UNIVERSITÉ DE LORRAINE

École Doctorale BioSE (Biologie-Santé-Environnement)

Thèse

Présentée et soutenue publiquement pour l'obtention du titre de

DOCTEUR DE L'UNIVERSITÉ DE LORRAINE

Mention : « Sciences de la Vie et de la Santé »

par **Hervé LABORDE-CASTÉROT**

Évaluation de l'effet des interventions en santé : intérêt des études observationnelles et méthodes d'analyse pour maîtriser le biais d'indication

Le 9 décembre 2016

Membres du jury :

Rapporteurs :	**Mme Mariette MERCIER**	**Professeure émérite, Université de Franche-Comté, Besançon**
	M. Benoît MARIN	**PU-PH, Université de Limoges, Limoges**
Examinateurs :	**M. Patrick BROCHARD**	**PU-PH, Université de Bordeaux, Bordeaux**
	Mme Nathalie THILLY	**PU-PH, Université de Lorraine, Nancy, Directrice de thèse**
	Mme Nelly AGRINIER	**MCU-PH, Université de Lorraine, Nancy, Co-directrice de thèse**

EA 4360 APEMAC « Maladies chroniques, santé perçue et processus d'adaptation. Approches épidémiologiques et psychologiques » - Université de Lorraine - École de Santé Publique - Faculté de Médecine - 9 avenue de la Forêt de Haye - CS 50184 - 54505 VANDŒUVRE LÈS NANCY Cedex

REMERCIEMENTS

Mes très sincères remerciements vont aux membres du jury :

À **Madame le Docteur Nelly AGRINIER**, co-directrice de thèse,

Je te remercie d'avoir co-dirigé mon travail, avec ce poil à gratter que tu sais semer pour faire germer des discussions épineuses. Toujours avec bienveillance et encouragements.

À **Monsieur le Professeur Patrick BROCHARD**, examinateur,

Vous me faîtes le grand honneur de présider ce jury. Je tiens à vous dire que je n'oublierai jamais cette année passée à Bordeaux, qui m'a tant apporté à la fois sur les plans scientifique et humain. Soyez assuré de ma profonde reconnaissance et de mon profond respect.

À **Monsieur le Professeur Benoît MARIN**, rapporteur,

Je vous remercie de l'honneur que vous me faîtes d'avoir accepté de juger ces travaux. Soyez assuré de ma respectueuse gratitude.

À **Madame le Professeur Mariette MERCIER**, rapporteur,

Je suis très honoré que vous ayez accepté de juger mon travail, d'abord comme membre du Comité de thèse puis comme rapporteur. Soyez assurée de ma gratitude et de mon profond respect.

À **Madame le Professeur Nathalie THILLY**, directrice de thèse,

Les circonstances dans lesquelles que t'ai sollicitée initialement étaient pour le moins singulières. Mon parcours par la suite ne l'a pas moins été... Malgré cela, tu m'as accordé puis renouvelé ta confiance et je t'en suis très reconnaissant. Je te remercie infiniment de m'avoir accompagné pendant ces trois ans. C'est bien ma bonne étoile qui m'a dirigé vers Nancy.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ACRONYMES

ATE	*Average treatment effect*
ATT	*Average treatment among treated*
CER	*Comparative effectiveness research*
cmRCT	*Cohort multiple randomised controlled trial*
CCS	*Comprehensive cohort study*
ECR	Essai contrôlé randomisé
hdPS	*high dimensional propensity score*
IC	Intervention complexe
IOM	*Institute Of Medicine*
IPTW	*Inverse ponderation treatment weighting*
MRC	*Medical Research Council*
PRPT	*Partially randomised preference trial*
SDiff	*Standardised difference*
SP	Score de propension
VI	Variable instrumentale

INTRODUCTION

Selon Contandriopoulos et coll., « évaluer consiste à porter un jugement de valeur sur une intervention en mettant en œuvre un dispositif permettant de fournir des informations valides sur cette intervention de façon à ce que les acteurs concernés soient en mesure de prendre position sur l'intervention et de construire un jugement qui puisse se traduire en actions » (Contandriopoulos, Champagne et al. 2000). Dans le domaine de la santé, l'épidémiologie évaluative propose un ensemble de méthodes qui visent à produire de telles informations valides dans l'objectif de démontrer ou vérifier l'efficacité d'interventions pour améliorer la santé des individus (Breart and Bouyer 1991). Plus précisément, l'évaluation de recherche consiste à démontrer (ou non) l'existence d'une relation entre la mise en œuvre de ressources et/ou de pratiques et l'obtention de résultats exprimés le plus souvent sous forme d'effet sur l'état de santé. Pour cela, on compare généralement le résultat obtenu dans un groupe de sujets soumis à l'intervention que l'on désire évaluer, par rapport au résultat obtenu dans un groupe de sujets non soumis à cette intervention (autre intervention ou absence d'intervention), appelé groupe témoin. Aujourd'hui, la méthodologie de référence pour effectuer cette comparaison est l'essai contrôlé randomisé (ECR).

Comme nous l'exposerons dans la première partie de ce mémoire de thèse, si l'ECR permet de porter un jugement causal valide entre l'intervention mise en œuvre et les résultats obtenus, il n'est pas dénué d'un certain nombre de limites qui appellent à recourir à des méthodologies d'évaluation alternatives. Parmi elles, les études observationnelles comparatives ont l'inconvénient d'être sujettes à des biais, dont en particulier le biais d'indication, qui est un biais de confusion particulier résultant de l'attribution non aléatoire de l'intervention. Aussi la mise en œuvre d'études observationnelles pour l'évaluation des interventions en santé

nécessite d'appliquer des méthodes d'analyse appropriées pour maîtriser ce biais d'indication et obtenir des conclusions valides, utilisables pour construire un jugement sur l'intervention. Ce travail présentera les différentes méthodes d'analyse disponibles et développera leurs forces et faiblesses relatives.

Dans la seconde partie, les méthodes d'analyse seront testées dans trois exemples d'évaluation d'interventions à partir de données de pratiques courantes, recueillies dans le cadre de deux études observationnelles de cohorte.

Ainsi, ce travail de thèse explorera sur le plan à la fois théorique et pratique, l'intérêt, les apports et les limites des méthodes d'analyse applicables pour l'évaluation des effets des interventions en santé dans les études observationnelles.

A. PREMIÈRE PARTIE : ASPECTS THÉORIQUES SUR L'ÉVALUATION DE L'EFFET DES INTERVENTIONS EN SANTÉ

A.1. Les forces de la méthodologie de référence, l'essai contrôlé randomisé

Les principes méthodologiques de l'évaluation des effets des interventions en santé visent à limiter la survenue de biais qui conduiraient à des conclusions erronées.

Par ses caractéristiques, l'ECR est considéré comme le type d'étude de référence, qui permet de tirer les conclusions causales les plus fortes entre l'intervention évaluée et les résultats observés. En effet, l'ECR constitue un cadre expérimental qui permet de garantir les deux principes méthodologiques de base de l'évaluation :

- la nécessité d'une comparaison : dans un essai contrôlé, l'expérimentateur choisit, parmi l'ensemble des individus participants, ceux qui peuvent bénéficier de l'intervention. Cela lui permet de constituer un groupe témoin (non soumis à l'intervention) auquel sera comparé le groupe exposé. En l'absence de groupe témoin, il serait impossible de dire si une modification de l'état de santé des individus exposés est bien liée à l'intervention, ou si elle est la conséquence d'une autre cause concomitante ou encore si elle correspond à une évolution spontanée ;
- l'attribution aléatoire de l'intervention. Le choix du groupe de comparaison est fondamental : les groupes exposés et témoins doivent être comparables en tous points au début de l'étude, c'est-à-dire qu'ils doivent avoir un même niveau de risque initial par rapport au critère de jugement mesuré. Dans l'ECR, cette comparabilité des groupes est assurée par la randomisation : les sujets qui vont recevoir l'intervention sont tirés au sort parmi l'ensemble des participants à l'étude. C'est le hasard qui garantit la comparabilité des individus exposés et témoins pour l'ensemble des

variables mesurées et non mesurées. La méthode de référence est la randomisation individuelle, où l'affectation de chaque participant au groupe exposé ou témoin est décidée aléatoirement.

