Réanimation (2012) 21:109-116
DOI 10.1007/s13546-011-0422-4

Comparaison de l'efficacité de deux thérapeutiques en l'absence de randomisation : intérêts et limites des méthodes utilisant les scores de propension

Advantages and limitations of propensity score methods to analyze non-randomized clinical trials

E. Gayat · R. Porcher

Reçu le 17 juillet 2011 ; accepté le 15 novembre 2011

Résumé Le score de propension est défini comme la probabilité d'un sujet de recevoir un traitement spécifique conditionnellement à ses caractéristiques observées. Les méthodes utilisant le score de propension ont été de plus en plus utilisées dans la littérature médicale durant les dix dernières années. Toutefois, la qualité de l'utilisation de ces méthodes et du rapport des études les utilisant pourrait être améliorée, en particulier dans la littérature de réanimation. Après une présentation de la théorie des scores de propension, illustrée d'un exemple, cette mise au point propose des recommandations pour les investigateurs dans le but d'améliorer la qualité du rapport des études utilisant des méthodes de score de propension. ***Pour citer cette revue : Réanimation 21 (2012).***

Mots clés Score de propension · Propension · Appariement · Méthodologie · Réanimation

Abstract The propensity score (PS), defined as a patient's probability of receiving a specific treatment conditional on the observed covariates, is a method that leads to control the bias associated with the non-randomly assigned treatment allocation in observational studies. PS methods have been increasingly used in the last 10 years. The quality of reporting PS in the intensive care medicine literature should be improved. This review aims to present the theory of PS, illustrated with an example and provides recommendations to the investigators in order to improve the reporting of PS analyses. ***To cite this journal: Réanimation 21 (2012).***

Keywords Propensity score · Propensity · Matching · Methodology · Intensive Care

E. Gayat · R. Porcher (✉)
Inserm U717, F-75010 Paris, France
e-mail : raphael.porcher@paris7.jussieu.fr

Épidémiologie clinique et biostatistique, UMR-S717,
université Paris-Diderot, Sorbonne Paris-Cité,
F-75010 Paris, France

R. Porcher
Département de biostatistique et d'informatique médicale,
hôpital Saint-Louis, 1, avenue Claude-Vellefaux,
F-75010 Paris, France

Introduction

L'essai thérapeutique randomisé (ETR) est considéré comme le *gold standard* pour l'évaluation de l'efficacité d'un médicament, d'un dispositif médical ou encore d'une stratégie de prise en charge. Certains statisticiens recommandaient encore récemment aux comités de lecture des journaux scientifiques de ne pas accepter de publier des études observationnelles compte tenu des nombreux biais auxquelles elles étaient associées [1]. Toutefois, l'ETR peut être difficile à envisager ou à mettre en place dans différentes situations : quand il s'agit d'une affection très rare (difficulté de recrutement), dans les situations d'extrême urgence ou quand le tirage au sort peut être considéré comme non éthique [2]. De surcroît, la validité externe d'un ETR peut être moindre que celle d'une étude observationnelle [3–5]. En effet, dans un ETR, seuls sont recrutés les sujets qui, d'une part, répondent aux critères prédéfinis d'inclusion et de non-inclusion et qui, d'autre part, acceptent de participer, ce qui peut introduire un biais de sélection rendant la population incluse dans l'essai non représentative de la population cible. Par ailleurs, le processus complexe d'inclusion dans un essai randomisé qui peut varier selon le centre, le médecin ou le patient peut limiter la confiance que l'on peut avoir dans les résultats d'un ETR et ainsi en limiter grandement l'applicabilité dans la pratique quotidienne.

Dans les études observationnelles, les investigateurs ne contrôlent pas l'allocation du traitement, ce qui peut entraîner d'importantes différences entre les groupes de patients

traités et non traités, à la fois en ce qui concerne leurs caractéristiques observées mais aussi leurs caractéristiques non observées. Il a été clairement établi que ces différences entraînent un biais dans l'estimation de l'effet du traitement [6]. Si une méthode statistique permettait de contrôler ce biais, cela pourrait relancer l'intérêt des études observationnelles dans différentes situations. Les deux principales stratégies permettant de contrôler les biais de sélection inhérents aux études observationnelles sont :

- l'ajustement de l'effet du traitement, qui est basé sur la relation entre les variables pronostiques et le critère de jugement ;
- la stratégie consistant à modéliser la probabilité d'être traité, cette dernière reposant sur la relation entre les variables pronostiques et le processus d'allocation du traitement.

Le score de propension est défini comme la probabilité d'un sujet de recevoir un traitement spécifique conditionnellement à ses caractéristiques observées. Rosenbaum et Rubin ont montré que, lors de l'analyse de données observationnelles, le fait de conditionner sur le score de propension permettait d'obtenir une estimation non biaisée de l'effet du traitement sous certaines conditions [7].

