UNIVERSITE PARIS DESCARTES

Ecole doctorale Pierre 393 Louis de Santé Publique à Paris :
Epidémiologie et Sciences de l'Information Biomédicale

Utilisation du score de propension dans les études observationnelles évaluant une procédure chirurgicale.

Présenté par Guillaume LONJON

Thèse de doctorat d'Epidémiologie Clinique

Thèse dirigée par le Pr Isabelle BOUTRON et le Dr Raphael PORCHER

Centre de Recherche Epidémiologique et statistiques Sorbonne Paris Cité
Equipe méthodes de l'évaluation thérapeutique des maladies chroniques

Présentée et soutenue publiquement le 17 mars 2017

Devant le jury composé de:

Pr Isabelle BOUTRON (directrice de thèse)
Pr Bertrand DOUSSET
Pr Paul LANDAIS
Pr Laurence MEYER (rapporteur)
Pr Rémy NIZARD
Pr Hugues PASCAL-MOUSSELARD (rapporteur)

Remerciements

A Madame le Professeur Isabelle Boutron,
Je te remercie d'avoir dirigé ce travail. Je te remercie aussi de m'avoir aidé et accompagné pendant toutes ces années. Tes qualités d'enseignement et d'encadrement sont pour moi un exemple à suivre. J'espère que nous pourrons continuer à travailler ensemble dans le futur. Sois assurée de mon immense estime et de mon admiration.

A Monsieur le Docteur Raphaël Porcher.
Je te remercie d'avoir dirigé ce travail. En encadrant ce projet tu as été bien plus que la caution statistique de ce travail. Je te remercie aussi pour ta disponibilité. J'espère pouvoir encore te déranger dans les années à venir. Sois assuré de mon immense estime et de mon admiration.

A Monsieur le Professeur Hugues Pascal-Mousselard,
Je te remercie d'avoir accepté d'être le rapporteur de ce travail qui peut paraitre bien loin de notre spécialité. Je te remercie pour cette tâche et pour tout le reste. J'espère que nous pourrons travailler ensemble sur des projets cliniques traitant de chirurgie rachidienne. Sois assuré de mon immense estime et de mon admiration.

A Madame le Professeur Laurence Meyer,
Je vous remercie d'avoir accepté d'être le rapporteur de ce travail. Vos critiques et commentaires ont permis de l'améliorer. Soyez assurée de mon immense estime.

A Monsieur le Professeur Rémy Nizard,
Je te remercie d'avoir accepté de juger ce travail. Je tiens aussi à te remercier de m'avoir permis de rencontrer Le Professeur Philippe Ravaud et Isabelle Boutron. Sans toi je ne serais pas là. Je te remercie pour tout ce que tu m'as apporté, tu es un exemple scientifique et chirurgical. Sois assuré de mon immense estime et de mon admiration.

A Monsieur le Professeur Bertrand Dousset,
Je vous remercie d'avoir accepté de juger ce travail. J'espère pouvoir travailler avec vous dans les années à venir. Soyez assuré de mon immense estime.

A Monsieur le Professeur Paul Landais,
Je vous remercie d'avoir accepté de juger ce travail et d'être venu de loin pour ça. Soyez assuré de mon immense estime.

A ceux qui m'ont aidé dans ce travail, et dans le reste de ma vie professionnelle.
Je vous remercie car ma vie professionnelle est un épanouissement et c'est grâce à vous.

A mes proches, ma famille, ma femme et mes enfants.
Vous êtes ma raison de vivre et la force qui me fait avancer. Je vous aime.

Résumé

La chirurgie a évolué dans les dernières décennies. Les techniques ont été rapidement modifiées et sont de plus en plus proposées. L'évaluation des techniques chirurgicales est donc devenue un enjeu de santé publique. Cependant l'évaluation des procédures chirurgicales par des essais contrôlés randomisés est compliquée. Les preuves apportées par des études observationnelles sont donc des alternatives même si celles-ci sont entachées de biais de confusion. L'utilisation du score de propension dans l'analyse statistique des études observationnelles est une méthode qui permet de réduire ce biais et d'augmenter la validité interne de ces études. Le score de propension (SP) est défini pour un patient comme la probabilité de recevoir le traitement évalué, en fonction de ses caractéristiques avant traitement (caractéristiques de base). Ainsi après utilisation du score de propension, les distributions des caractéristiques de base du groupé traité et non traité doivent être similaires. Cette méthode est de plus en plus utilisée en particulier pour évaluer des interventions chirurgicales.

Les objectifs de ce travail étaient 1) de comparer l'effet traitement estimé par les études observationnelles utilisant un score de propension et l'effet traitement estimé par des ECR en chirurgie ; 2) de faire une description précise de l'utilisation du score de propension dans les études observationnelles évaluant des interventions chirurgicales.
Notre premier travail repose sur une étude méta-épidémiologique. Cette étude incluait 31 questions cliniques qui avaient été traitées dans la littérature par au moins une étude observationnelle utilisant un score de propension et au moins un essai contrôlé randomisé. Après extraction des effets traitement de chacun des 2 types d'études nous avons regroupé les résultats pour l'ensemble des questions cliniques. Nous n'avons pas retrouvé en moyenne de différence statistiquement significative pour l'estimation de l'effet traitement entre les ECR et les études observationnelles avec SP. Cependant, nous avons pu constater des variations dans les 2 sens pouvant parfois être importantes pour certaines questions cliniques.

Dans la deuxième partie de cette thèse, nous avons réalisé une revue systématique méthodologique des études observationnelles évaluant une intervention chirurgicale utilisant un score de propension. Sur les 652 articles sélectionnés entre 1980 et 2014, nous avons montré une évolution importante et rapide du nombre d'articles au cours du temps, passant de moins de 10 articles par an avant 2000 à plus de 200 depuis 2013. Dans l'analyse plus fine de 129 articles récents, nous avons pu mettre en avant des limites méthodologiques (analyse avec score de propension mal détaillée, variables incluses dans le score de propension non rapportées, pas ou peu d'information sur les données manquantes, non-respect du principe de l'analyse en intention de traiter). Ainsi, des recommandations pour améliorer la planification, la réalisation, l'analyse et l'écriture du rapport des études observationnelles avec SP ont pu être faite.

Les études observationnelles avec score de propension sont de plus en plus fréquentes. Leurs avantages par rapport aux essais contrôlés randomisés dans le domaine de la chirurgie en font une alternative intéressante. En revanche, des limites méthodologiques existent. Certaines limites peuvent être corrigées facilement, d'autres méritent des analyses plus approfondies.

Abstract

Surgery has changed in the last decade. Surgical techniques have advanced and are being used more frequently. Assessment of surgical procedures has become a public health priority, but evaluating a surgical procedure in randomized controlled trials (RCTs) is challenging. Observational study is an alternative, despite the presence of confusion bias. Use of propensity score (PS) analysis in an observational study is a way to limit confounding bias. The PS is defined for each participant as the probability of receiving the treatment, given baseline covariates. With the assumption of no unmeasured confounders, a treated and an untreated patient with the same PS can be considered as if they had been randomly assigned to each group. PS analysis has been increasingly used in the last 10 years, specifically in studies of surgery.

In this project, I aimed to 1) compare treatment effect estimates from non-randomized studies with PS analysis and RCTs and 2) describe and assess the reporting and potential bias of PS analysis used in a sample of published observational studies assessing surgical procedures.

In a first part, a meta-epidemiological study of 31 clinical questions in surgery revealed no statistically significant difference in treatment effect estimates between non-randomized studies with PS adjustment and RCTs, although the treatment effect estimates varied widely between the two study designs.

In a second part, from 652 reports selected, I systematically assessed the use of PS analysis in observational studies evaluating surgical procedures. The number of observational studies evaluating a surgical procedure with PS analysis increased from < 10 before 2000 to > 200 after 2013. However, the use and reporting of PS analysis in observational studies raises important concerns related to constructing the PS, accounting for missing data and cross-overs and reporting all features necessary to reproduce the analysis.

Observational studies with PS analysis of surgery are increasing in frequency and their use can be relied upon for evidence when RCTs are not possible. Specific methodological issues and weaknesses in reporting exist. Some limitations should be easily correctable, but some need more assessment.

Laboratoire d'accueil :

Centre de Recherche Epidémiologie et Statistiques Sorbonne Paris Cité (CRESS)

CRESS INSERM UMR-1153.

Equipe METHODS : Méthodes en Évaluation Thérapeutique des Maladies Chroniques

Hôpital Hôtel Dieu

1 place du parvis Notre Dame

75 004 Paris, France

Publications liées à la thèse

Lonjon G, Boutron I, Trinquart L, Ahmad N, Aim F, Nizard R, Ravaud P. Comparison of Treatment Effect Estimates From Prospective Nonrandomized Studies With Propensity Score Analysis and Randomized Controlled Trials of Surgical Procedures. Ann Surg. 2013 Oct 3;

Lonjon G, Boutron I, Trinquart L, Ahmad N, Aim F, Nizard R, Ravaud P. Reply to Letter: "Equivalence Approach in Meta-epidemiological Studies Can Be Misleading."Ann Surg. 2014 Mar 25;

Lonjon G, Porcher R, Ergina P, Fouet M, Boutron I. Potential Pitfalls of Reporting and Bias in Observational Studies With Propensity Score Analysis Assessing a Surgical Procedure: A Methodological Systematic Review. Ann Surg. 26 mai 2016 ;

Table des matières

Liste des figures

Abréviations

SP: Score de propension

ECR: Essai Contrôlé Randomisé

IC : Intervalle de Confiance

FDA : Food and Drug Administration

I Introduction et objectifs

II Article 1: Comparaison de la mesure de l'effet traitements entre les études observationnelles utilisant un score de propension et les ECR évaluant une procédure chirurgicale

III Article 2 : Description et Sources de biais des Etudes observationnelles utilisant une analyse avec score de propension évaluant une procédure chirurgicale. Revue systématique

IV Discussion et perspectives

V Conclusion

VI Réferences

I Introduction et objectifs

1.1 La place de la chirurgie dans l'arsenal thérapeutique

Aux Etats-Unis en 2012, on estimait à 28 millions le nombre d'interventions chirurgicales réalisées chaque année. Les interventions chirurgicales représentaient un quart des séjours hospitaliers et la moitié des dépenses hospitalières (342 milliards de dollars). Sur cette même année, 2 500 américains sur 100 000 ont eu une intervention chirurgicale (1). En France, 6,5 millions interventions chirurgicales sont réalisées chaque année (2). Le nombre d'actes et les coûts engendrés font de la chirurgie un enjeu majeur de santé publique.

Outre l'importance actuelle de la chirurgie en termes de chiffres, l'évolution des techniques a été extraordinaire au cours des 50 dernières années. Des techniques qui paraissaient irréalisables ou complètement futuristes il y a quelques décennies, sont devenues « monnaie courante ». L'avènement des transplantations d'organe est un exemple typique de la révolution qu'a vécu la chirurgie (3). Actuellement, de façon un peu moins spectaculaire mais tout aussi fascinante, c'est le développement des techniques robotiques qui marque l'histoire de la chirurgie (4).

Cependant, l'évaluation de ces techniques chirurgicales soulève de nombreuses questions. Au siècle dernier, lorsque de nouvelles interventions étaient développées, celles-ci étaient rapidement utilisées en pratique courante sans avoir été évaluées de façon rigoureuse. Cette situation était acceptable à l'heure où l'effet des interventions chirurgicales était considérable (5,6). Des exemples comme la prothèse de hanche qui permettait de remarcher, la plastie mitrale qui guérissait de l'insuffisance cardiaque ou encore la transplantation qui sauvait la vie sont autant d'exemples probants.

Actuellement, le gain d'efficacité des nouvelles techniques est beaucoup plus faible (7). Des études de niveau de preuve élevé sont nécessaires pour démontrer l'efficacité de ces techniques (8–10).

1.2 Place des essais contrôlés randomisés en chirurgie.

1.2.1 Principes généraux

L'ECR est la méthode de référence de l'évaluation thérapeutique (11). L'objectif des ECR est de créer et de maintenir des groupes comparables tout au long de l'étude et que la seule différence entre les groupes réside dans l'administration ou non du traitement expérimental. La méthodologie des ECR a pour objectif de limiter les biais afin d'estimer l'effet traitement réel d'une intervention (Figure 1).

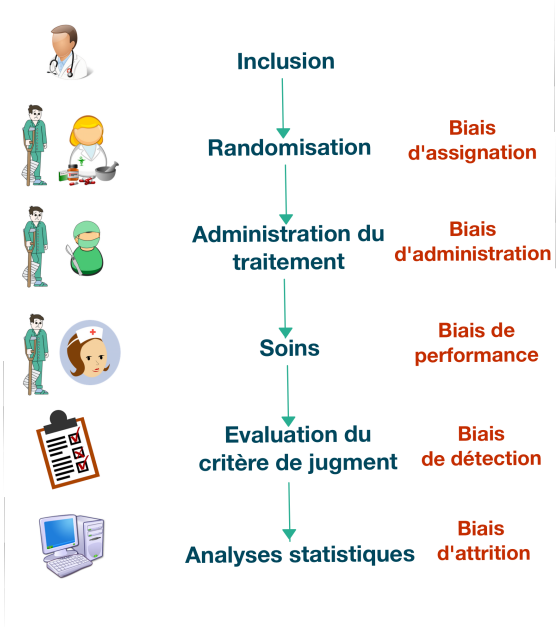
La méthodologie des ECR est bien codifiée. Elle repose sur une randomisation qui permet de créer des groupes comparables pour les facteurs de confusion connus (comme l'âge, le sexe, les comorbidités, etc.), mais aussi pour des facteurs de confusion non connus. Afin de limiter les biais d'assignation, la randomisation doit être aléatoire (méthode pour générer la séquence de randomisation informatisée) et non prédictible (respect de l'assignation secrète) (12,13).

L'aveugle du patient et du thérapeute permet de limiter les biais lors de la prise en charge et le suivi des patients (« biais de performance »). L'aveugle des évaluateurs est particulièrement important lorsque le critère de jugement est subjectif afin de limiter les biais de « détection » (14).

Il est enfin essentiel de maintenir la comparabilité des groupes au moment de l'analyse et de limiter les biais d'attrition (15). Afin de limiter ces biais, il est indispensable de réaliser une analyse en intention de traiter. Le principe de l'analyse

en « intention de traiter » consiste à analyser un patient dans l'essai littéralement « tel qu'on avait l'intention de le traiter », c'est à dire d'une part tous les patients randomisés sont analysés et d'autre part ils sont analysés dans le groupe dans lequel ils avait été randomisés quel que soit la traitement reçu. Ce principe a pour objectif de préserver les bénéfices de la randomisation.

Figure 1 : Schéma du déroulement d'un ECR avec les différents acteurs et les possibles biais



Plusieurs études méta-épidémiologiques (12,14,16–18) ont montré que l'estimation de l'effet traitement est plus favorable au traitement expérimental en l'absence de randomisation adéquate, en l'absence d'aveugle et en l'absence d'analyse en intention de traiter. Dans l'étude de Wood par exemple, regroupant 146 méta-analyses et plus de 1300 articles, l'absence de l'aveugle entrainait une surestimation de 25% de l'effet traitement (OR=0,75 [0,61 to 0,93]) lorsque le critère de jugement est un critère subjectif (12). Sur une étude méta-épidémiologique de 310 ECR, l'absence d'analyse

en intention de traiter avait tendance à surestimer l'effet traitement de 17 % (OR= 0,83 [0,71-0,96]) (18).

Une limite importante des ECR est la validité externe (19–23). La faible validité externe des ECR est fréquemment citée comme raison pour ne pas transposer les résultats des ECR dans la pratique clinique (24,25).

De nombreux travaux ont montré que les patients qui participent aux essais randomisés ne sont pas représentatifs des patients traités en pratique clinique courante (26–30). Steg et col. ont montré, à partir des données du registre multinational GRACE incluant de façon consécutive 8469 patients ayant un infarctus du myocarde, que seuls 66% des patients étaient éligibles dans un ECR et 17% des patients éligibles étaient inclus dans ces essais. Le pronostic des patients inclus était beaucoup plus favorable que le pronostic des patients traités en pratique courante (30).

Cette faible représentativité peut être due à l'utilisation abusive des critères de non inclusions ou d'exclusion. Dans une revue systématique de 283 ECR, 53% des critères de non inclusions n'étaient pas justifiés (31).

L'organisation logistique du recrutement peut aussi limiter la validité externe. Dans un essai de chirurgie sur l'efficacité de l'endartèriectomie carotidienne, les modalités de recrutement très standardisées limitaient la représentativité des patients (32). Ainsi, il a été montré que les patients arrivant à l'étape des critères d'éligibilité ne représentent que 10% de la population d'intérêt et que la population finalement randomisée ne représente que 0,001% de la population d'intérêt (33).

Outre la sélection des patients, la sélection des médecins et des centres a aussi un impact sur la validité externe. Les centres et les médecins participant aux ECR sont souvent sélectionnés sur des critères liés à leur expertise et à leur volume d'activité

dans le domaine (34). Or le volume d'activité est associé au succès de l'intervention et à l'effet traitement (35).

1.2.2 Particularités et limites des ECR en chirurgie.

La méthodologie des ECR a été développée essentiellement dans le contexte de l'obtention d'une autorisation de mise sur le marché pour des traitements pharmacologiques et l'utilisation des ECR pour l'évaluation d'interventions chirurgicales pose des problèmes méthodologiques spécifiques que nous allons détailler. (36,37).

1.2.2.1. Obstacles à la randomisation : la préférence des chirurgiens et des patients

La préférence du chirurgien est probablement la première difficulté à la mise en place d'un ECR pour évaluer une procédure chirurgicale. Bien souvent, une technique chirurgicale (par exemple, la prothèse totale de hanche, PTH) réalisée par un chirurgien est le fruit d'un héritage d'une technique transmise (ex : PTH par voie antérieure ou par voie postérieure) par des maitres à laquelle s'ajoute des modifications personnelles (ex : choix du type d'implants, types d'instruments, variations techniques), et dont le tout est modulé par des résultats scientifiques. Cette technique est alors considérée par le chirurgien comme la meilleure pour le patient. Dans ce contexte, il est difficile pour le chirurgien de reconsidérer ses certitudes et de remettre dans les mains du hasard le choix entre deux techniques (38). La réticence du chirurgien est encore plus importante dans le cadre d'une comparaison avec une intervention non chirurgicale qu'il ne réalisera pas ou ne maitrisera pas.

La préférence du patient est aussi un obstacle à la randomisation. Le patient qui va voir un chirurgien, place sa confiance en lui. Cette confiance est souvent exacerbée dans le cadre de la chirurgie, car le traitement parait plus dangereux et le traitement est directement lié au chirurgien. Ainsi la randomisation est difficile à proposer et le risque de refus par le patient est important. Celui-ci est encore plus important dans le contexte de comparaison d'une procédure chirurgicale avec un traitement non chirurgical. Dans l'ECR "SPORT" (The Spine Patient Outcome Research Trials) (39), qui comparait une chirurgie de cure de hernie discale à un traitement médical dans le cadre de lombo-radiculalgie, le taux de contamination était important. Dans cet essai, 30% des patients tirés au sort dans le bras « traitement médical » étaient finalement traités par chirurgie. Cette contamination était probablement le reflet de la préférence du chirurgien et des patients et de son influence sur l'administration du traitement.

1.2.2.2. Expérience du chirurgien et courbe d'apprentissage

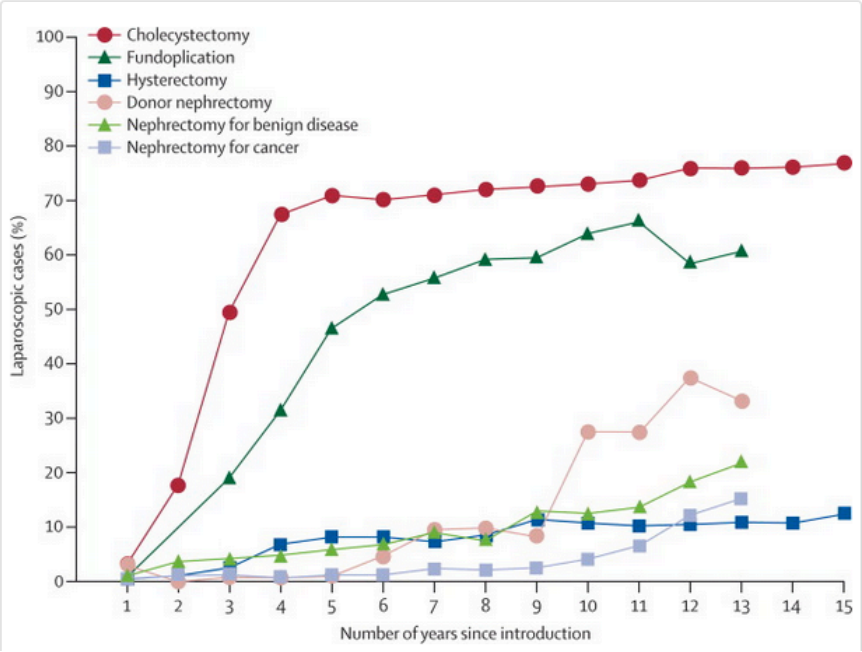
L'influence du chirurgien au moment de l'intervention est fondamentale, si bien que le chirurgien est difficilement dissociable de l'intervention. Son expérience influence les résultats de la chirurgie (40). Par ailleurs, son expérience évolue au cours du temps (41), et elle peut être responsable d'une modification de l'intervention entre le début et la fin d'un essai et pose des problèmes méthodologiques. Ceci est d'autant plus vrai lors de l'évaluation d'une nouvelle technique en comparaison à une ancienne où la courbe d'apprentissage de l'ancienne technique est peut- être déjà atteinte. Dans ce cas, il y a un risque de sous-estimer l'effet traitement de la nouvelle technique que le chirurgien ne maitrise pas encore complètement.

De plus, comme nous l'avons évoqué plus haut, une expertise trop importante du chirurgien va avoir un impact sur la validité externe de l'ECR (35).

1.2.2.3. Diffusion d'une technique : concept de point de bascule

Une autre difficulté est liée au moment opportun pour évaluer une nouvelle intervention chirurgicale. Si dans la théorie, il est conseillé de randomiser dès le premier patient, dans le cadre d'une intervention chirurgicale cet adage est discutable. Outre les problèmes liés à la courbe d'apprentissage que nous venons de voir (42), des modifications techniques sont fréquentes au début du développement d'une nouvelle technique et une évaluation trop prématurée pourrait répondre à tort à une infériorité de la nouvelle technique et freiner son développement. A l'inverse une évaluation trop tardive, (43) quand la procédure est déjà trop largement diffusée devient compliquée. En effet, il existe un moment où l'adoption de la technique est trop large pour que son évaluation par randomisation puisse se faire. On parle à ce moment de « point de bascule » (Figure 2) (44).

Figure 2: Courbe de la diffusion dans le temps de la méthode laparoscopique pour des chirurgies faisables par laparoscopie et par laparotomie (43).



1.2.2.4. *Difficulté de l'aveugle et problèmes liés au placebo.*

L'aveugle est souvent difficile à mettre en place dans les essais évaluant une procédure chirurgicale (45–47). Dans une revue systématique méthodologique, l'aveugle était considéré comme faisable dans 42% des cas pour le patient, 12% des cas pour le chirurgien, et 34% des cas pour l'évaluateur.(48)

La difficulté de l'aveugle en chirurgie est notamment liée au placebo. Car si le placebo est relativement facile à réaliser dans le cadre d'un essai pharmaceutique, celui-ci est compliqué dans le cadre d'une procédure chirurgicale.

Certains ECR évaluant des procédures chirurgicales ont utilisé des placebos chirurgicaux qui consistaient en une simulation de l'intervention (49). Dans une étude comparant le lavage arthroscopique versus placebo dans les douleurs arthrosiques de genou (50), une simulation de l'intervention a été mise en place. Elle se faisait au bloc opératoire, avec une préparation préopératoire et des incisions cutanées de 1 cm sans geste intra-articulaire. A noter que pour le groupe placebo, des tranquillisants et un masque avec de l'oxygène ont été utilisés pour simuler l'anesthésie. Dans cet exemple, même si ce placebo n'était pas très risqué, celui-ci avait une certaine morbidité, soulevant le problème éthique de ce type de procédé (17).

Outre les problèmes techniques d'une simulation d'intervention, le problème théorique d'acceptation par les patients est une autre limite dans ce choix méthodologique.

Les « Prospective Open Blinded Endpoint Studies » ont été proposées afin de limiter les biais lorsque l'aveugle des patients et thérapeutes est impossible. La méthode se base sur une évaluation centralisée et externalisée du critère de jugement avec des évaluateurs en aveugle du traitement alloué. Ce type de méthode d'évaluation se base

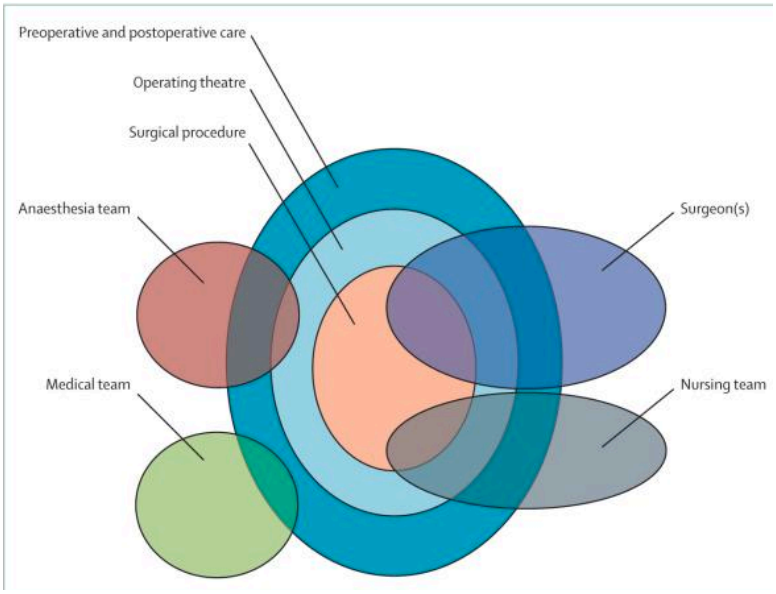
généralement sur des enregistrements vidéos, audios ou des photographies (51).