Certaines caractéristiques additionnelles permettent d'accroître la qualité méthodologique de l'ECR pour le jugement de causalité. Si la randomisation rend comparables les individus exposés et les témoins à l'inclusion dans l'étude, elle ne garantit pas qu'ils le demeurent pendant le déroulement de l'étude. Les individus des groupes comparés doivent en effet avoir les mêmes comportements de santé et bénéficier des mêmes interventions autres que celle évaluée tout au long de l'étude. Dans le cas contraire, les différences observées à la fin de l'étude ne pourront pas être imputées à la seule différence dans l'exposition à l'intervention. Deux moyens permettent de maintenir la comparabilité des groupes dans le temps :

- l'évaluation en aveugle : l'évaluation est réalisée en l'absence de connaissance, par les différents participants de l'étude (patient, médecin et évaluateur), du statut exposé/témoin des patients, afin d'éviter des modifications systématiques de leurs comportements qui biaiseraient les résultats (Boutron, Tubach et al. 2004). L'aveugle permet d'éviter la survenue d'un biais de performance (différence systématique dans l'administration de l'intervention : intervention réalisée différemment, meilleur suivi, meilleure observance, prescription de co-interventions…) et d'un biais de détection (différence systématique dans l'évaluation du résultat : appréciation du critère de jugement influencée par la connaissance de l'intervention, d'autant plus qu'il s'agit d'un critère subjectif). On parle de simple aveugle lorsque seuls les individus inclus n'ont pas connaissance de leur appartenance au groupe exposés ou témoins, de double aveugle lorsque, en plus, les personnes assurant le suivi des individus n'en n'ont pas connaissance, et de triple aveugle si les personnes qui évaluent les résultats ignorent

également le statut des individus par rapport à l'exposition. Dans l'évaluation des traitements médicamenteux, le choix d'un comparateur placebo indiscernable du médicament évalué permet habituellement d'assurer une évaluation en double aveugle. L'évaluation en aveugle est en revanche plus difficile, voire impossible, à mettre en œuvre pour les interventions non médicamenteuses. Lorsque l'aveugle n'est pas réalisable (on parle d'essai ouvert), des précautions doivent être prises : assurer un suivi identique entre les groupes, recueillir des informations sur les co-interventions reçues, choisir un critère de jugement aussi objectif que possible ou le faire évaluer par un observateur ignorant le traitement administré (lecture aveugle) ;

- l'analyse en intention de traiter : des écarts au protocole de l'étude (non-respect de l'intervention attribuée aléatoirement) sont fréquents, et peuvent être en rapport avec l'effet de l'intervention évaluée. Par exemple, un individu peut interrompre l'intervention qui lui a été allouée, parce qu'il en ressent des effets indésirables ou au contraire parce qu'il se sent guéri. Si on exclut ces individus de l'analyse, on détruit la comparabilité initiale des groupes et on introduit un biais d'attrition. L'analyse en intention de traiter est un choix *a priori* selon lequel tous les patients randomisés seront suivis jusqu'à la fin de l'essai et analysés dans leur groupe de randomisation quels que soient leur comportement au cours de l'étude (écarts au protocole).
- en outre, il est recommandé de choisir un unique critère de jugement principal, éventuellement associé à des critères de jugement secondaires, pour éviter l'inflation du risque d'erreur alpha induite par une multiplicité des comparaisons statistiques (risque de conclure à tort à l'efficacité de l'intervention).

Aujourd'hui, la méthode de référence en évaluation des nouveaux médicaments est l'ECR individuel en double aveugle. Par analogie, cette méthodologie a été étendue à

l'évaluation de toutes les interventions en santé. Elle permet de tirer les conclusions les plus solides en termes d'inférence causale, sa validité interne est excellente, mais n'est pas sans soulever parfois des difficultés – notamment lorsqu'il s'agit d'évaluer des interventions non médicamenteuses –, que nous aborderons dans le chapitre suivant.

A.2. Le nécessaire recours à des méthodologies alternatives

A.2.1. Les limites de l'essai contrôlé randomisé

Si l'ECR individuel en double aveugle est aujourd'hui considéré comme la méthode de référence pour l'évaluation de l'efficacité d'une intervention en santé, il n'est pas exempt d'un certain nombre de limites organisationnelles, de principe et/ou éthiques, qui légitiment le recours à des méthodes d'études alternatives, en complément ou en substitution à l'expérimentation. Black a ainsi recensé quatre principales raisons de mise en défaut de l'ECR, certaines étant spécifiques de l'ECR individuel en double aveugle, et d'autres concernent plus largement l'ensemble des méthodes expérimentales (Black 1996) :

- a : l'expérimentation peut ne pas être appropriée.

  (i) Si l'événement d'intérêt (le critère de jugement) est rare, la réalisation d'un ECR nécessitant un suivi prospectif est difficile, car elle implique l'inclusion d'un nombre de sujets important pour obtenir une puissance statistique suffisante. C'est notamment le cas pour l'étude des effets indésirables rares des interventions, qui sont habituellement des critères de jugement secondaires des essais, dont les effectifs sont calculés pour mettre en évidence un effet d'amplitude donnée sur un critère de jugement principal. De façon analogue, la nécessité d'importants effectifs rend difficile l'élaboration d'un ECR pour évaluer l'effet d'interventions de prévention d'événements de santé rares. Black cite l'exemple de l'évaluation de la position de sommeil des nourrissons en prévention de la mort subite, qui aurait nécessité un ECR incluant plusieurs centaines de milliers d'enfants pour répondre à la question.

  (ii) L'expérimentation, et plus largement les études prospectives, ne sont pas adaptées à l'évaluation d'interventions pour lesquelles le critère de jugement doit être mesuré longtemps après la mise en œuvre de l'intervention. Les

études observationnelles rétrospectives ont dans cette situation une faisabilité bien supérieure à l'ECR.

(iii) Certains contextes d'évaluation ne permettent pas la mise en œuvre de méthodes expérimentales dont la validité interne est la meilleure. Le double aveugle, qui permet de maintenir la comparabilité des groupes au cours d'une expérimentation, n'est pas toujours réalisable, en particulier pour les interventions non médicamenteuses qui impliquent la participation active des acteurs, effecteurs ou cibles de l'intervention (chirurgie, rééducation, interventions psycho-sociales…). Selon leurs croyances et préférences, les participants peuvent s'impliquer plus ou moins dans la mise en œuvre de l'intervention, influençant dans un sens ou dans l'autre les bénéfices potentiels de celle-ci. Les précautions méthodologiques supplémentaires évoquées plus haut (assurer un suivi identique entre les groupes, recueillir des informations sur les co-interventions reçues, choisir un critère de jugement aussi objectif que possible ou le faire évaluer par un observateur ignorant le traitement administré, p.9) ne permettent pas nécessairement de limiter le biais généré par l'absence d'aveugle.

- <u>b : l'expérimentation peut être impossible.</u>

(i) Lorsque l'intervention à évaluer est déjà mise en place, il n'est pas toujours possible d'envisager son évaluation par un ECR. D'abord pour des raisons pratiques, il peut être impossible d'identifier une population non-exposée au sein de laquelle il serait envisageable d'affecter aléatoirement l'intervention. Ensuite, pour des raisons éthiques, lorsque l'intervention est considérée comme une bonne pratique apportant un bénéfice important (malgré l'absence

d'évaluation) et qu'il n'est pas envisageable de constituer un groupe témoin sans intervention. Dans ces situations, les études observationnelles menées dans un premier temps pourraient permettre de générer une incertitude sur ce bénéfice, et rendre secondairement acceptable l'ECR, mais ce cas de figure est rarement rencontré en pratique.

(ii) Lorsque l'intervention ne peut raisonnablement pas être attribuée aléatoirement. Par exemple, si l'on veut savoir si les grands établissements de santé obtiennent de meilleurs résultats que les petits, l'orientation aléatoire des patients vers différentes structures de soins situées à des distances variables serait vraisemblablement inacceptable pour les patients.

(iii) Les expérimentations avec randomisation individuelle ne sont pas réalisables lorsque l'intervention ne peut pas être dirigée spécifiquement vers un individu, mais qu'elle est réalisée au bénéfice d'un groupe d'individus (par exemple, un programme d'éducation à la santé organisé dans une classe). L'expérimentation reste possible mais avec une randomisation au niveau des groupes d'individus (randomisation par grappes ou clusters). Cette alternative permet de répondre aux difficultés pratiques liées au ciblage de l'intervention et de réduire le risque de contamination entre les groupes, mais elle confronte à des difficultés méthodologiques supplémentaires (*cf.* paragraphe A.3.1(i), p.20).

(iv) Les interventions dans le champ de la santé sont innombrables et comprennent la plupart du temps de nombreuses composantes. Par exemple, l'effet d'un geste chirurgical est dépendant de l'ensemble des autres maillons d'une chaîne de soins à laquelle il appartient, comme la consultation pré-anesthésique, l'anesthésie, les soins post-opératoires immédiats ou les soins post-opératoires

ambulatoires. Pour des raisons pratiques, toutes ces composantes ne peuvent faire l'objet d'une évaluation expérimentale. Le plus souvent, des ECR sont mis en œuvre pour évaluer les composantes principales des interventions, faisant l'impasse sur d'autres composantes jugées secondaires. De mise en œuvre plus aisée, les études observationnelles permettent d'explorer des champs de recherche délaissés par les méthodes expérimentales, et notamment le rôle respectif de chaque composante de l'intervention.

(v) Rarement, des obstacles politiques ou législatifs peuvent empêcher la réalisation d'une expérimentation. Black cite l'exemple de la kératotomie radiaire qui était déjà pratiquée par les ophtalmologistes aux États-Unis avant qu'on envisage d'en faire l'évaluation par une expérimentation. Cette pratique aurait alors été reconsidérée comme expérimentale par les assureurs des praticiens, qui auraient majoré leurs cotisations. L'opposition des ophtalmologistes a abouti à l'abandon du projet d'évaluation expérimentale de la kératotomie radiaire. De tels obstacles sont probablement anecdotiques, sauf dans certains domaines comme celui de la santé au travail, où des enjeux sociaux ou économiques peuvent faire obstacle à la réalisation d'une expérimentation.

