de SP et à l'évaluation de l'équilibre obtenus après appariement sur le SP [42].

Bien que le choix des variables à inclure soit principalement guidé par les connaissances cliniques, l'aide à la sélection basée sur une mesure d'équilibre prenant en compte la force d'association avec le critère de jugement des variables et minimisant l'influence des variables instrumentales dans l'évaluation globale finale de l'équilibre obtenu peut aider les chercheurs dans la sélection du modèle final le plus parcimonieux possible. La mesure globale d'équilibre proposée dans ce travail de thèse peut constituer une aide précieuse pour améliorer la qualité de réalisation des études observationnelles et contribuer à la mise en place de recommandations de bonne pratique dans le domaine de l'analyse de données observationnelles.

7 APPENDICES

7.1 Appendice A

Soit un critère d'exposition binaire A, représentant deux groupes de traitements, un groupe de sujets traités (A=1) et non traités (A=0) d'une expérience; un critère de jugement, Y et un vecteur de variables W de k covariables distinctes.

La valeur absolue des différences de moyennes standardisées, également appelée la taille d'effet de Cohen, est la différence des moyennes entre les groupes divisée par l'écart-type intra-groupe pour chacune des k variables continues

$$d_i = \frac{(\mu_i^{A=1} - \mu_i^{A=0})}{\sigma_i} \; ; \; i = 1, \ldots, k \quad (1)$$

Elles sont estimées comme le ratio des différences de moyennes de l'échantillon pour chaque covariable entre le groupe de traitement et le groupe contrôle divisé par l'écart-type commun estimé sur l'échantillon global par :

$$d_i = \frac{(p_i^{A=1} - p_i^{A=0})}{\sqrt{\frac{\left(p_i^{A=1}\left(1-p_i^{A=1}\right) + p_i^{A=0}\left(1-p_i^{A=0}\right)\right)}{2}}} \qquad (2)$$

Pour les k covariables binaires, la différence standardisée est définie comme :

$$d_i = \frac{(p_i^{A=1} - p_i^{A=0})}{\sqrt{\frac{\left(p_i^{A=1}\left(1-p_i^{A=1}\right) + p_i^{A=0}\left(1-p_i^{A=0}\right)\right)}{2}}} \; ; \; i = 1, \ldots, k$$

avec $p_i^{A=1}$ et $p_i^{A=0}$ les prévalences des variables dichotomiques des sujets traités et non traités, respectivement.

La première mesure du déséquilibre proposée est la moyenne absolue des différences de moyennes standardisées estimée par :

$$\widehat{M}_{ASAM} = \frac{1}{k}\sum_{i=1}^{k}\left|\hat{d}_i\right| \qquad (3)$$

La seconde mesure proposée est la distance de Mahalanobis estimée comme :

$$\widehat{M}_{ASAM} = (\overline{W}^{A=1} - \overline{W}^{A=0})' \, \Sigma^{-1} \, (\overline{W}^{A=1} - \overline{W}^{A=0}) \quad (4)$$

avec $\overline{W}^{A=1}$ et $\overline{W}^{A=0}$ les vecteurs des moyennes de la variable du groupe traité et non traité et Σ la matrice de variance/covariance des variables.

Pour la mesure d'équilibre pondérée proposée, chaque différence de moyenne standardisée est pondérée en fonction de sa valeur pronostique :

$$W\hat{B}M = \frac{1}{k}\sum_{i=1}^{k}\left|\widehat{\omega}_i \cdot \hat{d}_i\right| \quad (5)$$

avec $_i$ le poids prenant en compte l'association entre la covariable W_i et le critère de jugement. Le coefficient obtenu spécifiquement pour chaque covariable W_i est obtenu du maximum de vraisemblance d'une régression logistique multivariée de Y sur W_i :

$$\widehat{\omega}_i = \hat{\beta}_i \cdot SD_i \qquad (6)$$

avec SD_i l'écart-type de l'échantillon pour la covariable W_i dans la cohorte originale. Pour un critère de jugement binaire, le poids proposé est directement dérivé de l'estimation des coefficients standardisés obtenus par un modèle de régression logistique multivarié du critère de jugement sur les covariables.