1.2.2.5. *Difficultés liées à la complexité de l'intervention*

Une intervention doit être reproductible. Dans le cadre d'un essai pharmaceutique, l'intervention est facile à décrire et à reproduire. Une procédure chirurgicale est une procédure complexe. L'intervention dépasse le cadre de la technique chirurgicale car elle allie un temps d'anesthésie, des soins préopératoires, un temps chirurgical et des soins post-opératoires (médicaments, pansement, rééducation, consignes) (Figure 3). Dans cette succession thérapeutique, chaque étape est au moins aussi importante que l'intervention chirurgicale elle-même.

Pourtant, cette intervention est généralement insuffisamment décrite. Dans une revue systématique de 158 articles, des détails sur la procédure chirurgicale étaient décrit dans seulement 87%, le protocole anesthésique dans 35%, le protocole pré-chirurgicale dans 15% et le protocole de soin postopératoire dans 50% des cas (52).

Figure 3 : Schéma de procédure complexe. (53)



Pour une bonne reproductibilité, les étapes doivent être détaillées de façon exhaustive et des précisions doivent être apportées sur leur standardisation. Le protocole de l'étude doit contenir des informations sur le matériel utilisé, le degré de souplesse autorisé, l'opérateur, la durée opératoire, et le lieu de chaque étape. Cependant cela reste en pratique difficile à faire et souvent bien mal rapporté (52,54).

Par ailleurs en fonction de la question posée, l'intervention peut être très standardisée (étude d'efficacité) ou au contraire laissée à la discrétion du chirurgien et des thérapeutes (étude pragmatique). Quelle que soit la question, l'intervention doit être décrite précisément pour permettre l'interprétation des résultats.

1.2.2.6. Difficulté liée au côté irréversible de la chirurgie

Dans le cadre des ECR de pharmacologie, la prise d'un médicament est bien souvent réversible et même si des effets secondaires apparaissent, ils disparaissent fréquemment à l'arrêt du médicament. Dans le cadre d'un ECR en chirurgie comparant 2 techniques chirurgicales, comme la prothèse de disque et l'arthrodèse lombaire (55) , si la supériorité de l'un des 2 traitements est réelle, les patients ayant reçu le moins bon traitement ne pourront plus recevoir le meilleur.

1.2.2.7. *Absence de contraintes technico-règlementaires*

Dans le cadre d'une évaluation d'une procédure chirurgicale, les autorités sanitaires sont moins exigeantes pour l'évaluation des procédures chirurgicales que pour des traitements médicamenteux. Par exemple la FDA (Food an Drug Administration) aux Etats-Unis demande en théorie deux ECR montrant une efficacité par rapport au traitement de référence pour accepter leur mise sur le marché d'un traitement

médicamenteux. Dans ce contexte, ce qui pourrait être un avantage est en fait un inconvénient, car comme les règles sont plus laxistes, les dispositifs médicaux sont souvent sous-évalués avant leur mise en route. De plus, il n'existe pas de règle pour les changements ou modification de technique (ex : évaluation de la robotique dans la chirurgie). Ainsi beaucoup de procédures chirurgicales sont évaluées de façons moins rigoureuses que des traitements médicamenteux.

1.3 Place des études observationnelles pour l'évaluation de la chirurgie

1.3.1 Principes généraux

Les études observationnelles ont une place très importante dans l'évaluation thérapeutique en particulier en chirurgie. Il est généralement admis que les études observationnelles sont nécessaires lorsque la randomisation est inutile (effet de l'intervention très important), inappropriée (critère de jugement rare ou tardif) ou impossible (problème éthique) (56).

Les études observationnelles permettent aussi une évaluation de l'effet du traitement dans des conditions proches de la « vraie vie » (26,30).

A l'heure de la révolution numérique où les informations collectées sont de plus en plus nombreuses notamment grâce aux outils connectés, mais aussi grâce à l'accès aux données administratives et aux données des dossiers médicaux informatisés et des entrepôts de données, les études observationnelles vont occuper une place encore plus importante.

La FDA a notamment changé ses recommandations en demandant aux investigateurs

d'associer des preuves basées sur des données provenant de la vie de tous les jours, aux preuves apportées par les ECR (57) .

Cependant l'apport des études observationnelles en termes de quantité de données disponibles et de validité externe se fait au prix d'une moins bonne validité interne.

1.3.2 Biais des études observationnelles

Comme pour les ECR, il existe plusieurs types de biais dans les études observationnelles (58,59). Comme pour les biais évalués dans les ECR, un biais peut être défini comme la différence systématique entre l'effet traitement estimé par l'étude et l'effet traitement réel.

Sterne, Hernan et col. (60) ont proposé de conceptualiser le biais dans les études observationnelles en comparant la méthodologie de l'étude observationnelle avec la méthodologie d'un essai randomisé hypothétique parfait (« target trial »). C'est à dire, un ECR incluant les mêmes patients et dont la méthodologie est parfaite (large ECR réalisé avec assignation secrète, sans contamination, avec une procédure d'aveugle des différents intervenants et sans données manquantes).

L'évaluation des biais des études observationnelles est alors facilitée en considérant l'étude observationnelle comme étant une étude qui tente de mimer un ECR « parfait ». C'est la notion de l'essai « cible ».

Les principaux biais dans ce type d'études selon cette approche sont : le biais de confusion, le biais de sélection des participants, le biais de classification de l'intervention, le biais de déviation par rapport à l'intervention prévue, le biais dû aux données manquantes et le biais de sélection des résultats.

1.3.2.1. *Biais de confusion*

A la différence des ECR, les caractéristiques de bases des patients différent généralement en fonction du groupe de traitement. Un biais de confusion existe quand les caractéristiques des variables de confusion ne sont pas similaires entre les 2 groupes et que ces différences ne sont pas prises en compte dans l'analyse statistique. Des exemples classiques de facteur de confusion sont l'âge, l'IMC, des pathologies pré-excitantes, des comorbidités ou encore l'environnement socio-économique. L'évaluation de ce biais est fondamentale dans les études observationnelles, mais comme certains facteurs de confusion ne sont pas mesurables ou mesurés, il peut persister des biais de confusion malgré des analyses solutions statistiques poussées.

1.3.2.2. *Biais de sélection des participants*

Un biais de sélection des participants existe quand des patients sont sélectionnés alors que cette sélection n'aurait pas été possible dans un ECR. Dans une étude observationnelle rétrospective évaluant la prévention de la non fermeture de l'arc neural par acide folique, seules les patientes dont la grossesse était mené à terme étaient sélectionnées (61). Dans le cadre d'un ECR, les patientes qui auraient eu un arrêt de grossesse après la randomisation auraient été suivies et analysées selon les principes de l'analyse en intention de traiter.

1.3.2.3. *Biais de classification de l'intervention*

Des biais peuvent survenir si le statut de l'intervention est mal classé. L'exemple typique est l'étude de cohorte rétrospective où les patients sont classés en fonction de l'intervention reçue et non l'intervention prévue. Dans un exemple de chirurgie, où des investigateurs veulent analyser les effets de la laparoscopie versus la laparotomie,

le fait de classer les patients en fonction du traitement reçu entraine un biais de classification (un nombre non négligeable de laparotomie peut être le résultat de conversions peropératoires d'une laparoscopie). Dans un ECR parfait, due à l'analyse en intention de traiter, ces patients seraient analysés dans leur groupe de randomisation.

1.3.2.4. Biais de déviation par rapport à l'intervention prévue

Les biais de déviation par rapport à l'intervention prévue, comme les biais de performance pour les ECR, sont dus à une différence de prise en charge entre les 2 groupes de traitement. Cette différence est souvent influencée par la connaissance du traitement administré. Ces biais sont limités par l'aveugle des patients et des thérapeutes, comme dans les ECR pharmacologiques.

1.3.2.5. Biais dus à des données manquantes.

Le biais dû aux données manquantes, est légèrement différent du biais d'attrition décrit dans les ECR car il considère aussi les données manquantes à l'origine. Ces données manquantes peuvent être des facteurs de confusion ou le type d'intervention reçu. Dans un ECR, les caractéristiques de bases ainsi que le type d'intervention sont généralement présents.

1.3.2.6. Biais de mesure du critère de jugement

Les biais de mesures des critères de jugement existent si les critères de jugement sont mal classés ou mesurés avec une erreur. Cela correspond aux biais de détection des ECR. Ce biais renvoie principalement à la présence ou non de l'aveugle des

évaluateurs. Cependant, dans le cadre des critères objectifs, le biais est considéré comme faible car non dépendant de l'évaluateur.

1.3.2.7. Biais de sélection des résultats.

Les biais de sélection des résultats rapportés sont plus larges que les biais de sélection des critères de jugement des ECR. Il comprend en plus de la sélection des critères de jugement comme pour les ECR, une sélection possible des analyses et une sélection possible de résultats dans des sous-groupes de population.

1.3.3 Estimation de l'effet traitement par les études observationnelles

Plusieurs études méta-épidémiologiques ont comparé l'effet traitement estimé dans des études observationnelles et dans les essais randomisées (62–69). Ces études consistent à identifier une collection de méta-analyses avec des essais randomisés et des études observationnelles et de comparer l'effet traitement estimé par ces 2 types d'étude. Benson et al ont identifiés 19 revues systématiques dans lesquelles ils retrouvaient au moins un essai randomisé et une étude observationnelle. Ces 19 comparaisons, regroupaient 53 études observationnelles et 83 ECR. Les résultats étaient méta-analysés pour chaque type d'étude au sein de chaque comparaison. Dans 2 comparaisons sur 19, les résultats apportés par les études observationnelles et les ECR différaient. Dans un cas, il y avait une tendance à la surestimation de l'effet traitement par l'étude observationnelle, et dans l'autre cas c'était une tendance à la surestimation de l'effet traitement par l'ECR.

Dans une étude similaire, Concato et al (67) ont identifié des revues systématiques publiées dans 5 journaux généralistes de facteur d'impact élevé. Ils retrouvaient 5 revues systématiques, regroupant 55 ECR et 44 études observationnelles. Pour évaluer une différence, les auteurs méta-analysaient, au sein de chacun des 5 comparaisons, les résultats pour l'ensemble des études observationnelle et des ECR. Les auteurs ne retrouvaient pas de différence d'effet traitement pour les 5 comparaisons, cependant au sein de chaque comparaison, les estimations d'effet traitement étaient plus dispersées pour les ECR que pour les études observationnelles.

Une analyse systématique de toutes les études comparant l'effet traitement estimé par les études randomisées et les études observationnelles a été réalisé en 2003(70). Dans cette revue systématique, huit articles comparant l'effet traitement estimé par les ECR et les études observationnelles ont pu être identifiés : 7 concernaient des traitements médicaux et 1 concernait une intervention psychologique. Cette revue systématique retrouvait des résultats contradictoires. En effet, plusieurs études concluaient à des différences d'estimation d'effet traitement entre les études observationnelles et les ECR, sans qu'une orientation soit clairement définie en faveur de l'un ou l'autre type d'étude (65,68,69,71). Au final, Deeks et al. considéraient qu'il pouvait exister des différences significatives entre les 2 types d'études sans pouvoir affirmer une surestimation de l'effet traitement par l'un ou l'autre type d'études.

1.4 L'analyse avec score de propension

Nous avons vu que les études observationnelles avaient une place importante en évaluation thérapeutique. Cependant ce type d'étude souffre toujours des problèmes

de validité interne, et notamment du problème de biais de confusion. L'analyse utilisant un score de propension est une méthode qui permet de réduire ce biais et d'augmenter la validité interne des études observationnelles.

Le score de propension (SP) est défini pour un patient comme la probabilité de recevoir le traitement évalué en fonction de ses caractéristiques avant traitement (caractéristiques de base). Ainsi après l'analyse avec score de propension, les distributions des caractéristiques de base des groupes traité et non traité doivent être similaires. Il existe quatre approches différentes d'analyse par score de propension : l'ajustement, l'appariement, la stratification et la pondération.

1.4.1 Le concept

L'analyse avec SP permet de mimer – ou d'émuler - un ECR dans le contexte d'une étude observationnelle.

Le SP a été défini par Rosenbaum et Rubin, dès 1983,(72) comme la probabilité de recevoir un traitement en fonction de caractéristiques de base. Il est compris entre 0 et 1. Plus il tend vers 1 et plus la probabilité de recevoir le traitement expérimental est importante. Dans les études observationnelles le véritable SP n'est pas connu, et il doit être estimé à partir des données observées, généralement par un modèle de régression logistique, dans lequel le groupe de traitement est régressé en fonction des caractéristiques de base.

Rosenbaum et Rubin ont alors montré qu'il était possible d'estimer sans biais l'effet du traitement en conditionnant l'analyse sur le SP. Cela n'est cependant possible que si toutes les variables de confusion sont incluses dans le modèle de SP. Les variables de confusion influencent à la fois le critère de jugement et l'attribution du traitement.

1.4.2 Les différentes méthodes d'utilisation du SP

Il existe 4 méthodes d'utilisation du SP (appariement, stratification, ajustement et pondération inverse).

1.4.2.1. Appariement sur le SP

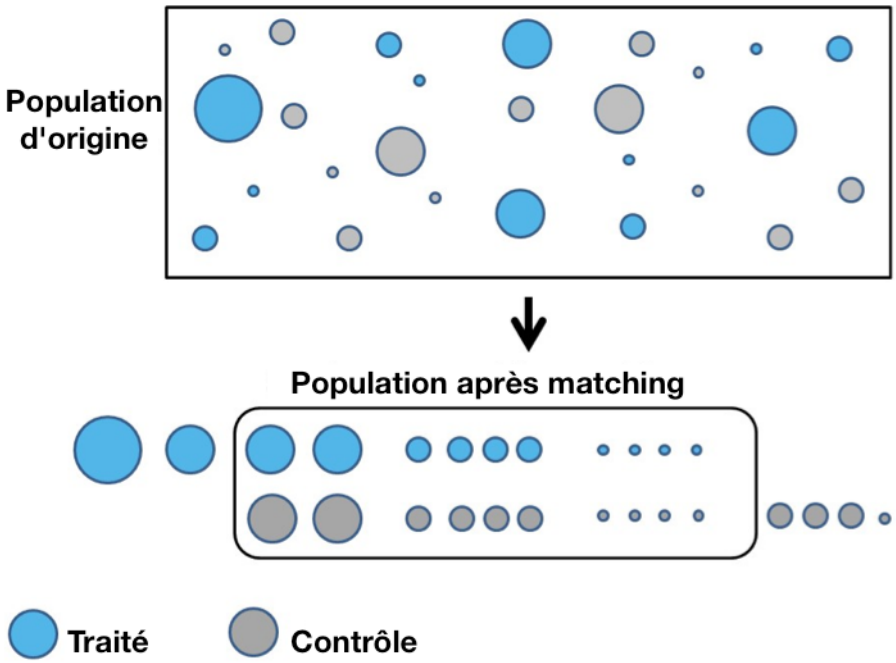
L'appariement sur le score de propension cherche à former des paires (ou des groupes) de patients qui ont le même score de propension (Figure 4). (73) Généralement, un patient traité est apparié avec un patient contrôle (1:1), cependant il est possible d'apparier plusieurs patients contrôles avec un patient traité (1:2,1:3,1:4,…), ou le contraire. Les patients appariés diffèrent donc juste par le traitement, même s'ils ont la même probabilité de recevoir le traitement. On peut alors considérer que pour les 2 patients l'attribution du traitement s'est faite de manière aléatoire, comme pour un ECR. Une approche légèrement différente a aussi été décrite. Elle consiste à apparier de façon variable un ou plusieurs sujets contrôle avec un sujet traité. Cette approche permet une meilleure réduction du biais que la méthode à nombre fixe (74). Cependant cette méthode est moins facile de compréhension pour le clinicien car la méthode diffère du modèle de l'ECR avec randomisation un pour un.

Une fois que les paires ont été formées, l'effet du traitement peut être directement estimé en comparant les critères de jugement des 2 groupes.

Dans le contexte de l'appariement, la variance de l'estimation de l'effet du traitement (et par conséquent le test de cet effet) doit être calculée avec des méthodes appropriées aux échantillons dépendants (75,76) , même si cela reste débattu et critiqué (77). Des études récentes de simulation sur des scenarii différents ont montré que les résultats étaient plus proches de la réalité dans le cas d'une analyse sur

échantillon dépendant (78,79).

Figure 4 : Schéma de la construction des paires dans l'analyse avec SP.



La formation des paires répond à une stratégie qui doit être définie. A chaque étape, il existe plusieurs possibilités (remise, optimisation des paires, distance entre le SP de 2 individus à l'intérieur d'une paire)

La première option à choisir est l'appariement des sujets contrôle avec ou sans remise. Quand on utilise une méthode sans remise, un patient du groupe contrôle apparié à un patient traité n'est plus disponible pour être apparié à un autre patient traité qui aurait un SP similaire. A contrario, quand on utilise une méthode avec remise, un patient contrôle peut être apparié plusieurs fois. (80)

La seconde option est la méthode de choix du patient contrôle entre l'option dite « gloutonne » ou l'option optimale. Dans la méthode dite « gloutonne », le premier

patient contrôle qui a un SP similaire à celui du patient traité est utilisé pour créer une paire. Dans la méthode optimale, le but est de choisir le meilleur contrôle pour chacun des patients traités de manière à réduire l'écart global entre les 2 SP à l'intérieur d'une paire. Si dans la théorie, la méthode optimale paraît supérieure à la méthode « gloutonne », dans les faits les études de simulation n'ont pas pu mettre en évidence de véritable différence (81).

La troisième option est la façon dont on définit la similarité des SP entre un patient traité et un patient contrôle. Dans les faits, les SP des patients au sein d'une paire ne sont pas exactement égaux, ils sont proches. Une distance (un « caliper » en anglais) maximale à ne pas dépasser entre le SP du patient traité et du patient contrôle doit être pré-spécifiée. Tous les patients du groupe traité et du groupe contrôle qui n'auront pas pu être appariés à la fin de la procédure en respectant cette distance maximale seront exclus de l'analyse.

Il n'y a pas de méthode clairement définie pour choisir cette distance. Dans la littérature médicale, plusieurs mesures ont été proposées (82,83). Cochrane et Rubin, (84), proposaient dès 1973, une distance qui était dépendante de l'écart-type du logit de la variable de confusion. En utilisant cette règle avec une distance inférieure ou égale à 0.2 écart-type du logit du PS , la quasi-totalité du biais de confusion était réduit (85)

1.4.2.2. Stratification sur le SP

Les participants sont regroupés et analysés à l'intérieur de strates créées en fonction de la valeur du SP. La méthode classique consiste à diviser l'échantillon en 5 strates égales en utilisant les quintiles du SP. Avec cette méthode, il a été montré sur un exemple que 90% du biais été réduit (86). Augmenter le nombre de strates augmente

la performance de la méthode, mais sans que le bénéfice soit réel en terme de pertinence statistique .(87)

Avec cette méthode, la distribution des caractéristiques de bases des participants sera similaire pour le groupe traité et le groupe contrôle à l'intérieur de chaque strate.

L'effet traitement est estimé au sein de chaque strate par une comparaison directe des critères de jugement, puis, un peu comme dans une méta-analyse, les mesures d'effet traitement de chaque strate sont combinées pour donner une mesure globale de l'effet traitement. Afin de limiter les biais, la mesure de l'effet traitement peut être pondérée en fonction du nombre de sujets dans chaque strate.

1.4.2.3. Pondération par l'inverse de la probabilité de recevoir le traitement.

La pondération par l'inverse de la probabilité de recevoir le traitement est une méthode qui vise à recréer une pseudo-population dans laquelle les caractéristiques des patients des deux groupes sont équilibrées en se servant du SP. Pour chaque individu, on commence par calculer le SP, puis à partir du SP un facteur de pondération est calculé, qui correspond à l'inverse de la probabilité que le patient ait réellement reçu le traitement. Dans le groupe traité le facteur est égal à 1/SP, et dans le groupe non traité le facteur est égale à 1/(1-SP). Un estimateur pondéré de la différence entre les groupes de patients est ensuite utilisé. Dans cette méthode, il est donné plus de poids aux sujets qui n'ont pas reçu le traitement qu'ils auraient dû recevoir (Exemple 1).

Exemple 1 :
Sujet 1 : Il a une forte probabilité de recevoir le traitement en regardant les caractéristiques de base (SP = 0,80) et il a effectivement reçu le traitement (w=1/SP) **w=1/0,8= 1,25**
Sujet 2 : Il a une forte probabilité de recevoir le traitement en regardant les caractéristiques de base (SP = 0,80), mais il n'a pas reçu le traitement (w=1/(1-SP)) **w=1/(1-0,8)=5**
Sujet 3 : Il a une faible probabilité de recevoir le traitement en regardant les caractéristiques de base (SP = 0,30), et il n'a pas reçu le traitement (w=1/(1-SP)) **w=1/(1-0,3)= 1,43**
Sujet 4 : Il a une faible probabilité de recevoir le traitement en regardant ces caractéristiques de base (SP = 0,30), mais il a reçu le traitement (w=1/SP) **w=1/(1-0,8)=3,33**
Conclusion : Les sujets 2 et 4 ont plus de poids que les sujets 1 et 3 pour des SP similaires**.**

1.4.2.4. Ajustement

La quatrième méthode utilise une méthode de régression du critère de jugement en fonction de l'assignation au traitement en ajustant sur le SP. Le modèle de régression dépend du type de critère de jugement. Pour les critères de jugement continus, on utilise une régression linéaire. Pour les critères de jugement binaire, on utilise traditionnellement une régression logistique.

1.4.2.5. Comparaisons des quatre méthodes

Tout d'abord, en fonction du type de méthodes, la réponse apportée n'est pas exactement la même.

En effet dans le cadre de la stratification et de la pondération inverse, le résultat est l'effet traitement moyen sur l'ensemble de la population (« Average treatment effect », ATE), alors que dans l'appariement le résultat est l'effet traitement moyen chez les sujets traités (« Average treatment effect on the treated », ATT). L'ATE correspond à répondre à la question « quel serait l'effet moyen si le traitement était étendu à l'ensemble de la population? », alors que l'ATT correspond à répondre à la question « quel est l'effet du traitement pour les sujets qui l'ont reçu? ».

On notera cependant qu'une définition différente des poids permet d'estimer l'ATT y compris par pondération (dans ces cas les patients traités ont un poids de 1, et les patients contrôle ont un poids égal à PS/(1-PS)). Si ATE et ATT peuvent être proches, elles peuvent aussi diverger de façon importante. Il est donc important d'avoir ces notions à l'esprit dans le cadre des études observationnelles avec SP.

Plusieurs études ont montré que la méthode d'appariement équilibrait mieux les caractéristiques de base dans chaque groupe que les méthodes de stratification ou d'ajustement (78,88,89). En revanche le débat l'efficacité de la méthode d'appariement et la méthode de pondération par l'inverse est complexe (90,91).

A noter que les méthodes de stratification, d'appariement et de pondération séparent la méthode de construction des groupes et leurs analyses. La séparation de la méthode de construction des groupes et d'analyses des résultats, permet comme dans un ECR de créer des groupes équilibrés et ensuite d'analyser les critères de jugement. L'absence de séparation de la méthode de création des groupes de l'analyse du critère

de jugement, peut poser des problèmes de fond car les auteurs peuvent être amenés à modifier les paramètres du modèle si les résultats ne sont pas satisfaisants. La réalisation d'un plan statistique précis, qui sera ensuite testé et modifié si nécessaire, tout en restant en aveugle du critère de jugement est une force méthodologique de ce type d'étude.

1.4.3. Evaluation de la similarité des groupes

Le vrai SP est un score d'équilibrage. Dans les conditions du vrai SP, la distribution des caractéristiques de base est similaire dans les 2 groupes « traité » et « contrôle ». Malheureusement dans les conditions d'étude observationnelle, le vrai SP n'est pas connu ; il est estimé par un calcul à partir des données de l'étude.

Ainsi, il est important de tester la capacité du SP calculé à lisser les biais de confusion.

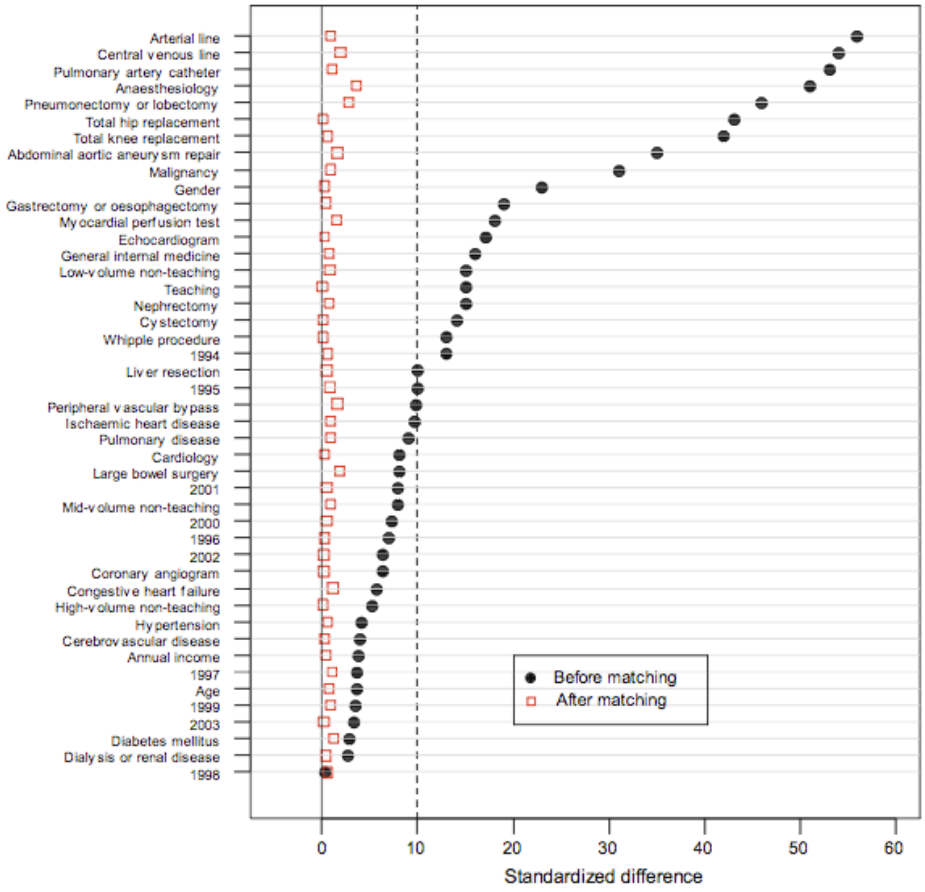
Si celle-ci n'est pas bonne, on peut alors remettre en question le modèle du SP, le modifier ou alors considérer que l'analyse par SP ne permettra pas diminuer suffisamment les biais de confusion.