Les études cliniques impliquant des méthodes utilisant les scores de propension sont de plus en plus nombreuses dans la littérature médicale [8–12] (Fig. 1) ; cette tendance s'observe également dans les journaux spécialisés de réanimation [13].

Les objectifs de cette mise au point sont :

- d'exposer de façon didactique la théorie des scores de propension ;
- d'illustrer leur utilisation dans la littérature de réanimation à partir d'un exemple ;
- d'apporter quelques conseils sur l'utilisation des scores de propension dans la pratique de la recherche clinique.

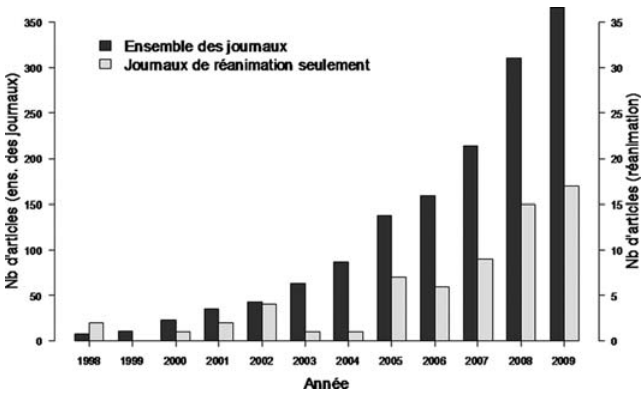

Fig. 1 Nombre d'articles publiés rapportant des études basées sur les scores de propension au sein de la littérature médicale entre 1998 et 2009 (d'après Gayat et al. [13], avec permission)

Principes des scores de propension

Bases théoriques

On note X les caractéristiques observées de chaque sujet et Z l'indicatrice de traitement (Z valant 1 si le sujet a reçu le traitement d'intérêt et 0 s'il est considéré comme dans le groupe témoin). X est un vecteur qui peut inclure un grand nombre de caractéristiques décrivant le sujet. Le score de propension, classiquement noté $e(X)$, peut être défini pour chaque individu comme la probabilité qu'il avait de recevoir le traitement d'intérêt, étant donné ses caractéristiques, que ce sujet ait ou non effectivement reçu le traitement [7]. La théorie du score de propension repose sur trois hypothèses majeures :

- l'allocation du traitement à chaque individu a été faite de façon indépendante d'un individu à l'autre (en terme statistique, cela se traduit par : « les Z sont indépendants conditionnellement à X ») ;
- tous les facteurs de confusion ont été mesurés (c'est-à-dire que toutes les variables associées à l'allocation du traitement et au devenir du patient sont connues et mesurées) ;
- l'allocation du traitement est « fortement ignorable » (traduction de l'anglais *strongly ignorable treatment assignement*).

Cela signifie que, une fois tenu compte des caractéristiques du patient, la réponse au traitement peut dépendre du traitement lui-même mais pas du fait qu'il ait été décidé de donner le traitement à ce patient en particulier. Il est à noter qu'on considère trois types de variables caractérisant un sujet : les facteurs de confusion potentiels qui sont les variables associées à l'allocation du traitement mais pas nécessairement au pronostic, les vrais facteurs de confusion qui sont les variables associées à la fois à l'allocation du traitement et au pronostic et enfin les non-confondeurs qui sont les variables ni associées à l'allocation du traitement ni associées au pronostic. Même s'il est difficile de déterminer a priori à quelle catégorie appartient telle ou telle caractéristique, les variables à inclure dans le modèle de score de propension, pour que celui-ci soit valide, sont les vrais facteurs de confusion [10,14].

Sous les hypothèses susmentionnées, un patient traité et un patient non traité ayant le même score de propension peuvent être considérés comme ayant été alloués à l'un des deux groupes par tirage au sort, c'est le concept de « pseudorandomisation » [7]. En effet, dans un ETR, si les deux groupes sont équilibrés, un sujet inclus a la même probabilité d'appartenir à l'un des deux groupes, et on peut ainsi considérer que le score de propension est égal à 50 %.

De nombreux autres travaux ont par ailleurs confirmé que l'utilisation des scores de propension pour l'analyse de données observationnelles permettait d'obtenir une estimation

non biaisée ou moins biaisée de l'effet du traitement, que le critère de jugement soit binaire, continu ou censuré et qu'il s'agisse d'un effet marginal (effet du traitement dans une population où les caractéristiques des sujets sont distribuées selon une loi de probabilité) ou conditionnel (effet du traitement à caractéristiques d'un individu données) [15–23].