(vi) Enfin, à l'évidence, une expérimentation n'est pas réalisable lorsque les intervenants clé de l'intervention refusent d'y participer. Black évoque le cas de chirurgiens convaincus de l'efficacité d'une intervention et refusant de la remettre en question par une expérimentation.

- c : l'expérimentation peut être inadéquate, en ce que l'expérimentation se déroule dans des conditions très particulières : l'expérimentateur se place dans des conditions

optimales pour démontrer l'efficacité d'une intervention, qui ne sont pas les conditions observées dans la pratique courante. On oppose ainsi l'*'efficacy'* d'une intervention, qui est son effet théorique dans un cadre expérimental optimal et bien défini, à son *'effectiveness'*, qui est son effet réellement observé lorsqu'on met en œuvre l'intervention de façon courante dans une population (Barreto 2005). Dans ce contexte, l'ECR pose le problème de la validité externe de ses résultats, c'est-à-dire que les conclusions tirées de l'expérimentation ne seront pas nécessairement transférables dans la pratique courante. Trois principaux facteurs font obstacle à la généralisation des résultats d'un ECR à la pratique courante :

(i) Les professionnels de santé qui participent aux études expérimentales ne sont pas représentatifs de leurs pairs. Ils ont le plus souvent une expertise dans le domaine étudié et manifestent un intérêt particulier pour l'intervention mise en œuvre.

(ii) Les patients recrutés dans les essais ne sont pas représentatifs de l'ensemble des patients. Ils sont sélectionnés sur des critères stricts qui excluent une part significative des patients rencontrés dans la pratique clinique habituelle (par exemple, les patients inclus sont souvent plus jeunes et présentent peu de comorbidités significatives).

(iii) Le contexte de l'expérimentation n'est pas représentatif de la pratique courante. En effet, les patients bénéficient pendant l'étude d'une attention particulière (qu'ils reçoivent ou non l'intervention évaluée), d'un suivi rapproché dans le cadre de protocoles stricts qui tendent à maximiser l'observance. De plus, le comportement des participants à un essai peut être modifié par la seule raison qu'ils se savent observés : c'est l'effet Hawthorne (Gale 2004).

- d : l'expérimentation peut ne pas être nécessaire : lorsque l'effet de l'intervention est très important, l'influence éventuelle de facteurs de confusion peut être considérée comme négligeable. Cette situation est aujourd'hui rare. Historiquement, l'efficacité d'antibiotiques sur des maladies infectieuses mortelles a ainsi pu être montrée sans expérimentation. Un autre exemple est l'évaluation de l'immobilisation orthopédique des fractures osseuses. Sur un ton humoristique, Smith et Pell ont souligné qu'aucun ECR n'avait évalué le bénéfice du parachute dans la chute libre (Smith and Pell 2003).

Parmi cette liste de motifs s'opposant à la réalisation d'un ECR, il existe un certain nombre d'obstacles d'ordre pratique qui sont théoriquement surmontables. Cependant, ils ne pourraient l'être qu'au prix d'une mobilisation de ressources d'une telle importance que la faisabilité de l'essai en serait largement compromise (par exemple, à la condition que l'expérimentation soit réalisée sur un échantillon de plusieurs dizaines de milliers d'individus, ou avec une durée de suivi très longue).

A.2.2. La perspective singulière des interventions complexes en santé

L'ECR a été conçu pour évaluer l'efficacité des thérapeutiques médicamenteuses. Or, les interventions en santé peuvent être de nature très variée et être administrées à un niveau individuel ou collectif. Il peut s'agir par exemple de chirurgie, de psychothérapie, d'éducation thérapeutique, d'un réseau de soins pluridisciplinaire spécialisé dans une maladie, de rééducation, d'un programme de prévention, d'un programme favorisant l'accès aux soins, d'une politique de réduction de risques environnementaux, etc. Ces interventions non

médicamenteuses s'opposent à l'administration d'un médicament par leur complexité, justifiant leur qualification d'interventions complexes (IC). Le cadre méthodologique des IC proposé par le *Medical Research Council* prend en compte les différentes dimensions de cette complexité (Craig, Dieppe et al. 2008) : (i) le nombre de composantes de l'intervention et leurs interactions, entre elles et avec le contexte environnemental ; (ii) le nombre et la difficulté des comportements à adopter par les bénéficiaires de l'intervention et ceux qui la délivrent ; (iii) le nombre de groupes ou de niveaux organisationnels ciblés ; (iv) le nombre et la variabilité des résultats ; (v) le degré de flexibilité ou d'adaptation autorisée de l'intervention. L'évaluation des IC soulève l'insuffisance d'une approche limitée aux résultats et invite à s'intéresser également aux processus, en mettant en œuvre de nouvelles approches méthodologiques (Ridde and Haddad 2013).

Les recommandations du *MRC* sur l'évaluation des IC mettent ainsi en avant la nécessité de prendre en compte les différents éléments de l'intervention en s'intéressant non seulement aux résultats, mais également au contexte, à la mise en œuvre et aux mécanismes (Craig, Dieppe et al. 2008) :

- le contexte : il s'agit d'identifier les éléments extérieurs à l'intervention qui peuvent avoir un impact sur la mise en œuvre de l'intervention, les mécanismes et/ou les effets. Une bonne connaissance du contexte permet d'interpréter les résultats observés et de discuter leur généralisation dans des contextes différents ;
- la mise en œuvre : comment l'intervention a-t-elle été effectivement délivrée ? L'évaluation de la fidélité de mise en œuvre de l'intervention par rapport au protocole initial implique de prendre en compte deux aspects : la réalisation des activités prévues et la participation réelle de la population cible.
- les mécanismes d'impact : les mécanismes par lesquels l'intervention produit ou non ses effets doivent être étudiés. La longueur et la complexité de la chaîne causale reliant

l'intervention et ses effets incitent généralement à recourir à plusieurs critères de jugement répartis tout au long de cette chaîne.

Cette prise en compte de toute la complexité de l'intervention permet d'interpréter correctement les résultats observés :

- en cas de résultats négatifs : faire la différence entre une intervention inefficace et une intervention qui n'a pas été correctement mise en œuvre (et dans ce cas, en comprendre les raisons pour y remédier) ;
- en cas de résultats positifs : comprendre comment l'intervention produit ses effets et comment elle peut être optimisée.

De plus, cette évaluation de l'ensemble du processus permet de discuter la transférabilité des résultats dans un environnement différent (Cambon, Minary et al. 2012).

Pour l'évaluation des résultats, l'ECR individuel est encouragé (Craig, Dieppe et al. 2008). Cependant, les IC constituent un objet d'évaluation particulier qui confronte habituellement l'évaluateur aux limites de la méthodologie classique de l'ECR, notamment parce que l'aveugle ou la randomisation individuelle ne sont pas possibles. Des alternatives méthodologiques sont nécessaires (*cf.* paragraphe A.3.1, p.20). En outre, le cadre méthodologique d'évaluation des IC invite à identifier plusieurs critères de jugement à différents niveaux des processus et à compléter l'approche quantitative par des méthodes qualitatives et/ou mixtes.

L'esprit de cette démarche d'évaluation des IC promue par le *MRC* est partagé par l'approche '*realist'* utilisée en sciences sociales, qui cherche à comprendre ce qui fonctionne, comment, dans quelles conditions et pour qui, en s'intéressant au contexte, aux mécanismes et aux résultats, plutôt que répondre trivialement à la question de savoir si l'intervention marche (Connelly 2007).

A.3. Les méthodologies alternatives

La méthodologie classique de l'ECR, avec randomisation individuelle et en double aveugle, n'est donc pas toujours possible ou souhaitable, en particulier pour l'évaluation des résultats d'une intervention non médicamenteuse. Des alternatives, expérimentales et non expérimentales, existent. Nous aborderons d'abord les études expérimentales dérivées de l'ECR, puis les études observationnelles.

A.3.1. Les alternatives expérimentales randomisées

De nombreuses adaptations de la méthodologie classique de l'ECR (avec randomisation individuelle et double aveugle) ont été proposées pour répondre à certaines de ses limites. Elles demeurent des méthodes expérimentales randomisées (en totalité ou en partie).

(i) Méthodes alternatives à la randomisation individuelle

Lorsque la randomisation n'est pas réalisable à l'échelle individuelle, ou lorsqu'on craint une contamination trop importante entre les groupes, la randomisation peut être effectuée à l'échelle d'un groupe d'individus : on parle alors d'essai randomisé en grappes ('*cluster-randomised trial'*). Les conséquences méthodologiques ne sont pas neutres : la randomisation par grappes nécessite, à puissance égale, d'inclure davantage d'individus que la randomisation individuelle, complexifie l'analyse statistique, et est moins performante (les individus des groupes exposé et témoin ne seront pas nécessairement comparables).

