Ann. Fr. Med. Urgence (2021) 11:25-32
DOI 10.3166/afmu-2020-0235

Clés de compréhension du score de propension à l'usage du clinicien

Keys to Propensity Score Understanding for Clinicians

C. Claustre · C. El Khoury · L. Fraticelli

Reçu le 20 septembre 2019 ; accepté le 14 mai 2020

Résumé Les études observationnelles, en l'absence de biais de sélection, présentent l'avantage de refléter la pratique en situation réelle et d'être moins contraignantes (contrainte éthique, faisabilité) que les essais cliniques randomisés. Les scores de propension sont de plus en plus souvent utilisés dans les études observationnelles afin de corriger les biais de confusion (14 occurrences anglophones Pubmed dans le titre ou l'abstract en 2000, 448 en 2010 et 3 388 en 2018). Cette méthode permet de se rapprocher d'une interprétation causale des effets observés comme il serait possible de le faire dans un essai clinique randomisé. Peu d'articles décrivent leurs conditions d'utilisation et d'interprétation induisant des défauts dans leur mise en œuvre, leur interprétation et leur présentation dans les articles. Nous proposons dans cet article une synthèse pragmatique, à l'usage du clinicien, présentant les avantages et limitations associés à l'utilisation des scores de propension lors d'inférences causales. Notre objectif est de donner aux cliniciens les clés de compréhension et d'interprétation des scores de propensions. Nous développerons tout au long de cet article un exemple fictif fondé sur des données simulées. Nous présenterons la création d'un score de propension, son utilisation selon quatre méthodes (appariement, stratification, pondération inverse et ajustement), sa validation ainsi que des règles pour présenter les résultats issus de cette méthode dans un article scientifique afin de garantir les règles de reproductibilité des résultats.

Mots clés Score de propension · Inférence causale · Méthodologie · Facteurs de confusion

Abstract Observational studies in the absence of selection bias reflect real life practices and have fewer constraints (ethical concerns, feasibility) than randomised clinical trials. Propensity score methods are increasingly used in observational studies to adjust for confounding and perform causal inferences (14 English Pubmed occurrences in the title or the abstract in 2000, 448 in 2010 and 3388 in 2018). This method allows interpretations of effects to be closer to causal inference like in a randomised clinical trial. There is a lack of articles describing propensity scores assumptions and interpretation rules. This leads to inadequacies in their use and in their reporting in scientific articles. We propose in this article a pragmatic synthesis, for the use of clinicians, presenting the advantages and limitations associated with the use of propensity scores for causal inferences. Our objective is to give clinicians keys to understanding and interpret propensity scores. We will develop an example based on simulated data to demonstrate the creation of the propensity score, four of its uses for covariate adjustment (matching, stratification, inverse probability weighing and adjustment) and its validation as well as reporting guidelines to ensure reproducibility of results in a scientific article.

Keywords Propensity score · Causal inference · Methodology · Confusion factors

C. Claustre (✉) · C. El Khoury · L. Fraticelli
Réseau des Urgences de la Vallée du Rhône (RESUVal), hôpital Lucien-Hussel, F-38200 Vienne, France
e-mail : c.claustre@resuval.fr

C. El Khoury
Structure d'urgence et unité de recherche clinique, médipôle, F-69100 Villeurbanne, France

HESPER EA 7425, université Claude-Bernard–Lyon-I, F-69008 Lyon, France

L. Fraticelli
EA 4129 P2S Parcours Santé Systémique, université Claude-Bernard–université Lyon-I, F-69372 Lyon cedex 08, France

Introduction

Les études observationnelles, en l'absence de biais de sélection, présentent l'avantage de refléter la pratique en situation réelle et d'être moins contraignantes (contrainte éthique, faisabilité) que les essais cliniques randomisés (ECR). Bien que ces études ne permettent pas de conclure avec certitude à des

liens de causalité entre un événement et une exposition (traitement), elles apportent une estimation de leur association. Des méthodes statistiques peuvent être mises en place afin de corriger les biais de confusion. Pour cela, les scores de propension (SP) sont de plus en plus souvent utilisés (14 occurrences anglophones Pubmed dans le titre ou l'abstract en 2000, 448 en 2010 et 3 388 en 2018).