En pratique, le score de propension n'est pas connu et doit être estimé. Pour ce faire, le modèle le plus couramment utilisé est le modèle logistique, où la variable à expliquer est l'allocation du traitement, et les variables explicatives sont les caractéristiques du patient.

Techniques utilisant les scores de propension

Les quatre techniques les plus communes utilisant un score de propension sont :

- l'appariement ;
- la stratification ;
- l'ajustement ;
- la pondération, cette dernière ayant été décrite plus récemment.

Appariement sur le score de propension (en anglais : matching *ou* matched sampling*)*

L'appariement est la technique la plus usuelle ; elle consiste à créer une paire constituée d'un patient témoin et d'un patient traité ayant des caractéristiques semblables. Même s'il peut sembler aisé d'apparier des patients, il est souvent difficile de trouver des sujets qui ont l'ensemble de leurs caractéristiques identiques, voire simplement similaires. L'appariement sur le score de propension permet de résoudre cette difficulté en permettant de contrôler un grand nombre de caractéristiques simultanément via l'appariement sur une variable unique, en l'occurrence le score de propension [24–27]. Cette analyse conduit cependant à exclure les patients qui n'ont pas pu être appariés.

Stratification sur le score de propension (en anglais : stratification *ou* subclassification*)*

La stratification sur le score de propension consiste à estimer l'effet du traitement au sein de strates définies le plus souvent par les quintiles ou les déciles du score de propension.

Ajustement sur le score de propension (en anglais : adjustment*)*

Comme pour un modèle multivarié « classique », l'ajustement sur le score de propension consiste à introduire le score de propension comme une variable explicative dans un modèle de régression dans lequel la variable dépendante est le critère de jugement, et l'autre variable explicative le traitement reçu. Comme nous le discuterons plus loin, la performance d'un modèle de score de propension est évaluée par sa capacité à équilibrer les caractéristiques des sujets entre le groupe traité et le groupe non traité. Ainsi, le nombre de variables à inclure dans le modèle de régression n'est pas limité, et les indices de performance habituels, comme l'aire sous la courbe ROC (*receiver operating characteristic curve*) du modèle ou des tests d'adéquation de modèle, ne sont pas de bons critères d'évaluation [18,28].

Pondération par l'inverse du score de propension (en anglais : inverse-probability-of-treatment weighting *[IPTW])*

L'idée de pondérer les sujets traités et les sujets témoins par leur score de propension respectif dans le but de les rendre plus représentatifs de la population d'intérêt a été d'abord proposée par Rubin [29] puis plus récemment développée par Lunceford et Davidian [30]. Le poids affecté aux sujets réellement traités est l'inverse de leur score de propension et le poids affecté aux sujets réellement non traités est l'inverse de leur probabilité de ne pas être traité (c'est-à-dire 1 moins leur score de propension). Par rapport à l'appariement, cette technique présente l'avantage de conserver l'ensemble de l'échantillon initial dans l'analyse. En effet, la réduction de la taille de l'échantillon peut conduire à une baisse de la puissance statistique.

Cas particulier de l'appariement sur le score de propension

L'appariement sur le score de propension est la technique actuellement la plus utilisée et, comme déjà discuté précédemment, permet une estimation non biaisée de l'effet du traitement. Les performances des différentes méthodes d'appariement sur le score de propension ont fait récemment l'objet d'une étude [17].

Mise en œuvre pratique

Lorsqu'on choisit de faire un appariement sur le score de propension, il est nécessaire de fixer trois paramètres :

- l'équilibre de l'appariement ;
- le fait d'apparier avec ou sans remise ;
- le choix de l'algorithme d'appariement.

La plupart des études cliniques utilisent un appariement un pour un (1:1). Dans ce cas, des paires de sujets traité et non traité ayant un score de propension similaire sont constituées. Il est également possible d'apparier un patient traité à plusieurs patients non traités ou le contraire, mais cela est rarement fait en pratique.

Dans le cas d'un appariement sans remise, un sujet témoin (ou non traité) qui a déjà été apparié avec un sujet traité n'est plus disponible pour être apparié avec un autre sujet traité. En revanche, l'appariement avec remise peut conduire à ce qu'un sujet témoin soit apparié plusieurs fois à différents sujets traités. Ce type d'appariement avec remise, peu utilisé en pratique, peut créer des difficultés dans l'estimation de la variance de l'effet du traitement [31].