Les études avec randomisation séquentielle en grappes de type '*stepped wedge'* proposent une adaptation de la randomisation en grappes conventionnelle « en parallèle » (**Figure I**), lorsqu'il ne semble pas éthique qu'un groupe d'individus ne bénéficie pas de l'intervention

(Hemming, Haines et al. 2015). Au départ, aucune grappe ne bénéficie de l'intervention, puis celle-ci est mise en œuvre séquentiellement, grappe après grappe, dans un ordre aléatoire. Cette approche améliore la faisabilité de l'étude, car la mise en œuvre de l'intervention est progressive, et est plus acceptable sur le plan éthique que les grappes parallèles car l'ensemble des grappes mettront en place l'intervention, à un moment ou un autre de l'étude. Toutefois, les résultats peuvent être biaisés par l'évolution temporelle spontanée du critère de jugement, les non-exposés étant en moyenne observés plus tôt que les exposés.

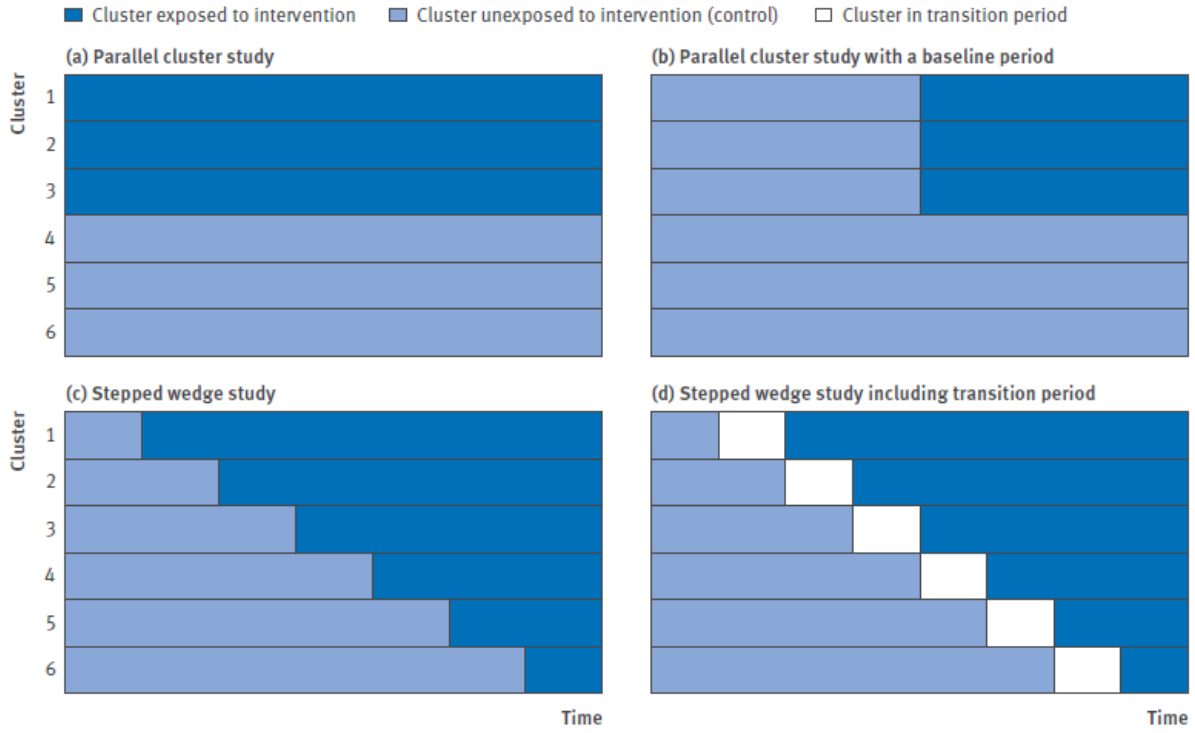


Figure I: Illustration des schémas d'essai avec randomisation en grappes. (a) et (b) Type classique avec grappes parallèles. (c) et (d) Type séquentiel '*stepped wedge*'. Tiré de (Hemming, Haines et al. 2015).

(ii) Méthodes permettant de pallier les difficultés liées à l'absence d'aveugle

Lorsqu'une expérimentation ne peut pas être réalisée en aveugle, la préférence du patient (ou de l'investigateur) pour l'une ou l'autre des interventions peut biaiser l'évaluation. En effet,

dans un essai ouvert, le comportement des individus est modifié lorsqu'ils ont une préférence forte pour l'une ou l'autre des interventions comparées : (i) ils peuvent refuser de participer à l'étude s'ils n'ont pas la garantie de bénéficier de l'intervention qu'ils préfèrent (à l'origine d'un biais de sélection) ; (ii) ils peuvent arrêter de participer à l'étude s'ils ne sont pas randomisés dans le groupe recevant l'intervention qu'ils préfèrent (risque de perte de la comparabilité initiale des groupes) ; (iii) s'ils participent à l'étude mais qu'ils ne bénéficient pas de l'intervention qu'ils préfèrent, ils peuvent être démotivés, avoir une observance moins bonne, et ressentir un effet de l'intervention moins important (possiblement à l'origine d'une sous-estimation de l'effet de l'intervention) ; (iv) inversement, s'ils participent à l'étude et qu'ils bénéficient de l'intervention qu'ils préfèrent, leur observance sera vraisemblablement meilleure et l'effet ressenti plus important (possiblement à l'origine d'une surestimation de l'effet de l'intervention).

Les '*preference trials'* prennent en compte la préférence du patient (ou de l'investigateur) pour limiter les biais inhérents à l'absence d'aveugle (Torgerson and Sibbald 1998). Ce type d'essais ajoute au schéma classique de la comparaison de deux groupes randomisés (intervention A *versus* intervention B) une comparaison des groupes d'individus recevant une intervention attribuée non pas aléatoirement, mais selon leur préférence. Ces études prennent ainsi en compte d'une manière ou d'une autre la préférence des individus dans la constitution des groupes. Lors de l'analyse, l'effet de l'intervention est évalué séparément dans les groupes où l'intervention a été attribuée aléatoirement et ceux où elle a été attribuée selon la préférence, afin d'évaluer l'influence de la préférence dans les résultats. Ces études ont l'avantage d'associer les forces d'un ECR à l'étude de l'influence de la préférence. Cependant, elles ont l'inconvénient d'être complexes et coûteuses à mettre en œuvre, et posent des problèmes méthodologiques pour l'analyse des groupes non randomisés.

Il existe de nombreuses variantes de '*preference trials'* qui donnent une latitude décisionnelle variable aux individus dans le choix de l'intervention qu'ils reçoivent, à des étapes différentes du protocole de l'étude. Les principaux schémas d'études de type '*preference trials'* sont les suivants :

- *Comprehensive cohort study –CCS–* (Olschewski, Schumacher et al. 1992) ou *Partially randomised preference trial –PRPT–* (Brewin and Bradley 1989) : une cohorte d'individus éligibles et consentant à participer à une étude est constituée (**Figure II**). Il est proposé à tous de bénéficier d'une intervention attribuée aléatoirement. Certains individus acceptent (ceux qui n'ont pas de préférence forte pour une intervention), d'autres non. L'ensemble de la cohorte est suivie, que les sujets aient accepté ou non la randomisation. Les sujets qui ont refusé la randomisation, c'est-à-dire ceux qui ont une préférence forte pour l'une ou l'autre des interventions, bénéficient de l'intervention qui a leur préférence.