Pour tester la capacité du modèle, plusieurs méthodes existent, notamment en fonction des types d'utilisation du SP (92–96). Cependant toutes ces méthodes n'ont de sens et de valeur si et seulement si on considère que l'ensemble des variables de confusion sont prises en compte dans le SP (72).

Comparer la similarité des groupes traité et non traité peut commencer en comparant les moyennes ou médianes de distributions des variables continues, et les pourcentages de distribution des variables catégorielles.

Pour plus de finesse de discernement, la méthode des différences standardisées peut être utilisée. Elle compare les différences en unité de l'écart-type combiné sur les deux groupes. Cette méthode n'est pas influencée par la taille de l'échantillon et permet une harmonisation pour des variables qui ne sont pas dans les mêmes unités. Bien qu'il n'y ait pas de limite universelle, une limite de 0,1 (différence entre les moyennes inférieures ou égales à un dixième de l'écart-type) a été considérée comme acceptable (97). Il s'agit de la limite la plus souvent utilisée. Pour une visualisation plus claire du lissage des différences, des méthodes graphiques ont été proposées (Figure 5)

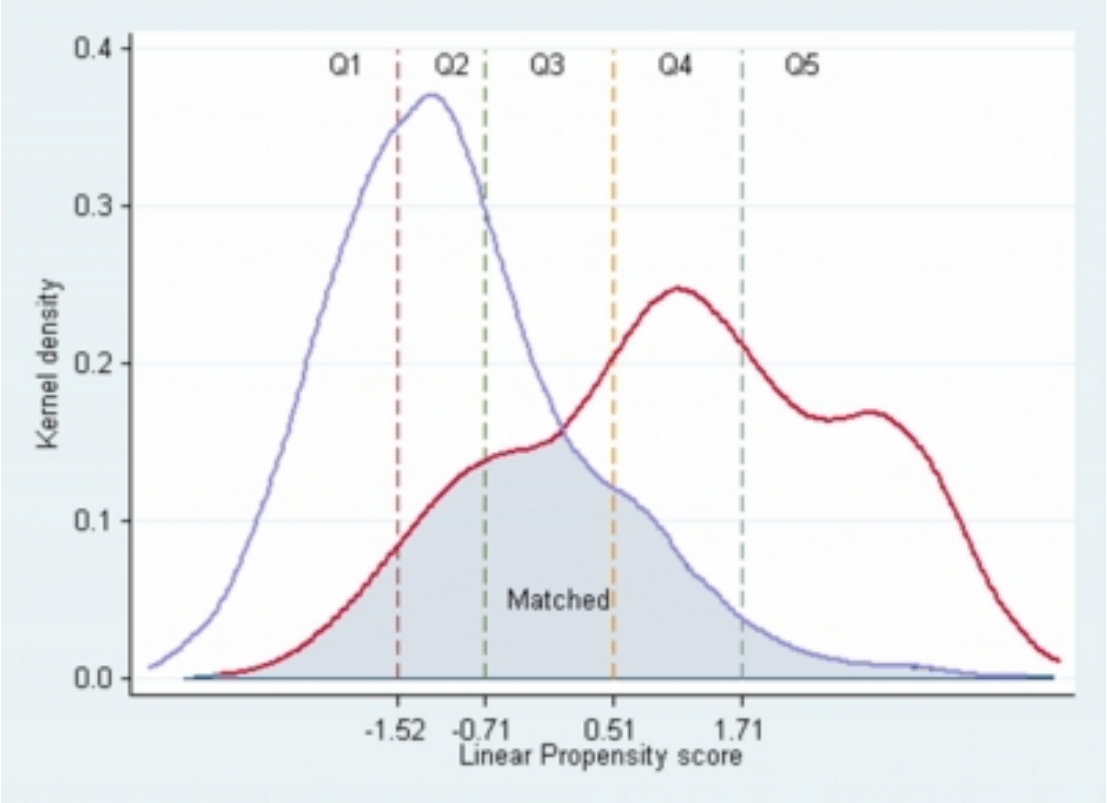
Figure 5 : Exemple de représentation graphique des différences standardisées avant et après appariement (98).



Rosenbaum et Rubin ont initialement décrit une méthode d'approche itérative où le modèle est modifié en ajoutant ou en enlevant des variables tant que des différences persistent entre les deux groupes sur les caractéristiques de base (99).

La représentation des SP pour chaque groupe (figure 6), est plus intéressante pour avoir une notion des zones de chevauchement que pour tester la qualité du modèle.

Figure 6 : Distribution des scores de propension dans la population traitée (ligne bleue) et non traitée (ligne rouge) et la zone de chevauchement en bleu correspondant à la population d'étude (100) .



Plusieurs auteurs ont aussi utilisé des méthodes plus courantes de tests statistiques (101–103). Cependant ce type d'approche a été largement critiqué (92,97). La raison principale est que le degré de signification est directement lié à la taille de l'échantillon. Or dans le cadre de la méthode d'appariement l'échantillon analysé est

toujours plus petit que l'échantillon de départ. Dans ce contexte une différence qui existait dans l'échantillon de départ, peut ne plus être détectée simplement du fait de la réduction de la taille de l'échantillon.

Une autre méthode largement diffusée (104), est l'utilisation de la statistique « C » . Cependant cette méthode est à éviter car elle a montré sa faible capacité a discriminé un bon modèle d'un mauvais (105,106).

1.4.4. Sélection des variables

Il n'y a pas de consensus dans la littérature pour savoir précisément quelles variables il faut inclure dans le modèle du SP. Cependant, l'essence même de l'analyse avec SP réside dans le choix des variables et dans la présence de toutes les variables de confusion. Il existe plusieurs façons de faire. La première consiste à intégrer toutes les caractéristiques de bases mesurées dans le modèle du SP (modèle dit" non parcimonieux"). La seconde approche consiste à ne sélectionner que certaines caractéristiques de bases mesurées (modèle "parcimonieux"). Les dangers de la première méthode sont d'avoir trop de variables et d'avoir un modèle sur-paramétré, rendant des résultats peu fiables (107). Il a aussi été montré que certaines variables, comme les variables dites « instrumentales », peuvent être délétères au modèle. Ces variables « instrumentales » sont fortement liées à l'attribution du traitement mais pas directement au critère de jugement. (108). Ces variables auront donc un poids très important dans le modèle de SP, et les équilibrer se fera au détriment d'autres variables de poids moindre dans le modèle. Or l'équilibre des distributions d'une variable instrumentale entre les groupes n'entraîne aucune réduction du biais (puisque cette variable n'est pas liée au critère de jugement), alors que les déséquilibres induits

sur les autres variables vont entraîner un biais.

Dans la deuxième méthode, le danger est d'oublier une variable de confusion (88).

Il est aussi important de n'intégrer que les variables qui sont des caractéristiques de base et ne pas intégrer des variables qui surviennent après l'assignation au traitement.

Une autre subtilité existe dans la sélection des variables à inclure dans le score. Nous avons vu plusieurs stratégies en fonction des variables mesurées. Cependant, des variables de confusion non mesurées peuvent manquer dans la base de données. Dans ce contexte aucune statistique ou test ne pourra permettre de le détecter. Seule une bonne connaissance du sujet permet aux lecteurs d'être critiques de l'analyse.

1.4.5 Comparaison de l'analyse avec SP avec les autres méthodes d'ajustement

Historiquement, les méthodes de régression classiques étaient plus utilisées que les méthodes se basant sur le SP. Cependant il semblerait que les méthodes avec SP soient plus efficaces (109,110).

En effet, la méthode d'évaluation du modèle est simple et se fonde uniquement sur la similarité des 2 groupes. A contrario, il est un peu plus délicat de préciser si un modèle d'ajustement classique est bien paramétré.

Par ailleurs, comme pour un ECR, les groupes peuvent être créés avec la plus grande similarité avant que l'analyse du critère de jugement soit faite, ce qui n'est pas le cas de la méthode d'ajustement sur le SP.

Pour les évènements rares, les méthodes avec SP sont beaucoup plus flexibles. En effet dans le cas de méthode d'ajustement classique, il est recommandé de ne pas ajuster sur plus d'une variable pour 10 évènements observés. (111) Ainsi, s'il y a peu

d'événement, le nombre de variables prises en compte par la méthode de régression classique sera limité, ce qui n'est pas de cas avec les méthodes avec SP.

Enfin, il sera possible d'analyser le degré de chevauchement des caractéristiques de base. S'il existe des différences importantes sur ces caractéristiques entre les groupes avant ajustement, l'étude du recouvrement des SP (voir figure 7) permet d'adapter l'analyse. Il a alors la possibilité de réduire l'effectif pour comparer ce qui est comparable, ou de ne pas réaliser l'analyse en considérant que la population d'étude serait alors tellement réduite qu'elle ne serait plus représentative de la population d'origine. Dans une analyse ajustée, il est facile de passer à côté d'un manque de chevauchement des distributions des variables ; le modèle conduit alors à estimer l'effet du traitement par extrapolation, ce qui est problématique.

1.5 Objectifs

Les objectifs de ce travail étaient :

1) de comparer l'effet traitement estimé par les études observationnelles utilisant un score de propension et l'effet traitement estimé par des ECR en chirurgie

2) de faire une description précise de l'utilisation du score de propension dans les études observationnelles évaluant des interventions chirurgicales.

I Introduction et objectifs

II Article 1: Comparaison de la mesure de l'effet traitements entre les études observationnelles utilisant un score de propension et les ECR évaluant une procédure chirurgicale

III Article 2 : Description et Sources de biais des Etudes observationnelles utilisant une analyse avec score de propension évaluant une procédure chirurgicale. Revue systématique

IV Discussion et perspectives

V Conclusion

VI Réferences

II Comparaison de la mesure de l'effet traitement entre les études observationnelles utilisant un score de propension et les ECR évaluant une procédure chirurgicale.

Un résumé de la méthodologie et des principaux résultats est présenté ci-dessous. La description détaillée de la méthodologie et des résultats de ce travail, sont présentés dans l'article publié en anglais (112) (section 2.6). Dans le cadre de la publication, une lettre à l'éditeur écrite par une équipe anglo-saxonne critiquant nos choix statistiques (113), ainsi que notre réponse ont été publiées (114) (section 2.7).

2.1 Introduction

Le nombre de publications d'études observationnelles comparatives utilisant un score de propension augmente de façon exponentielle (82,115). Cependant leur capacité à estimer l'effet traitement est remise en question.

Récemment deux études méta-épidémiologiques dans le domaine de la cardiologie interventionelle et de la chirurgie cardiaque ont comparé l'estimation de l'effet traitement des ECR et des études observationnelles utilisant un score de propension (116) (117). Ces deux études concluaient qu'en moyenne les deux types d'études estimaient de façon similaire l'effet traitement mais que, dans certains exemples, des différences importantes d'estimation de l'effet traitement pouvaient exister.

Le but de notre étude était, à l'instar des études précédentes, de comparer l'estimation de l'effet traitement entre les ECR et les études observationnelles utilisant un SP mais dans le cadre de l'évaluation thérapeutique d'une procédure chirurgicale.

2.2 Méthodes

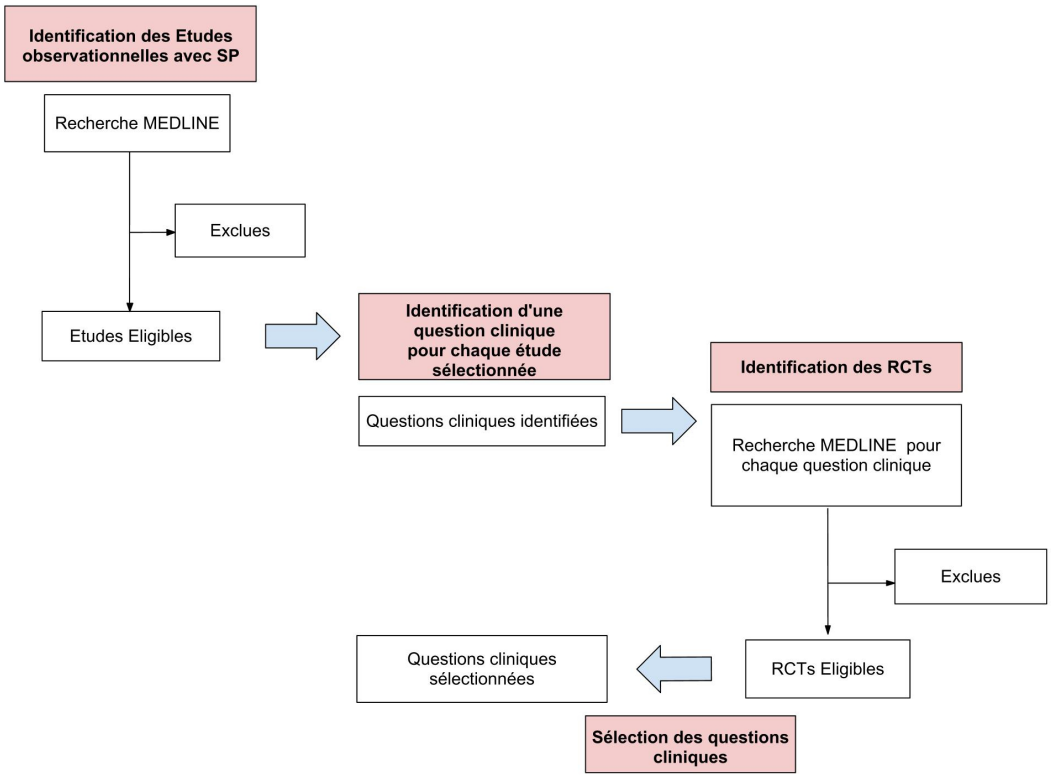
2.2.1. Plan expériemental

Nous avons réalisé une étude méta-épidémiologique.

2.2.2. Identification des études randomisées et observationnelles

Pour comparer les mesures d'effet traitement nous avons procédé en 2 étapes (Figure 7). Dans un premier temps, nous avons identifié toutes les études observationnelles utilisant une analyse avec SP évaluant une procédure chirurgicale indexées dans PubMed. Dans un deuxième temps nous avons identifié tous les ECR qui répondaient aux mêmes questions cliniques que les études observationnelles sélectionnées.

Figure 7 : Processus d'identification des Etudes observationnelles avec SP et ECR répondant aux mêmes questions cliniques.



L'identification des études observationnelles a été réalisée selon les standards des revues systématiques, à savoir une recherche par mots clés la plus large possible sur la base de données PUBMED. Une fois que les études observationnelles ont été sélectionnées, nous avons identifié des questions cliniques, définies par une population, une intervention, un contrôle et un critère de jugement. Pour chaque question clinique, une nouvelle recherche systématique d'ECR a été réalisée. A noter que pour une question clinique donnée, plusieurs études observationnelles et/ou plusieurs études pouvaient être retrouvées.

2.2.3. Extraction des données

Nous avons systématiquement collecté 1) les caractéristiques générales de l'étude (année de publication, pays d'investigation, spécialité chirurgicale, nombre de centre, nombre de chirurgien, taille de l'échantillon), 2) des caractéristiques plus spécifiques des études observationnelles (le type et le nombre de variables inclus dans le score de propension et la méthode utilisée (appariement, ajustement, stratification, pondération inverse)).

2.2.4. Mesure de l'effet traitement et comparaison

Pour chaque question clinique, nous avons identifié 1 critère de jugement dichotomique objectif et 1 critère de jugement dichotomique subjectif. Pour chaque critère de jugement, nous avons extrait la table 2x2 rapportant le nombre d'évènements dans chaque groupe et le nombre de patient par groupe. Si cette table n'était pas disponible nous extrayions directement l'effet traitement rapporté par l'article (OR ou RR).

A partir des données rapportées dans l'article, nous avons calculé pour chaque type d'étude (étude observationnelle avec SP et ECR) et chaque question clinique une mesure relative de l'effet traitement (OR). Ensuite les mesures relatives de l'effet traitement ont été comparées au sein des différentes questions cliniques sous la forme d'un Ratio d'Odd Ratio (ROR). Les ROR de toutes les questions cliniques ont ensuite été méta-analysés.

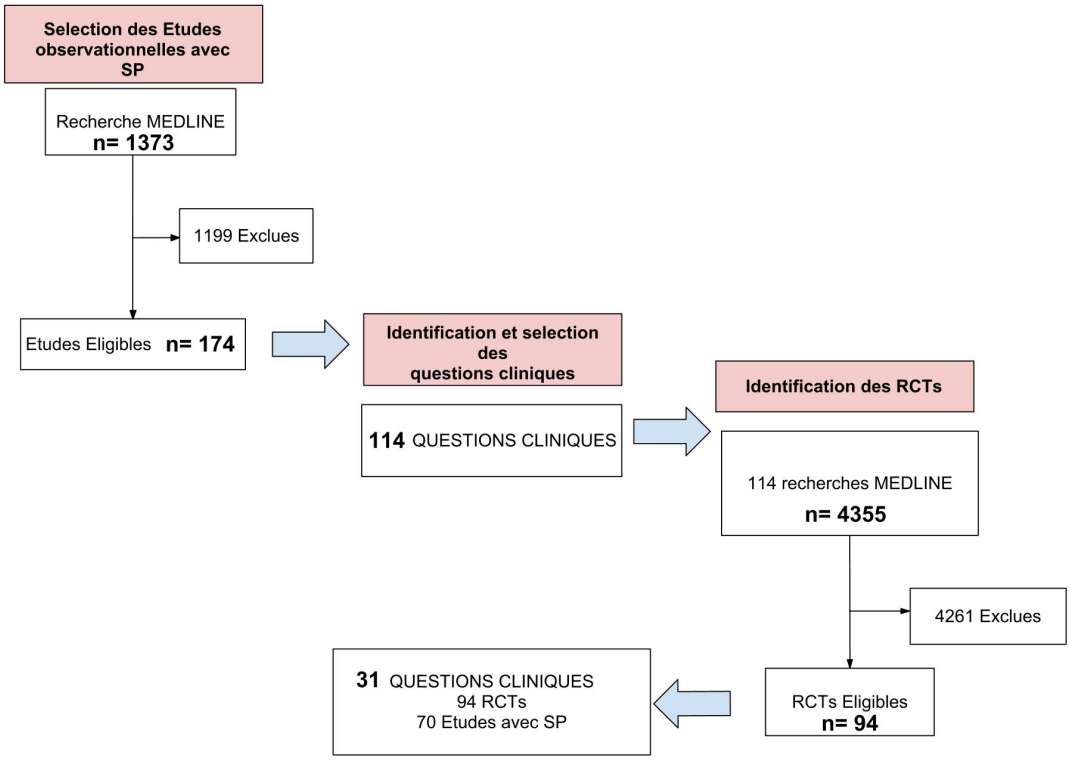
2.3 Résultats

2.3.1. Identification des études

Parmi les 1373 citations retrouvées sur PUBMED, 174 articles rapportant une étude observationnelle avec un score de propension ont été sélectionnées. Ces 174 articles concernaient 114 questions cliniques différentes.

Pour chacune de ces 114 questions cliniques, nous avons systématiquement recherché des ECR répondant à la même question clinique. Pour 31 questions cliniques, nous avons retrouvé au moins un ECR (Figure 8). Au final, les 31 questions cliniques regroupaient 94 publications d'ECR et 70 publications d'études observationnelles avec un SP.

Figure 8 : Résultats des du processus d'identification des différents types d'études



La médiane du nombre d'ECR par question clinique était de 2 (Q1-Q3: 1–3, min-max: 1-22) et la médiane du nombre d'études observationnelles était de 1 (Q1-Q3: 1–2, min-max: 1-14).

Sur les 31 questions cliniques, 22 rapportaient au moins un critère de jugement objectif, et 26 au moins un critère subjectif. Les questions cliniques concernaient majoritairement la chirurgie cardiaque (n=22).

Le nombre médian de participant était de 2,049 (Q1-Q3: 828–6,461) pour les études observationnelles et de 179 (Q1-Q3: 99–341) pour les ECR.

2.3.2. Comparaison des effets traitements de façon qualitative (figure 9)

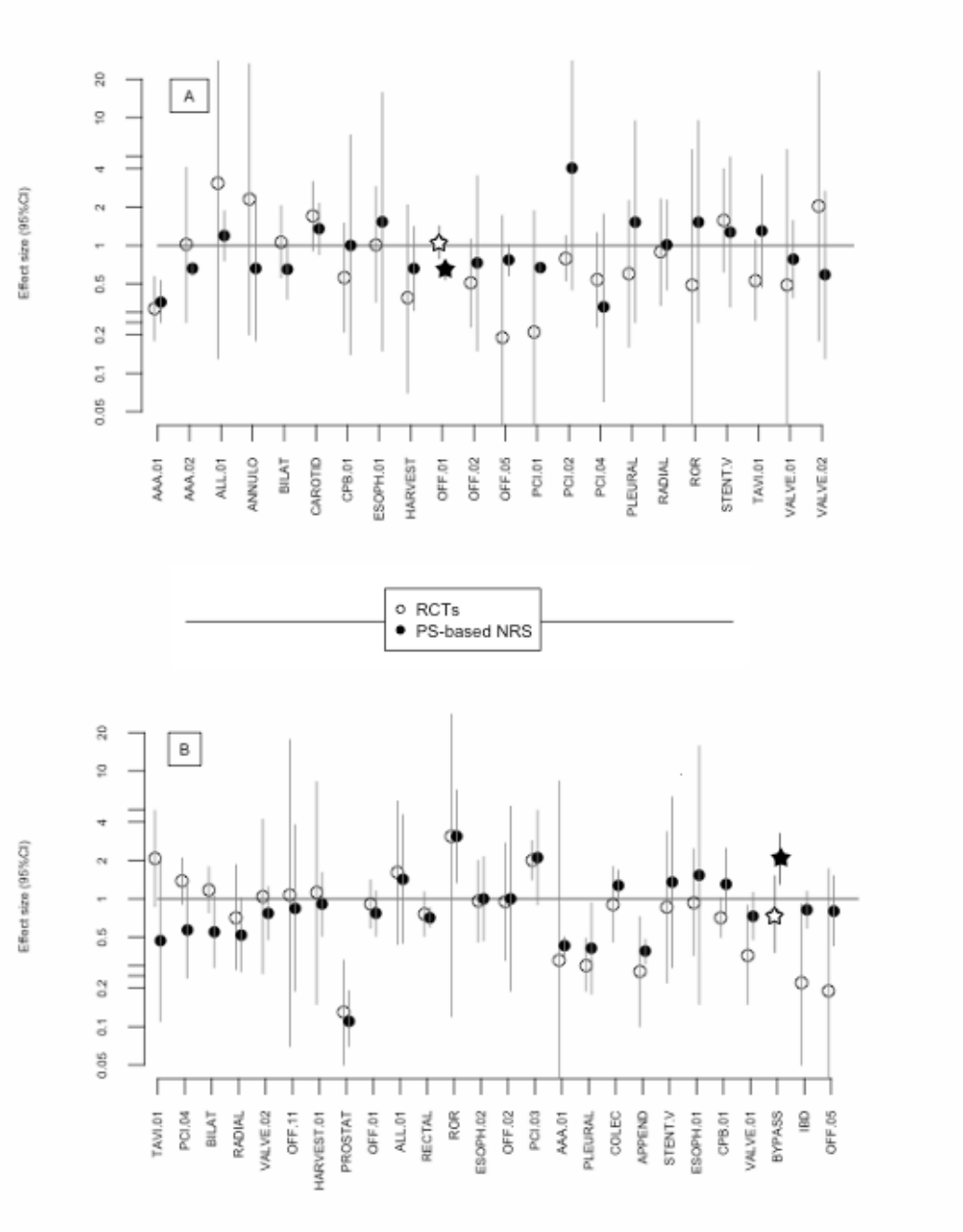
Pour les critères objectifs, dans 9 cas sur 22 (40%) les études observationnelles avec score de propension estimaient un effet traitement plus important que les ECR et dans 13 cas c'était l'inverse. A noter que pour seulement une question clinique, cette

différence (estimation plus importante de l'effet traitement par l'ECR) était statistiquement significative.

Pour les critères subjectifs, dans 11 cas sur les 26 (42%), les études observationnelles avec score de propension estimaient un effet traitement plus important que les ECR et dans 15 cas c'était l'inverse (58%). Cette différence (estimation plus importante de l'étude observationnelle) était statistiquement significative pour une seule question clinique.

Figure 9 : Représentation graphique de l'estimation de l'effet traitement (rond) avec leur intervalle de confiance (trait) de chaque type d'étude, pour les différentes questions cliniques. Les ronds pleins représentent l'estimation de l'effet traitement pour les études observationnelles et les ronds vides pour les ECR.

A) Représentation des critères de jugement objectifs ; B) Représentation des critères de jugement subjectif. Les représentations avec étoiles sont les questions cliniques où la différence est statistiquement significative.