Le SP est la probabilité pour chaque patient d'être exposé en fonction de ses caractéristiques (démographie, facteurs de risques). Il permet de rendre « artificiellement » comparables les patients exposés et non exposés en fonction de leurs caractéristiques. Ce qui revient en d'autres termes à réduire fortement l'effet des biais de confusion.

Peu d'articles décrivent les conditions d'utilisation et d'interprétabilité des SP, induisant des défauts dans la mise en œuvre des méthodes ou dans l'interprétation des résultats [1]. Nous proposons ici une synthèse pragmatique, à l'usage du clinicien, présentant les avantages et limitations associés à l'utilisation des SP lors d'inférences causales utilisant des données non expérimentales.

Quand utiliser les scores de propension ?

Les SP sont utilisés pour mettre en évidence l'effet d'une exposition sur un événement. L'exposition d'intérêt peut être variée (un traitement, un facteur de risque). Dans la majorité des cas, une exposition binaire et unique est utilisée (traitement oui/non, facteur de risque oui/non). Le SP est difficilement applicable à la comparaison de trois traitements ou à l'étude simultanée de plusieurs facteurs de risque.

Deux hypothèses sont nécessaires à leur utilisation [2] :

- l'indépendance conditionnelle qui est respectée si tous les facteurs de confusion sont observés et ajustés. Elle est difficile à vérifier en pratique ;
- le support commun qui suppose que les patients exposés et non exposés sont suffisamment similaires pour être rendus comparables par un ajustement.

Le SP sera a priori mal spécifié si ces hypothèses ne sont pas vérifiées. Ces hypothèses ne sont pas propres au SP, mais doivent également être respectées pour les autres méthodes d'analyses multivariées. Si ces hypothèses sont vérifiées, l'effet de l'exposition sur l'événement s'interprète « toutes choses étant égales par ailleurs », c'est-à-dire en considérant toutes les autres variables comme constantes entre les patients exposés et non exposés.

L'ampleur des effets estimés par les SP et les autres méthodes d'analyses multivariées est similaire. Le SP présente des avantages et inconvénients par rapport aux autres méthodes multivariées [3–6] (Tableau 1). Les principaux avantages du SP sont : un biais possiblement plus faible sur les petites populations (< 7 événements par covariables) [5], la visualisation de l'ampleur du biais corrigé et l'interprétation proche de celle d'un ECR.

Tableau 1 Comparaison des méthodes fondées sur les scores de propension et des méthodes multivariées classiques

Critère	Scores de propension	Autres méthodes multivariées
Conditions	Indépendance conditionnelle et support commun	Indépendance conditionnelle, support commun, indépendance des variables d'ajustement et hypothèses propres à chaque modèle
Complexité du modèle	Le SP peut inclure de nombreuses variables et interactions	Le modèle ne doit pas être trop complexe en termes de nombre de variables, interactions, etc. afin d'éviter un surajustement
Design	La création du SP et l'estimation de l'effet de l'exposition sont réalisées en deux étapes séparées Tous les facteurs de confusion sont résumés en un seul score	La création de l'analyse et l'estimation de l'effet de l'exposition en une seule étape
Puissance	Possiblement plus puissant lorsque le nombre d'événements par variable est inférieur à 7	Possiblement moins puissant lorsque le nombre d'événements par variable est inférieur à 7
Validation	Hypothèses vérifiables à l'exception de l'indépendance conditionnelle Diminution du biais directement observable	Difficile de valider les hypothèses et d'estimer l'importance de la correction du biais
Interprétation	Interprétation proche de celle d'un essai clinique	Interprétation propre à chaque modèle
Comparaison de plus de deux expositions	Difficile	Possible avec une régression multinomiale