On distingue deux algorithmes d'appariement : l'appariement séquentiel (en anglais : *greedy matching*) et l'appariement optimal. L'appariement séquentiel consiste à sélectionner au hasard un sujet traité et à lui adjoindre ensuite le sujet témoin ayant le score de propension le plus proche. Ce sujet témoin est sélectionné même s'il aurait pu servir comme meilleur témoin pour le patient traité suivant, au sens d'un score de propension plus proche. Dans le cas d'un appariement optimal, les paires de sujets témoins et traités sont formées de façon à minimiser globalement la différence de score de propension intrapaire. Dans ce cas, une paire formée au préalable peut être défaite si une autre combinaison peut entraîner une diminution de la différence globale.

Les procédures d'appariement les plus couramment utilisées sont la technique du plus proche voisin au sein d'un intervalle de tolérance sur le score de propension (en anglais : *caliper*) et l'appariement de la cinquième à la première décimale (en anglais : *5 to 1 digit matching*). La première consiste à apparier un sujet traité au sujet témoin ayant le score de propension le plus proche avec une différence maximale autorisée de score de propension entre les membres de la paire. La différence maximale est généralement définie comme une fraction de l'écart-type (ET) du score de propension ; la valeur la plus souvent utilisée est 0,2 ET, même s'il semble qu'utiliser une différence maximale de 0,05 ET puisse conduire à de meilleurs résultats, au prix d'une réduction de la taille de l'échantillon apparié [17]. Si on utilise la seconde approche, le sujet témoin à apparier au sujet traité est d'abord recherché parmi les sujets ayant un score similaire à la cinquième décimale. S'il n'y en a pas, on recherche parmi ceux ayant un score de propension égal à la quatrième décimale, et ainsi de suite jusqu'à la première décimale. Les sujets traités et non traités qui restent non appariés à l'issue de la procédure sont donc écartés de l'analyse. Cette dernière méthode a souvent été utilisée dans la littérature médicale, malgré le fait que ses performances n'aient jamais été évaluées de façon rigoureuse, contrairement à l'appariement au sein d'un intervalle de tolérance.

Évaluation de la qualité de l'appariement sur le score de propension

Concernant l'évaluation de la qualité de l'appariement sur le score de propension en lui-même, il a été clairement démontré que l'aire sous la courbe ROC ou un test d'adéquation de modèle n'étaient pas des moyens adéquats pour évaluer la capacité d'un score de propension donné à correctement équilibrer les caractéristiques des groupes témoin et traité au sein d'un échantillon apparié sur le score de propension [31,32]. La mesure de l'adéquation du modèle ou de la capacité de discrimination n'est pas non plus utilisable pour détecter l'absence d'un vrai facteur de confusion dans le modèle de score de propension [28].

En effet, l'objectif de l'appariement sur le score de propension est d'obtenir des groupes de sujets traités et non traités équilibrés en termes de variables pronostiques. Ainsi, une mesure du succès de l'appariement doit préférentiellement reposer sur une mesure du déséquilibre entre les caractéristiques de ces deux groupes de patients. Il a été proposé d'utiliser la différence standardisée (d), qui est définie comme :

$$d = \frac{100 \times (\overline{x}_{traités} - \overline{x}_{témoin})}{\sqrt{\frac{s^2_{traités} + s^2_{témoin}}{2}}}$$

avec $\overline{x}_{traités}$, $s^2_{traités}$, $\overline{x}_{témoin}$, $s^2_{témoin}$ les moyennes et variances respectivement dans le groupe traité et dans le groupe témoin. On estime que l'appariement est un succès si la différence standardisée résiduelle après appariement est limitée pour l'ensemble des confondeurs. Une valeur de d inférieure ou égale à 10 % a été déterminée comme acceptable, de façon toutefois empirique [33].

Analyse du critère principal de jugement après appariement sur le score de propension

Le modèle à utiliser pour analyser le critère de jugement dépend d'abord du type de critère dont il s'agit (continu, binaire ou censure). Toutefois, l'appariement sur le score de propension peut induire un certain degré de corrélation entre les sujets appartenant à la même paire, ce qui peut influencer l'estimation de la variance de l'effet du traitement. Ainsi, il est recommandé de tenir compte de l'appariement dans l'analyse, en utilisant par exemple la différence intrapaire, un modèle à effets aléatoires ou un estimateur robuste de la variance [34–36].

Un exemple issu de la littérature : effet des inotropes sur la survie à court terme des patients hospitalisés pour insuffisance cardiaque aiguë

Il s'agit d'une étude effectuée à partir des données d'un registre international, appelé *Acute Heart Failure Global Survey of Standard Treatment* (ALARM-HF) qui a recueilli les données cliniques, de traitement et de survie intrahospitalière de près de 5 000 patients admis pour insuffisance cardiaque aiguë [37].