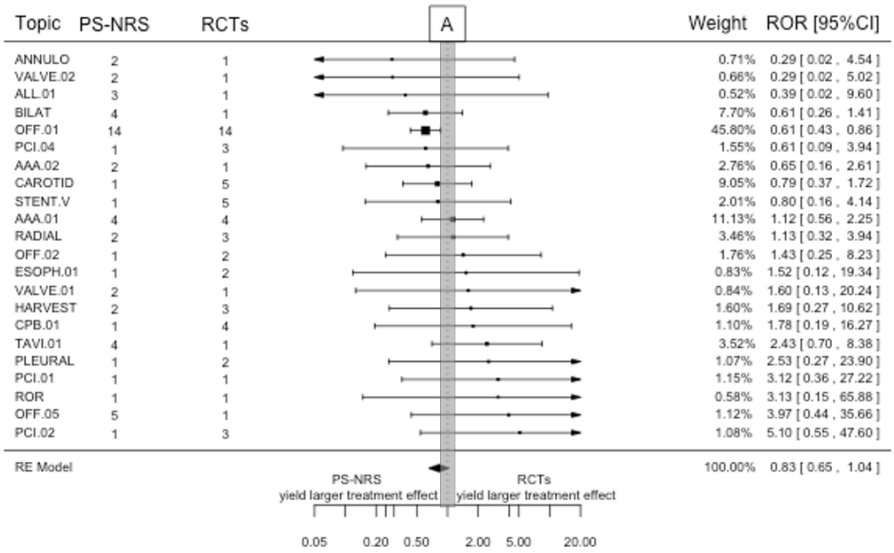


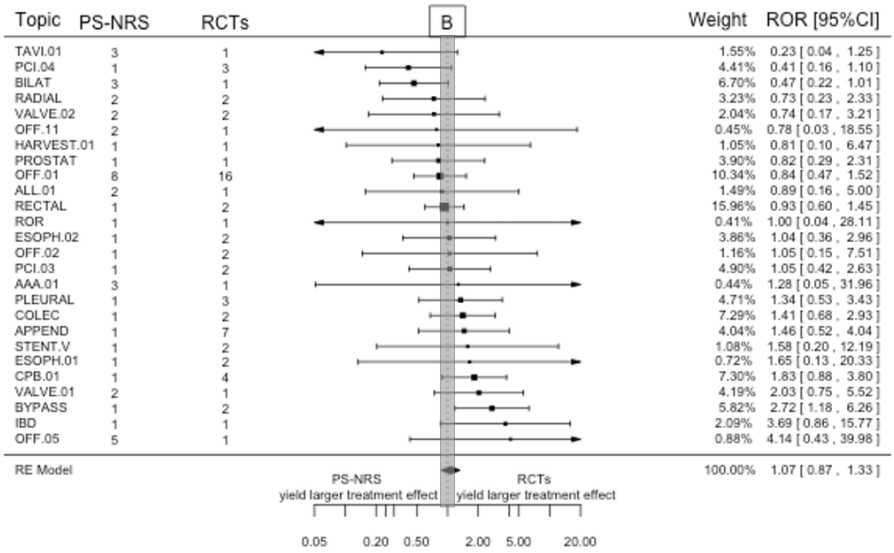
Comparaison des effets traitements de façon quantitative (figure 10)

Pour l'analyse méta-épidémiologique, le ROR combiné pour les critères de jugement objectifs était de 0.83 ([95% CI 0.65–1.04], p=0.11, I^2= 0%; variance entre les questions cliniques = 0.00). A noter que dans cette méta-analyse, une question clinique (comparaison dans le cadre des pontages coronariens de la présence ou non d'une Circulation Extra Corporelle) représentait 46 % du poids global de l'analyse.

Pour les critères de jugement subjectifs, le ROR combiné était égal à 1.07 ([0.87–1.33], p=0.52, I^2= 8%; variance entre les questions clinique = 0.03).

Figure 10 : Diagramme en forêt représentant les résultats de la méta-analyse. Chaque ligne représente une question clinique. A) Représentation des critères de jugement objectifs ; B) Représentation des critères de jugements subjectifs.





2.4 Discussion

Dans cette étude, nous n'avons pas retrouvé en moyenne de différence statistiquement significative (ROR global) pour l'estimation de l'effet traitement entre les ECR et les études observationnelles avec SP. Cependant, nous avons pu constater des variations dans les 2 sens pouvant parfois être importantes.

Ces résultats sont consistants avec la série d'études méta-épidémiologiques, comparant les effets traitements des ECR et études observationnelles (66,68,118–121). Les résultats sont aussi cohérents avec ceux de deux études méta-épidémiologiques récentes comparant les traitement des ECR avec des études observationnelles utilisant un score de propension (116,117). La première réalisée par Kuss et al., comparait dans le cadre d'une question définie (comparaison dans le cadre des pontages coronariens de la présence ou non d'une CEC), 10 critères de jugement différents. (116) La seconde réalisée par Dahabreh et al., regroupait dans le cadre d'une pathologie définie (le syndrome coronarien aigu), 17 comparaisons de traitements possibles de cette pathologie (dilatation, traitement médical, pontage) (117) .

Nos résultats doivent être cependant interprétés avec précaution. Dans les faits, le vrai effet traitement n'est pas connu et plusieurs facteurs peuvent affecter l'effet traitement aussi bien pour les ECR que pour les études observationnelles avec SP. Dans notre étude, une grande différence existait dans la qualité méthodologique des ECR. Par exemple seul 33% des ECR décrivaient une méthode d'aveugle. De la même manière, une grande hétérogénéité existait pour les études observationnelles. La taille de l'échantillon variait de 161 à 322 892.

Dans le cadre de question clinique avec une seule étude observationnelle ou un seul ECR, il est compréhensible de comprendre la force de ces facteurs sur l'estimation de

l'effet traitement et donc expliquer les grandes différences pouvant exister entre les deux types d'études dans un sens comme dans l'autre.

Notre étude présente de nombreuses limites. Notre recherche est limitée à une seule base de données (MEDLINE). Cependant, dans le contexte de cette étude, il est peu probable que le choix de la base de données utilisée puisse biaiser les résultats. Nous avons concentré nos recherches uniquement sur les études avec critère de jugement binaire et les résultats ne sont donc pas généralisables aux critères de jugement continus ou censurés. Cette étude méta-épidémiologique ne tient pas compte de la qualité méthodologique des essais randomisés et des études observationnelles incluses dans l'analyse.

2.5 Conclusion

Dans notre étude, il n'existe pas de différence statistiquement significative dans l'estimation de l'effet traitement entre ECR et études observationnelles avec SP. Cependant, compte tenu du nombre limité d'études, il faut rester prudent dans l'interprétation de ces résultats car pour certaines questions cliniques on observait une différence importante entre l'estimation de l'effet traitement par l'ECR et l'étude observationnelle avec SP.

2.6 Article publié

Comparison of Treatment Effect Estimates From Prospective Nonrandomized Studies With Propensity Score Analysis and Randomized Controlled Trials of Surgical Procedures

Guillaume Lonjon, MD,† Isabelle Boutron, MD, PhD,*‡§¶ Ludovic Trinquart, MSc,*‡§¶*
Nizar Ahmad, MD, PhD,‡ Florence Aim, MD,*† Rémy Nizard, MD, PhD,† and Philippe Ravaud, MD, PhD*‡§¶||*

Objective: We aimed to compare treatment effect estimates from NRSs with PS analysis and RCTs of surgery.

Background: Evaluating a surgical procedure in randomized controlled trials (RCTs) is challenging. Nonrandomized studies (NRSs) involving use of propensity score (PS) analysis to limit bias are of increasing interest.

Design: Meta-epidemiological study.

Methods: We systematically searched MEDLINE via PubMed for all prospective NRSs with PS analysis evaluating a surgical procedure. Related RCTs, addressing the same clinical questions, were systematically retrieved. Our primary outcome of interest was all-cause mortality. We also selected 1 subjective outcome. We calculated the summary odds ratios (OR) for each study design, the ratio of OR (ROR) between the designs and the summary ROR across clinical questions. An ROR < 1 indicated that the experimental intervention is more favorable in NRSs with PS analysis than RCTs.

Results: We retrieved 70 reports of NRSs with PS analysis and 94 related RCTs evaluating 31 clinical questions, of which 22 assessed all-cause mortality and 26 a subjective outcome. The combined ROR for all-cause mortality was 0.83 (95% confidence interval: 0.65–1.04). For subjective outcomes, the combined ROR was 1.07 (0.87–1.33).

Conclusions: There was no statistically significant difference in treatment effect between NRSs with PS analysis and RCTs. Prospective NRSs with suitable and careful PS analysis can be relied upon as evidence when RCTs are not possible.

Keywords: comparative effectiveness research, meta-epidemiological study, nonrandomized study, propensity score, randomized controlled trial

(*Ann Surg* 2013;00:1–8)

Randomized controlled trials (RCTs) are usually considered the criterion standard for therapeutic evaluation. However, in the field of surgery, RCTs present several methodological and practical challenges.[1–4] Patient recruitment is usually difficult because of patient and surgeon preference. Furthermore, most RCTs evaluate only the efficacy of the intervention in a small number of selected centers, with highly trained surgeons performing a standardized procedure in a selected population. These trials do not usually provide information on the effectiveness of the intervention, that is, the effect of the intervention applied to a wider population in a more representative context.

Recently, several initiatives such as comparative effectiveness research have highlighted the need to identify which treatments work in a real-world setting to perform pragmatic trials[5] and to use data from nonrandomized studies (NRSs).[6–8] However, unlike RCTs, with patient characteristics balanced in both groups, NRSs exhibit significant risk of confounding bias.[9] Propensity score (PS) analyses are statistical techniques for dealing with confounding bias in observational studies. Patient, context, and provider characteristics are used to calculate a PS for each individual (ie, the probability that a subject will receive an intervention).[10] These scores are then used to adjust the analysis or to create a matched cohort of patients. Under some assumptions, PS analysis can be used to reduce bias in observational studies.[11] Moreover, previous findings showed that PS analysis could produce estimates that were less biased, more robust, and more precise than the logistic regression estimates when there were few events per confounding variables.[12] Therefore, PS analysis has been increasingly used in the last 10 years, specifically in studies of surgery.[13–16]

Previous meta-epidemiological studies compared treatment effect estimates from randomized studies and NRSs.[17–24] However, these studies did not focus on NRSs with PS analysis evaluating a wide range of surgical procedures.

We aimed to perform a meta-epidemiological study comparing treatment effect estimates from NRSs with PS analysis and RCTs of surgical procedures.

METHODS

To compare treatment effect estimates from NRSs with PS analysis and RCTs, our meta-epidemiological study involved 2 main parts. First, we performed a systematic review by searching MEDLINE via PubMed for reports of NRSs of surgery that involved PS analysis. Second, for each selected report, we searched MEDLINE to identify all reports of RCTs answering the same clinical question (Fig. 1; the PRISMA check list is in the Supplemental Digital Content, available at http://links.lww.com/SLA/A456).

Study Selection

Search for NRSs with PS analysis

We systematically searched MEDLINE via PubMed to identify reports of NRSs with PS analysis of surgery with the following strategy "Surgical Procedures, Operative"[Mesh] AND (Propensity OR "Propensity Score"[Mesh]) AND (versus[tiab] OR comparison[tiab] OR compared[tiab] OR "Comparative Study"[All Fields] OR "Comparative Study"[Publication Type]) with a limitation to studies published in English and no time restriction.

From the *Unité INSERM U738; †Assistance Publique-Hôpitaux de Paris, Hôpital Lariboisière, Service d'orthopédie et traumatologie; ‡Assistance Publique-Hôpitaux de Paris, Hôpital Hôtel-Dieu, Centre d'Epidémiologie Clinique; §Université Paris Descartes—Sorbonne Paris Cité; ¶French Cochrane Centre, Paris, France; and ||Columbia University, Mailman School of Public Health, NY.

Disclosure: This study received source from Equipe "Espoirs de la Recherche" par la Fondation pour la Recherche Médicale (FRM). All the authors declare no support from any organization for the submitted work; no financial relationships with any organization that might have an interest in the submitted work in the previous 3 years; and no other relationships or activities that could appear to have influenced the submitted work.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.annalsofsurgery.com).

Reprints: Guillaume Lonjon, MD, Centre d'Epidémiologie Clinique, Hôpital Hôtel-Dieu, 1 place du parvis Notre Dame, 75004 Paris, France. E-mail: dr.guillaume.lonjon@gmail.com.

ISSN: 0003-4932/13/00000-0001
DOI: 10.1097/SLA.0000000000000256

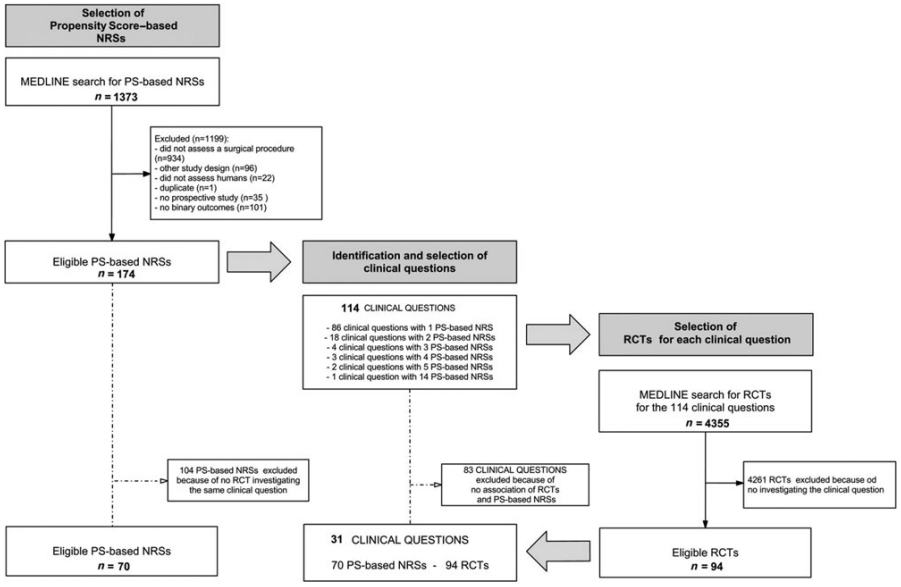


FIGURE 1. Selection of reports of NRSs involving propensity scores (PS-based NRSs) and RCTs of surgery by clinical question.

One of us, with surgical and methodological expertise, screened the title, abstract and full text of reports to identify eligible studies. Prespecified eligibility criteria were nonrandomized comparative studies assessing a surgical procedure in humans, with prospective recruitment and data collection (ie, prospective cohort studies and administrative databases), and involving PS analysis with a binary outcome.

Because we wanted to focus on a homogenous sample of surgical procedures, we excluded reports of interventional procedures such as percutaneous coronary intervention or gastrointestinal endoscopy not performed by surgeons. Surgical procedures could be compared to any comparator (eg, surgical procedure, interventional procedures, pharmacological treatment, physiotherapy, usual care, placebo).

Identification of clinical questions and search for related RCTs

For each eligible NRS with PS analysis, one reviewer identified a clinical question defined by a population, an intervention, and a comparator and then sorted the studies by clinical question. Then, we systematically searched PubMed for RCTs addressing the same clinical questions. We used a sensitive search strategy that combined the Cochrane Highly Sensitive Search Strategy for RCTs (Appendix 1, available at http://links.lww.com/SLA/A453) and keywords related to the condition and intervention. We limited our search to reports published in English 5 years before and after the publication date of the selected NRSs. For cases of 2 or more NRSs per clinical question, we restricted our search to 5 years before the oldest publication and 5 years after the most recent publication.[25] One reviewer screened the title, abstract and full text of reports to identify the RCTs. Another reviewer, with surgical and methodological expertise, assessed whether the selected RCTs indeed addressed the same clinical question as the NRSs. Any discrepancies were discussed with a third reviewer.

Data Extraction and Quality Assessment

A standard data collection form was developed for each study design and preliminarily tested by 2 reviewers. One reviewer extracted all data. A second reviewer independently duplicated data extraction related to risk of bias and collection of outcomes. All discrepancies were discussed to achieve consensus. Interobserver reproducibility was calculated by Cohen kappa coefficient (K). The median K was 0.67 [quartile 1–3 (Q1–Q3): 0.57–0.78].

General characteristics

For each study, we recorded the year of publication, location of study, surgical specialty, period of recruitment, number of centers, number of participating surgeons, and reporting of sample size calculation. For NRSs, we determined whether the data came from a cohort study or an administrative database and whether the study had the same period of recruitment for the experimental and control groups. We also extracted specific information on PS analysis: the type and number of covariates included in the PS and the PS technique used (matching, stratification, inverse probability of treatment weighting, or adjustment).

We assessed the methodological quality of RCTs by using the Cochrane Risk of Bias (RoB) tool.[26] For NRSs, we focused on the same domains of bias as assessed by the RoB tool for RCTs, except for bias due to confounding because we aimed to determine whether the use of PS analysis could limit the risk of such bias.

- Performance bias, that is, bias due to departures from intended interventions. Because patient blinding is not possible in NRSs, we assessed whether contaminations could bias treatment effect estimates. For this purpose, we determine whether participants remained with the original intervention without contamination by the other intervention and we evaluated whether contamination was sufficiently low to avoid bias.
- Detections bias, that is, bias in taking measurements. We rated studies with an objective outcome as having a low risk of bias and studies with a subjective outcome as having high risk of bias unless outcome assessors were blinded.
- Attrition bias, that is, bias due to missing data. We rated studies as having low risk of bias if the proportion of missing data was low, with balanced number and reasons for missing data across intervention groups.

Outcomes selection and extraction of results

Our primary outcome of interest was all-cause mortality. In fact, there is some evidence that the risk of bias is higher for subjective outcomes than all-cause mortality.[27,28]

We also selected 1 subjective outcome such as patient-reported outcomes—physician-assessed disease outcomes and measures combined from several outcomes with at least 1 subjective outcome (eg, major cardiac and cerebrovascular event). Subjective outcomes were selected first by order of frequency (ie, the outcome reported for the greatest number of studies) and if necessary, by number of events (ie, the outcome reported with the greatest number of events).

For both NRSs with PS analysis and RCTs, we extracted a reported 2×2 table summarizing the number of patients with the outcome and the total number of patients in each group. If a 2×2 table was not available, we extracted the treatment effect estimates reported in the article [odds ratio (OR) or relative risk (RR)] and the associated 95% confidence interval (CI).

Statistical Analyses

Estimation of treatment effect estimates for each clinical question by study design

Treatment effects were estimated by ORs. Endpoints were recoded so that an OR < 1 indicated a beneficial effect of the experimental intervention. For reports of NRSs with PS analysis giving only treatment effect estimates (OR or RR) but no 2×2 table (ie, with stratification or adjustment on PS), we reconstructed a 2×2 table according to the method described by Di Pietrantonj.[29] When required, we used a continuity correction method (ie, adding to all cells a factor proportional to the reciprocal of the size of the contrasting study group) to take into account zero cell counts in 1 group only.[30] If no events occurred in all arms, the study was excluded from the analysis. With more than 1 study per study type (NRSs or RCTs), we calculated combined ORs and 95% CIs by using both a random-effects model[31] and a fixed-effect model.

To evaluate concordance between treatment effect estimates of NRSs with PS analysis and RCTs, we plotted side by side the summary ORs for randomized trials and NRSs with PS analysis in each clinical question.[17] We evaluated whether the difference in the 2 summary ORs for the same clinical question was larger than what would be expected by chance alone.

Meta-epidemiological analysis

We used a meta-epidemiological analysis to estimate the combined difference in treatment effect estimates between NRSs with PS analysis and RCTs by a 2-step method.[32] For each clinical question, we estimated the ratio of the treatment effect for NRSs with PS analysis to that for RCTs, the ratio of ORs (ROR). Then, we derived the log ROR as log(OR of NRS with PS analysis) − log(OR of RCT) and its standard error as SE(log ROR) = $\sqrt{}$[SE(OR of NRS with PS analysis)^2 + SE(OR of RCT)^2].[33] We estimated a combined ROR and 95% CI across clinical questions by a random-effects meta-analysis model. An ROR < 1 indicated that NRSs with PS analysis yielded larger treatment effect estimates than RCTs. Heterogeneity of RORs across the different clinical questions was assessed by the I^2 statistic and by estimating the variance between clinical questions. The square root of the latter is the estimated SD of underlying RORs across clinical questions. We plotted the ROR for each clinical question and the combined ROR on a forest plot. We performed 2 separate analyses for all-cause mortality and subjective outcomes.

Sensitivity analyses

In sensitivity analyses, we restricted the meta-epidemiological analyses to several subpopulations. First, we limited the analysis to NRSs with PS matching. Second, we limited the analysis to NRSs with PS analysis that recruited patients in the treatment and control groups simultaneously. Then, we restricted the analyses to different groups of selected RCTs (RCTs at low risk of bias for allocation concealment, random sequence generation, and incomplete outcome data).

Analyses involved use of R software (online at http://www.R-project.org, the R Foundation for Statistical Computing, Vienna, Austria). A 2-tailed $P < 0.05$ was considered statistically significant.

RESULTS

Selection of Studies

From the 1373 citations retrieved from MEDLINE, we selected 174 reports of NRSs with PS analysis that addressed 114 clinical questions (Fig. 1). The MEDLINE search of related RCTs yielded 4355 citations, from which we selected 94 reports of RCTs matching 70 reports of NRSs with PS analysis and addressing 31 clinical questions. These questions involved cardiac surgery (n = 22), digestive surgery (n = 5), thoracic surgery (n = 2), urology (n = 1) and vascular surgery (n = 1). The median number of NRSs by clinical question was 1 (Q1–Q3: 1–2, min–max: 1–14) and the median number of RCTs was 2 (Q1–Q3: 1–3, min–max: 1–22). Of these 31 questions, 22 assessed all-cause mortality (in-hospital or 30-day mortality for 21) and 26 a subjective outcome, mainly physician-assessed disease outcomes (eg, myocardial infarction, stroke, wound infection, return to surgery) (Table 1).

Characteristics of NRSs With PS Analysis and RCTs

The median number of participants was 2049 (Q1–Q3: 828–6461) for NRSs with PS analysis and 179 (Q1–Q3: 99–341) for RCTs (Table 2). The sample size calculation was given in 3% of NRSs reports and 46% of RCT reports.

For RCTs, risk of bias varied by domain: 50% of reports had a low risk of selection bias, 3% low risk of performance bias, 33% low risk of detection bias, and 91% low risk of attrition bias.

For NRSs with PS analysis, 90% used an administrative database (Table 2). Most reports of these studies (81%) described recruiting patients in the treatment and control groups simultaneously, 9% declared that recruitment was not simultaneous and 10% did not provide sufficient information. The number of covariates included in the PS model was mentioned in 71% of reports and the median number of included covariates was 14 (Q1-Q3: 10–20). The type of PS technique was matching in 75% of reports. The risk of bias varied by domain. For performance bias, as expected in all studies, patients and care providers were not blinded. When evaluating the rate of contamination, 91% of reports did not provide any

TABLE 1. Summary of 31 Clinical Questions Studied in NRSs Involving Propensity Scores (PS-Based NRSs) and RCTs (Text in Bold Explain the Choice of Clinical Question Code)

Clinical Question Code	Population (Subpopulation)	Intervention	Comparator	Speciality	All-Cause Mortality	PS-Based NRSs/RCTs	Subjective Outcome	PS-Based NRSs /RCTs
AAA.01	Abdominal Aortic Aneurysm (no ruptured)	Endovascular repair	Open repair	Cardiac surgery	Death perioperation	4/4	Acute renal failure	3/1
AAA.02	Abdominal Aortic Aneurysm (ruptured)	Endovascular repair	Open repair	Cardiac surgery	Death perioperation	2/1	*None*	
ALL.01	CABG	All arterial revascularization	Conventional revascularization	Cardiac surgery	Death perioperation	3/1	MI	3/1
Annulo	CABG + Mitral regurgitation	CABG + Annuloplasty	CABG alone	Cardiac surgery	Death perioperation	2/1	*None*	
Append	Appendectomy	Laparoscopic	Open	Digestive surgery	None		Wound infection	1/7
Bilat	CABG	Bilateral internal thoracic arteries	Single internal thoracic artery	Cardiac surgery	Death perioperation	4/1	Reoperation for bleeding	3/1
Bypass	Gastric bypass	Laparoscopic	Open	Digestive surgery	None		Postoperative complication	1/2
Carotid	Carotid stenosis	Stent	Endarterectomy	Vascular surgery	Death perioperation	1/5	Stroke	2/5
Colec	Colectomie	Laparoscopic	Open	Digestive surgery	None		Wound infection	1/2
CPB.01	CABG	Miniaturized CPB	Conventional CPB	Cardiac surgery	Death perioperation	1/4	New AF	1/4
Esoph.01	Esophagectomy anastomosis	Stapled	Hand sewn	Thoracic surgery	Death perioperation	1/2	Anastomosis leak	1/2
Esoph.02	Esophagectomy anastomosis	Stapled	Hand sewn	Thoracic surgery	None		Cardiac complication	1/1
Harvest.01	Vein harvesting for CABG	Mini invasif	Conventional	Cardiac surgery	Death perioperation	2/3	Reoperation for bleeding	1/1
IBD	Inflammatory Bowel Disease resection	Laparoscopic	Open	Digestive surgery	None		Postoperative complication	1/1
OFE.01	CABG	Off-pump (beating heart)	On-pump	Cardiac surgery	Death perioperation	14/14	MI	8/16
OFE.02	CABG (High risk patient)	Off-pump (beating heart)	On-pump	Cardiac surgery	Death perioperation	1/2	Stroke	1/2
OFE.05	CABG (Emergency)	Off-pump (beating heart)	On-pump	Cardiac surgery	Death perioperation	5/1	Reoperation for bleeding	5/1
OFE.11	CABG (All arterial revascularization)	Off-pump (beating heart)	On-pump	Cardiac surgery	None		Stroke	2/1
PCI.01	Acute coronary syndrome (Elderly)	Percutaneous Coronary Intervention	CABG	Cardiac surgery	Death perioperation	1/1	None	
PCI.02	Acute coronary syndrome (Multi-vessel disease)	Percutaneous Coronary Intervention	CABG	Cardiac surgery	Death perioperation	1/3	None	
PCI.03	Acute coronary syndrome (Diabetes)	Percutaneous Coronary Intervention	CABG	Cardiac surgery	None		MACCE 1Y	1/2
PCI.04	Acute coronary syndrome (Unprotected left main artery)	Percutaneous Coronary Intervention	CABG	Cardiac surgery	Death at 1 year	1/3	MACCE 1Y	1/3
Pleural	CABG	Preservation of pleural integrity	No preservation	Cardiac surgery	Death perioperation	1/2	Pleural effusion	1/3
Prostat	Prostatectomy	Mini-invasif	Open (retropubic approach)	Urology	None		Transfusion	1/1
Radial	CABG	Radial artery graft	Venous artery graft	Cardiac surgery	Death perioperation	2/3	Stroke	2/2
Rectal	Resection of rectal cancer	Laparoscopic	Open	Digestive surgery	None		Post op complication	1/2
ROR	Mitral valve surgery	Replacement	Repair	Cardiac surgery	Death perioperation	1/1	Reoperation for bleeding	1/1
Stent.v	Aortic valve surgery	Stentless valve	Stented valve	Cardiac surgery	Death perioperation	1/5	None	
TAVI.01	Aortic valve replacement	Transcatheter aortic valve implantation	Conventional surgery	Cardiac surgery	Death perioperation	4/1	Stroke	3/1
Valve.01	Valve surgery (Aortic)	Minimally invasive approach	Full sternotomy	Cardiac surgery	Death perioperation	2/1	Transfusion	2/1
Valve.02	Valve surgery (Mitral)	Minimally invasive approach	Full sternotomy	Cardiac surgery	Death perioperation	2/1	Reoperation for bleeding	2/2

AF indicates atrial fibrillation; CABG, coronary artery bypass graft; CPB, cardio pulmonary bypass; MACCE, major adverse cardiac and cerebrovascular event; MI, myocardial infarction.

TABLE 2. Characteristics of PS-Based NRSs and RCTs.