SP = Score de propension

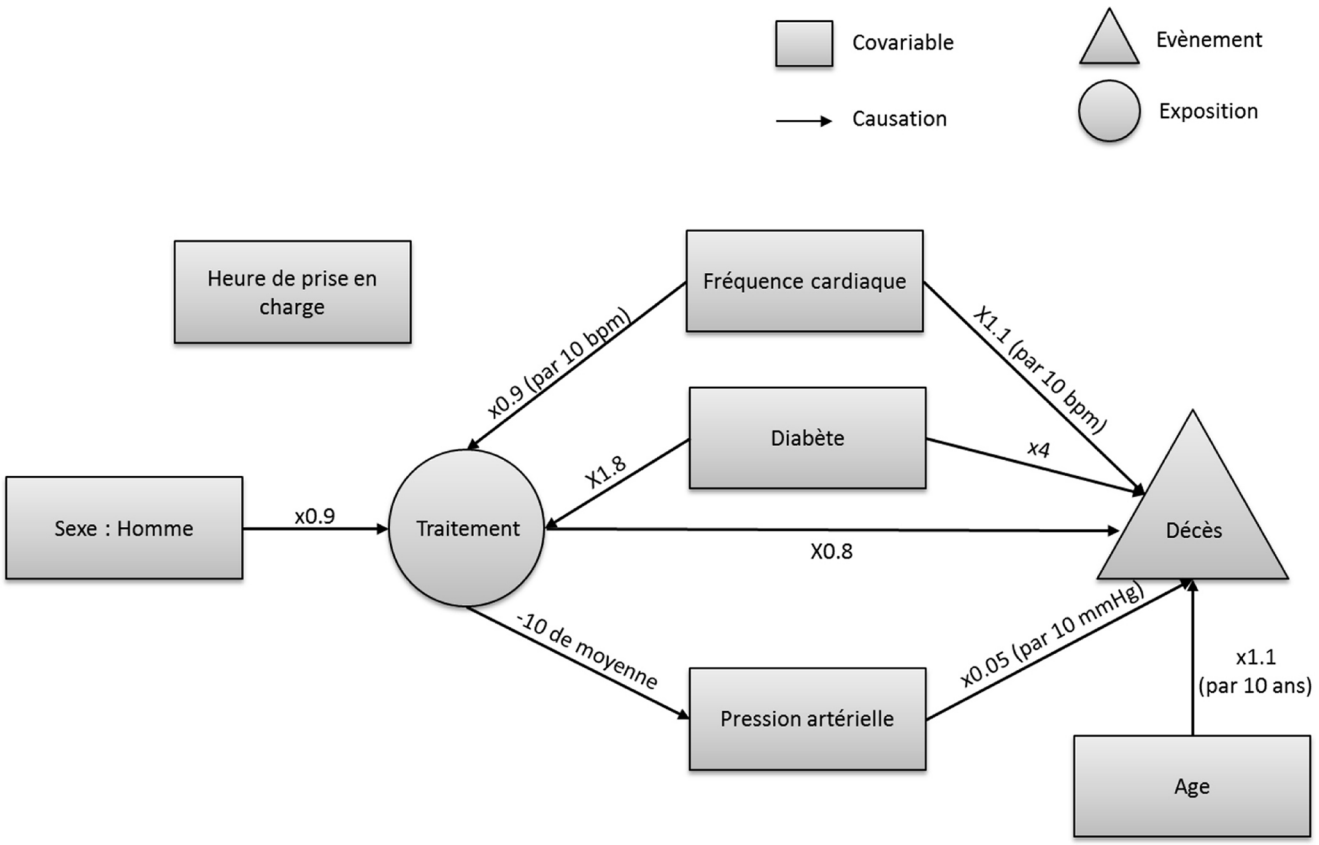

Fig. 1 Structure de la population simulée (n = 4 000)

Dans tous les cas, seuls les facteurs de confusion connus et pris en compte dans le SP sont ajustés. Une interprétation causale entre l'exposition et l'événement est donc impossible, car il peut rester des variables de confusion non observées. Un ECR rigoureux et suffisamment grand reste le meilleur moyen d'ajuster sur tous les facteurs de confusion observés ou non [7]. Les SP et les autres analyses multivariées ne corrigent que les biais de confusion et non les autres types de biais (sélection, classement).