7.2 Appendice B

Soit Y un critère de jugement binaire, A une exposition au traitement binaire, et un vecteur de 11 covariables : quatre confondeurs $W_{1,\ldots,4}$, associés à la fois avec le critère de jugement et l'exposition au traitement, trois variables instrumentales $W_{5,6,7}$, trois facteurs pronostiques $W_{8,9,10}$ et une variable non associée à l'exposition au traitement ni au critère de jugement W_{11}.

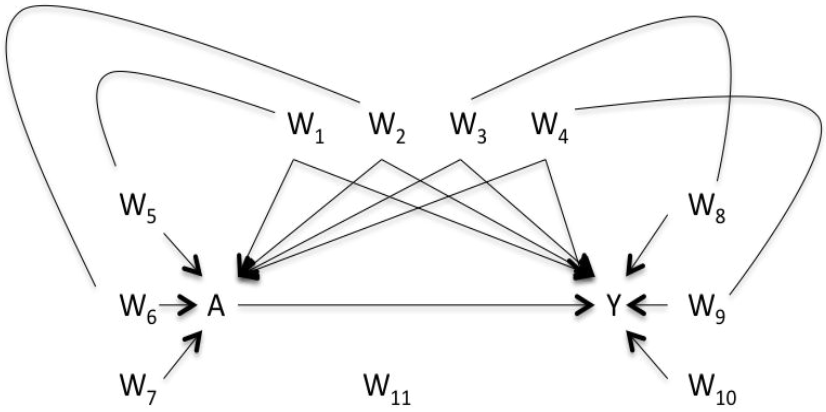
Pour chaque échantillon de données simulé, les 11 covariables W_i (i=1,..., 11) ont été générées à partir d'une distribution normale de moyenne zéro et de variance un. Premièrement, huit covariables (V_i, i=1,..., 11) ont été générées comme des variables aléatoires indépendantes et normales. Secondairement, huit autres covariables (W_i, i=1...6,8,9) ont été générées comme une combinaison linéaire de V_i, i=1...6,8,9. En plus, trois covariables (W_7, W_{10} et W_{11}) ont été générées comme des variables aléatoires indépendantes et normalement distribuées. Dans cette seconde étape, des corrélations entre certaines covariables ont été introduites, avec des coefficients variant entre 0.1 et 0.2. Pour atténuer la magnitude des coefficients de corrélation, une dichotomisation a été réalisée (basée sur la moyenne de chaque covariable). Au final, ces variables (W_1, W_3, W_5, W_6, W_8 et W_9) ont été considérées comme binaire en échelle. Les coefficients utilisés pour générer les données sont:

	Vrai modèle de SP (β)	**Modèle du critère de jugement (α)**
Intercept	0	0.8
Coefficient 1	0.2	1.2
Coefficient 2	0.4	0.8
Coefficient 3	0.8	0.4
Coefficient 4	-1.2	0.2
Coefficient 5	-1.2	-1.2
Coefficient 6	-1.2	-1.2
Coefficient 7	-1.8	-1.2

La matrice de corrélation pour la génèse des données et le graphique acyclique orienté correspondant sont présentés ci-dessous:

		Vrais confondeurs				**Prédicteurs de A**			**Prédicteurs de Y**		
		W_1	W_2	W_3	W_4	W_5	W_6	W_7	W_8	W_9	W_{10}
Vrais confondeurs	W_1	1									
	W_2	0	1								
	W_3	0	0	1							
	W_4	0	0	0	1						
Prédicteurs de A	W_5	0.1	0	0	0	1					
	W_6	0	0.2	0	0	0	1				
	W_7	0	0	0	0	0	0	1			
Prédicteurs de Y	W_8	0	0	0.1	0	0	0	0	1		
	W_9	0	0	0	0.1	0	0	0	0	1	
	W_{10}	0	0	0	0	0	0	0	0	0	0

Abréviations: A (Traitement), Y (critère de jugement)