Un effet potentiellement délétère des inotropes dans cette indication a été décrit dans la littérature, toutefois cela était alors principalement basé sur des avis d'experts [38–41]. Par ailleurs, il semblait difficile, compte tenu du manque d'ambivalence prévisible des cliniciens quant au choix d'administrer au non un inotrope en cas d'insuffisance cardiaque aiguë, d'envisager un ETR. Ainsi, l'efficacité des inotropes a été testée au sein du registre ALARM-HF en utilisant un appariement sur le score de propension.

Dans cet exemple, le critère principal de jugement était la survie intrahospitalière. La Figure 2 représente les dix variables pour lesquelles la différence standardisée avant appariement sur le score de propension était la plus grande. On constate que, de façon attendue, les inotropes ont été administrés aux patients présentant la gravité la plus importante (pression artérielle plus basse, classe NYHA plus élevée, âge plus élevé, présence de signes d'hypoperfusion périphérique). L'utilisation d'un inotrope était d'ailleurs associée à une mortalité intrahospitalière plus importante (Fig. 3A).

La question était alors de savoir si la surmortalité associée à l'utilisation d'un inotrope était due au fait que le traitement était administré aux patients ayant le pronostic le plus mauvais, à un effet délétère propre de l'inotrope ou à une association des deux.

Un score de propension a alors été développé à l'aide d'un modèle logistique. Un appariement 1 pour 1, sans remise, utilisant la technique du plus proche voisin dans un intervalle de 0,2 ET du score de propension (sur l'échelle du modèle logistique) a été effectué. L'appariement a permis d'obtenir un équilibre satisfaisant entre le groupe traité par inotrope et le groupe témoin (Fig. 2). L'estimation de l'effet du traitement au sein de cet échantillon apparié a mis en évidence la persistance d'une surmortalité des patients recevant un inotrope par rapport aux patients n'en recevant pas (Fig. 3B). On peut ainsi conclure qu'en comparant des patients pris en charge pour insuffisance cardiaque aiguë ayant la même probabilité de recevoir un traitement par inotrope compte tenu de leurs caractéristiques, un effet délétère propre aux inotropes a été mis en évidence par rapport a un groupe témoin de patients n'ayant pas reçu d'inotropes.

Ce résultat doit donc conduire tout au moins à une utilisation prudente des inotropes dans cette indication et pourrait être le point de départ à la planification d'un ETR.

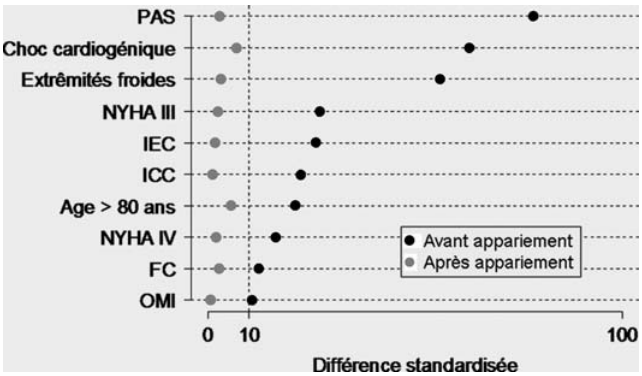

Fig. 2 Représentation graphique de la différence standardisée avant et après appariement sur le score de propension Seules les dix variables, avec les différences standardisées avant appariement, les plus importantes sont représentées. On constate que, malgré des différences très importantes avant appariement, l'utilisation d'un score de propension a permis d'obtenir une balance « satisfaisante », c'est-à-dire une différence standardisée inférieure a 10 %, pour l'ensemble de ces caractéristiques.
PAS : pression artérielle systolique ; NYHA : New York Heart Association ; IEC : inhibiteur de l'enzyme de conversion ; ICC : insuffisance cardiaque chronique ; FC : fréquence cardiaque ; OMI : œdème des membres inférieurs.

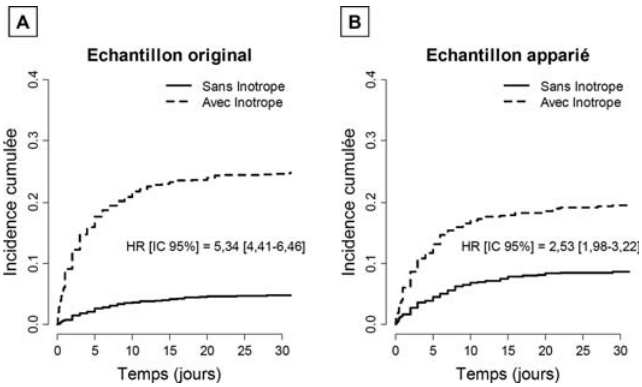

Fig. 3 Courbes d'incidence cumulées pour la mortalité intrahospitalière des patients ayant reçu un inotrope et/ou un vasopresseur versus celle des patients n'en ayant pas reçu
On constate que la surmortalité intrahospitalière observée des patients ayant reçu un inotrope et/ou un vasopresseur est plus faible mais toujours présente après appariement sur le score de propension. Ce résultat implique que la surmortalité observée était en partie expliquée par le processus d'allocation du traitement (la gravité des patients ayant reçu le traitement était plus importante) mais qu'il existe un effet délétère propre de l'utilisation d'inotropes et/ou de vasopresseurs chez les patients présentant une insuffisance cardiaque aiguë. HR : *hazard ratio*, rapport des risques instantanés