Exemple : Nous cherchons à mettre en évidence l'effet d'un traitement sur le décès dans une population observée. Nous disposons de 4 000 patients, de leur âge, sexe, heure de prise en charge, fréquence cardiaque, diabète, pression artérielle, traitement et décès. Les règles utilisées pour la simulation sont résumées en figure 1. Les patients traités sont plus souvent diabétiques et ont une fréquence cardiaque plus élevée. En analyse univariée, la mortalité est similaire entre les patients traités et non traités (18,10 vs 16,60 %, p = 0,22). L'odds ratio (OR) de décès estimé par une régression logistique univariée est de 1,11 : [0,94 ; 1,31] ; p = 0,21 pour les patients traités. Ce résultat peut s'interpréter comme : les patients qui ont reçu le traitement ont un odd de décès 1,11 fois plus élevé que ceux qui ne l'ont pas reçu (résultat non significatif).

Comment utiliser les scores de propension

Création du score

Le SP est classiquement estimé par une régression logistique modélisant l'exposition. Le SP de chaque patient est la probabilité d'exposition prédite par le modèle [2].

Le choix des variables utilisées affecte le biais et la variance de l'estimation de l'effet de l'exposition.

- Le biais est l'écart entre l'estimation de l'effet de l'exposition et son effet réel. Il entraîne une mauvaise estimation de l'effet de l'exposition ;
- la variance est la dispersion de l'estimation. Elle augmente l'ampleur des intervalles de confiance.

Les méthodes classiques de sélection de variables par maximisation d'un critère (Critère d'information d'Akaike) ne sont pas recommandées [8]. Ces méthodes optimisent la prédiction de l'exposition et pourraient donc ne pas inclure une variable de confusion faiblement associée à l'exposition mais fortement à l'événement.

Tableau 2 Effet de l'inclusion des variables au score de propension

Variable associée à l'exposition	Variable associée à l'événement	Inclure	Biais	Variance
Non	Non	On peut	=	=
Oui	Non	Il ne faut pas	=	↑
Non	Oui	Il faut	↓	=
Oui	Oui	Il faut	↓	↑

Les résultats sur données simulées indiquent les principes suivants pour la sélection de variables [8,9] :

- toutes les variables associées à l'événement devraient être incluses dans le SP ;
- les variables associées uniquement à l'exposition devraient ne pas être incluses dans le SP ;
- les variables associées ni à l'événement ni à l'exposition peuvent ne pas être incluses.

L'effet de chaque cas sur le biais et la variance sont résumés en tableau 2. En pratique, il n'est pas toujours possible de déterminer quelles variables sont liées à l'événement ou à l'exposition. Il est préférable d'inclure une variable à tort que d'oublier d'inclure une variable de confusion. Dans le premier cas, le risque est de ne pas pouvoir conclure à un effet. Dans le second cas, le risque est de conclure à tort.

Exemple (suite) : Le SP est estimé à partir de la régression logistique suivante sur les données simulées :
Traitement (0/1) ~ Fréquence cardiaque + Diabète + Âge
Le sexe n'est pas inclus, car il est indépendant du décès. Bien qu'il n'affecte pas l'attribution du traitement, l'âge est inclus, car il affecte le décès. La pression artérielle n'est pas incluse, car elle est influencée par le traitement. L'heure de prise en charge n'est associée ni au traitement ni au décès et n'est donc pas incluse (Fig. 1).

Utilisation du score

Les quatre méthodes d'utilisation des SP les plus fréquentes pour ajuster l'effet d'une exposition sont : l'appariement, la stratification, la pondération inverse et l'ajustement (Fig. 2) [2]. Ces quatre méthodes ont été appliquées à des cas pratiques et comparées dans la littérature [10,11].

Appariement : Des paires de patients exposés/non exposés sont formées de façon à avoir un SP proche au sein de chaque paire. Les patients ainsi appariés sont similaires dans leurs caractéristiques. On parle de population appariée. Les patients au sein de la population appariée ne sont pas indépendants. L'effet de l'exposition devra donc être estimé à l'aide de méthodes prenant en compte la non-indépendance (régression logistique conditionnelle, test du signe de Wilcoxon) [12]. L'interprétation des résultats est proche de celle d'un ECR. Il existe plusieurs approches pour l'appariement (appariement un à plusieurs, *optimal matching*, *full matching*) [13–15].