Exposition au traitement: A
Vrais confondeurs: W_1,W_2,W_3,W_4
Facteurs pronostics: W_8,W_9,W_{10}

Critère de jugement: Y
Prédicteurs du traitement: W_5,W_6,W_7
Sans lien avec A/Y: W_{11}

La probabilité d'exposition au traitement, qui est le vrai SP, était générée comme une fonction des covariables W_i:

$$Pr(A = 1|W_i) = f(W_i, \beta)$$

avec f correspondant à une fonction linéaire et additive sur une échelle de rapport de côte logarithmique, garantissant que le vrai SP soit compris entre 0 et 1 et que l'exposition au traitement soit en moyenne de 0.4. Le vecteur A était issu d'une distribution de Bernoulli avec une probabilité fixée par $f(W_i, \beta)$.

Similairement, le critère de jugement binaire Y était généré à partir d'une combinaison linéaire de A et W_i:

$$Pr(Y = 1|A, W_i) = f(A, W_i, \gamma, \alpha)$$

avec f un modèle de régression logistique, garantissant une vraie probabilité de survenue du critère de jugement comprise entre 0 et 1 et que la probabilité de survenue de Y soit en moyenne de 0.5. L'effet du traitement γ était fixée à -0.4. Le vecteur Y était issu d'une distribution de Bernoulli avec une probabilité fixée par $f(A, W_i, \gamma, \alpha)$.

8 BIBLIOGRAPHIE

1 Carlson MDA, Morrison RS. Study Design, Precision, and Validity in Observational Studies. *Journal of Palliative Medicine* 2009;12:77.

2 Pearl J. *Causality: models, reasoning and inference*. Cambridge Univ Press 2000.

3 Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology* 1992;3:143–55.

4 Greenland S. Causal Analysis in the Health Sciences. *Journal of the American Statistical Association* 2000;95:286–9.

5 Imbens GW. Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review. *The Review of Economics and Statistics* 2004;86:4–29.

6 Hernan MA. A definition of causal effect for epidemiological research. *Journal of Epidemiology & Community Health* 2004;58:265–71.

7 Rosenberger WF, Lachin JM. *Randomization in Clinical Trials: Theory and Practice*. John Wiley & Sons 2015.

8 Altman DG. Randomisation. *BMJ* 1991;302:1481–2.

9 Hernan MA. Estimating causal effects from epidemiological data. *Journal of Epidemiology & Community Health* 2006;60:578–86.

10 Greenland S. Randomization, statistics, and causal inference. *Epidemiology* 1990;1:421–9.

11 Greenland S, Robins JM. Identifiability, exchangeability and confounding revisited. *Epidemiol Perspect Innov* 2009;6:4.

12 Greenland S, Robins JM, Pearl J. Confounding and Collapsibility in Causal Inference. *Statist Sci* 1999;14:29–46.

13 Rubin DB. Bayesian Inference for Causal Effects: The Role of Randomization. *Ann Statist* 1978;6:34–58.

14 Moher D, Schulz KF, Altman D, *et al.* The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *JAMA* 2001;285:1987–91.

15 Rochon PA, Gurwitz JH, Sykora K, *et al.* Reader's guide to critical appraisal of cohort studies: 1. Role and design. *BMJ* 2005;330:895–7.

16 Rosenbaum PR, Rubin DB. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* 1983;70:41.

17 Rubin DB. Assignment to Treatment Group on the Basis of a Covariate. *ETS Research Bulletin Series* 1976;1976:i-20.

18 Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 1974;66:688–701.

19 Rosenbaum PR, Rubin DB. Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *Journal of the American Statistical Association* 1984;79:516–24.

20 Pearl J. An Introduction to Causal Inference. *Int J Biostat* 2010;6.

21 VanderWeele TJ. Concerning the Consistency Assumption in Causal Inference: *Epidemiology* 2009;20:880–3.

22 Pearl J. Causal diagrams for empirical research. *Biometrika* 1995;82:669–88.

23 Shrier I, Platt RW. Reducing bias through directed acyclic graphs. *BMC Med Res Methodol* 2008;8:70.

24 Hernán MA, Hernández-Díaz S, Werler MM, *et al.* Causal Knowledge as a Prerequisite for Confounding Evaluation: An Application to Birth Defects Epidemiology. *Am J Epidemiol* 2002;155:176–84.