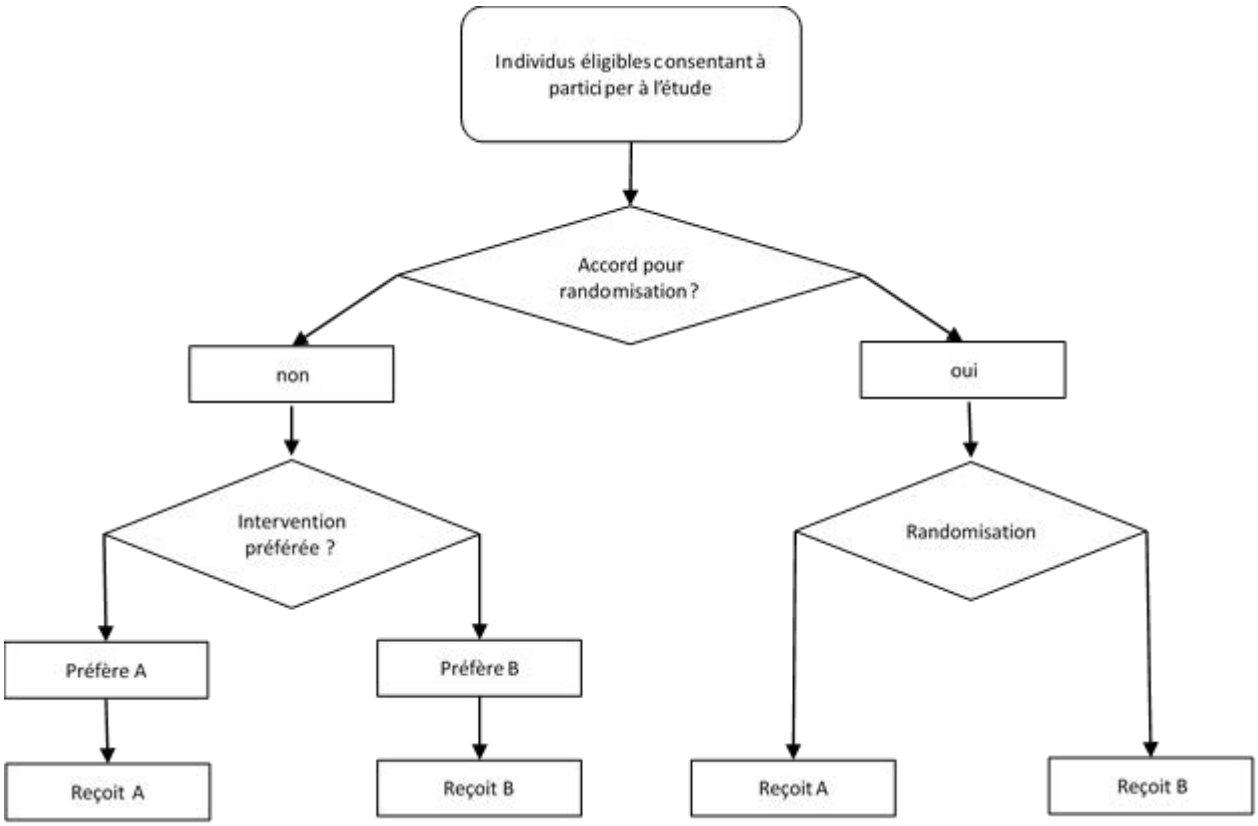


Figure II : Schéma d'étude de type *Comprehensive cohort study*.

- Schéma de Zelen (**Figure III**)(Zelen 1990) : il peut être utilisé pour comparer une nouvelle intervention à une intervention témoin déjà couramment mise en œuvre. L'originalité de cette méthode est que les individus sont randomisés avant même d'avoir consenti à participer à l'étude. Les individus randomisés dans le groupe témoin vont recevoir l'intervention courante. Ceux randomisés dans le groupe intervention seront interrogés sur leur consentement à recevoir la nouvelle intervention. En cas de refus, ils recevront l'intervention courante mais seront analysés dans le groupe bénéficiant de la nouvelle intervention (analyse en intention de traiter). La réalisation de la randomisation avant la sollicitation des individus et même l'absence de consentement de certains participants rend ce schéma contestable sur le plan éthique.

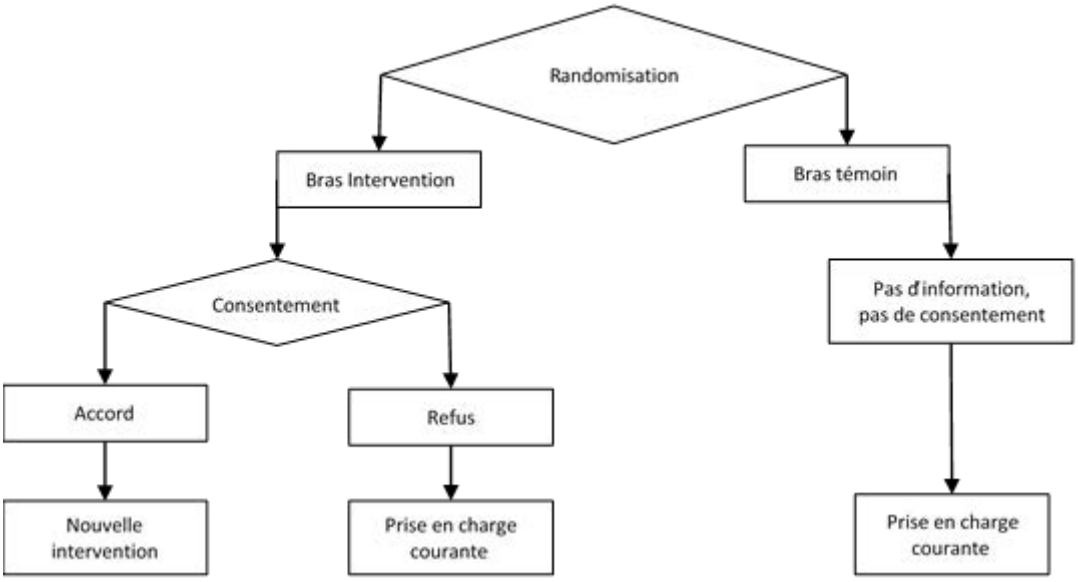


Figure III : Schéma d'étude de type Zelen.

- Schéma de Wennberg (**Figure IV**)(Salkind 2010) : les individus d'une étude sont randomisés dans deux groupes, l'un où l'intervention est attribuée aléatoirement, l'autre où elle l'est selon la préférence des individus. À la différence de la *CCS* et du

schéma de Zelen, l'avantage de cette méthode est que l'ensemble des individus sont consentants et randomisés.

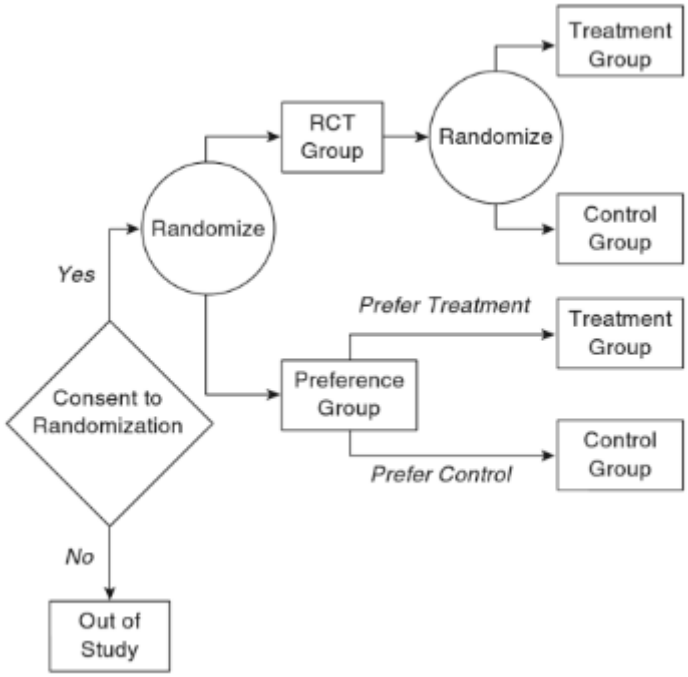


Figure IV : Schéma d'étude de type Wennberg. Tiré de (Salkind 2010).

- *Cohort multiple randomised controlled trial –cmRCT–* (**Figure V**)(Relton, Torgerson et al. 2010) : cette méthode est utilisable lorsqu'on veut comparer une nouvelle intervention à l'intervention couramment pratiquée. À partir d'une cohorte préexistante de patients présentant la maladie d'intérêt et suivis régulièrement, certains patients sont tirés au sort pour participer à un ECR évaluant une nouvelle intervention. Seuls les patients à qui l'intervention nouvelle à évaluer est proposée vont être informés de l'ECR. Ils seront comparés aux autres patients de la cohorte. Cette approche facilite le recrutement de patients représentatifs de la pratique courante et permet d'évaluer plusieurs interventions par différents *cmRCT* à partir de la même cohorte.

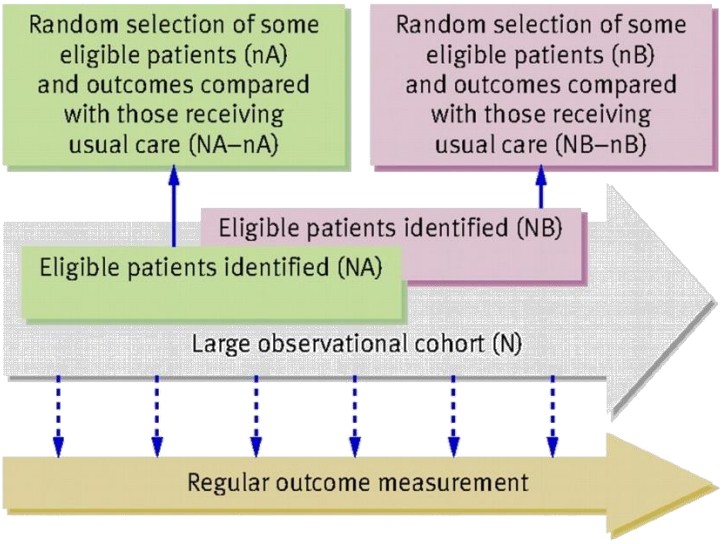


Figure V : Schéma d'étude de type *cohort multiple randomised controlled trial*. A partir d'une cohorte N, réalisation d'un ECR pour évaluer l'intervention A, puis d'un autre ECR pour évaluer l'intervention B. Tiré de (Relton, Torgerson et al. 2010).