Stratification : La population est séparée en sous-groupes de SP proches. Par exemple, une séparation de la population en cinq groupes sur les quantiles du SP permettrait de réduire le biais de 90 % [2]. L'augmentation du nombre de strates diminue le biais mais augmente la variance des estimations. L'effet de l'exposition est estimé de façon univariée au sein de chaque strate et peut être moyenné entre les strates pour obtenir un effet global. Cette méthode se rapproche d'une méta-analyse d'essais cliniques quasi randomisés.

Pondération inverse : Un « poids » est attribué à chaque patient. Les patients exposés sont pondérés par *1/SP*, et les non-exposés par *1/(1-SP)*. L'effet de l'exposition sur l'événement est estimé par une régression pondérée qui accorde plus d'importance aux patients de pondération forte.

Ajustement : Le SP est directement utilisé comme une variable d'ajustement dans une régression, à la place de toutes les covariables sélectionnées pour sa construction.

Les quatre méthodes sont appliquées aux données simulées, et les résultats résumés en tableau 3. Dans le cas de l'appariement, de la pondération inverse et de l'ajustement, les OR estimés après ajustement sont significatifs et relativement proches, entre 0,78 et 0,71. L'utilisation du SP a permis de réduire les biais de confusion qui faussaient l'analyse univariée dans laquelle l'OR estimé était 1,11. Dans le cas de la stratification, le biais de confusion persiste dans un sous-groupe, et la puissance est insuffisante dans trois des sous-groupes. Cependant, l'effet moyenné sur les cinq strates, bien que non significatif, va dans le sens d'un effet protecteur du traitement (0,87 [0,62 ; 1,22] ; $p = 0,43$).

Validation du score

Les indicateurs habituels de validité des régressions logistiques (AUC, erreur de cross-validation) sont parfois utilisés dans la littérature, mais ne sont pas appropriés pour tester la validité du SP [16]. Le SP peut en effet prédire correctement l'exposition tout en omettant une variable de confusion importante qui n'a pas été intégrée dans le modèle.