25 Rubin DB. On principles for modeling propensity scores in medical research. *Pharmacoepidemiol Drug Saf* 2004;13:855–7.

26 D'Agostino RB. Tutorial in biostatistics: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med* 1998;17:2265–2281.

27 Rosenbaum PR. Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. In: *Matched Sampling for Causal Effects*. Cambridge University Press 2006.

28 Lee BK, Lessler J, Stuart EA. Weight Trimming and Propensity Score Weighting. *PLoS One* 2011;6.

29 Setoguchi S, Schneeweiss S, Brookhart MA, *et al.* Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiol Drug Saf* 2008;17:546–55.

30 Westreich D, Lessler J, Funk MJ. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology* 2010;63:826–33.

31 Pirracchio R, Petersen ML, van der Laan M. Improving propensity score estimators' robustness to model misspecification using super learner. *Am J Epidemiol* 2015;181:108–19.

32 Brookhart MA, Schneeweiss S, Rothman KJ, *et al.* Variable selection for propensity score models. *Am J Epidemiol* 2006;163:1149–56.

33 Patrick AR, Schneeweiss S, Brookhart MA, *et al.* The implications of propensity score variable selection strategies in pharmacoepidemiology: an empirical illustration. *Pharmacoepidem Drug Safe* 2011;20:551–9.

34 Pearl J. Invited Commentary: Understanding Bias Amplification. *Am J Epidemiol* 2011;174:1223–7.

35 Myers JA, Rassen JA, Gagne JJ, *et al.* Myers et al. Respond to "Understanding Bias Amplification." *American Journal of Epidemiology* 2011;174:1228–9.

36 Mortimer KM, Neugebauer R, van der Laan M, *et al.* An application of model-fitting procedures for marginal structural models. *Am J Epidemiol* 2005;162:382–8.

37 Myers JA, Rassen JA, Gagne JJ, *et al.* Effects of adjusting for instrumental variables on bias and precision of effect estimates. *Am J Epidemiol* 2011;174:1213–22.

38 Pearl J. On a class of bias-amplifying variables that endanger effect estimates. *arXiv preprint arXiv:12033503* Published Online First: 2012.

39 Hill J. Discussion of research using propensity-score matching: comments on "A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003" by Peter Austin, Statistics in Medicine. *Stat Med* 2008;27:2055-2061-2069.

40 Austin PC. The Relative Ability of Different Propensity Score Methods to Balance Measured Covariates Between Treated and Untreated Subjects in Observational Studies. *Med Decis Making* Published Online First: 14 August 2009.

41 Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Statistics in Medicine* 2007;26:734–53.

42 Ali MS, Groenwold RHH, Belitser SV, *et al.* Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: a systematic review. *Journal of Clinical Epidemiology* 2015;68:122–31.

43 Austin PC, Mamdani MM, Stukel TA, *et al.* The use of the propensity score for estimating treatment effects: administrative versus clinical data. *Stat Med* 2005;24:1563–78.

44 Austin PC. The performance of different propensity score methods for estimating marginal odds ratios. *Stat Med* 2007;26:3078–94.

45 Pirracchio R, Carone M, Rigon MR, *et al.* Propensity score estimators for the average treatment effect and the average treatment effect on the treated may yield very different estimates. *Stat Methods Med Res* Published Online First: 6 November 2013.

46 Ali MS, Groenwold RHH, Klungel OH. Propensity score methods and unobserved covariate imbalance: comments on "squeezing the balloon." *Health Serv Res* 2014;49:1074–82.

47 Stürmer T, Joshi M, Glynn RJ, *et al.* A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *Journal of clinical epidemiology* 2006;59:437.

48 Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine* 2008;27:2037–49.