- *Balanced incomplete block design* (**Figure VI**)(Trietsch, Leffers et al. 2014). Cette appellation est parfois utilisée pour désigner un schéma expérimental permettant d'évaluer simultanément deux interventions destinées à deux problèmes de santé différents mais relativement proches. Par exemple, on veut évaluer deux programmes de kinésithérapie différents destinés à deux affections rhumatologiques A et B. L'échantillon étudié est constitué à la fois d'individus atteints de la maladie A et d'autres atteints de la maladie B. Chaque intervention va constituer un bras de traitement, dans lequel des patients atteints des deux affections vont être randomisés. Ainsi, tous les patients vont bénéficier d'une intervention, ce qui doit permettre de limiter les biais liés à la préférence et au contexte expérimental (effet Hawthorne). Ce schéma est particulièrement complexe à mettre en œuvre et a des indications limitées.

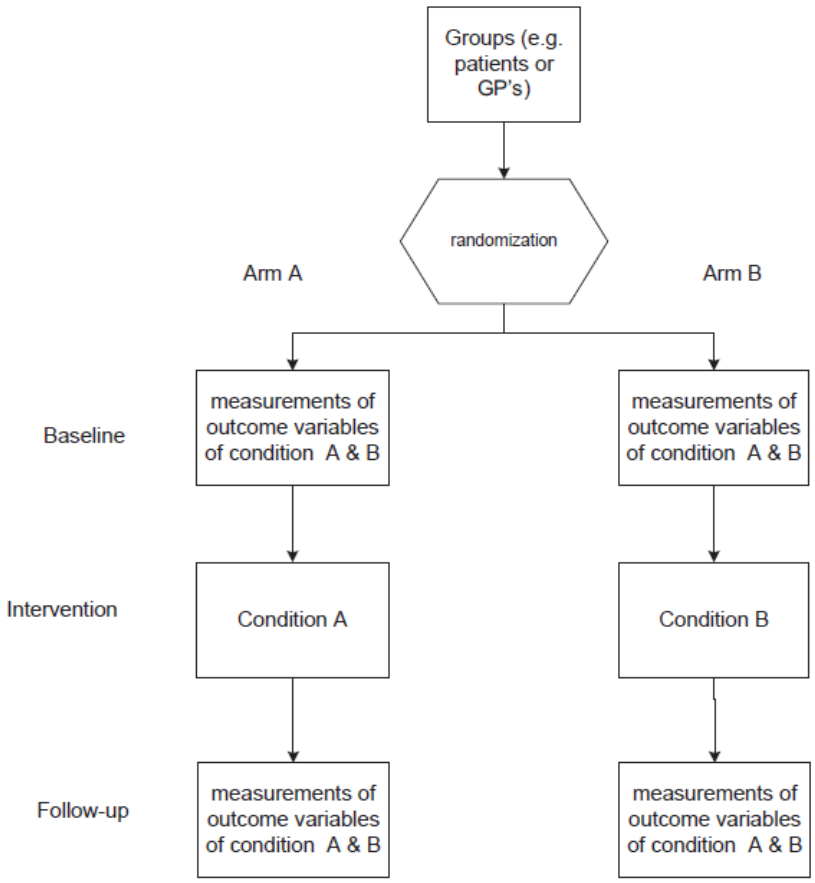


Figure VI : Schéma d'étude parfois qualifié de type *balanced incomplete block*. Tiré de (Trietsch, Leffers et al. 2014).

(iii) Méthodes pour rapprocher le contexte expérimental de la pratique courante

Les '*pragmatic trials'* se rapprochent des conditions de la pratique courante, dans une démarche dite pragmatique par opposition à l'attitude plus conventionnelle dite explicative ('*explanatory trials'*) qui privilégie la validité interne (Godwin, Ruhland et al. 2003, Treweek and Zwarenstein 2009). Il s'agit bien d'ECR, mais ces essais pragmatiques incluent des populations hétérogènes pas ou peu sélectionnées et accordent plus de souplesse dans la mise en œuvre de l'intervention.

Dans l'ensemble, ces méthodologies expérimentales alternatives sont souvent complexes à mettre en œuvre, puisqu'elles conservent les difficultés pratiques de réalisation d'un ECR, en y ajoutant d'autres contraintes. Certains schémas ont des indications limitées, d'autres confrontent à des problèmes éthiques. Aucun n'apporte de réponse à l'ensemble des limites de l'ECR individuel en double-aveugle. Ces schémas poursuivent simultanément deux objectifs, la validité interne et la validité externe, qui se comportent comme des vases communicants, si bien que le compromis final obtenu entre '*efficacy'* et *'effectiveness'* n'est pas toujours très clair (Marchal, Westhorp et al. 2013). *In fine*, l'utilisation de ces adaptations de l'ECR reste aujourd'hui rare.

A.3.2. Les alternatives observationnelles comparatives (quasi-expérimentales)

La critique de la faible validité externe de l'ECR classique est au cœur du concept de '*Comparative Effectiveness Research'* (*CER*) qui a émergé à la fin des années 2000 (Sox 2010). La *CER* a été d'abord définie par l'*Institute Of Medicine (IOM)* comme « la production et la synthèse de données permettant de comparer les bénéfices et risques de différentes méthodes pour prévenir, diagnostiquer, traiter et surveiller des maladies, ou d'améliorer la délivrance des soins » (IOM 2009), puis par le *Federal Coordinating Council of CER* comme « la conduite et la synthèse des recherches comparant les bénéfices et risques d'interventions variées et de stratégies pour prévenir, diagnostiquer, traiter et suivre des états de santé en conditions réelles de soins » (Conway and Clancy 2009). Selon l'*IOM*, l'objectif de la *CER* est « d'assister les utilisateurs, les effecteurs, les financeurs et les décideurs du système de soins afin que soient prises des décisions qui améliorent les soins à la fois au niveau

individuel et au niveau populationnel » (IOM 2009). Il s'agit de privilégier l'*effectiveness* plutôt que l'*efficacy*. Pragmatiquement, l'idée de base de la *CER* est de pallier les limites de l'ECR par des données d'évaluation complémentaires issues de la pratique courante. Si l'exploitation de grandes bases de données médico-administratives est souvent mise en avant (Hershman and Wright 2012), cette démarche encourage plus largement la réalisation d'études observationnelles comparatives (encore appelées méthodes quasi-expérimentales)(Concato, Lawler et al. 2010).

Dans le cadre l'évaluation des interventions en santé, il existe deux grands types d'études observationnelles comparatives : les études « exposés – non exposés » ou de cohorte, et les études « cas – témoins ». Dans l'étude « exposés – non exposés », certains individus vont être exposés à l'intervention, les autres non. L'exposition des sujets à l'intervention n'est pas issue du choix d'un expérimentateur, mais est le fruit d'un processus décisionnel spontané qui n'est pas connu de l'évaluateur. À la fin de l'étude, on compare l'état de santé des sujets exposés à l'intervention à celui des sujets qui n'y ont pas été exposés. Dans l'étude « cas – témoins », on identifie d'abord des individus présentant la maladie d'intérêt (les cas) et d'autres qui en sont exempts (les témoins). Ensuite, on compare le taux d'exposition à l'intervention dans les deux groupes. Par exemple, si une intervention préventive est moins prévalente chez les témoins que chez les cas, c'est qu'elle a produit l'effet escompté (sous réserve d'éventuels biais). Cette méthode est appropriée quand l'évènement de santé étudié est rare.

Depuis une quinzaine d'années, on note un regain d'intérêt pour les expérimentations dites naturelles (‘*natural experiment’*), illustrant, comme la *CER*, le besoin de produire des données d'évaluation en conditions réelles, et donc encourageant la recherche observationnelle comparative (Petticrew, Cummins et al. 2005, Craig, Cooper et al. 2012). Les expérimentations naturelles n'ont pas de définition précise. Si l'attribution de l'intervention n'est pas forcément naturelle, c'est-à-dire étrangère au fait de l'homme, elle n'est en tout cas

pas le fait de l'évaluateur. Ce dernier ne contrôle pas l'administration de l'intervention, mais mesure la variation de l'exposition pour en estimer les effets. L'intervention est une modification de l'environnement des individus, qui est spontanée (« naturelle ») ou décidée à un niveau macroscopique, souvent politique, et qui affecte largement une zone géographique ou une population prédéterminée. Par exemple, on peut étudier ainsi les effets sur la santé induits par l'augmentation du prix du paquet de cigarette ou la pollution atmosphérique. Les expérimentations naturelles permettent d'évaluer un type particulier d'interventions pour lesquelles une approche expérimentale randomisée n'est pas envisageable. Elles sont appropriées lorsque les effets attendus sont importants et avec une latence courte, et nécessitent de disposer d'un système de mesure fiable de l'exposition et du critère de jugement sur une large population.

A.4. Les problèmes méthodologiques spécifiques posés par les études observationnelles

Toute étude d'évaluation peut être sujette à des biais, erreurs systématiques qui altèrent la validité de ses résultats. Il existe trois grands types de biais : le biais de sélection, qui concerne la constitution de l'échantillon étudié ; le biais de classement (également appelé biais de mesure ou d'information), relatif à la mesure de l'exposition ou du critère de jugement ; le biais de confusion, qui perturbe la mesure de l'association entre l'exposition et le critère de jugement. Lorsqu'on étudie l'effet d'une intervention en situation observationnelle, ces trois types de biais peuvent être rencontrés, mais le biais de confusion est celui qui mérite une attention particulière. De plus, les résultats d'études observationnelles de cohorte peuvent être faussés par la survenue d'un biais spécifique, le biais de temps immortel, qui concerne la définition du facteur d'exposition étudié et des périodes d'exposition, et qui est un biais de classement ou de sélection selon les cas.