Characteristics of Trials	PS-Based NRSs (n = 70)	RCTs (*n* = 94)
Median date of publication, y (Q1,Q3)	2008 (2005,2010)	2008 (2005,2010)
Speciality		
Cardiac surgery	53/70 (76)	65/94 (69)
Vascular surgery	9/70 (13)	10/94 (11)
Digestive surgery	5/70 (7)	14/94 (15
Thoracic surgery	2/70 (3)	4/94 (4)
Urology	1/70 (1)	1/94 (1)
Location of study		
North America	43/70 (61)	13/94 (14)
Western Europe	23/70 (33)	48/94 (51)
North America and Western Europe	2/70 (3)	9/94 (9)
Other	2/70 (3)	24/94 (26)
No. centers		
Single	27/70 (39)	33/94 (35)
Multiple	38/70 (54)	56/94 (60)
Not reported	5/70 (7)	5/94 (5)
No. surgeons		
Single	4/70 (6)	2/94 (2)
Multiple	45/70 (64)	43/94 (45)
Not reported	2/701 (30)	49/94 (53)
Same period of recruitment for intervention and control groups	57/70 (81)	94/94 (100)
Type of study		
Prospective	7/70 (10)	94/94 (100)
Administrative database	63/70 (90)	0/94
Median date of period of recruitment, month-years (Q1,Q3)	11–2002 (12–1999,08–2005)	10–2003 (12–2000,10–2005)
Median time of recruitment, month (Q1,Q3)	60 (35, 86)	39 (27, 61)
Median of average age, y (Q1,Q3)	65.5 (62.5, 68.25)	63.5 (60.50, 67.50)
Median of average of male ratio, % (Q1,Q3)	73 (59,81)	74 (64,84)
Median no. participants, *n* (Q1,Q3)	2049 (828, 6,461)	179 (99, 341)
Sample size calculation	2/70 (3)	43/94 (46)
Risk of bias (low risk of bias)		
Selection bias (allocation concealment)	—	47/94 (50)
Performance bias	6/70 (9)	4/94 (3)
Detection bias for all-cause mortality outcomes for subjective outcomes	56/56 (100) 2/50 (4)	60/60 (100) 25/65 (38)
Attrition bias	66/70 (92)	85/94 (91)

Data are number (%) unless otherwise indicated.
Q1 indicates quartile 1; Q3, quartile 3.PRISMA Checklist

details, but, when reported, all studies were considered to have low risk of bias because the rate of contamination was less than 10% and the analysis was intention to treat. For detection bias, 3% of reports mentioned a blinded assessment of the subjective outcome. Finally, 92% of reports had a low risk of attrition bias.

Treatment Effect Estimates Between NRSs With PS Analysis and RCTs

The comparison between ORs for NRSs with PS analysis and RCTs for each clinical question is in Figure 2. For all-cause mortality, 9 clinical questions (40%) yielded a larger treatment effect for non-randomized studies with PS analysis and 13 (60%) a larger treatment effect for RCTs, and 1 clinical question (off-pump vs on-pump coronary artery bypass grafting) in 22 showed differences in treatment effects beyond chance, with more favorable results for NRSs with PS analysis. For subjective outcomes, 11 clinical questions (45%) yielded a larger treatment effect for NRSs with PS analysis and 13 (55%) a larger treatment effect for RCTs, and 1 clinical question (laparoscopic vs open gastric bypass) in 26 showed differences in treatment effects beyond chance, with more favorable results for NRSs with PS analysis.

The meta-epidemiological analysis is in Figure 3. For all-cause mortality, the combined ROR was 0.83 [(95% CI: 0.65–1.04), $P = 0.11$, $I^2 = 0\%$; variance between clinical questions = 0.00]. For this outcome, the clinical question comparing off-pump and on-pump coronary artery bypass grafting in the general population (clinical question code: OFF.01) had a relative weight of 46% in the meta-analysis. For subjective outcomes, the combined ROR was 1.07 [(0.87–1.33), $P = 0.52$, $I^2 = 8\%$; variance between clinical questions = 0.03].

We used a fixed-effect model to estimate summary ORs for NRSs with PS analysis and RCTs for each clinical question. For all-cause mortality, none of 22 clinical questions yielded a statistically significant difference (instead of 1 statistically significant difference with random-effects summary ORs). The combined ROR was 0.91 (95% CI: 0.74–1.12, $P = 0.37$). For subjective outcomes, the results were the same: 1 clinical question (laparoscopic vs open gastric bypass) in 26 showed differences in treatment effects beyond chance. The combined ROR was 1.08 (0.87–1.34, $P = 0.16$).

Sensitivity Analyses

Sensitivity analyses showed robust results. When restricting the analysis to NRSs with PS matching, the combined ROR was 0.84 (0.65–1.09) for all-cause mortality (Appendix 2, available at http://links.lww.com/SLA/A454). Results were similar after restricting the analysis to RCTs at low risk of bias for the different domains of the RoB tool (Appendix 3, available at http://links.lww.com/SLA/A455). Similarly, when restricting the analyses to NRSs with PS analysis that

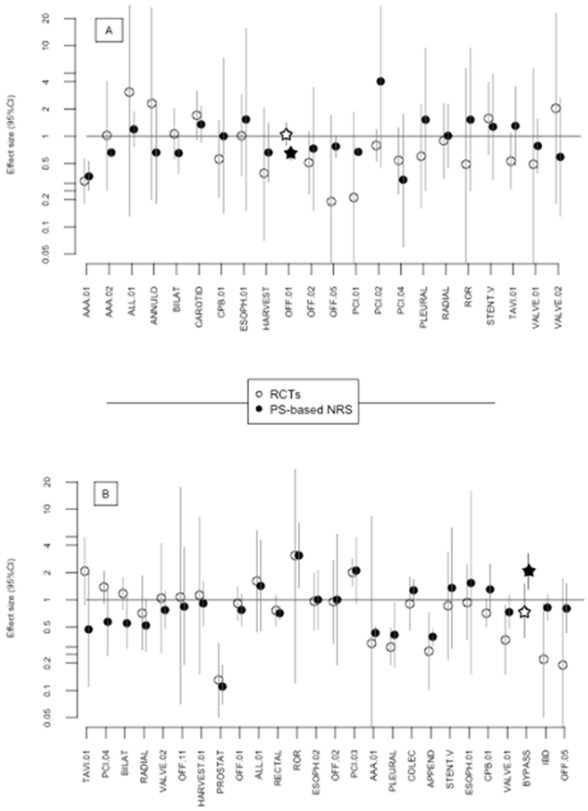


FIGURE 2. Summary ORs and 95% CIs for PS-based NRSs and RCTs for all-cause mortality (A) and subjective outcomes (B). Data with stars indicate the clinical questions for which the difference between PS-based NRS and RCT ORs was beyond chance.

recruited patients in the treatment and control groups simultaneously, the ROR was 0.87 (0.64–1.18) for all-cause mortality.

DISCUSSION

In this meta-epidemiological study comparing treatment effect estimates between NRSs with PS analysis and RCTs for a broad range of unselected clinical questions in surgery, we found no statistically significant difference in treatment effect between NRSs with PS adjustment and RCTs. However, there is an important variability of treatment effect estimates between the 2 study designs. For 45% of clinical questions, treatment effect estimates were larger for NRSs with PS analysis than for RCTs and conversely for 55% estimates were larger for RCTs than for NRSs with PS analysis.

Previous work compared treatment effect estimates between NRSs without PS analysis and RCTs.[17–21,34] These studies were diverse in terms of NRS design (prospective, retrospective studies) and outcomes considered (efficacy, safety). In a study of harms of medical interventions in RCTs and NRSs of more than 4000 subjects, the increase in relative risk and absolute risk differed more than twofold between RCTs and NRSs for more than half of the topics, with no clear predilection for RCTs or NRSs in estimating greater relative risk.[17] More recently, a meta-epidemiological study in the field of safety showed no difference between the 2 designs in risk estimates for an intervention but showed disagreement in findings between RCTs and NRSs in one-third of the selected topics.[34] However, these studies did not specifically focus on NRSs with PS analysis.

Recently, some researchers tried to explore this issue in meta-epidemiological studies.[23,24] Kuss et al compared treatment effect estimates from NRSs with PS analysis and RCTs for 10 outcomes of 1 specific clinical question (off-pump vs on-pump coronary artery bypass grafting in the general population).[24] Dahabreh et al compared the results of NRSs with PS analysis and RCTs for 17 topics covering various therapeutic interventions for acute coronary syndrome. The authors found substantial differences between the 2 study designs in effect estimates for some topics.[23]

Our results need to be interpreted with caution. In fact, we do not know the true treatment effects, and several important

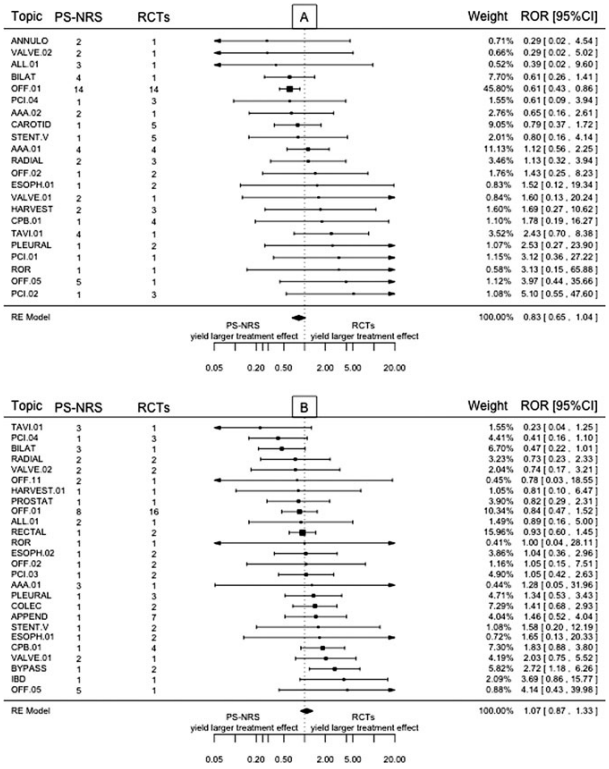


FIGURE 3. Forest plot of ratio of ORs and 95% CIs between PS-based NRSs and RCTs for all-cause mortality (A) and subjective outcomes (B). RE indicates random effect

factors could affect treatment effect estimates for both NRSs with PS analysis and RCTs. Reported results of RCTs with flaws in design, conduct, and analysis could be biased.[35,36] Similarly, the population, centers, surgeons, and standardization of the intervention can affect treatment effect estimates for both study designs. In our analysis, RCTs were heterogeneous in design (location of study, number of centers, methodology), sample size (20–4753), and quality of reporting. One third of RCTs involved single-center studies. Only 50% of RCT reports described an adequate method for sequence generation and allocation concealment and only 33% reported blinded evaluation of outcomes. NRSs with PS analysis were heterogeneous in terms of sample size (median 2049, range 161–322,892) and type of PS technique. In all, 90% were analyses of administrative databases. The use of such databases raises an important issue related to the availability of all relevant covariates included in the PS. In fact, administrative databases are not designed to answer a specific scientific question and important true confounders may not be systematically recorded. PS analysis could be biased if important prognostic factors are not accounted for in the PS model.[37,38] For example, tobacco use may not be recorded in administrative databases but is an essential prognostic factor that should be taken into account when studying the effect of lung cancer treatment.[39]

Our study has some limitations. First, we searched in only 1 literature database, with a restriction on language and without double selection process; nonetheless, our aim was not to be exhaustive but to provide a representative sample of NRSs with PS analysis indexed in MEDLINE, and we see no reason why NRSs with PS analysis and RCTs on the same topic should be indexed differently in MEDLINE and other databases. Second, we did not identify meta-analyses of NRSs with PS analysis and RCTs. We first searched for NRSs with PS analysis for practical reasons (ie, NRSs with PS analysis are rarely included in meta-analyses). This method has been used in other contexts.[19,21,23,24] To ensure that NRSs with PS analysis and RCTs answered the same clinical question, 2 reviewers with surgical expertise independently evaluated the homogeneity of studies, without discrepancies. Third, we assessed only binary outcomes and therefore, our results cannot be generalized to studies of continuous and censured outcomes.

This work has important implications for practice and research. In fact, evidence-based surgery is widely based on evidence

from NRSs and we need appropriate methods for limiting confounding bias. PS analysis could be an interesting technique for dealing with such bias. However, more research is needed to better understand discrepancies in treatment effect estimates between RCTs and NRSs.

CONCLUSIONS

There was no statistically significant difference in treatment effect between NRSs with PS analysis and RCTs. Prospective NRSs with suitable and careful PS analysis can be relied upon as evidence when RCTs are not possible.

ACKNOWLEDGMENTS

The contributions of all the authors were as follows: G.L. searched for and selected the trials, extracted and analyzed the data, interpreted the results, and drafted the manuscript. I.B. conceived the study, selected the trials, extracted and analyzed the data, interpreted the results, and drafted the manuscript. She is the guarantor. L.T. conceived the study, analyzed the data, interpreted the results, and drafted the manuscript. N.A. extracted the data and interpreted the results. F.A. extracted the data and interpreted the results. R.N. conceived the study, interpreted the results, and drafted the manuscript. P.R. conceived the study, interpreted the results, and drafted the manuscript. All authors, external and internal, had full access to all of the data (including statistical reports and tables) in the study, and they take responsibility for the integrity of the data and the accuracy of the data analysis.

REFERENCES

1. McCulloch P, Altman DG, Campbell WB, et al. No surgical innovation without evaluation: the IDEAL recommendations. *Lancet*. 2009;374:1105–1112.
2. Ergina PL, Cook JA, Blazeby JM, et al. Challenges in evaluating surgical innovation. *Lancet*. 2009;374:1097–1104.
3. Barkun JS, Aronson JK, Feldman LS, et al. Evaluation and stages of surgical innovations. *Lancet*. 2009;374:1089–1096.
4. Boutron I, Moher D, Altman DG, et al. Extending the CONSORT statement to randomized trials of nonpharmacologic treatment: explanation and elaboration. *Ann Intern Med*. 2008;148:295–309.
5. Commitee on Comparative Research Prioritization. *Institute of Medicine Initial National Priorities for Comparative Effectiveness Research*. Washington, DC: National Academies Press; 2009.
6. McCulloch P, Taylor I, Sasako M, et al. Randomised trials in surgery: problems and possible solutions. *BMJ*. 2002;324:1448–1451.
7. Patsopoulos NA. A pragmatic view on pragmatic trials. *Dialogues Clin Neurosci*. 2011;13:217–224.
8. Kunz LM, Yeh RW, Normand S-LT. Comparative effectiveness research: does one size fit all? *Stat Med*. 2012;31:3062–3065.
9. Hemmila MR, Birkmeyer NJ, Arbabi S, et al. Introduction to propensity scores: a case study on the comparative effectiveness of laparoscopic vs open appendectomy. *Arch Surg*. 2010;145:939–945.
10. Cook EF, Goldman L. Performance of tests of significance based on stratification by a multivariate confounder score or by a propensity score. *J Clin Epidemiol*. 1989;42:317–324.
11. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70:41–55.
12. Cepeda MS, Boston R, Farrar JT, et al. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am J Epidemiol*. 2003;158:280–287.
13. Kuss O, von Salviati B, Börgermann J. Off-pump versus on-pump coronary artery bypass grafting: a systematic review and meta-analysis of propensity score analyses. *J Thorac Cardiovasc Surg*. 2010;140:829–835.
14. Austin PC. Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: a systematic review and suggestions for improvement. *J Thorac Cardiovasc Surg*. 2007;134:1128–1135.
15. Cao C, Manganas C, Ang SC, et al. Video-assisted thoracic surgery versus open thoracotomy for non-small cell lung cancer: a meta-analysis of propensity score-matched patients. *Interact Cardiovasc Thorac Surg*. 2013;16:244–249.
16. Tleyjeh IM, Kashour T, Zimmerman V, et al. The role of valve surgery in infective endocarditis management: a systematic review of observational studies that included propensity score analysis. *Am Heart J*. 2008;156:901–909.
17. Ioannidis JPA, Haidich A-B, Pappa M, et al. Comparison of evidence of treatment effects in randomized and nonrandomized studies. *JAMA*. 2001;286:821–830.
18. Concato J, Lawler EV, Lew RA, et al. Observational methods in comparative effectiveness research. *Am J Med*. 2010;123:e16–e23.
19. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *N Engl J Med*. 2000;342:1878–1886.
20. Shikata S, Nakayama T, Noguchi Y, et al. Comparison of effects in randomized controlled trials with observational studies in digestive surgery. *Ann Surg*. 2006;244:668–676.
21. Papanikolaou PN, Christidi GD, Ioannidis JPA. Comparison of evidence on harms of medical interventions in randomized and nonrandomized studies. *CMAJ*. 2006;174:635–641.
22. Deeks JJ, Dinnes J, D'Amico R, et al. Evaluating non-randomised intervention studies. *Health Technol Assess*. 2003;7:iii–x, 1–173.
23. Dahabreh IJ, Sheldrick RC, Paulus JK, et al. Do observational studies using propensity score methods agree with randomized trials? A systematic comparison of studies on acute coronary syndromes. *Eur Heart J*. 2012;33:1893–1901.
24. Kuss O, Legler T, Börgermann J. Treatments effects from randomized trials and propensity score analyses were similar in similar populations in an example from cardiac surgery. *J Clin Epidemiol*. 2011;64:1076–1084.
25. Patsopoulos NA, Ioannidis JP. The use of older studies in meta-analyses of medical interventions: a survey. *Open Med*. 2009;3:e62–e68.
26. Higgins J, Green S. *Cochrane Handbook for Systematic Reviews of Interventions, Version 5.1.0*. Oxford, England: The Cochrane Collaboration; 2011.
27. Wood L, Egger M, Gluud LL, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. *BMJ*. 2008;336:601–605.
28. Savović J, Jones HE, Altman DG, et al. Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. *Ann Intern Med*. 2012;157:429–438.
29. Di Pietrantonj C. Four-fold table cell frequencies imputation in meta analysis. *Stat Med*. 2006;25:2299–2322.
30. Sweeting MJ, Sutton AJ, Lambert PC. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. *Stat Med*. 2004;23:1351–1375.
31. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials*. 1986;7:177–188.
32. Sterne JAC, Jüni P, Schulz KF, et al. Statistical methods for assessing the influence of study characteristics on treatment effects in "meta-epidemiological" research. *Stat Med*. 2002;21:1513–1524.
33. Altman DG, Bland JM. Interaction revisited: the difference between two estimates. *BMJ*. 2003;326:219.
34. Golder S, Loke YK, Bland M. Meta-analyses of adverse effects data derived from randomised controlled trials as compared to observational studies: methodological overview. *PLoS Med*. 2011;8:e1001026.
35. Gluud LL. Bias in clinical intervention research. *Am J Epidemiol*. 2006;163:493–501.
36. Dechartres A, Boutron I, Trinquart L, et al. Single-center trials show larger treatment effects than multicenter trials: evidence from a meta-epidemiologic study. *Ann Intern Med*. 2011;155:39–51.
37. 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;26:734–753.
38. Brookhart MA, Schneeweiss S, Rothman KJ, et al. Variable selection for propensity score models. *Am J Epidemiol*. 2006;163:1149–1156.
39. Rubin DB. Estimating causal effects from large data sets using propensity scores. *Ann Intern Med*. 1997;127:757–763.

2.7 Réponse de la lettre à l'éditeur publiée

Equivalence Approach Is More Appropriate for Comparison of Treatment Effect Estimates

To the Editor:

We read with interest the study of Lonjon and colleagues[1] in which effects estimates from propensity score–based nonrandomized studies are compared with those from randomized controlled trials (RCTs) of the same surgical procedure.

As the authors discuss, there are many situations where an RCT is unfeasible and evidence of effect must be sought from observational studies. Although propensity scoring can reduce the effect of observed variables, it does not account for unobserved variables, which are presumed to be equally distributed between the treatment and control groups in RCTs. It is important, therefore, to know how RCTs and propensity score–based analyses compare.

The authors conclude that there was "no statistically significant difference in treatment effect between nonrandomized studies with propensity score analysis and RCTs." Although this is true, the authors have performed a set of significance tests for superiority, when an assessment of equivalence would have been more appropriate. It is long recognized that absence of evidence is not evidence of absence.[2]

What is important is that the difference in effect size (odds ratios) between the 2 trial designs falls within a range that can be considered equivalent. This is a value judgment. A generous allowance may be a ratio of odds ratios (RORs) of 0.5 to 2.0. As is seen in Figure 1, of "clinical question" groups reporting all-cause mortality, none achieved this and only one reporting subjective outcomes did.

However, the authors should highlight that the summary measure for RORs falls within a symmetrical equivalence margin of ROR = 0.7–1.5 and ROR = 0.8–1.3 for all-cause and subjective outcomes, respectively. It could be concluded that although equivalence was demonstrated only in 1 of 48 individual clinical question outcomes, it may be reasonable to suggest that within the limitations of this study, equivalence in outcomes exists.

Disclosure: The authors declare no conflicts of interest.

ISSN: 0003-4932/14/26202-e0067
DOI: 10.1097/SLA.0000000000000598

Ewen M. Harrison, FRCS
Clinical Surgery
Royal Infirmary of Edinburgh
University of Edinburgh
Edinburgh, United Kingdom

Aneel Bhangu, MB ChB, MRCS
Department of Colorectal Surgery
Royal Marsden Hospital
London, United Kingdom

Olivia Swann, MB ChB, MRes
Centre for Immunity, Infection and Evolution
University of Edinburgh
Edinburgh, United Kingdom

Stephen J. Wigmore, FRCS
Clinical Surgery
Royal Infirmary of Edinburgh
University of Edinburgh
Edinburgh, United Kingdom
ewen.harrison@ed.ac.uk

REFERENCES

1. Lonjon G, Boutron I, Trinquart L, et al. Comparison of treatment effect estimates from prospective nonrandomized studies with propensity score analysis and randomized controlled trials of surgical procedures. *Ann Surg*. 2014;259:18–25.
2. Altman DG, Bland JM. Statistics notes: absence of evidence is not evidence of absence. *BMJ*. 1995;311:485.

Reply:

We thank Harrison and colleagues for their comment regarding the comparison of effects estimates from propensity score–based nonrandomized studies and randomized controlled trials of the same surgical procedure.[1] The authors proposed to use an equivalence approach for interpreting the results of meta-epidemiological studies. To our knowledge, such approach has never been used to analyze results of meta-epidemiological studies.[2]

Choosing a margin of equivalence is challenging. In clinical trials, the margin must be smaller than or equal to "the smallest value that would represent a clinically meaningful difference, or the largest value that would represent a clinically meaningless difference."[3] The CONSORT statement for noninferiority and equivalence randomized trials mentions that the margin of noninferiority should be prespecified and justified on clinical grounds.[4] However, in meta-epidemiological studies, clinical grounds do not exist and the margins proposed by Harrison and colleagues defined a posteriori are questionable. The authors considered that the smallest and largest values that would represent a meaningful difference would be a ratio of odds ratio of 0.5 and 2.0, respectively. Therefore, an odds ratio of 1 (no difference between the experimental and control interventions in the odds of death) would be considered similar to an odds ratio of 0.5 or 2 (a 2-fold increase or decrease in the odds of death), which is obviously not acceptable. If the authors chose a less generous and more acceptable margin (eg, a margin of 0.9–1.1), we would not be able to conclude equivalence between the 2 designs.[4]

In conclusion, use of an equivalence approach in meta-epidemiological studies is a very interesting approach, particularly when comparing 2 different types of designs. However, such analysis should be prespecified and more research is needed to determine the most appropriate margin and avoid misleading conclusions.

Disclosure: The authors declare no conflicts of interest.