49 Cochran WG, Rubin DB. Controlling Bias in Observational Studies: A Review. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)* 1973;35:417–46.

50 Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research* 2011;46:399–424.

51 Robins JM, Hernán MÁ, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000;11:550–560.

52 Austin PC. Goodness-of-fit diagnostics for the propensity score model when estimating treatment effects using covariate adjustment with the propensity score. *Pharmacoepidem Drug Safe* 2008;17:1202–17.

53 Ho DE, Imai K, King G, *et al.* Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis* 2006;15:199–236.

54 Imai K, King G, Stuart EA. Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the royal statistical society: series A (statistics in society)* 2008;171:481–502.

55 Senn S. Testing for baseline balance in clinical trials. *Statistics in medicine*

1994;13:1715–1726.

56 Flury BK, Riedwyl H. Standard Distance in Univariate and Multivariate Analysis. *The American Statistician* 1986;40:249–51.

57 Tritchler D. Interpreting the standardized difference. *Biometrics* 1995;51:351–3.

58 Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. Routledge 1988.

59 Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine* 2009;28:3083–107.

60 Fleiss JL, Levin B, Paik MC. *Statistical methods for rates and proportions*. 3rd ed. Hoboken, N.J: : J. Wiley 2003.

61 Austin PC. Using the Standardized Difference to Compare the Prevalence of a Binary Variable Between Two Groups in Observational Research. *Communications in Statistics - Simulation and Computation* 2009;38:1228–34.

62 Belitser SV, Martens EP, Pestman WR, *et al.* Measuring balance and model selection in propensity score methods: balance measure for propensity scores methods. *Pharmacoepidemiology and Drug Safety* 2011;20:1115–29.

63 Ali MS, Groenwold RHH, Pestman WR, *et al.* Propensity score balance measures in pharmacoepidemiology: a simulation study. *Pharmacoepidemiol Drug Saf* 2014;23:802–11.

64 Kuss O. The z-difference can be used to measure covariate balance in matched propensity score analyses. *Journal of Clinical Epidemiology* 2013;66:1302–7.

65 Rubin DB, Thomas N. Characterizing the Effect of Matching Using Linear Propensity Score Methods with Normal Distributions. *Biometrika* 1992;79:797.

66 Rubin DB, Thomas N. Matching Using Estimated Propensity Scores: Relating Theory to Practice. *Biometrics* 1996;52:249.

67 Rubin DB. Bias Reduction Using Mahalanobis-Metric Matching. *Biometrics*

1980;36:293.

68 Franklin JM, Rassen JA, Ackermann D, *et al.* Metrics for covariate balance in cohort studies of causal effects. *Statist Med* 2014;33:1685–99.

69 Stephens MA. Use of the Kolmogorov-Smirnov, Cramer-Von Mises and Related Statistics Without Extensive Tables. *Journal of the Royal Statistical Society Series B (Methodological)* 1970;32:115–22.

70 Smirnov N. Table for Estimating the Goodness of Fit of Empirical Distributions. *Ann Math Statist* 1948;19:279–81.

71 Silverman BW. *Density Estimation for Statistics and Data Analysis*. CRC Press 1986.

72 Wand MP, Jones MC. *Kernel Smoothing*. CRC Press 1994.

73 Pestman WR. *Mathematical Statistics*. Walter de Gruyter 2009.

74 Thompson JW. A Note on the Lévy Distance. *Journal of Applied Probability* 1975;12.

75 Iacus SM, King G, Porro G. Multivariate Matching Methods That Are Monotonic Imbalance Bounding. *Journal of the American Statistical Association* 2011;106:345–61.

76 Westreich D, Cole SR, Funk MJ, *et al.* The role of the c-statistic in variable selection for propensity score models. *Pharmacoepidemiology and Drug Safety* 2011;20:317–20.

77 Perkins SM, Tu W, Underhill MG, *et al.* The use of propensity scores in pharmacoepidemiologic research. *Pharmacoepidemiol Drug Saf* 2000;9:93–101.

78 Stuart EA, Lee BK, Leacy FP. Prognostic score–based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. *Journal of Clinical Epidemiology* 2013;66:S84–S90.e1.