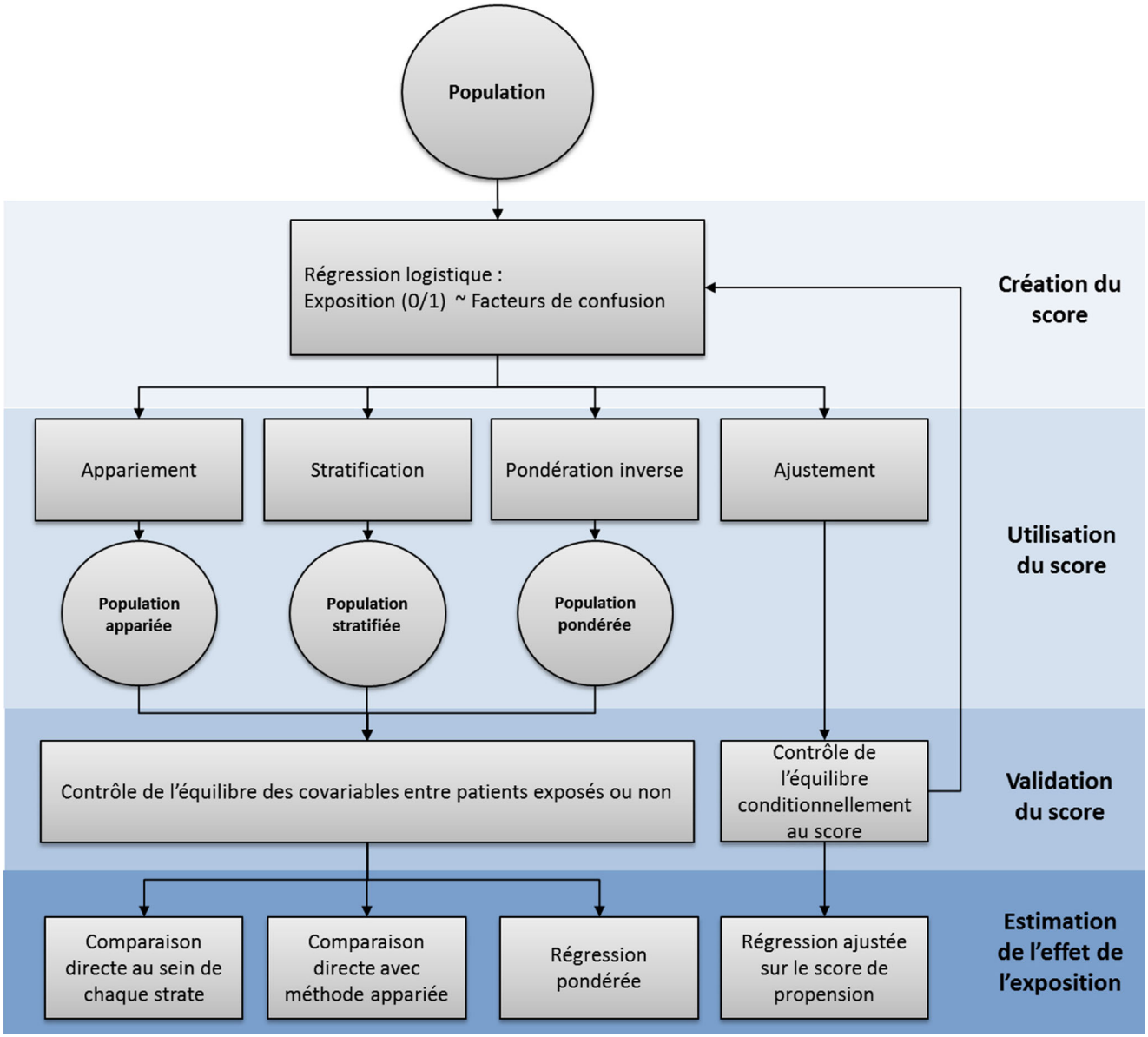

Fig. 2 Utilisation des scores de propension

Dans le cas de la stratification, de la pondération inverse et de l'appariement, une nouvelle population est créée (Fig. 2). La validation du SP se fait en comparant au sein de cette population l'équilibre entre les covariables chez les patients exposés et non exposés.

Pour cela, la différence moyenne standardisée (DMS) ou différence moyenne standardisée absolue (DMSA) entre les groupes de patients est reportée pour chaque covariable. Cette mesure permet de rendre comparables des variables exprimées dans des unités différentes. Elle est préférée aux tests inférentiels (Chi^2), car ces derniers sont sensibles à la taille des populations. Un test inférentiel peut être significatif pour un écart négligeable si la population est grande et non significatif pour un écart fort si la population est petite. Une différence inférieure à 10 ou 15 % est considérée comme négligeable. S'il reste des écarts supérieurs, il est possible de complexifier le SP par l'ajout d'interaction ou d'effets polynomiaux. Une seconde approche consiste à ajuster une seconde fois les variables déséquilibrées au moment de l'estimation de l'effet de l'exposition (Fig. 2), on parle alors d'analyse « double robuste » [17]. Si le SP ne permet pas d'obtenir une population équilibrée, les groupes de patients sont possiblement trop différents et donc non ajustables [2].