DOI: 10.1097/SLA.0000000000000597

Guillaume Lonjon, MD
Unité INSERM U738
Paris, France
Assistance Publique-Hôpitaux de Paris
Hôpital Lariboisière
Service d'orthopédie et traumatologie
Paris, France

Isabelle Boutron MD, PhD
Ludovic Trinquart, MSc
Unité INSERM U738
Paris, France
Assistance Publique-Hôpitaux de Paris
Hôpital Hôtel-Dieu
Centre d'Epidémiologie Clinique
Paris, France
Université Paris Descartes-Sorbonne Paris Cité
Paris, France
French Cochrane Centre
Paris, France

Nizar Ahmad, MD, PhD
Unité INSERM U738
Paris, France
Assistance Publique-Hôpitaux de Paris
Hôpital Hôtel-Dieu
Centre d'Epidémiologie Clinique
Paris, France

Florence Aim, MD
Unité INSERM U738
Paris, France
Assistance Publique-Hôpitaux de Paris
Hôpital Lariboisière
Service d'orthopédie et traumatologie
Paris, France

Rémy Nizard, MD, PhD
Assistance Publique-Hôpitaux de Paris
Hôpital Lariboisière
Service d'orthopédie et traumatologie
Paris, France

Philippe Ravaud, MD, PhD
Unité INSERM U738
Paris, France

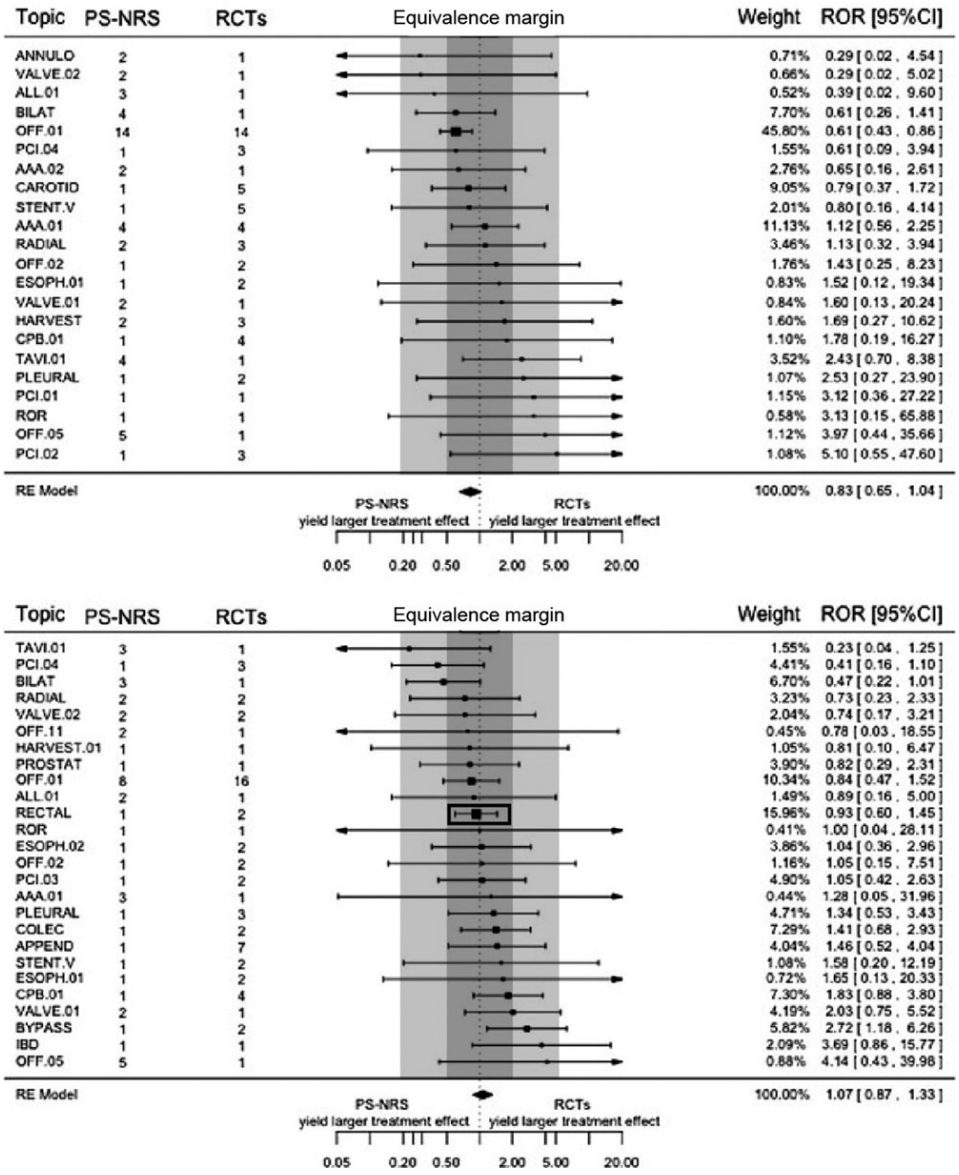


FIGURE 1. Dark and light gray boxes represent equivalence margins of ROR = 0.5–2.0 and ROR = 0.2–5.0, respectively. The clinical question where equivalence for ROR = 0.5–2.0 is demonstrated is marked with a box. CI indicates confidence interval; NRS, nonrandomized studies; PS propensity score. Modified from Lonjon et al.[1]

Assistance Publique-Hôpitaux de Paris
Hôpital Hôtel-Dieu
Centre d'Epidémiologie Clinique
Paris, France
Université Paris Descartes-Sorbonne Paris Cité
Paris, France
French Cochrane Centre
Paris, France
Mailman School of Public Health
Columbia University
New York, NY
dr.guillaume.lonjon@gmail.com

REFERENCES

1. Lonjon G, Boutron I, Trinquart L, et al. Comparison of treatment effect estimates from prospective nonrandomized studies with propensity score analysis and randomized controlled trials of surgical procedures. *Ann Surg*. 2014;259:18–25.
2. Savović J, Jones HE, Altman DG, et al. Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. *Ann Intern Med*. 2012;157:429–438.
3. Le Henanff A, Giraudeau B, Baron G, et al. Quality of reporting of noninferiority and equivalence randomized trials. *JAMA*. 2006;295:1147–1151.
4. Piaggio G, Elbourne DR, Pocock SJ, et al. Reporting of noninferiority and equivalence randomized trials: extension of the CONSORT 2010 statement. *JAMA*. 2012;308:2594–2604.

I Introduction et objectifs

II Article 1: Comparaison de la mesure de l'effet traitements entre les études observationnelles utilisant un score de propension et les ECR évaluant une procédure chirurgicale

III Article 2 : Description des Etudes observationnelles utilisant une analyse avec score de propension évaluant une procédure chirurgicale. Revue systématique méthodologique

IV Discussion et perspectives

V Conclusion

VI Réferences

III Description des études observationnelles utilisant une analyse avec score de propension évaluant une procédure chirurgicale. Revue systématique méthodologique.

Un résumé de la méthodologie et des principaux résultats est présenté ci-dessous. La description détaillée de la méthodologie et des résultats de ce travail, sont présentés dans l'article publié en anglais (122) (section 3.6).

3.1 Introduction.

Compte tenu de l'intérêt des études observationnelles avec SP pour l'évaluation des interventions chirurgicales, nous avons réaliser une revue systématique méthodologique pour

1) décrire l'évolution dans le temps de l'utilisation des études observationnelles utilisant un SP pour évaluer une procédure chirurgicale

2) évaluer sur un échantillon restreint d'études observationnelles utilisant un SP, la qualité des rapports et les biais potentiels.

3.2 Matériel et méthodes.

3.2.1. Type d'étude

Nous avons réalisé une revue systématique méthodologique.

3.2.2. Sélection des articles

Pour décrire l'évolution dans le temps de l'utilisation des études observationnelles utilisant un SP pour évaluer une procédure chirurgicale, nous avons identifié, à partir d'une recherche la plus large possible sur PUBMED, tous les articles rapportant une étude observationnelle utilisant une analyse avec SP et évaluant une procédure chirurgicale (pas de limite dans le temps, date de la recherche : fin juillet 2014).

Pour analyser les méthodes utilisées et la présentation de ces méthodes et des résultats, nous avons identifié un échantillon de publications récentes (1 Août 2013 au 31 juillet 2014).

3.2.3. Extraction de données

Nous avons collecté les données suivantes sur

- les caractéristiques générales de l'étude (année, type de comparaison, nombre de centre, présence d'un statisticien dans les auteurs, caractère retro ou prospectif, etc.).
- les caractéristiques méthodologique de l'étude (présence d'un statisticien dans les auteurs, caractère retro ou prospectif, l'identification ou non d'un critère de jugement principal)
- les caractéristiques évoquant un biais (données manquantes, mauvaise classification de l'intervention par conversion, analyse en intention de traiter)
- La méthode de construction du SP (types de variables incluses dans le score, justification du choix de ces variables, etc.) et la méthode d'utilisation du SP (appariement, ajustement, stratification pondération inverse)
- Pour les études avec SP utilisant la méthode d'appariement, nous avons collecté la présence ou l'absence des 4 informations techniques les plus

importantes (ratio, remise, choix de la paire, type d'analyse statistiques) permettent de critiquer la méthodologie (79,123). La représentativité de la population d'analyse par rapport à la population d'origine, ainsi que la méthode d'évaluation de la méthode d'appariement ont aussi été analysées.

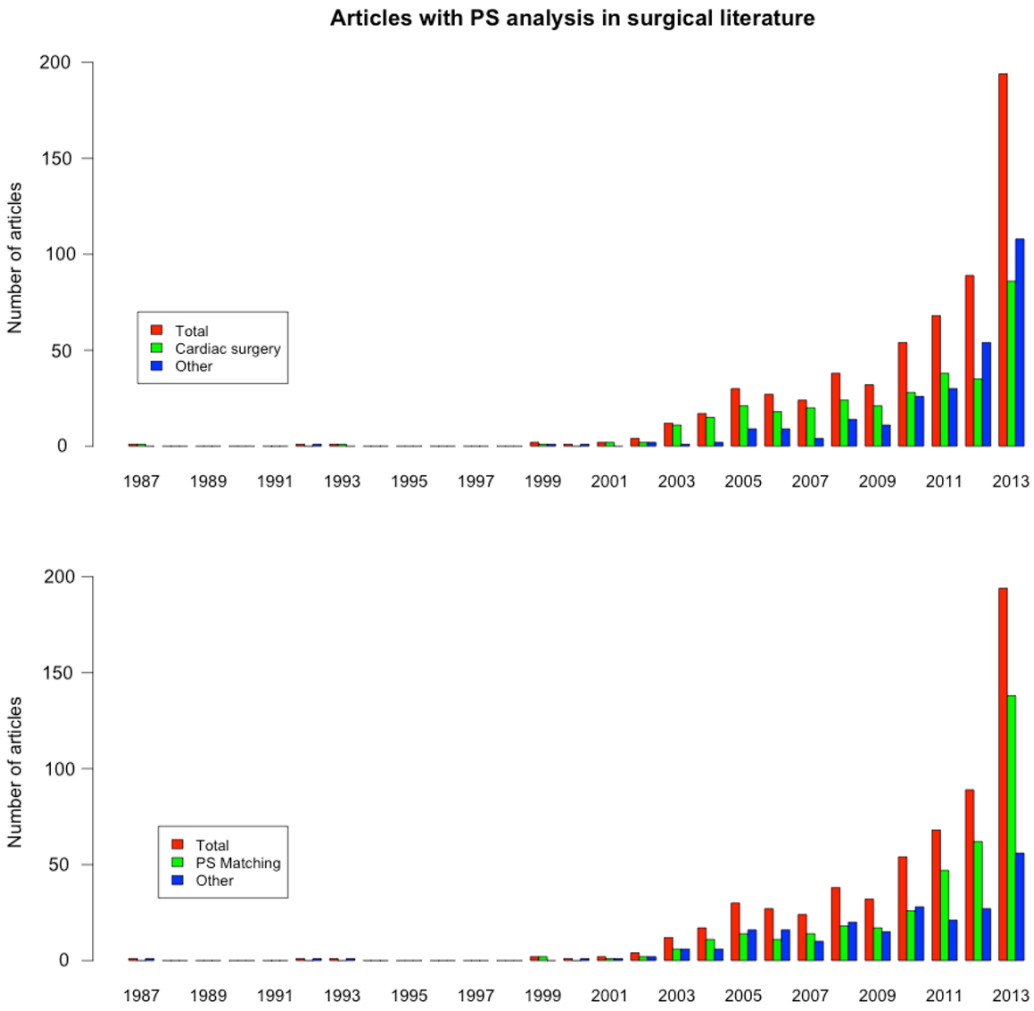
3.3 Résultats

3.3.1. Evolution du nombre de publications d'études observationnelles avec S en fonction de la spécialité chirurgicale et du type de procédure.

Parmi les 2436 citations retrouvées, 652 ont été sélectionnées. La figure 11 représente l'évolution globale du nombre d'articles publiés entre 1980 et 2014.

L'augmentation du nombre d'articles au cours des dernières années est importante. Sur la partie haute de la figure 11 est rapportée la répartition entre les spécialités (chirurgie cardiaque/ autres spécialités). Initialement la spécialité de chirurgie cardiaque était largement prépondérante, mais une répartition plus homogène entre les spécialités ces dernières années est constatée. Sur la partie basse de la figure 11 est rapportée la répartition en fonction du type de méthode d'utilisation du SP (appariement / autres méthodes). La méthode d'appariement est de plus en plus fréquente.

Figure 11 : Représentation graphique du nombre de publication des études observationnelles avec SP, en fonction de la spécialité chirurgicale (chirurgie cardiaque/ autres) et de la méthode utilisée (appariement /autres)



3.3.2. Description des méthodes utilisées

Parmi les 129 articles référencés dans la dernière année, 83% était des analyses rétrospectives de bases de données. Le nombre moyen de participants par étude était de 904 (Q1-Q3: 348–4133, min-max: 25-407070). Un statisticien ou un service de statistique était cité dans 53% des cas.

Quatre-vingt dix-sept articles décrivaient une méthode d'appariement, 19 une méthode d'ajustement, 7 une méthode de pondération inverse et 6 une méthode de stratification.

Pour construire le score, 20% des articles ne précisaient pas les variables incluses dans le score et 77% ne donnaient pas de justification pour le choix des variables. Neuf pourcent des articles, rapportaient des données manquantes et 14% des conversions.

Pour l'analyse des articles décrivant une méthode d'appariement (n=97), 10% des articles rapportaient les 4 points clés de la construction d'une méthode d'appariement. Pour la représentativité de la population finale, aucun article ne représentait dans un même tableau les 2 populations ou ne comparait les caractéristiques de ces 2 populations. La comparabilité des groupes « traité » et « contrôle » était étudiée dans 86 % des cas, et dans 20% des cas avec une méthode recommandée.

3.4 Discussion

Dans cette étude nous avons montré qu'il y a une augmentation importante du nombre de publication des études observationnelles utilisant un SP. Cependant, les méthodes utilisées sont variées et insuffisamment décrites.

Nos résultats sont cohérents avec ceux de la littérature (79,82,83,124). Dans la revue systématique publiée en 2007, Austin évoquait déjà les problèmes d'analyse de la comparabilité des groupes par des méthodes non recommandées. De la même manière Gayat et al, dans une analyse de la littérature en anesthésie et réanimation, retrouvait des lacunes dans le rapport de ce type d'étude. Plus récemment, une étude spécifique sur les variables incluses dans le score, trouvait que 65% des articles ne rapportaient pas de justification de la sélection des variables (124). Ces chiffres associés aux nôtres (77%) soulèvent une des limites de l'analyse avec SP, à savoir le choix des variables à inclure dans le score.

A l'issue de ce travail, nous avons proposé des recommandations pour améliorer la planification, la réalisation, l'analyse et l'écriture des articles rapportant des études observationnelles avec SP. Ces recommandations mettent en avant que

1) l'aide d'un statisticien paraît indispensable pour guider le clinicien dans les différents choix qu'il à faire.
2) un plan d'analyse statistique doit être au préalable décidé et écrit dans un protocole. Ce plan d'analyse devra préciser les variables qui devront être utilisées pour le calcul du SP, et justifier le choix de ces variables.
3) Dans l'article, l'utilisation d'un SP dans l'analyse statistique doit être énoncée dans le titre. La méthodologie de la construction du SP et de la méthode utilisée, doit être rapportée. Les caractéristiques de la population de base et de la population d'analyse doivent être rapportées. La comparaison des groupes « traité » et « non traité » doit être présentée pour chaque variable en utilisant une analyse de la différence de moyenne standardisée.

3.5 Conclusion

Les études observationnelles utilisant un score de propension pour évaluer une procédure chirurgicale sont de plus en plus nombreuses, mais l'utilisation de cette méthode et la présentation des articles utilisant cette méthode dans les études observationnelles évaluant des interventions chirurgicales n'est pas approprié et doit être amélioré.

3.6 Article publié

Potential Pitfalls of Reporting and Bias in Observational Studies With Propensity Score Analysis Assessing a Surgical Procedure

A Methodological Systematic Review

Guillaume Lonjon, MD,†‡ Raphael Porcher, PhD,*‡§ Patrick Ergina, MD,¶ Mathilde Fouet, MD,*‡ and Isabelle Boutron, PhD*‡§*

Objective: To describe the evolution of the use and reporting of propensity score (PS) analysis in observational studies assessing a surgical procedure.

Background: Assessing surgery in randomized controlled trials raises several challenges. Observational studies with PS analysis are a robust alternative for comparative effectiveness research.

Methods: In this methodological systematic review, we identified all PubMed reports of observational studies with PS analysis that evaluated a surgical procedure and described the evolution of their use over time. Then, we selected a sample of articles published from August 2013 to July 2014 and systematically appraised the quality of reporting and potential bias of the PS analysis used.

Results: We selected 652 reports of observational studies with PS analysis. The publications increased over time, from 1 report in 1987 to 198 in 2013. Among the 129 reports assessed, 20% (n = 24) did not detail the covariates included in the PS and 77% (n = 100) did not report a justification for including these covariates in the PS. The rate of missing data for potential covariates was reported in 9% of articles. When a crossover by conversion was possible, only 14% of reports (n = 12) mentioned this issue. For matched analysis, 10% of articles reported all 4 key elements that allow for reproducibility of a PS-matched analysis (matching ratio, method to choose the nearest neighbors, replacement and method for statistical analysis).

Conclusions: Observational studies with PS analysis in surgery are increasing in frequency, but specific methodological issues and weaknesses in reporting exist.

Keywords: bias, observational study, propensity score, surgery, systematic review

(*Ann Surg* 2016;xx:xxx–xxx)

Comparative effectiveness research is designed to inform health care decisions by providing evidence of the effectiveness, benefits, and harms of different treatment options.[1] Comparative effectiveness research encompasses a broad array of study types, including randomized trials and observational studies.

From the *INSERM, UMR 1153, Centre of Research in Epidemiology and Statistics Sorbonne Paris Cité (CRESS), METHODS Team, Paris, France; †Orthopaedic Department, Assistance Publique-Hôpitaux de Paris, European Hospital Georges Pompidou, Paris, France; ‡Medical school, Paris Descartes university, Sorbonne Paris cité, Paris, France; §Centre of research in epidemiology and statistics Sorbonne Paris Cité, Hopital Hotel dieu, place du Parvis notre dame, Paris, France; and ¶Department of Surgery, McGill University Health Centre, Montreal, Canada.

Reprints: Guillaume Lonjon, MD, INSERM, UMR 1153, Centre of Research in Epidemiology and Statistics Sorbonne Paris Cité (CRESS), METHODS team, Paris, France. E-mail: dr.guillaume.lonjon@gmail.com.

Disclosure: The authors report no conflicts of interest.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.annalsofsurgery.com).

ISSN: 0003-4932/14/26105-0821

DOI: 10.1097/SLA.0000000000001797

Randomized controlled trials (RCTs) are considered the gold standard for therapeutic evaluation.[2,3] These studies test the effect of an intervention under carefully controlled conditions and limit the risk of bias. However, assessing surgical procedures in RCTs raises several methodological and practical challenges for the surgical research community.[4,5] Many surgeons view RCTs to be too difficult and impractical to undertake and at worst, irrelevant to their practice because of concerns about generalizability.[6] Surgical procedures are often evaluated in observational studies.[7] However, bias, particularly confounding bias, can lead to overestimated treatment effects with this kind of study.[8]

Methods such as propensity score (PS) analysis have been proposed to deal with confounding bias in observational studies. Such analysis aims to mimic randomization. The PS is defined for each participant as the probability of receiving the treatment, given baseline covariates. With the assumption of no unmeasured confounders, a treated and an untreated patient with the same PS can be considered as if they had been randomly assigned to each group.[9]

Previous findings from simulated data showed that PS analysis could produce estimates that were less biased, more robust, and more precise than with multivariable analysis.[10,11] A recent meta-epidemiologic study comparing treatment effect estimates from observational studies with PS analysis and from RCTs for the same clinical question in the field of surgery revealed no statistically significant difference.[12] However, the use and reporting of PS analysis for studies may not be appropriate[13,14] and specific bias could occur.[15,16]

In this project, we aimed to (1) describe the evolution of the use of PS analysis in observational studies assessing surgical procedures and (2) assess the reporting and potential bias of PS analysis used in a sample of recently published observational studies assessing surgical procedures.

METHODS

We performed a methodological systematic Review.

In a first step, we identified all PubMed reports of observational studies using PS analysis to evaluate a surgical procedure and described the evolution of such studies over time by surgical area and type of PS analysis used. In a second step, we selected a sample of recently published reports of observational studies using PS analysis to evaluate a surgical procedure (ie, published from August 2013 to July 2014) and systematically appraised the quality of reporting and potential bias of PS analysis used.

Evolution of the Use of PS Analysis in Observational Studies Assessing Surgical Procedures

We systematically searched MEDLINE via PubMed to identify all English-language reports of observational studies using PS analysis to evaluate a surgical procedure (search date July 2014) with the following search strategy: ((("Surgical Procedures, Operative" [Mesh])) AND ((Propensity) OR ("Propensity Score"[Mesh])))

TABLE 1. Definition and Explanation of the 4 Principal Types of PS Analysis[22]

Matching: PS-matching analysis is the most common type of PS analysis, probably because it is easy to understand and reliable. PS matching involves creating 2 groups of participants, one receiving the treatment of interest (treated subjects) and the other not (untreated subjects), by matching participants with identical or close PSs. The PS-matching analysis can then approximate that of a randomized trial by directly comparing outcomes between the 2 groups, using methods that account for the paired nature of the data. In some studies, PS matching has been shown to be more reliable than other types of PS analysis, such as stratification or adjustment. To construct a matching analysis, several features have to be decided:

Matching ratio: The matching balance can be ''one-to-one'' (1:1) (ie, pairs of treated and untreated subjects with a similar PS), many-to-one (2:1, 3:1, 4:1, etc), or many-to-many (3:2, 5:3, etc). An unfixed ratio of many-to-many matching is also possible. Asymmetric ratios allow for maximizing the number of participants.

Use of replacement: In matching without replacement, once an untreated subject is selected to be matched to a treated subject, he/she is no longer available as a potential match for subsequent treated subjects. In matching with replacement, this subject can subsequently be matched to other treated subjects.

Matching algorithm: Many algorithms have been proposed to form matched pairs without replacement. They can be divided generally into 2 types: ''greedy'' matching and optimal matching. Greedy matching consists of selecting a random treated subject and then matching it to the first nearest untreated subject. This untreated subject is selected even if the subject would be better matched with a subsequent treated subject. Optimal matching consists of forming pairs of treated and untreated subjects to minimize the total within-pair differences in the PS.

Acceptable match choice: In the most basic version of PS-matching analysis, only the treated patient whose PS does not differ from a prespecified value (caliper distance) is considered a potential match for the subject. The most frequently used value is 0.2 SD of the logit of the PS. Treated and untreated subjects that remain unmatched are then discarded.

Stratification: This technique involves separating participants into distinct groups based on their PSs (ie, strata). Five strata are commonly used, although increasing the number can reduce bias. The treatment effect could be estimated within each stratum or could be pooled across strata to provide a global treatment effect estimate.

Adjustment: For this approach, a multivariable model is developed whereby the study outcome serves as the dependent variable and the treatment group and PS as predictor variables. This approach allows the investigator to estimate the outcome associated with the treatment of interest while adjusting for the probability of receiving that treatment.

Inverse probability of treatment weighting: In this technique, PSs are used to calculate statistical weights for each individual so as to create a sample in which the distribution of potential confounding factors is independent of exposure, thereby allowing for an unbiased estimate of the relationship between treatment and outcome. In practice, the IPTW technique weights each treated subject by the inverse of the PS and each control subject by the inverse of ''1 minus'' the PS, or the probability of not being treated.

IPTW indicates inverse probability of treatment weighting.

AND ((((versus[tiab] OR comparison[tiab]) OR compared[tiab]) OR ''Comparative Study''[All Fields]) OR ''Comparative Study''[Publication Type]). Two of us (GL, MF) screened the retrieved titles, abstracts, and full texts to identify eligible studies. Any discrepancies were discussed between the 2 reviewers. Prespecified eligibility criteria were reports of observational studies using PS analysis and assessing surgical procedures in humans.

A surgical procedure was defined as an act performed by a surgeon that involves physically changing body tissues and organs through direct or indirect manual manipulation such as cutting, suturing, abrading, or the use of lasers.[17] Interventional procedures such as percutaneous coronary or endoscopy were excluded. The surgical procedure could be compared with any comparator (surgical procedure, interventional procedures, usual care, or drugs). For all of these reports, we recorded the year of publication, the surgical area, and the type of PS analysis used (matching, adjustment, stratification, or weighting). Details and specificity of each type of PS analysis are reported in Table 1.

Reporting and Potential Bias of PS Analysis Used in a Recently Published Sample

From the reports selected in the first step, we selected those published between August 1, 2013 and July 31, 2014 and extracted specific information. A standard data-collection form (Appendix 1, http://links.lww.com/SLA/B29) was developed and preliminarily tested by 4 reviewers (GL, IB, RP, MF) with 10 articles. One reviewer extracted all data (GL). A second reviewer (MF) independently duplicated data extraction for 30% of the selected reports. Discrepancies were discussed to achieve consensus. Interobserver reproducibility was calculated with the Cohen kappa coefficient (K). From 35 categorical variables (2–12 categories) extracted, the median K was 0.87 (quartile 1–3: 0.66–1).

General Characteristics of the Study

We recorded the type of comparator (surgery, nonsurgical procedure, usual care/no treatment, drugs), number of participating centers and participating care providers, study design, and sample size. We also assessed whether a primary outcome was clearly identified by the authors in the methods section, whether a description of the interventions was described (ie, at least some details of the intervention reported) and whether the period of recruitment was similar for the 2 groups.

Key Methodological Components of PS Analysis

The following methodological components were assessed.

First, we determined whether a statistician or epidemiologist (ie, affiliated with a statistics, epidemiology, or public health department) was listed among authors, cited in the text, or cited in the acknowledgements section. In fact, appropriate use of a PS analysis supposes a high level of statistical expertise and would need specific support.

Second, we recorded the type of PS analysis used and whether the PS analysis was considered the main analysis. In fact, there are 4 different ways of using PS analysis: PS-matching, stratification on the PS, covariate adjustment with the PS, and inverse probability of treatment weighting with the PS. Details and specificities of each type of PS analysis are reported in Table 1.

Third, we recorded whether and how authors reported the variables used to construct the PS. We also recorded how these variables were identified (ie, by preexisting knowledge or data driven via stepwise variable selection algorithms, for instance). For multicenter studies, we recorded whether the center was included as a covariate in the PS. In fact, one crucial aspect of PS analysis is the identification of variables used to construct the PS. Omitting a true confounder variable (ie, a variable associated both with the treatment

and the outcome) or a prognostic variable in the PS may produce substantially biased estimates.[18] Consequently, this information must be reported to allow readers to detect inappropriate analysis by omission of important confounders. We also recorded whether a prespecified protocol of the PS analysis was available at a trial registry, for example.