Utilisation en recherche clinique des scores de propension

Nous proposons quelques recommandations simples visant à aider à la fois les investigateurs potentiellement intéressés par les scores de propension et les lecteurs d'articles

rapportant une étude ayant utilisé une technique impliquant les scores de propension. Nous insistons encore une fois particulièrement sur l'appariement sur le score de propension, parce que, d'une part, il s'agit de la technique la plus populaire et, d'autre part, il a été démontré qu'elle était la plus performante dans différentes situations [15,16,19].

Il est important que les auteurs expliquent comment ils ont constitué leur modèle de score de propension, en particulier les variables incluses doivent être détaillées et leur choix doit être justifié. En effet, les auteurs doivent démontrer qu'ils ont fait le maximum afin d'identifier et de recueillir tous les facteurs de confusion potentiels, car omettre un facteur de confusion entraînera un biais dans l'estimation de l'effet du traitement. Il faut ensuite vérifier quel moyen ont utilisé les auteurs pour évaluer et valider leur analyse. Comme discuté plus haut, les statistiques globales de discrimination ou de qualité d'ajustement n'ont pas d'intérêt particulier ici. De même, en cas d'appariement, l'objectif étant d'obtenir des groupes de patients traités et non traités équilibrés sur leurs caractéristiques, il est primordial que cet équilibre soit rapporté et correctement évalué. Une bonne façon d'évaluer l'équilibre est de calculer la différence standardisée de chaque variable entre le groupe témoin et le groupe traité, et ce, avant et après appariement ; une représentation graphique similaire à la Figure 2 permet de faire une synthèse lisible pour le lecteur. Enfin, il est important que le lecteur soit capable d'estimer l'impact qu'a eu l'analyse par score de propension sur l'estimation de l'effet du traitement ; il est donc intéressant de rapporter aussi les résultats de l'analyse brute du critère de jugement. Des analyses de sensibilité aux choix qui ont été faits dans la construction et l'utilisation du score de propension sont aussi souhaitables. Nous proposons une synthèse de ces recommandations sous la forme d'une *check-list* (Tableau 1) qui peut à la fois être utile aux lecteurs et aux rédacteurs d'articles rapportant une analyse impliquant les scores de propension. À titre d'exemple, nous avons appliqué cette *check-list* à l'article pris en exemple ci-dessus.

Par ailleurs, la Figure 4 est une tentative de synthèse des différentes étapes de planification et d'analyse d'une étude observationnelle utilisant une méthode de score de propension.

Tableau 1 Proposition de grille de lecture d'un article rapportant une étude utilisant les scores de propension

Éléments à vérifier	**Application à la référence [37]**
Type de cohorte	Rétrospective
Type de données	Registre
Type d'exposition	Traitement par inotrope et/ou vasopresseur en cas d'insuffisance cardiaque aiguë
Critère principal de jugement	Survie intrahospitalière (critère censuré)
Nombre de patients	
Dans le groupe « expérimental »	1 617
Dans le groupe « témoin »	3 256
Développement du score de propension	
Nombre de variables incluses	30
Choix des variables justifié	Oui
Modèle utilise	
Technique utilisée	Appariement
En cas de stratification, nombre de strates	–
En cas d'ajustement, modèle utilisé	–
En cas d'appariement	
Procédure utilisée	Plus proche voisin avec un intervalle de tolérance égal à 0,2 ET de logit (SP)
Balance	1/1
Avec remise	Non
Nombre de patients par groupe	954/954
Méthode utilisée pour évaluer le déséquilibre entre les groupes	Différence standardisée
Analyse du critère de jugement	
Modèle utilise	Modèle de Cox
Prise en compte de l'appariement (le cas échéant)	Oui (utilisation d'un estimateur robuste de la variance)

SP : score de propension ; ET : erreur standard.