A.4.1. Le biais de confusion

Un biais de confusion se produit quand un facteur associé à l'exposition (sans qu'il en soit une conséquence) est également un déterminant indépendant du résultat mesuré, et que ce facteur n'est pas identiquement réparti entre les deux groupes étudiés. En d'autres termes, la confusion crée une association apparente entre l'exposition et le critère de jugement, alors qu'il n'existe pas de relation de cause à effet entre eux. Dans l'expérimentation, la randomisation permet de s'assurer que les individus exposés et les témoins sont comparables en tous points, excepté l'exposition, et donc de s'affranchir d'un biais de confusion. En revanche, dans une étude observationnelle, par définition, l'intervention évaluée n'est pas allouée aléatoirement, mais en fonction de caractéristiques diverses (du patient, du soignant

ou de l'environnement). Ces caractéristiques, pas nécessairement mesurées ni même connues de l'évaluateur, peuvent avoir un effet sur le critère de jugement et donc biaiser les résultats de l'évaluation. On désigne habituellement ce biais de confusion comme le biais d'indication. Le biais d'indication est donc un type particulier de biais de confusion qui survient lorsque l'intervention est allouée en fonction d'une (ou plusieurs) caractéristique(s) qui est (sont) liée(s) au critère de jugement.

D'un point de vue terminologique strict, le terme de biais d'indication devrait être réservé à la situation où c'est la pathologie justifiant la mise en œuvre de l'intervention qui constitue en elle-même un facteur de confusion (Salas, Hofman et al. 1999). Cependant, ce terme est usuellement utilisé pour désigner plus globalement plusieurs sous-types de biais de confusion générés par l'allocation de l'intervention par un processus spontané qui fait intervenir de façon complexe le patient, le personnel soignant et le système de santé (Walker 1996, Salas, Hofman et al. 1999, Brookhart, Sturmer et al. 2010) :

- La confusion induite par la gravité de la pathologie : la prescription de l'intervention est souvent liée à la gravité de la maladie prise en charge, l'intervention étant réservée aux formes les plus graves de la maladie, ou, au contraire, aux formes débutantes ;
- La confusion induite par le pronostic de la pathologie : ce biais est assez proche du précédent, sauf que dans ce cas l'indication de l'intervention est portée non pas sur la gravité actuelle de la pathologie, mais sur son évolution prévisible, compte tenu de nombreux facteurs tels que l'âge ou l'existence de comorbidités ;
- La confusion induite par le pronostic vital : les patients dont la durée de vie est limitée par des problèmes de santé importants vont moins fréquemment se voir prescrire des interventions à visée préventive. Ou, si ces interventions étaient réalisées, elles pourraient être interrompues à l'approche de la fin de vie. Ce phénomène peut générer une surestimation de l'effet des interventions de prévention ;

- La confusion induite par le statut fonctionnel ou la déficience cognitive : dans ce cas, le facteur de confusion n'est pas la pathologie ciblée par l'intervention évaluée, mais l'état général des individus. En effet, les sujets ayant une diminution de leur autonomie fonctionnelle au quotidien ou un déficit cognitif ont souvent un accès restreint au système de santé. Elles peuvent de fait être plus difficilement ciblées par l'intervention, alors qu'elles sont plus vulnérables. Ce biais peut donc entraîner une surestimation de l'effet de l'intervention ;
- La confusion induite par un facteur de risque partagé par la maladie cible de l'intervention et une autre maladie, lorsque la présence de ce facteur de risque influence la probabilité de recevoir l'intervention : soit E une intervention proposée à des individus présentant la maladie A1. À la fin de l'étude d'évaluation de E, on observe une incidence plus importante de la maladie A2 dans le groupe intervention, comparativement au groupe témoin, suggérant que E provoque la maladie A2. En fait, il n'y a pas de lien causal entre E et A2. C'est l'existence d'un facteur de risque commun U des maladies A1 et A2, également associé à E, qui explique l'association statistique entre E et A2 : la présence du facteur U chez un individu augmente la probabilité qu'il reçoive E, les individus du groupe intervention ont donc un risque initial augmenté de présenter la maladie A2 ;
- La confusion induite par l'utilisateur sain ou la bonne observance : les individus qui font le choix de bénéficier d'une intervention (en particulier préventive) ont souvent des comportements de vie globalement plus sains que les autres (par exemple ils adhèrent davantage aux programmes de dépistage, font plus d'activité physique ou encore consomment moins d'alcool). Il en est de même des individus qui suivent fidèlement les prescriptions faites. Là encore, l'effet de l'intervention peut être surestimé ;

- La confusion induite par l'accès au système de santé : les facteurs qui influencent l'accès à l'intervention évaluée sont bien souvent associés à de nombreux autres facteurs économiques, sociaux, éducatifs, culturels… qui influencent directement la santé.

On remarque ainsi que certains facteurs de confusion sont liés à la pathologie ciblée par l'intervention, ce sont les facteurs qui paraissent habituellement les plus évidents, mais bien d'autres facteurs de confusion n'ont pas de lien spécifique avec cette pathologie (caractéristiques des individus ou de leur environnement).

Les méthodes utilisables pour limiter le biais de confusion seront développées dans le chapitre 0 (p.38). Il est important de souligner que les facteurs de confusion peuvent être connus de l'évaluateur, qu'ils soient mesurés ou non au cours de l'étude, ou inconnus. Par conséquent, même si le biais généré par des facteurs de confusion mesurés est maîtrisé, un biais résiduel est possible du fait de facteurs de confusion non mesurés. Des méthodes telles que les analyses de sensibilité ('*sensitivity analyses'*), permettent d'estimer l'impact potentiel d'un facteur de confusion non mesuré (Greenland 1996).

En outre, le biais d'indication ne doit pas être confondu avec deux autres biais de nature différente, le biais protopathique et le biais de sélection (Salas, Hofman et al. 1999). Le terme de biais d'indication est parfois utilisé abusivement pour désigner le biais protopathique Le biais protopathique se produit lorsqu'une intervention est réalisée en réponse à un symptôme qui est en fait un signe d'une pathologie encore non diagnostiquée. On peut alors conclure à tort que l'intervention est la cause de la pathologie. On ne peut pas ici parler de biais d'indication, le biais étant généré par l'antériorité des premiers signes de la pathologie par rapport à l'exposition (Walker 1996). Quant au biais de sélection, il est bien différent du biais

d'indication, puisqu'il résulte des modalités de constitution de la population de l'étude, et non des modalités d'allocation de l'exposition à l'intervention évaluée au sein de cette population.

A.4.2. Le biais de temps immortel

Identifié initialement dans les années 1970, le biais de temps immortel suscite un fort regain d'intérêt dans les années 2000 suite aux travaux de Suissa, révélant que de nombreuses études de cohorte en pharmaco-épidémiologie étaient sujettes à ce biais, aboutissant à des conclusions faussement favorables à l'intervention évaluée (Suissa 2008).

Le biais de temps immortel peut apparaître lors des analyses de survie, c'est-à-dire lorsque que le critère de jugement est le temps de survenue d'un événement, alors que l'exposition à l'intervention peut survenir au cours du suivi. Le temps immortel est une période du suivi d'un individu de la cohorte pendant laquelle l'événement d'intérêt (non récidivant) n'a pas pu se produire. Lorsque l'exposition d'un sujet à l'intervention survient au cours de son suivi, le temps entre le début du suivi et le début de l'intervention est dit immortel, car le sujet a nécessairement dû ne pas présenter l'événement d'intérêt pendant cette période de temps (**Figure VII**, page suivante). Cette période particulière de non-exposition à l'intervention génère un biais, dit biais de temps immortel, si elle n'est pas prise en compte correctement dans l'analyse (c'est-à-dire si elle est considérée comme une période exposée ou si elle est exclue du suivi). En effet, la période de temps immortel donne un avantage de survie artificiel aux individus du groupe exposé. On peut ainsi conclure à tort à un effet de l'intervention, ou surestimer son effet réel.

Ce biais a été mis en évidence pour la première fois dans deux études concluant à l'amélioration de la survie par la transplantation cardiaque (Gail 1972). Le bénéfice de l'intervention disparaissait en prenant en compte la période d'attente pré-transplantation, qui

est une période de temps immortel, conférant un avantage de survie artificiel aux sujets transplantés. En 2007, Suissa a illustré l'importance du biais de temps immortel en identifiant ce biais dans 20 études observationnelles récentes concluant à un effet important d'une intervention (c'est-à-dire permettant de réduire de 25 à 50% la morbidité ou la mortalité)(Suissa 2007). Dans le même travail, il a ré-analysé les données d'une étude de cohorte de 3315 patients atteints de bronchite chronique, dont la première analyse effectuée sans précaution par rapport au biais de temps immortel, suggérait l'efficacité de deux médicaments sur la mortalité à 1 an, avec une réduction significative de l'ordre de 25%. En ré-analysant les données en considérant correctement le temps immortel, plus aucun effet sur la mortalité n'était mis en évidence.