Fourth, we determined whether authors reported some information about missing data and how they handled these data. We particularly checked whether missing data were reported for the baseline characteristics, which are used to construct the PS. In observational studies, missing data are common,[19] and calculating PSs including variables with a high rate of missing data but not the methods to handle them could lead to biased results.[20,21]

Fifth, we assessed the accounting for crossover by surgical conversion. Observational studies with PS analysis aim to mimic RCTs in which the analysis should be conducted according to the intent-to-treat (ITT) principle, a strategy for analyzing data in which all participants are included in the group to which they were assigned, whether or not they received the intervention allocated. ITT analysis avoids overoptimistic estimates of the efficacy of an intervention.[22,23] Such analysis is particularly important in surgery when a conversion from one intervention to another is possible. Consequently, we systematically determined whether a per-operative

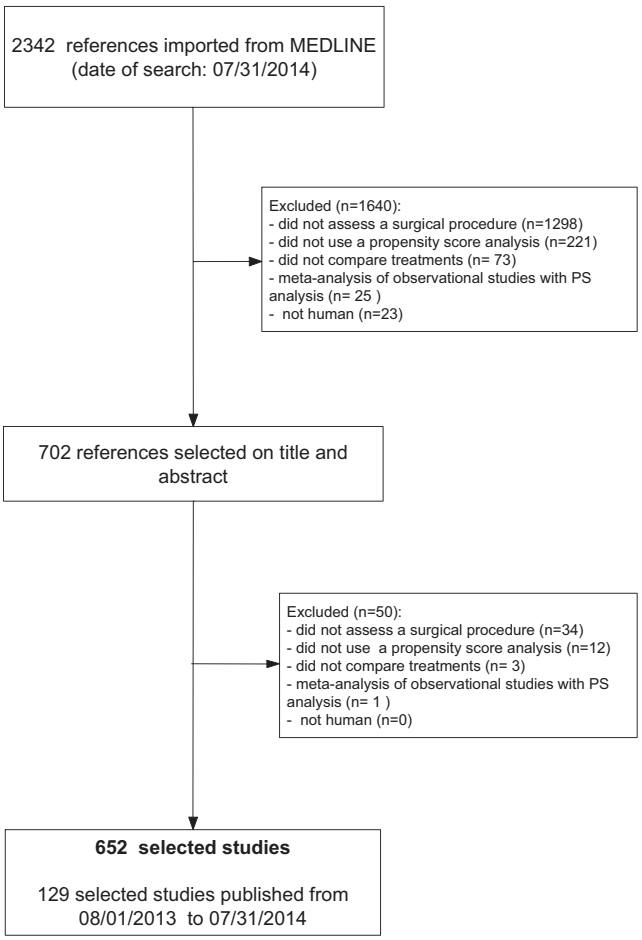


FIGURE 1. Flow of studies in the review.

conversion from 1 arm to the other was possible, and when possible, if the authors reported how they planned to handle these crossovers in the analysis. We considered that a per-operative conversion was possible if the study compared 2 surgical procedures performed by the same surgeon and when the surgeon could decide to change the treatment during the procedure (eg, laparoscopy for appendectomy converted to laparotomy).

Reporting of PS-matched Analyses

PS matching is the most common technique of PS analysis.[12,13,24] This technique relies on matching patients who received the experimental treatment to those who received the comparator by similar or identical PSs. The analysis of the PS-matched sample should compare outcomes by using methods that account for the paired nature of the data. For matched analyses, we checked whether information to allow reproducibility and critical appraisal of the analysis was reported.

First, in the construction and comparison of pairs, 4 features should be reported to allow for reproducibility (see details in Table 1): (1) the matching ratio (1/1, 1/2, 1/3, etc.), (2) the method of selecting the nearest match, (3) whether matching was performed with replacement or not, and (4) whether the statistical analysis accounted for paired data, which is recommended.[22,23,25,26]

Second, to evaluate the external validity and representativeness of the final sample, we recorded whether the initial sample and the postmatching sample baseline characteristics were reported. We also extracted the sample size included in the PS analysis and calculated the related decrease in sample size due to matching. Depending on the matching ratio and the overlap between the population in the intervention and control groups, the loss in sample size could be important.

Third, we examined how the group comparability was reported and assessed. We recorded whether authors reported baseline characteristics for the experimental and control groups after matching and whether and how they assessed imbalance. A substantial difference between experimental and control groups after matching indicates that residual confounding bias remains. Two main ways to evaluate this imbalance are standardized differences and statistical significance testing. However, the latter approach has been criticized because significance levels depend on the study sample size.[27,28]

Statistical Analysis

Descriptive statistics (median, Q1–Q3) were used for continuous variables and number (%) for categorical variables. Analyses involved use of R (http://www.R-project.org, The R Foundation for Statistical Computing, Vienna, Austria).

RESULTS

Evolution of the Use of PS Analysis in Observational Studies Assessing Surgical Procedures

From the 2436 citations retrieved from MEDLINE, we selected 652 reports of observational studies with PS analysis (Fig. 1). The number of observational studies with PS analysis increased over time, from 17 reports in 2004 to 198 in 2013. The

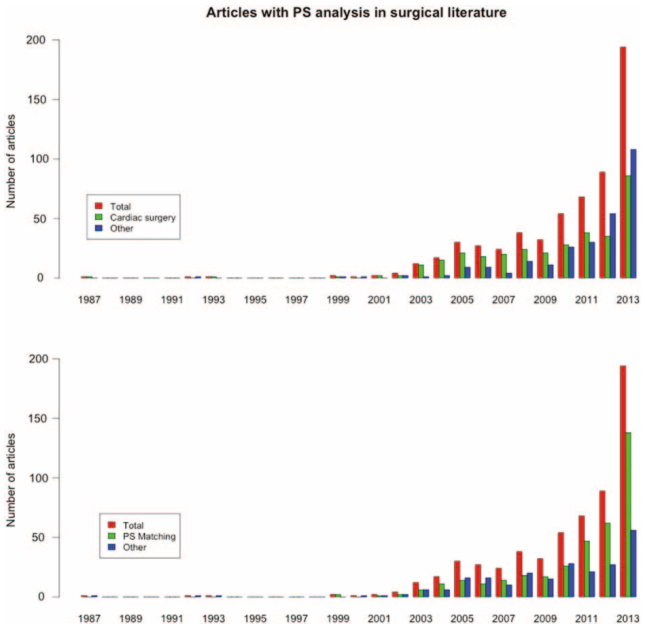


FIGURE 2. Evolution over time of the surgical area and types of propensity-score analysis.

TABLE 2. Characteristics and Quality of Reporting of Observational Studies of Surgery Involving PS analysis

Characteristics of Trials	n = 129
Specialty	
Cardiac surgery	56 (43)
Digestive surgery	26 (20)
Thoracic surgery	14 (11)
Urology	9 (7)
Orthopedic surgery	8 (6)
Gynecology	4 (3)
Vascular surgery	4 (3)
Ear-nose-throat	4 (3)
Other	5 (4)
Type of comparison	
Surgery vs surgery	89 (69)
Surgery vs nonsurgical procedure	21 (16)
Surgery vs conventional care	13 (10)
Surgery vs drugs	6 (5)
More than 2 groups	11 (9)
No. centers	
Single	56 (43)
Multiple	73 (57)
No. care providers	
Single	2 (2)
Multiple	127 (98)
Type of study	
Prospective cohort study	5 (4)
Administrative database	26 (21)
Register	18 (14)
In-house prospective database	62 (48)
Retrospective study	16 (13)
Median no. participants, n (Q1–Q3) [min–max]	904 (348–4133) [25–407070]
Period of recruitment for intervention and control groups	
Similar	117 (92)
Different	8 (6)
Not reported	3 (2)
Funding sources reported	28 (22)
Description of interventions reported	73 (57)
Primary outcome defined	77 (60)

Data are no. (%) unless otherwise indicated.
Q1 indicates quartile 1; Q3, quartile 3.

surgical area evolved over time. In 2004, cardiac surgery represented 90% of studies but only 44% in 2013. Among the reports of observational studies with PS analysis, the proportion with PS matching increased from 47% in 2004 to 71% in 2013 (Fig. 2).

Reporting and Potential Bias of PS Analysis Used in a Recently Published Sample

We systematically appraised the 129 reports published between August 2013 and July 2014. This sample represented 20% of all reports of observational studies with PS analysis.

General Study Characteristics

The principal comparison was "surgery vs surgery" (Table 2). For most articles (83%, n = 106), the study design was a retrospective analysis of prospectively collected data (administrative database, registry, in-house database). Only 5 articles were of full prospective cohorts (4%). The median number of participants was 904 (Q1–Q3: 348–4133, min–max: 25–407070). A primary outcome was reported in 60% reports (n = 77). Most (92%, n = 117) described recruiting patients in the treatment and control groups simultaneously.

TABLE 3. Characteristics and Quality of Statistical Analyses

Characteristics of Trials	n = 129
Statistical department or expert in methodology reported	69 (53)
Type of PS analysis	
Matching	97 (75)
Adjustment	19 (16)
Inverse weighting on PS	7 (5)
Stratification	6 (4)
PS analysis considered as the main analysis	97 (75)
Details of variables included in PS reported	105 (80)
Method to choose variable included in the PS	
Data-driven	18 (14)
A priori defined	11 (9)
Not reported	100 (77)
Median no. variables included in PS, n (Q1–Q3) [min–max]	12 (7–18) [3–41]
Rate of missing data reported	11 (9)
Method to handle missing data reported	9 (7)
Conversion possible during the procedure by the same care provider	69 (53)
Cases of conversion mentioned	9/69 (13)
In case of conversion, analysis	
in the original arm	3/9 (33)
in the other arm	6/9 (67)

Data are no. (%) unless otherwise indicated.
Q1 indicates quartile 1; Q3, quartile 3.

Key Methodological Components of PS Analysis

A statistician was reported among the authors in half of the articles (53%, n = 69) (Table 3). PS analysis was the main analysis in 75% (n = 97), and the main type of PS analysis was matching (75%, n = 97). A prespecified protocol of the PS analysis was never mentioned or accessible. For constructing the PS, 20% of reports (n = 24) did not detail the covariates included in the PS and 77% (n = 100) did not report a justification for these covariates in the PS. For the 59 reports of multicenter studies that reported details on covariates, 17 (29%) included a covariate considering the center effect. The median number of included covariates was 12 (Q1–Q3: 7–18, min–max: 3–41). The rate of missing data for potential covariates was reported in 11 articles (9%); 9 reports (7%) gave a method to handle these data.

We determined that a crossover by conversion with the same care provider was possible in 53% reports (n = 69); 60 did not mention the issue (86%), and 9 (14%) reported at least 1 conversion. In these 9 reports, 3 reported an ITT analysis (33%).

Reporting of PS-matched Analyses

Of the 97 reports with PS-matched analysis, only 9 (10%) reported all 4 key elements that allow for reproducibility of a PS-matched analysis (matching ratio, method to choose the nearest neighbors, replacement and method for statistical analysis) (Table 4). For the construction of pairs, 2 articles (2%) did not report the matching ratio (89% used 1:1 matching), 35 (36%) did not report the method used to choose the nearest match and 76 (78%) did not clearly state whether matching was performed with or without replacement. For the comparison of groups, the authors of 75 reports (77%) did not mention the test used. A paired analysis was described in 19 articles (20%).

TABLE 4. Characteristics and Reporting for Studies With PS Matching

Characteristics of Trials	n = 97
Ratio of matching	
1/1	86 (89)
Other	9 (9)
Not reported	2 (2)
Method for matching	
By caliper with caliper size reported	35 (36)
By caliper without caliper size reported	14 (15)
By digit (5−>1)	13 (13)
Not reported	35 (36)
Use of replacement	
With replacement	1 (1)
Without replacement	20 (21)
Not reported	76 (78)
Analysis used	
Paired	19 (20)
Other	3 (3)
Not reported	75 (77)
The 4 main features for reproducibility of the matching method reported*	9 (10)
Percentage of loss of sample size between initial population and matched population, n (Q1–Q3)	61 (50–77)
Presentation in a table of initial and matched population	0 (0)
Presentation in a table of interventional group and control group in matched population	87 (90)
Imbalance assessment between interventional group and control group in matched population	83 (86)
Method to analyze imbalance assessment	
Test and *P* value	60/83 (72)
Standardized difference	17/83 (20)
Both	6/83 (8)
Analysis used	
Paired	19 (20)
Other	3 (3)
Not reported	75 (77)

*(1) the matching ratio, (2) the method of selecting the nearest match, (3) whether matching was performed with replacement or not, and (4) whether the statistical analysis accounted for paired data.

For the evaluation of the external validity and representativeness of the final sample, no report compared or presented in a same table the baseline characteristics of the original and final sample. The median proportion of loss of sample size was 61% (Q1–Q3: 50–77) and 20% (Q1–Q3: 2–49) for treated groups and 72% (Q1–Q3: 56–89) for the control group.

For methods for assessing balance, 90% of reports (n = 87) described baseline characteristics of the matched treated and control groups. Imbalance was evaluated in 83 reports (86%) and by standardized differences in 17 of 83 (20%).

DISCUSSION

In this study, we systematically assessed the use of PS analysis in observational studies evaluating surgical procedures. Observational studies with PS analysis evaluating a surgical procedure increased in frequency over the years and concerned a large variety of surgical areas. However, the use and reporting of PS analysis in observational studies raises important concerns related to constructing the PS, accounting for missing data and crossovers and reporting all the features necessary to reproduce the analysis.

Our results are consistent with other studies of other fields.[14,25,29,30] In a systematic review of the medical literature between 1996 and 2003, Austin[14,29] showed that the distribution of baseline variables between treated and untreated subjects in the matched PS sample was never reported or explicitly documented with appropriate statistical methods. Gayat et al[13] found that the quality of reporting of PS in the intensive care and anesthesiology literature should be improved. A recent review specifically appraising covariate selection and imbalance found that 65% of articles did not explicitly report how variables were selected to construct the PS, and the use of preexisting knowledge was mentioned in <5% of reports.[30] Furthermore, the method for assessing imbalance was checked and reported in 60% of the reports, but only 25% described standardized differences.

Researchers are increasingly using observational studies with PS analysis to evaluate surgery. However, this type of study requires accounting for specific methodological issues to avoid an important risk of bias. These issues are related to the choice of variables included in the PS model, accounting for crossover and handling missing data.

PS analyses are reliable when there is no unmeasured confounder (ie, when all factors associated with both the choice of the intervention and the outcome of patients can be accounted for in the analysis). However, in reality, the identification of confounders is difficult, and missing an important confounder leads to biased results.[18] In our study, we showed that the construction of the PS was questionable. Overall, 20% of reports did not detail the baseline variables included in the PS and only 29% of reports of multicenter studies reported covariates to account for a center effect. These results highlight the problem of identifying and selecting covariates to include in the PS. The well-designed multicenter study by Castleberry et al[31] identified a large number of covariates (n = 26) but included no covariate accounting for a center effect. This example illustrates that even though the number of covariates is important, omission of a covariate accounting for the center effect in a multicenter study is questionable.

Moreover, 77% of reports did not report any justification to include these variables in the PS. To limit these biases, some authors recommend using all baseline variables available to construct the PS (ie, a nonparsimonious model).[18,32–34] However, this recommendation has been challenged because including some variables (ie, variables associated with the treatment but not the prognosis) could be deleterious.[35] Consequently, some authors have proposed specific methods for selecting these variables (ie, a parsimonious model).[36] Although the theory requires that all confounders be included in the PS model, in practice, consensus is lacking on which variables should be used in the PS model. Nevertheless, both the method used to select these variables and the variables themselves must be reported. Moreover, whatever the method chosen, if data on important confounders are not collected and are missing, the PS analysis will be biased. Preidentification of important covariates to include in the PS, by literature search or experts, could limit biases due to noncollected true confounders.

Another important issue relates to accounting for crossover, which can occur frequently in surgical trials. RCT results should be analyzed according to the ITT principle, with crossover patients analyzed in the group to which they were allocated, not the intervention they actually received. PS aims at mimicking RCTs, but there is risk that the actual intervention only would be recorded, whether it was a crossover or not. In our sample, we showed that crossover was almost never taken into account when it occurred. This issue needs to be considered when planning the study, and the intervention intended and actually performed should be reported.

Finally, missing data are frequent in large observational studies, particularly when data are not specially collected in the context of clinical practice (eg, administrative database) not for a research purpose.[37] Construction of PS with missing data for

TABLE 5. Examples of Reports With Detailed Methodology

Key Methodological Components	Explanation	Reports With Well-detailed Methodology (Ref)
Justification of covariate included in the score	Omitting a true confounder in the PS may produce substantially biased estimates. Consequently, information and justification of covariates included in the PS must be reported to allow readers to detect inappropriate analysis by omission of important confounders. Two main type of strategies exist: All baseline characteristics available (nonparsimonious model) are included in the score Covariates are selected from all baseline characteristics (parsimonious model). The selection could be based on statistical arguments (A) or on preexisting knowledge (B). ''When the number of treated (or untreated) patients is limited, there is currently no rule-of-thumb concerning the maximum number of covariates to be included in the PS, as with multiple regression models.''	(I) Schwann 2014[38] ''Preoperative risk factors and demographic and operative variables were entered in the propensity models irrespective of their significance (all factors in Table 1).'' (IIA) Albacker 2014[39] ''Logistic regression analysis was used to identify statistically significant preoperative differences. A parsimonious model was developed using bagging.'' (IIB) Jenkinson 2014[40] ''The factors considered to be the most important confounders also contributing to deep-infection risk were chosen for the propensity-score algorithm. [. . .] These factors were chosen, based on consensus among the investigators, as the factors most important for predicting later infection but also as those most divergent between the immediate and delayed-closure groups (Table II).''
ITT analysis	ITT analysis defined as all patients analyzed according to the intervention allocated not the intervention received, as recommended in RCTs, avoids overoptimistic estimates of the efficacy of an intervention. Such analysis is particularly important in surgery when a conversion from one intervention to another is possible.	Odermatt 2013[41] ''Every unplanned extension of an incision for any task other than retrieving the specimen was considered to be a conversion. Laparoscopic operations that had to be converted to open surgery were analyzed according to the intention-to-treat principle.''
Method to handle missing data	Calculating PSs including variables with a high rate of missing data could lead to biased results if observations with missing data are discarded. Other methods to handle missing data may limit bias. Two methods were frequent Create a category ''unknown'' and match on that new category Multiple imputation method	Castelberry 2013[42] ''Consistent with previously established methods specific to addressing missing data in propensity score calculations, a separate category for 'unknown' was created for missing data for nominal variables.'' Hu 2012[43] ''Missing data were infrequent (5% on any variable). We performed additional analyses using various missing data statistical approaches including multiple imputation and weighted estimating equations.''
Detail on the construction of matched pairs	In the construction of pairs, some features should be reported to allow for reproducibility: The matching ratio (1/1, 1/2, 1/3, etc) The method of selecting the nearest match Whether matching was performed with replacement or not	D'Onofrio 2013[44] ''A 1:1 match on the propensity score, without replacement, was performed using the psmatch2 procedure, with a conservative caliper width of 20% of the standard deviation of the log of propensity score.''
Imbalance assessment	A substantial difference between experimental and control groups after matching indicates that residual confounding remains. The recommended way to evaluate this imbalance is StDiff. A StDiff $< 10\%$ has been empirically found acceptable	Ravi 2014[45] ''We estimated standardised differences for all covariates before and after matching, with a standardised difference of 10% or more considered indicative of imbalance.''
Paired analysis	The statistical analysis of pair-matched data should account for pairing	Ghezzi 2013[46] ''Continuous outcomes were compared in the PS-matched groups using paired t-tests or Wilcoxon signed rank test as appropriate; differences in proportions were compared using the McNemar's test.''
Presentation of sample before and after matching	To evaluate the representativeness of the final sample, the presentation of the initial sample and the postmatching sample baseline characteristics should be reported	Ravi 2014[45] Table 1

PS indicates propensity score; RCT, randomized controlled trial; StDiff, standardized differences.

covariates is problematic and needs to rely on appropriate methods (Table 5). However, the rate of missing data and the method used to handle it were rarely reported in our sample. Future meta-epidemiologic studies evaluating the impact of these key methodological issues are needed.

We propose some recommendations when planning, conducting, and reporting a surgical observational study using PS analysis.

Recommendations for Planning and Conducting the Study

First, a statistician and methodologist should be involved because the conduct and interpretation of the results suppose a high level of statistical and methodological expertise. Second, all information related to the PS analysis should be prespecified in a protocol before performing any analysis. Particularly, all confounders should be identified, justified in the protocol, and recorded. Third, procedures to limit missing data should be planned and implemented. Fourth, to allow for an ITT analysis, data should be collected on the intervention planned and not just the intervention actually administered.

Recommendations for Reporting the Results

The authors should first follow the StrengThening the Reporting of Observational Studies in Epidemiology statement[38] for reporting. For special issues in surgical trials, authors should report (1) the percentage of missing data for each baseline variable and the method for handling missing data and (2) the method for analyzing crossovers (eg, ITT principle). The authors should also report specific information related to the PS analysis. Particularly, use of the PS analysis should be reported in the title and abstract. In methods, authors should report the type of PS analysis and methods to select variables and construct the PS.

For PS-matched analysis, authors should report in the methods section the method used to construct the matched population and to assess imbalance. In the results section, they should report (1) baseline data for the whole sample before and after matching to evaluate the representativeness of the final sample and (2) baseline data for the experimental and control groups after matching with standardized mean difference.

Examples of reports with detailed methodology are reported in Table 5.[31,39–46]

Limitations

One limitation is that we selected only articles indexed in MEDLINE, which may not be representative of all studies. However, MEDLINE is the largest database of biomedical journal articles and is the most common source used by physicians and policymakers to access study findings. Second, we evaluated the information reported in the articles, which may differ from what was actually done.

CONCLUSIONS

Observational studies with PS analysis of surgery are increasing in frequency, but specific methodological issues and weaknesses in reporting exist.

REFERENCES

1. Sox HC, Greenfield S. Comparative effectiveness research: a report from the Institute of Medicine. *Ann Intern Med*. 2009;151:203–205.
2. Byar DP, Simon RM, Friedewald WT, et al. Randomized clinical trials. Perspectives on some recent ideas. *N Engl J Med*. 1976;295:74–80.
3. Sacks H, Chalmers TC, Smith H Jr. Randomized versus historical controls for clinical trials. *Am J Med*. 1982;72:233–240.
4. Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?". *Lancet*. 2005;365:82–93.
5. Ergina PL, Cook JA, Blazeby JM, et al. Challenges in evaluating surgical innovation. *Lancet*. 2009;374:1097–1104.
6. Demange MK, Fregni F. Limits to clinical trials in surgical areas. *Clinics (Sao Paulo)*. 2011;66:159–161.
7. Feinstein AR. Epidemiologic analyses of causation: the unlearned scientific lessons of randomized trials. *J Clin Epidemiol*. 1989;42:481–489.
8. Stukel TA, Fisher ES, Wennberg DE, et al. Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. *JAMA*. 2007;297:278–285.
9. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70:41–55.
10. Cepeda MS, Boston R, Farrar JT, et al. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am J Epidemiol*. 2003;158:280–287.
11. Arbogast PG, Ray WA. Performance of disease risk scores, propensity scores, and traditional multivariable outcome regression in the presence of multiple confounders. *Am J Epidemiol*. 2011;174:613–620.
12. Lonjon G, Boutron I, Trinquart L, et al. Comparison of Treatment effect estimates from prospective nonrandomized studies with propensity score analysis and randomized controlled trials of surgical procedures. *Ann Surg*. 2014;259:18–25.
13. Gayat E, Pirracchio R, Resche-Rigon M, et al. Propensity scores in intensive care and anaesthesiology literature: a systematic review. *Intensive Care Med*. 2010;36:1993–2003.
14. Austin PC. Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: a systematic review and suggestions for improvement. *J Thorac Cardiovasc Surg*. 2007;134:1128–1135. e3.
15. Hade EM, Lu B. Bias associated with using the estimated propensity score as a regression covariate. *Stat Med*. 2014;33:74–87.
16. Austin PC, Grootendorst P, Normand S-LT, et al. Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a Monte Carlo study. *Stat Med*. 2007;26:754–768.
17. Cook JA. The challenges faced in the design, conduct and analysis of surgical randomised controlled trials. *Trials*. 2009;10:9.
18. Brookhart MA, Schneeweiss S, Rothman KJ, et al. Variable selection for propensity score models. *Am J Epidemiol*. 2006;163:1149–1156.
19. Norris CM, Ghali WA, Knudtson ML, et al. Dealing with missing data in observational health care outcome analyses. *J Clin Epidemiol*. 2000;53:377–383.
20. D'Agostino RB, Rubin DB. Estimating and using propensity scores with partially missing data. *J Am Stat Assoc*. 2000;95:749–759.
21. Mitra R, Reiter JP. Estimating propensity scores with missing covariate data using general location mixture models. *Statist Med*. 2011;30:627–641.
22. Heritier SR, Gebski VJ, Keech AC. Inclusion of patients in clinical trial analysis: the intention-to-treat principle. *Med J Aust*. 2003;179:438–440.
23. Danaei G, Tavakkoli M, Hernán MA. Bias in observational studies of prevalent users: lessons for comparative effectiveness research from a meta-analysis of statins. *Am J Epidemiol*. 2012;175:250–262.
24. Haukoos JS, Lewis RJ. The propensity score. *JAMA*. 2015;314:1637–1638.
25. Gayat E, Resche-Rigon M, Mary J-Y, et al. Propensity score applied to survival data analysis through proportional hazards models: a Monte Carlo study. *Pharm Stat*. 2012;11:222–229.
26. Austin PC. Comparing paired vs non-paired statistical methods of analyses when making inferences about absolute risk reductions in propensity-score matched samples. *Stat Med*. 2011;30:1292–1301.
27. 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*. 2009;29:661–677.
28. Franklin JM, Rassen JA, Ackermann D, et al. Metrics for covariate balance in cohort studies of causal effects. *Stat Med*. 2014;33:1685–1699.
29. Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Stat Med*. 2008;27:2037–2049.
30. Ali MS, Groenwold RHH, Belitser SV, et al. Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: a systematic review. *J Clin Epidemiol*. 2015;68:112–121.
31. Castleberry AW, Worni M, Osho AA, et al. Use of lung allografts from brain-dead donors after cardiopulmonary arrest andresuscitation. *Am J Respir Crit Care Med*. 2013;188:466–473.

32. Robins JM, Mark SD, Newey WK. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*. 1992;48:479–495.
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;26:734–753.
34. Austin PC. The performance of different propensity score methods for estimating marginal odds ratios. *Stat Med*. 2007;26:3078–3094.
35. Bhattacharya J, Vogt W. *Do Instrumental Variables Belong in Propensity Scores?* 2007. NBER Thechnical working paper n°343, Issued in september 2007. (DOI): 10.3386/t0343.
36. 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.
37. Twisk J, de Vente W. Attrition in longitudinal studies. How to deal with missing data. *J Clin Epidemiol*. 2002;55:329–337.
38. Elm E von, Altman DG, Egger M, et al. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *BMJ*. 2007;335:806–808.
39. Schwann TA, Tranbaugh RF, Dimitrova KR, et al. Time-varying survival benefit of radial artery versus vein grafting: a multiinstitutional analysis. *Ann Thorac Surg*. 2014;97:1328–1334.
40. Albacker TB, Blackstone EH, Williams SJ, et al. Should less-invasive aortic valve replacement be avoided in patients with pulmonary dysfunction? *J Thorac Cardiovasc Surg*. 2014;147:355–361. e5.
41. Jenkinson RJ, Kiss A, Johnson S, et al. Delayed wound closure increases deep-infection rate associated with lower-gradeopen fractures: a propensity-matched cohort study. *J Bone Joint Surg Am*. 2014;96:380–386.
42. Odermatt M, Miskovic D, Siddiqi N, et al. Short- and long-term outcomes after laparoscopic versus open emergency resection for colon cancer: an observational propensity score-matched study. *World J Surg*. 2013;37:2458–2467.
43. Hu JC, Gu X, Lipsitz SR, et al. Comparative effectiveness of minimally invasive vs open radical prostatectomy. *JAMA*. 2009;302:1557–1564.
44. D'Onofrio A, Rizzoli G, Messina A, et al. Conventional surgery, sutureless valves, and transapical aortic valve replacement: what is the best option for patients with aortic valve stenosis? A multicenter, propensity-matched analysis. *J Thorac Cardiovasc Surg*. 2013;146:1065–1070.
45. Ravi B, Croxford R, Austin PC, et al. The relation between total joint arthroplasty and risk for serious cardiovascular events in patients with moderate-severe osteoarthritis: propensity score matched landmark analysis. *BMJ*. 2013;347:f6187.
46. Ghezzi F, Fanfani F, Malzoni M, et al. Minilaparoscopic radical hysterectomy for cervical cancer: multi-institutional experience in comparison with conventional laparoscopy. *Eur J Surg Oncol*. 2013;39:1094–1100.