79 Hansen BB. The Prognostic Analogue of the Propensity Score. *Biometrika* 2008;95:481–8.

80 Bring J. How to Standardize Regression Coefficients. *The American Statistician* 1994;48:209–13.

81 Abadie A, Imbens GW. Bias-Corrected Matching Estimators for Average Treatment Effects. *Journal of Business & Economic Statistics* 2011;29:1–11.

82 Sekhon JS. Multivariate and propensity score matching software with automated balance optimization: the matching package for R. *Journal of Statistical Software, Forthcoming* Published Online First: 2008.

83 Sellick BA. Cricoid pressure to control regurgitation of stomach contents during induction of anaesthesia. *Lancet* 1961;2:404–6.

84 Neumar RW, Otto CW, Link MS, *et al.* Part 8: Adult Advanced Cardiovascular Life Support: 2010 American Heart Association Guidelines for Cardiopulmonary Resuscitation and Emergency Cardiovascular Care. *Circulation* 2010;122:S729–67.

85 Schwartz DE, Matthay MA, Cohen NH. Death and other complications of emergency airway management in critically ill adults. A prospective investigation of 297 tracheal intubations. *Anesthesiology* 1995;82:367–76.

86 Hartsilver EL, Vanner RG. Airway obstruction with cricoid pressure. *Anaesthesia* 2000;55:208–11.

87 Noguchi T, Koga K, Shiga Y, *et al.* The gum elastic bougie eases tracheal intubation while applying cricoid pressure compared to a stylet. *Can J Anaesth* 2003;50:712–7.

88 Société Française d'Anesthésie et de Réanimation. Intubation difficile. 2006.

89 Vanner RG, Asai T. Safe use of cricoid pressure. *Anaesthesia* 1999;54:1–3.

90 Cormack RS, Lehane J. Difficult tracheal intubation in obstetrics. *Anaesthesia* 1984;39:1105–11.

91 Langeron O, Birenbaum A, Le Saché F, *et al.* Airway management in obese patient. *Minerva Anestesiol* 2014;80:382–92.

92 Wang HE, Kupas DF, Paris PM, *et al.* Multivariate Predictors of Failed Prehospital Endotracheal Intubation. *Academic Emergency Medicine* 2003;10:717–724.

93 Brodsky JB, Lemmens HJM, Brock-Utne JG, *et al.* Morbid obesity and tracheal intubation. *Anesth Analg* 2002;94:732–736; table of contents.

94 Freund Y, Duchateau FX, Devaud ML, *et al.* Factors associated with difficult intubation in prehospital emergency medicine. *European Journal of Emergency Medicine* 2012;19:304–308.

95 Trethewy CE, Burrows JM, Clausen D, *et al.* Effectiveness of cricoid pressure in preventing gastric aspiration during rapid sequence intubation in the emergency department: study protocol for a randomised controlled trial. *Trials* 2012;13:17.

96 Ellis DY, Harris T, Zideman D. Cricoid Pressure in Emergency Department Rapid Sequence Tracheal Intubations: A Risk-Benefit Analysis. *Annals of Emergency Medicine* 2007;50:653–65.

97 Turgeon AF, Nicole PC, Trépanier CA, *et al.* Cricoid pressure does not increase the rate of failed intubation by direct laryngoscopy in adults. *Anesthesiology* 2005;102:315–9.

98 Wooldridge J. Should instrumental variables be used as matching variables. Michigan State University, MI 2009.

99 Dehejia RH, Wahba S. Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs. *Journal of the American Statistical Association* 1999;94:1053–62.

100 Caliendo M, Kopeinig S. Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys* 2008;22:31–72.

101 Petersen ML, Porter KE, Gruber S, *et al.* Diagnosing and responding to violations in the positivity assumption. *Stat Methods Med Res* 2012;21:31–54.

102 Krage R, van Rijn C, van Groeningen D, *et al.* Cormack-Lehane classification revisited. *British Journal of Anaesthesia* 2010;105:220–7.