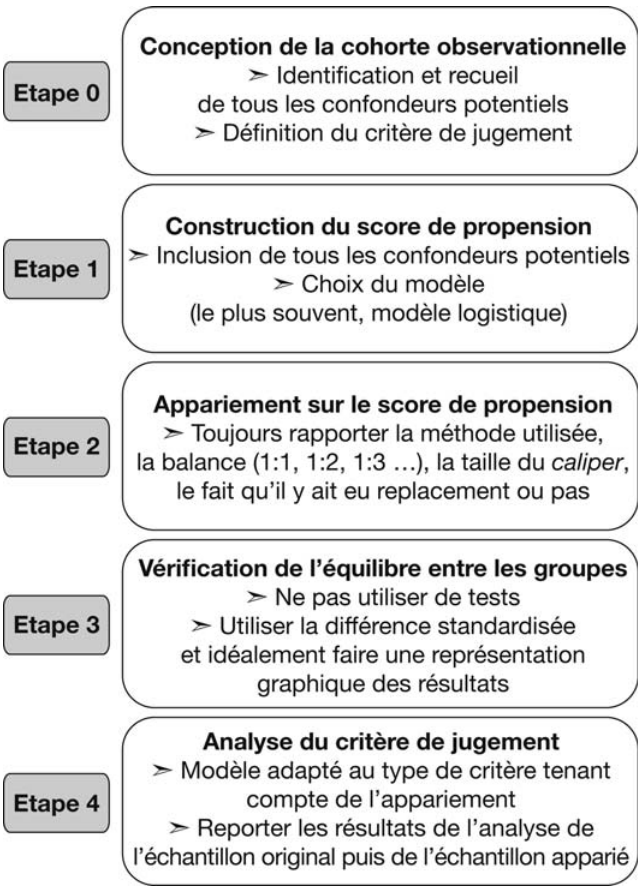

Fig. 4 Proposition de planification d'une étude observationnelle utilisant les méthodes de scores de propension (d'après Gayat et al. [13], avec permission)

Comme nous l'avons déjà souligné, la théorie des scores de propension est sous-tendue par le fait de mesurer tous les confondeurs. Cela explique pourquoi l'étape 0 est cruciale pour s'assurer que les confondeurs ont bien été identifiés et ainsi bien mesurés.

Discussion

Il a été démontré que l'utilisation croissante des scores de propension dans la littérature médicale était malheureusement associée à un certain degré de mésusage et à une hétérogénéité dans la qualité du rapport des études ayant utilisé cette méthode [8–13]. Même si les méthodes de score de propension permettent en théorie d'obtenir une estimation non biaisée de l'effet d'un traitement par l'analyse de données observationnelles, en permettant de recréer une situation de « quasi-randomisation », elles doivent être maniées avec précaution. En effet, il faut se rappeler les conditions de validité et les hypothèses qui sous-tendent la théorie des scores de propension. En particulier, pour être valide, il est indispensable que toutes les variables associées avec l'allocation du traitement aient été recherchées et incluses dans le modèle du score de propension.

Ainsi, tout comme un ETR, une étude observationnelle incluant une analyse utilisant les scores de propension doit idéalement être minutieusement planifiée avant sa réalisation, et l'identification préalable de tous les confondeurs potentiels est une étape clé de cette planification. Toutefois, l'application des méthodes de score de propension, même dans le cadre d'une étude observationnelle correctement planifiée, permet d'obtenir un équilibre entre les sujets traités et les sujets témoins uniquement sur leurs caractéristiques observées. En revanche, un ETR permet d'obtenir des groupes équilibrés pour l'ensemble de leurs caractéristiques, qu'elles soient observées mais aussi non observées.

Conflit d'intérêt : les auteurs déclarent ne pas avoir de conflit d'intérêt.

Références

1. Ellenberg JH (1994) Selection bias in observational and experimental studies. Stat Med 13:557–67
2. Rossi P, Freeman H (1993) A Systematic Approach. 5th Edition ed. Sage Publications, Inc, Newburgy Park, CA
3. Corrie P, Shaw J, Harris R (2003) Rate limiting factors in recruitment of patients to clinical trials in cancer research: descriptive study. BMJ 327:320–1
4. Fossa SD, Skovlund E (2002) Selection of patients may limit the generalizability of results from cancer trials. Acta Oncol 41:131–7
5. Guyatt GH, Sackett DL, Cook DJ (1994) Users' guides to the medical literature. II. How to use an article about therapy or prevention. B. What were the results and will they help me in caring for my patients? Evidence-Based Medicine Working Group. JAMA 271:59–63
6. Pocock SJ, Elbourne DR (2000) Randomized trials or observational tribulations? N Engl J Med 342:1907–9
7. Rosenbaum P, Rubin D (1983) The central role of the propensity score in observational studies for causal effect. Biometrika 70:41–55
8. Austin PC (2007) Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: a systematic review and suggestions for improvement. J Thorac Cardiovasc Surg 134:1128–35
9. Austin PC (2008) A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. Stat Med 27:2037–49
10. 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
11. Sturmer 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 59:437–47
12. Weitzen S, Lapane KL, Toledano AY, et al (2004) Principles for modeling propensity scores in medical research: a systematic literature review. Pharmacoepidemiol Drug Saf 13:841–53
13. Gayat E, Pirracchio R, Resche-Rigon M, et al (2010) Propensity scores in intensive care and anaesthesiology literature: a systematic review. Intensive Care Med 36:1993–2003