Tableau 3 Application de quatre méthodes d'utilisation du score de propension aux données simulées

Méthode	Appariement	Stratification	Pondération inverse	Ajustement
Ajustement des patients exposés et non exposés	1 733 paires sont créées à partir des 2 033 patients traités et 1 948 non traités à l'aide de la librairie « MatchIt » [19] sur R [20]. 333 patients traités sont exclus, car aucun patient non traité n'était suffisamment proche	Séparation de la population en 5 sous-groupes sur les quintiles du SP	Attribution d'un « poids » à chaque patient à partir du SP	L'effet du traitement est estimé par la régression logistique suivante : Décès (0/1) ~ Traitement + SP
Estimation de l'effet du traitement	Estimé par une régression logistique conditionnelle	Estimé par une régression logistique univariée par strate et estimation d'un effet moyen à l'aide de la librairie « metafor » [21] sur R [20]	Estimé par une régression logistique pondérée	
Odd de décès estimé	0,78 [0,66 ; 0,93] $p = 0,01$	Non significatif dans 3 groupes avec des OR allant de 0,71 à 0,81 et des p-valeurs allant de 0,10 à 0,32. Effet protecteur dans un groupe (0,63 [0,40 ; 0,96]) Effet délétère dans un groupe (1,64 [1,15 ; 2,33]) L'effet moyen sur les 5 groupes est de 0,87 [0,62 ; 1,22] ; $p = 0,43$	0,71 [0,63 ; 0,79] $p < 0,01$	0,72 [0,60 ; 0,87] $p < 0,01$

SP : score de propension

Exemple (suite) : Les différences standardisées sont représentées avant/après appariement (Fig. 3).

Dans le cas de l'ajustement, comparer directement l'équilibre de la population exposée et non exposée n'est pas possible. Austin propose une méthode pour évaluer l'équilibre conditionnellement au SP [18].

Comment présenter et interpréter les scores de propension ?

Il est essentiel de reporter les informations suivantes pour permettre la reproduction et l'évaluation de la qualité des résultats [1] :

- la méthode de calcul du SP (Modèle utilisé, variables à expliquer, variables explicatives sélectionnées, variables explicatives non retenues) ;
- l'utilisation du SP (appariement) ;
- les indicateurs de validité du SP (DMS/DMSA) (Fig. 3).

Dans le cas de l'appariement, deux informations complémentaires sont nécessaires [1] :

- les effectifs avant et après appariement ;
- les méthodes utilisées pour prendre en compte la non-ind épendance des individus.

L'interprétation des résultats reportés après ajustement sur le SP dépend de la méthode utilisée pour estimer l'effet de l'exposition (régression logistique, modèle de Cox,

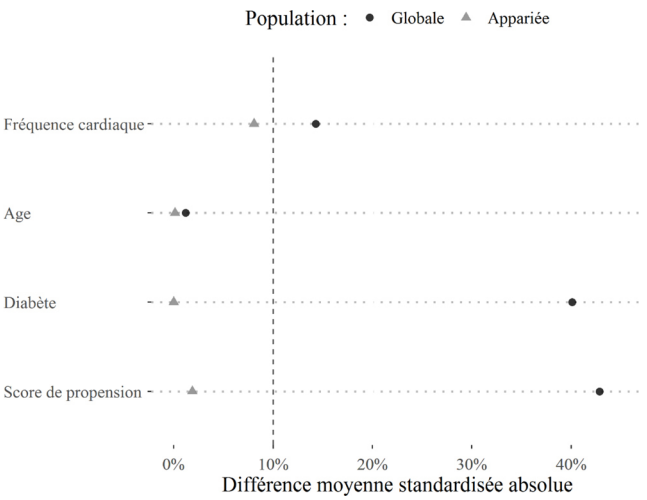

Fig. 3 Différences moyennes standardisées absolues entre les patients traités et non traités avant/après appariement Les différences moyennes standardisées absolues entre les patients traités et non traités sont reportées avant et après appariement. La différence est considérée comme négligeable si inférieure à 10 % d'écart. On observe qu'avant appariement (ronds noirs), les patients traités et non traités étaient différents sur la fréquence cardiaque et le diabète. Après appariement (triangles gris), toutes les différences sont de moins de 10 %, la population est donc équilibrée

régression linéaire). Dans tous les cas, l'effet estimé s'interprète comme l'effet de l'exposition toutes choses étant égales par ailleurs pour les variables ajustées.

Conclusion

Le SP permet de corriger une partie des biais de confusion dans les études observationnelles afin d'estimer l'association d'une exposition et d'un événement. Le SP peut s'utiliser de quatre façons, l'appariement, la stratification, la pondération inverse et l'ajustement.