I Introduction et objectifs

II Article 1: Comparaison de la mesure de l'effet traitements entre les études observationnelles utilisant un score de propension et les ECR évaluant une procédure chirurgicale

III Article 2 : Description et Sources de biais des Etudes observationnelles utilisant une analyse avec score de propension évaluant une procédure chirurgicale. Revue systématique

IV Discussion et perspectives

V Conclusion

VI Références

IV Discussion et perspectives

Compte tenu des limites des essais randomisés pour évaluer les interventions chirurgicales, du nombre croissant de données observationnelles disponibles avec l'accès aux bases de données administratives, aux données des outils connectés et au dossier médical informatisé, il y a un intérêt croissant pour les études observationnelles pour l'évaluation des interventions chirurgicales. L'analyse par SP pourrait permettre de limiter les biais de confusion et pourrait avoir un apport important dans ce contexte. Notre premier travail n'a pas mis en évidence de surestimation systématique de l'estimation de l'effet traitement dans les études observationnelles avec SP par rapport aux ECR. Même si l'interprétation de ces résultats doit être prudente compte tenu des possibles facteurs de confusion et de la taille de notre échantillon, cela renforce l'intérêt de cette méthode.

Cependant, nous travaux ont aussi montré que l'utilisation de cette méthode en chirurgie soulève de nombreuses questions méthodologiques avec en particulier le manque de transparence sur le choix des variables permettant le calcul du SP, la prise en compte des données manquante, l'absence d'évocation des facteurs de confusion qui ne peuvent pas être pris en compte car non collectés, la non prise en compte des contaminations ainsi que l'absence de support de statisticiens ou méthodologistes.

Les perspectives de ce travail vont s'intéresser à 2 problématiques, le choix des variables et le biais de sélection des résultats.

Problème du choix des variables

Le SP est calculé à partir des données de l'étude. Par conséquent, le SP calculé n'est qu'une approximation du « vrai » SP, qui lui garantit l'absence de biais de confusion.

La justesse de cette approximation du « vrai » SP, est liée à la présence ou non de l'ensemble des variables de confusion. Ainsi le choix des variables à inclure dans le calcul du SP est fondamental (107,125). Dans la crainte d'oublier une variable de confusion, certains auteurs proposent de sélectionner toutes les variables disponibles (méthode dite « non parcimonieuse »)(126). Cependant, ce type d'approche expose à d'autres problèmes : la sur-paramétrisation du modèle de calcul du SP dans le cas d'échantillon limité effectif (107) et le risque d'inclure des variables dites « instrumentales » (variable liée au choix du traitement mais pas au critère de jugement), délétères aux résultats (127).

Nos résultats ont montré montrent que peu d'articles rapportent la méthode de sélection des variables et les arguments guidant le choix.

De plus, dans la pratique courante outre le fait que l'identification des variables de confusions est souvent difficile (88,107,125,128), certaines variables de confusion peuvent ne pas avoir été collectées dans la base de données. Par exemple, dans l'étude multicentrique de Castleberry et al., toutes les variables correspondant aux caractéristiques de bases du patient ont été incluses dans le SP (n=26), cependant les auteurs n'ont pas pris en compte l'effet centre qui semble pourtant être une variable pouvant influencer les résultats (125). En l'absence de ces variables dans la base de données il sera donc impossible de s'approcher du « vrai » SP.

Nos travaux sur cette thématique vont avoir pour objectif de déterminer dans un échantillon d'études observationnelles avec SP publiée, la proportion d'étude qui n'incluent pas des facteurs de confusion essentiels dans le calcul du SP. Ce travail reposera sur une revue méthodologique dans des domaines d'interventions chirurgicales bien définies (chirurgie cardiaque, la chirurgie orthopédique, chirurgie viscérale). A partir d'un échantillon d'études observationnelles avec SP évaluant une

procédure chirurgicale, le nombre et le type de variables incluses dans le calcul du score de propension seront relevés.

A partir de cette revue, nous élaborerons des vignettes d'études qui seront soumises à l'avis d'expert du domaine. Les experts devront identifier les facteurs de confusion essentiels à prendre en compte dans le cadre de chaque étude présentée. Le résultat des vignettes permettre de déterminer le nombre d'études publiées qui ne prennent pas en compte des variables de confusion essentielles dans l'analyse par SP.

Ce travail sera réalisé en collaboration avec des équipes de chirurgiens de spécialité différente dont l'équipe du Dr Ergina, (Department of Surgery, McGill University Health Centre, Montreal, QC, Canada).

Biais de sélection des résultats

Dans le cadre des études observationnelles avec SP, la grande majorité des études sont rétrospectives, et l'analyse statistique pour répondre à l'hypothèse de travail n'est que rarement pré-spécifiée. Dans le cadre d'une étude observationnelle utilisant une analyse avec SP, cette analyse se surajoute régulièrement à une analyse brute et a d'autre analyse ajustée. Cette multiplication des analyses expose au risque de biais de sélection des résultats. Cette sélection peut être comparé au biais déjà bien étudié dans le cadre des ECR, consistant à une sélection des résultats entre les différents critères de jugement évalués (129–131).

Notre travail sur cette thématique aura pour objectif d'estimer le biais de présentation des résultats dans un échantillon d'études observationnelles avec SP.

Notre hypothèse de travail est qu'il existe un biais de sélection des résultats par sélection des analyses statistiquement significatives. Comme l'analyse avec SP offre une possibilité d'analyse de plus, les investigateurs de ce type d'étude peuvent être tentés de mettre en avant les résultats statistiquement significatifs pour en faciliter la publication.

Le travail consistera à étudier le nombre d'analyses réalisées, le type d'analyse (régressions multiples, score de propension) et les résultats des analyses mises en avant dans le résumé de l'étude.

La méthodologie s'appuierait ainsi sur une méthodologie de revue systématique méthodologique à partir d'un échantillon d'études observationnelles utilisant un SP. Une analyse descriptive de toutes les analyses rapportées dans la section « méthode », section « résultat » et section « résumé » permettra d'évaluer le biais.

I Introduction et objectifs

II Article 1: Comparaison de la mesure de l'effet traitements entre les études observationnelles utilisant un score de propension et les ECR évaluant une procédure chirurgicale

III Article 2 : Description et Sources de biais des Etudes observationnelles utilisant une analyse avec score de propension évaluant une procédure chirurgicale. Revue systématique

IV Discussion et perspectives

V Conclusion

VI Réferences

V Conclusion

Les études observationnelles avec SP sont de plus en plus fréquentes. Leurs avantages dans le domaine de la chirurgie en font une alternative intéressante et fiable aux ECR. Cependant, il est important de respecter les règles méthodologiques de l'utilisation de cette méthode et d'être transparent sur la réalisation des analyses et la présentation de tous les résultats (aide d'un statisticien, écrire un plan d'analyse en aveugle des données collectées, définir en aveugle des données collectées les variables à utiliser pour calculer le SP, ne pas multiplié les analyses statistiques, critiquer la population d'étude par rapport à la population d'origine).

I Introduction et objectifs

II Article 1: Comparaison de la mesure de l'effet traitements entre les études observationnelles utilisant un score de propension et les ECR évaluant une procédure chirurgicale

III Article 2 : Description et Sources de biais des Etudes observationnelles utilisant une analyse avec score de propension évaluant une procédure chirurgicale. Revue systématique

IV Discussion et perspectives

V Conclusion

VI Réferences

VI Références :

1. Most Frequent Operating Room Procedures Performed in U.S. Hospitals, 2003-2012 #186. n.d. (accessed January 4, 2017).

2. Conseil national de chirurgie. Activité chiurgicale en france en 2009. 2009.

3. Watson CJE, Dark JH. Organ transplantation: historical perspective and current practice. Br J Anaesth. 2012; 108 Suppl 1:i29-42.

4. Solis M. New Frontiers in Robotic Surgery: The latest high-tech surgical tools allow for superhuman sensing and more. IEEE Pulse. 2016; 7:51–5.

5. Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. Chest. 1989; 95:2S–4S.

6. Black N. Why we need observational studies to evaluate the effectiveness of health care. BMJ. 1996; 312:1215–8.

7. Glasziou P, Chalmers I, Rawlins M, et al. When are randomised trials unnecessary? Picking signal from noise. BMJ. 2007; 334:349–51.

8. Horton R. Surgical research or comic opera: questions, but few answers. Lancet. 1996; 347:984–5.

9. Salzman EW. Is Surgery Worthwhile?: Presidential Address. Arch Surg. 1985; 120:771–6.

10. Spodick DH. Numerators without denominators. There is no FDA for the surgeon. JAMA. 1975; 232:35–6.

11. The periodic health examination. Canadian Task Force on the Periodic Health Examination. Can Med Assoc J. 1979; 121:1193–254.

12. Wood L, Egger M, Gluud LL, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. BMJ. 2008; 336:601–5.

13. Schulz KF, Grimes DA. Allocation concealment in randomised trials: defending against deciphering. Lancet. 2002; 359:614–8.

14. Hróbjartsson A, Thomsen ASS, Emanuelsson F, et al. Observer bias in randomised clinical trials with binary outcomes: systematic review of trials with both blinded and non-blinded outcome assessors. BMJ. 2012; 344:e1119.

15. Tierney JF, Stewart LA. Investigating patient exclusion bias in meta-analysis. Int J Epidemiol. 2005; 34:79–87.

16. Moher D, Pham B, Jones A, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet. 1998; 352:609–13.

17. Boutron I, Tubach F, Giraudeau B, et al. Blinding was judged more difficult to achieve and maintain in nonpharmacologic than pharmacologic trials. J Clin Epidemiol. 2004; 57:543–50.

18. Abraha I, Cherubini A, Cozzolino F, et al. Deviation from intention to treat analysis in randomised trials and treatment effect estimates: meta-epidemiological study. BMJ. 2015; 350:h2445.

19. Black D. The limitations of evidence. J R Coll Physicians Lond. 1998; 32:23–6.

20. Hampton JR. Size isn't everything. Stat Med. 2002; 21:2807–14.

21. Evans JG. Evidence-based and evidence-biased medicine. Age Ageing. 1995; 24:461–3.

22. Charlton BG, Miles A. The rise and fall of EBM. QJM. 1998; 91:371–4.

23. Naylor CD. Grey zones of clinical practice: some limits to evidence-based medicine. Lancet. 1995; 345:840–2.

24. Messerli FH. Antihypertensive therapy: beta-blockers and diuretics-why do physicians not always follow guidelines? Proc (Bayl Univ Med Cent). 2000; 13:128-131-134.

25. Cabana MD, Rand CS, Powe NR, et al. Why don't physicians follow clinical practice guidelines? A framework for improvement. JAMA. 1999; 282:1458–65.

26. Iyngkaran P, Toukhsati SR, Thomas MC, et al. A Review of the External Validity of Clinical Trials with Beta-Blockers in Heart Failure. Clin Med Insights Cardiol. 2016; 10:163–71.

27. Spitzer E, Hadorn S, Zanchin T, et al. External validity of a contemporaneous primary percutaneous coronary intervention trial in patients with acute ST-elevation myocardial infarction: insights from a single-centre investigation. EuroIntervention. 2016; 12:1135–43.

28. Morin-Ben Abdallah S, Dutilleul A, Nadon V, et al. Quantification of the External Validity of Randomized Controlled Trials Supporting Clinical Care Guidelines: The Case of Thromboprophylaxis. Am J Med. 2016; 129:740–5.

29. Kennedy-Martin T, Curtis S, Faries D, et al. A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. Trials. 2015; 16:495.

30. Steg PG, López-Sendón J, Lopez de Sa E, et al. External validity of clinical trials in acute myocardial infarction. Arch Intern Med. 2007; 167:68–73.

31. Van Spall HGC, Toren A, Kiss A, et al. Eligibility criteria of randomized controlled trials published in high-impact general medical journals: a systematic sampling review. JAMA. 2007; 297:1233–40.

32. Rothwell PM, Eliasziw M, Gutnikov SA, et al. Analysis of pooled data from the randomised controlled trials of endarterectomy for symptomatic carotid stenosis. Lancet. 2003; 361:107–16.

33. Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?" Lancet. 2005; 365:82–93.

34. Halm EA, Lee C, Chassin MR. Is volume related to outcome in health care? A systematic review and methodologic critique of the literature. Ann Intern Med. 2002; 137:511–20.

35. Biau DJ, Halm JA, Ahmadieh H, et al. Provider and center effect in multicenter randomized controlled trials of surgical specialties: an analysis on patient-level data. Ann Surg. 2008; 247:892–8.

36. Chan A-W, Altman DG. Epidemiology and reporting of randomised trials published in PubMed journals. Lancet. 2005; 365:1159–62.

37. Boutron I, Tubach F, Giraudeau B, et al. Methodological differences in clinical trials evaluating nonpharmacological and pharmacological treatments of hip and knee osteoarthritis. JAMA. 2003; 290:1062–70.

38. McCulloch P, Kaul A, Wagstaff GF, et al. Tolerance of uncertainty, extroversion, neuroticism and attitudes to randomized controlled trials among surgeons and physicians. Br J Surg. 2005; 92:1293–7.

39. Weinstein JN, Tosteson TD, Lurie JD, et al. Surgical vs nonoperative treatment for lumbar disk herniation: the Spine Patient Outcomes Research Trial (SPORT): a randomized trial. JAMA. 2006; 296:2441–50.

40. A prospective analysis of 1518 laparoscopic cholecystectomies. The Southern Surgeons Club. N Engl J Med. 1991; 324:1073–8.

41. Hardoon SL, Lewsey JD, Gregg PJ, et al. Continuous monitoring of the performance of hip prostheses. J Bone Joint Surg Br. 2006; 88:716–20.

42. Grange P, Mulla M. Learning the “learning curve.” Surgery. 2015; 157:8–9.

43. Barkun JS, Aronson JK, Feldman LS, et al. Evaluation and stages of surgical innovations. Lancet. 2009; 374:1089–96.

44. Lilford RJ, Braunholtz DA, Greenhalgh R, et al. Trials and fast changing technologies: the case for tracker studies. BMJ. 2000; 320:43–6.

45. Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. Int J Surg. 2012; 10:28–55.

46. Day SJ, Altman DG. Statistics notes: blinding in clinical trials and other studies. BMJ. 2000; 321:504.

47. Noseworthy JH, Ebers GC, Vandervoort MK, et al. The impact of blinding on the results of a randomized, placebo-controlled multiple sclerosis clinical trial. Neurology. 1994; 44:16–20.

48. Michels KB, Rothman KJ. Update on unethical use of placebos in randomised trials. Bioethics. 2003; 17:188–204.

49. Wartolowska K, Collins GS, Hopewell S, et al. Feasibility of surgical randomised controlled trials with a placebo arm: a systematic review. BMJ Open. 2016; 6:e010194.

50. Moseley JB, O'Malley K, Petersen NJ, et al. A controlled trial of arthroscopic surgery for osteoarthritis of the knee. N Engl J Med. 2002; 347:81–8.

51. Taly AB, Sivaraman Nair KP, Murali T, et al. Efficacy of multiwavelength light therapy in the treatment of pressure ulcers in subjects with disorders of the spinal cord: A randomized double-blind controlled trial. Arch Phys Med Rehabil. 2004; 85:1657–61.

52. Jacquier I, Boutron I, Moher D, et al. The reporting of randomized clinical trials using a surgical intervention is in need of immediate improvement: a systematic review. Ann Surg. 2006; 244:677–83.

53. Ergina PL, Cook JA, Blazeby JM, et al. Challenges in evaluating surgical innovation. Lancet. 2009; 374:1097–104.

54. Glasziou P, Meats E, Heneghan C, et al. What is missing from descriptions of treatment in trials and reviews? BMJ. 2008; 336:1472–4.

55. Guyer RD, McAfee PC, Banco RJ, et al. Prospective, randomized, multicenter Food and Drug Administration investigational device exemption study of lumbar total disc replacement with the CHARITE artificial disc versus lumbar fusion: five-year follow-up. Spine J. 2009; 9:374–86.

56. Smith GCS, Pell JP. Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials. BMJ. 2003; 327:1459–61.

57. Kesselheim AS, Avorn J. New "21st Century Cures" Legislation: Speed and Ease vs Science. JAMA. 2017.

58. Hernán MA, Alonso A, Logan R, et al. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. Epidemiology. 2008; 19:766–79.

59. Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. Am J Epidemiol. 2016; 183:758–64.

60. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 2016; 355:i4919.

61. 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.

62. Sacks H, Chalmers TC, Smith H Jr. Randomized versus historical controls for clinical trials. Am J Med. 1982; 72:233–40.

63. Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. BMJ. 1998; 317:1185–90.

64. Britton A, McKee M, Black N, et al. Choosing between randomised and non-randomised studies: a systematic review. Health Technol Assess. 1998; 2:i–iv, 1-124.

65. MacLehose RR, Reeves BC, Harvey IM, et al. A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. Health Technol Assess. 2000; 4:1–154.

66. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. N Engl J Med. 2000; 342:1878–86.

67. Concato J, Shah N, Horwitz RI. Randomized, Controlled Trials, Observational Studies, and the Hierarchy of Research Designs. New England Journal of Medicine. 2000; 342:1887–92.

68. Ioannidis JPA, Haidich A-B, Pappa M, et al. Comparison of Evidence of Treatment Effects in Randomized and Nonrandomized Studies. JAMA: The Journal of the American Medical Association. 2001; 286:821–30.

69. Wilson DB, Lipsey MW. The role of method in treatment effectiveness research: evidence from meta-analysis. Psychol Methods. 2001; 6:413–29.

70. Deeks JJ, Dinnes J, D'Amico R, et al. Evaluating non-randomised intervention studies. Health Technol Assess. 2003; 7:iii–x, 1-173.

71. Britton A, McKee M, Black N, et al. Choosing between randomised and non-randomised studies: a systematic review. Health Technol Assess. 1998; 2:i–iv, 1-124.

72. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983.

73. Rosenbaum PR, Rubin DB. The bias due to incomplete matching. Biometrics. 1985; 41:103–16.

74. Ming K, Rosenbaum PR. Substantial gains in bias reduction from matching with a variable number of controls. Biometrics. 2000; 56:118–24.

75. Imbens G. Nonparametric estimation of average treatment effects under exogeneity: A review. The Review of Economics and Statistics. 2004.

76. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behav Res. 2011; 46:399–424.

77. Schafer JL, Kang J. Average causal effects from nonrandomized studies: a practical guide and simulated example. Psychol Methods. 2008; 13:279–313.

78. Austin PC. Some methods of propensity-score matching had superior performance to others: results of an empirical investigation and Monte Carlo simulations. Biom J. 2009; 51:171–84.

79. Gayat E, Resche-Rigon M, Mary J-Y, et al. Propensity score applied to survival data analysis through proportional hazards models: a Monte Carlo study. Pharm Stat. 2012.

80. Hill J, Reiter JP. Interval estimation for treatment effects using propensity score matching. Stat Med. 2006; 25:2230–56.

81. Gu XS, Rosenbaum PR. Comparison of Multivariate Matching Methods: Structures, Distances, and Algorithms. Journal of Computational and Graphical Statistics. 1993; 2:405–20.

82. Austin PC. Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: A systematic review and suggestions for improvement. The Journal of Thoracic and Cardiovascular Surgery. 2007; 134:1128–1135.e3.

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

84. 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.

85. Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. Pharm Stat. 2011; 10:150–61.

86. Rosenbaum PR, Rubin DB. Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. 2010.

87. Hullsiek KH, Louis TA. Propensity score modeling strategies for the causal analysis of observational data. Biostatistics. 2002; 3:179–93.

88. 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.

89. Austin PC, Mamdani MM. A comparison of propensity score methods: a case-study estimating the effectiveness of post-AMI statin use. Stat Med. 2006; 25:2084–106.

90. Austin PC, Stuart EA. The performance of inverse probability of treatment weighting and full matching on the propensity score in the presence of model misspecification when estimating the effect of treatment on survival outcomes. Stat Methods Med Res. 2015.

91. King G, Nielsen R. Why Propensity Scores Should Not Be Used for Matching. 2016.

92. 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. 2009; 29:661–77.

93. 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.

94. Joffe MM, Have TRT, Feldman HI, et al. Model Selection, Confounder Control, and Marginal Structural Models: Review and New Applications. The American Statistician. 2004; 58:272–9.

95. Morgan SL, Todd JJ. A Diagnostic Routine for the Detection of Consequential Heterogeneity of Causal Effects. Sociological Methodology. 2008; 38:231–81.

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

97. Imai K, King G, Stuart E. Misunderstandings Among Experimentalists and Observationalists about Causal Inference. Journal of the Royal Statistical Society, Series A. 2008; 171, part 2:481–502.

98. Austin PC. The performance of different propensity-score methods for estimating relative risks. J Clin Epidemiol. 2008; 61:537–45.

99. 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.

100. Wang Z. Propensity score methods to adjust for confounding in assessing treatment effects: bias and precision. The Internet Journal of Epidemiology. 2008; 7.

101. Chen H-Y, Wang Q, Xu Q-H, et al. Statin as a Combined Therapy for Advanced-Stage Ovarian Cancer: A Propensity Score Matched Analysis. Biomed Res Int. 2016; 2016.

102. Dai F, Meng S, Mei L, et al. Single-port video-assisted thoracic surgery in the treatment of non-small cell lung cancer: a propensity-matched comparative analysis. J Thorac Dis. 2016; 8:2872–8.

103. Bao F, Zhang C, Yang Y, et al. Comparison of robotic and video-assisted thoracic surgery for lung cancer: a propensity-matched analysis. J Thorac Dis. 2016; 8:1798–803.

104. 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. J Clin Epidemiol. 2006; 59:437–47.

105. Weitzen S, Lapane KL, Toledano AY, et al. Weaknesses of goodness-of-fit tests for evaluating propensity score models: the case of the omitted confounder. Pharmacoepidemiol Drug Saf. 2005; 14:227–38.

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

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

108. 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.

109. Cepeda MS, Boston R, Farrar JT, et al. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. Am J Epidemiol. 2003; 158:280–7.

110. Arbogast PG, Ray WA. Performance of Disease Risk Scores, Propensity Scores, and Traditional Multivariable Outcome Regression in the Presence of Multiple Confounders. American Journal of Epidemiology. 2011; 174:613–20.

111. Peduzzi P, Concato J, Kemper E, et al. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. 1996; 49:1373–9.

112. Lonjon G, Boutron I, Trinquart L, et al. Comparison of Treatment Effect Estimates From Prospective Nonrandomized Studies With Propensity Score Analysis and Randomized Controlled Trials of Surgical Procedures. Ann Surg. 2013.

113. Harrison EM, Bhangu A, Swann O, et al. Equivalence Approach Is More Appropriate for Comparison of Treatment Effect Estimates. Ann Surg. 2015; 262:e67.

114. Lonjon G, Boutron I, Trinquart L, et al. Reply to Letter: "Equivalence Approach in Meta-epidemiological Studies Can Be Misleading." Ann Surg. 2014.

115. Gayat E, Pirracchio R, Resche-Rigon M, et al. Propensity scores in intensive care and anaesthesiology literature: a systematic review. Intensive Care Med. 2010; 36:1993–2003.

116. Kuss O, Legler T, Börgermann J. Treatments effects from randomized trials and propensity score analyses were similar in similar populations in an example from cardiac surgery. J Clin Epidemiol. 2011; 64:1076–84.

117. Dahabreh IJ, Sheldrick RC, Paulus JK, et al. Do observational studies using propensity score methods agree with randomized trials? A systematic comparison of studies on acute coronary syndromes. Eur Heart J. 2012; 33:1893–901.

118. Concato J, Lawler EV, Lew RA, et al. Observational methods in comparative effectiveness research. Am J Med. 2010; 123:e16-23.

119. Shikata S, Nakayama T, Noguchi Y, et al. Comparison of Effects in Randomized Controlled Trials With Observational Studies in Digestive Surgery. Ann Surg. 2006; 244:668–76.

120. Papanikolaou PN, Christidi GD, Ioannidis JPA. Comparison of evidence on harms of medical interventions in randomized and nonrandomized studies. CMAJ. 2006; 174:635–41.

121. Golder S, Loke YK, Bland M. Meta-analyses of adverse effects data derived from randomised controlled trials as compared to observational studies: methodological overview. PLoS Med. 2011; 8:e1001026.

122. Lonjon G, Porcher R, Ergina P, et al. Potential Pitfalls of Reporting and Bias in Observational Studies With Propensity Score Analysis Assessing a Surgical Procedure: A Methodological Systematic Review. Ann Surg. 2016.

123. Austin PC. Comparing paired vs non-paired statistical methods of analyses when making inferences about absolute risk reductions in propensity-score matched samples. Stat Med. 2011; 30:1292–301.

124. Ali MS, Groenwold RHH, Belitser SV, et al. Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: a systematic review. J Clin Epidemiol. 2015; 68:112–21.

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

126. Rubin DB, Thomas N. Matching using estimated propensity scores: relating theory to practice. Biometrics. 1996; 52:249–64.

127. 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.

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

129. Dechartres A, Bond EG, Scheer J, et al. Reporting of statistically significant results at ClinicalTrials.gov for completed superiority randomized controlled trials. BMC Med. 2016; 14:192.

130. Ioannidis JP. Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. JAMA. 1998; 279:281–6.

131. Dwan K, Gamble C, Williamson PR, et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias - an updated review. PLoS ONE. 2013; 8:e66844.