14. Brookhart MA, Schneeweiss S, Rothman KJ, et al (2006) Variable selection for propensity score models. Am J Epidemiol 163:1149–56
15. Austin PC (2007) The performance of different propensity score methods for estimating marginal odds ratios. Stat Med 26:3078–94
16. Austin PC (2008) The performance of different propensity-score methods for estimating relative risks. J Clin Epidemiol 61:537–45
17. Austin PC (2009) Some methods of propensity-score matching had superior performance to others: results of an empirical investigation and Monte Carlo simulations. Biom J 51:171–84
18. Austin PC, Grootendorst P, Anderson GM (2007) 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 26:734–53
19. Austin PC, Grootendorst P, Normand SL, Anderson GM (2007) Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a Monte Carlo study. Stat Med 26:754–68
20. Hill J, Reiter JP (2006) Interval estimation for treatment effects using propensity score matching. Stat Med 25:2230–56
21. Rubin D, Thomas N (2000) Combining propensity score matching with additional adjustments for prognostic covariates. J Am Stat Assoc 95:573–85
22. Senn S, Graf E, Caputo A (2007) Stratification for the propensity score compared with linear regression techniques to assess the effect of treatment or exposure. Stat Med 26:5529–44
23. Kurth T, Walker AM, Glynn RJ, et al (2006) Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. Am J Epidemiol 163:262–70
24. Carpenter R (1977) Matching when covariables are normally distributed. Biometrika 64:299–307
25. Rubin D (1976) Matching methods that are equal percent bias reducing: some examples. Biometrics 35:417–46
26. Rubin D (1979) Using multivariate matched sampling and regression adjustment to control bias in observational studies. J Am Stat Assoc 74:318–24
27. Rubin D (1980) Bias reduction using Mahalanobis metric matching. Biometrics 36:293–8
28. Weitzen S, Lapane KL, Toledano AY, et al (2005) Weaknesses of goodness-of-fit tests for evaluating propensity score models: the case of the omitted confounder. Pharmacoepidemiol Drug Saf 14:227–38
29. Rubin D (2001) Using propensity scores to help design observational studies. Health Serv Out Res Method 2:169–88
30. Lunceford JK, Davidian M (2004) Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Stat Med 23:2937–60
31. Austin PC (2008) Assessing balance in measured baseline covariates when using many-to-one matching on the propensity-score. Pharmacoepidemiol Drug Saf 17:1218–25
32. 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(12):1202–17
33. Austin PC (2009) Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat Med 28(25):3083–107
34. Austin PC (2011) Comparing paired vs non-paired statistical methods of analyses when making inferences about absolute risk reductions in propensity-score matched samples. Stat Med 30(11):1292–301
35. Austin PC (2009) Type I error rates, coverage of confidence intervals, and variance estimation in propensity-score matched analyses. Int J Biostat 5(1). pii: article 13
36. Hill J (2008) 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 27:2055–61; discussion 66–9
37. Mebazaa A, Parissis J, Porcher R, et al (2011) Short-term survival by treatment among patients hospitalized with acute heart failure: the global ALARM-HF registry using propensity scoring methods. Intensive Care Med 37:290–301
38. Abraham WT, Adams KF, Fonarow GC, et al (2005) In-hospital mortality in patients with acute decompensated heart failure requiring intravenous vasoactive medications: an analysis from the Acute Decompensated Heart Failure National Registry (ADHERE). J Am Coll Cardiol 46:57–64
39. Dickstein K, Cohen-Solal A, Filippatos G, et al (2008) ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure 2008: the Task Force for the Diagnosis and Treatment of Acute and Chronic Heart Failure 2008 of the European Society of Cardiology. Developed in collaboration with the Heart Failure Association of the ESC (HFA) and endorsed by the European Society of Intensive Care Medicine (ESICM). Eur Heart J 29:2388–442
40. Singer M (2007) Catecholamine treatment for shock-equally good or bad? Lancet 370:636–7
41. Thackray S, Easthaugh J, Freemantle N, Cleland JG (2002) The effectiveness and relative effectiveness of intravenous inotropic drugs acting through the adrenergic pathway in patients with heart failure-a meta-regression analysis. Eur J Heart Fail 4:515–29