Les avantages des méthodes fondées sur le SP par rapport aux méthodes classiques sont multiples : un biais possiblement plus faible sur les petites populations, l'ampleur du biais corrigé est quantifiable, et l'interprétation est proche de celle d'un ECR.

Même en utilisant des méthodes appropriées à la correction du biais, une étude observationnelle ne peut pas remplacer un ECR rigoureux pour déduire des liens de cause à effets entre une exposition et un événement, car on ne peut jamais être sûr que tous les biais de confusion ont été totalement corrigés.

Liens d'intérêts : les auteurs déclarent que le Réseau des Urgences de la Vallée du Rhône (RESUVal) est financé par l'agence régionale de santé (ARS) Auvergne-Rhône-Alpes.

Références

1. Zakrison TL, Austin PC, McCredie VA (2018) A systematic review of propensity score methods in the acute care surgery literature: avoiding the pitfalls and proposing a set of reporting guidelines. Eur J Trauma Emerg Surg 44:385–95
2. Austin PC (2011) An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behav Res 46:399–424
3. Stürmer T, Joshi M, Glynn RJ, et al (2006) 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. J Clin Epidemiol 59:437.e1-437.e24
4. Shah BR, Laupacis A, Hux JE, Austin PC (2005) Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. J Clin Epidemiol 58:550–9
5. Cepeda MS, Boston R, Farrar JT, Strom BL (2003) Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. Am J Epidemiol 158:280–7
6. Biondi-Zoccai G, Romagnoli E, Agostoni P, et al (2011) Are propensity scores really superior to standard multivariable analysis? Contemp Clin Trials 32:731–40
7. Concato J, Shah N, Horwitz RI (2000) Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med 342:1887–92
8. Brookhart MA, Schneeweiss S, Rothman KJ, et al (2006) Variable selection for propensity score models. Am J Epidemiol 163:1149–56
9. Wyss R, Girman CJ, LoCasale RJ, et al (2013) Variable selection for propensity score models when estimating treatment effects on multiple outcomes: a simulation study. Pharmacoepidemiol Drug Saf 22:77–85
10. Austin PC, Mamdani MM (2006) A comparison of propensity score methods: a case-study estimating the effectiveness of post-AMI statin use. Stat Med 25:2084–106
11. Elze MC, Gregson J, Baber U, et al (2017) Comparison of propensity score methods and covariate adjustment: evaluation in 4 cardiovascular studies. J Am Coll Cardiol 69:345–57
12. Austin PC (2011) Comparing paired vs non-paired statistical methods of analyses when making inferences about absolute risk reductions in propensity-score matched samples. Statist Med 30:1292–301
13. Austin PC (2014) A comparison of 12 algorithms for matching on the propensity score. Statist Med 33:1057–69
14. Rassen JA, Shelat AA, Myers J, et al (2012) One-to-many propensity score matching in cohort studies. Pharmacoepidemiol Drug Saf 21:69–80
15. Stuart EA, Green KM, (2008) Using full matching to estimate causal effects in nonexperimental studies: examining the relationship between adolescent marijuana use and adult outcomes. Dev Psychol 44:395–406
16. Westreich D, Cole SR, Funk MJ, et al (2011) The role of the c-statistic in variable selection for propensity score models. Pharmacoepidemiol Drug Saf 20:317–20

17. Nguyen T-L, Collins GS, Spence J, et al (2017) Double-adjustment in propensity score matching analysis: choosing a threshold for considering residual imbalance. BMC Med Res Methodol 17:78
18. Austin PC (2008) Goodness-of-fit diagnostics for the propensity score model when estimating treatment effects using covariate adjustment with the propensity score. Pharmacoepidemiol Drug Saf 17:1202–17
19. Ho DE, Imai K, King G, Stuart EA (2011) MatchIt: Nonparametric preprocessing for parametric causal inference. J. Stat. Softw. 42:1–28
20. R Core Team (2019) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria
21. Viechtbauer W (2010) Conducting meta-analyses in R with the metafor package. J Stat Softw 36:1–48