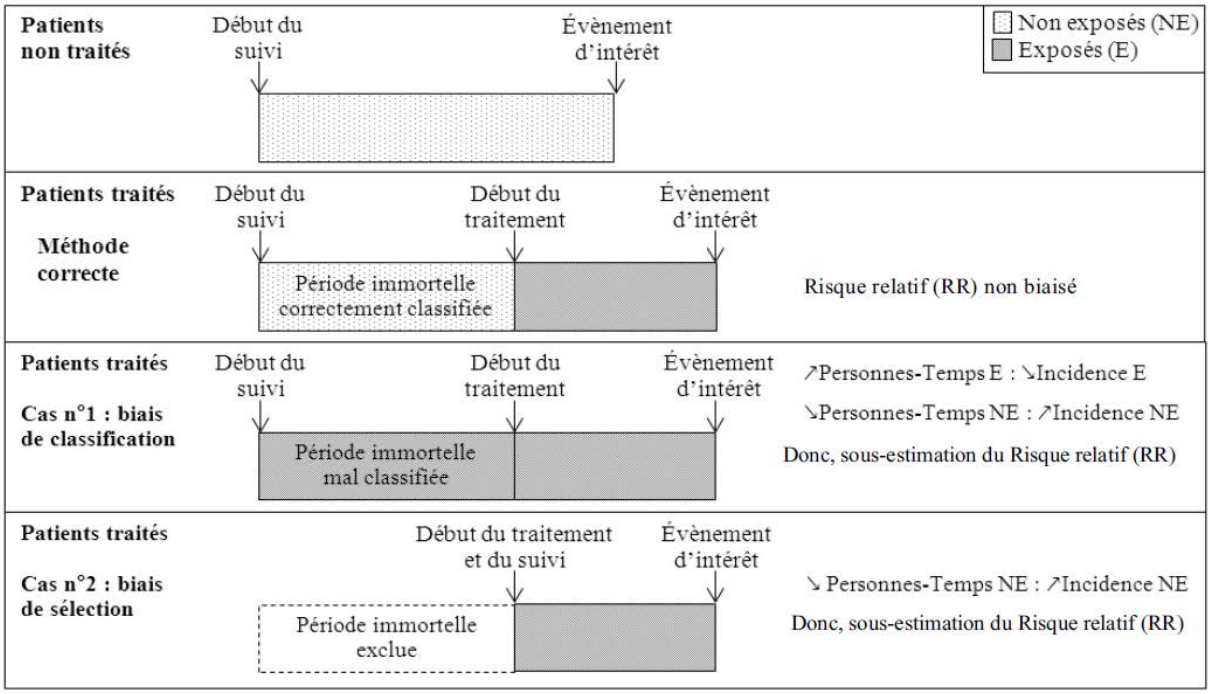


Figure VII : Illustration du biais de temps immortel. Tiré de (Faillie and Suissa 2015).

Ce biais est contrôlable et peut être paré par des choix méthodologiques appropriés. Deux solutions sont utilisables dans les situations les plus fréquentes (Suissa 2008). La première est

de considérer la période de temps immortel comme une période de non-exposition, en définissant l'exposition par une variable dépendante du temps. La seconde solution est de redéfinir la date de début de la période de suivi de sorte qu'à cette date, l'ensemble des individus classés comme exposés auront effectivement commencé à être exposés.

A.5. Les méthodes d'analyse pour maîtriser le biais d'indication dans les études observationnelles

Dans une étude observationnelle, la prise en compte des facteurs de confusion peut être réalisée à l'un ou l'autre des deux moments suivants du déroulement de l'étude :

- *A priori*, lors de l'élaboration du protocole de l'étude : des méthodes permettent de limiter l'existence de facteurs de confusion dès la constitution de l'échantillon. Elles permettent d'obtenir des groupes qui se ressemblent pour les facteurs de confusion retenus, afin de limiter le biais de confusion qu'ils génèrent. L'influence de ces facteurs sur les résultats est neutralisée et ne peut plus être étudiée ;
- *A posteriori*, lors de l'analyse des données : des méthodes d'analyse permettent de maîtriser le biais de confusion induit par la dissemblance des groupes.

Dans le cadre de cette thèse, nous nous plaçons en situation d'évaluation par une étude observationnelle prospective (exposés – non exposés ou cohorte), et notre propos portera sur les méthodes *a posteriori* de contrôle du biais de confusion. Nous envisagerons d'abord les méthodes que nous qualifierons de « conventionnelles », c'est-à-dire d'utilisation ancienne et habituelle, puis les méthodes dites « novatrices », d'utilisation plus récente.

Les méthodes *a priori* et les autres types d'études (telles que les études cas-témoin et ses adaptations – '*case-crossover design*', '*case-time control design*' – (Uddin, Groenwold et al. 2016), la méthode des doubles différences, la régression sur discontinuité…) ne seront pas abordées, du fait d'un champ d'application plus restreint.

A.5.1. Les méthodes d'analyse conventionnelles

A.5.1.1. La stratification dans l'analyse

Lorsque les groupes exposés et témoins diffèrent pour une variable qui constitue un facteur de confusion, la stratification permet de créer des sous-échantillons (strates) au sein desquels tous les individus seront similaires (ou proches) pour la variable considérée. L'analyse est ensuite effectuée dans chaque strate, limitant le biais de confusion généré par cette variable. Par exemple, si l'âge est un facteur de confusion, la population peut être découpée par tranches d'âge de 5 ans, puis une comparaison des individus exposés à l'intervention et des témoins est effectuée dans chaque strate. De mise en œuvre facile, cette méthode d'analyse ne permet de prendre en compte qu'un nombre limité de facteurs de confusion (si N facteurs de confusion binaires, il y a 2^N strates à analyser). S'il existe davantage de facteurs de confusion, comme c'est le cas le plus souvent, la méthode devient difficile à mettre en œuvre. L'indication de la stratification dans l'analyse pour limiter le biais de confusion est donc restreinte.

A.5.1.2. L'appariement dans l'analyse

L'appariement est la constitution de paires d'individus exposés/témoins similaires pour une ou plusieurs covariables. Les individus des deux groupes de comparaison ainsi formés sont donc comparables pour les variables utilisées lors de l'appariement. Les techniques d'analyse statistique utilisées ensuite doivent prendre en compte cet appariement. L'avantage de cette technique est sa facilité de mise en œuvre (ainsi que la facilité de vérification de la comparabilité des groupes pour les variables d'appariement). Cependant, elle ne permet de prendre en compte qu'un petit nombre de facteurs de confusion. En effet, il est d'autant plus difficile de former des paires correspondant aux critères d'appariement, que ces critères sont nombreux ou exigeants. Lorsqu'il existe un faible recouvrement des variables d'appariement

entre exposés et témoins, l'appariement peut entraîner une perte importante d'effectif, parfois rédhibitoire. Cette restriction de l'échantillon analysé peut entraîner, outre une perte de puissance statistique, un biais de sélection si cet échantillon n'est pas représentatif de la cohorte entière. D'autres désavantages de cette méthode sont qu'il n'est pas possible d'analyser l'effet des facteurs utilisés pour l'appariement, et qu'il existe un risque de sur-appariement (constitution de paires trop similaires, gommant les différences entre exposés et témoins).

A.5.1.3. L'ajustement multivarié

L'ajustement multivarié est l'approche analytique la plus communément utilisée pour maîtriser le biais de confusion dans les études de cohorte. Les méthodes basées sur l'ajustement multivarié ne cherchent pas à redresser le déséquilibre des covariables entre les groupes comparés, mais à corriger les effets potentiels de ce déséquilibre en réalisant une estimation conditionnelle de l'effet de l'intervention pour un niveau constant de chaque covariable. La maîtrise du biais de confusion passe par une modélisation statistique de la relation entre l'exposition, les facteurs de confusion et le critère de jugement. Différents modèles sont utilisables en fonction de la question posée, le choix du modèle dépendant de divers paramètres, mais plus particulièrement du type de critère du jugement (qualitatif, quantitatif, durée). Par exemple, on pourra utiliser une régression logistique pour un critère de jugement qualitatif, ou un modèle des risques proportionnels de Cox pour un critère de jugement de type délai de survenue de l'événement. Le principal avantage de ces méthodes est qu'il est possible d'inclure plusieurs facteurs de confusion dans la modélisation, et d'en étudier les effets propres. Le nombre de facteurs de confusion n'est toutefois pas illimité. En effet, la prise en compte d'un nombre important de variables de confusion par rapport au nombre d'évènements survenus augmente le risque d'erreurs dans la modélisation, qui peut

être un sur-ajustement (modélisation des particularités de l'échantillon – le bruit de fond –, et non de l'information utile), un sous-ajustement (mauvaise prédiction du modèle) ou une modélisation paradoxale (mise en évidence d'une association dans la direction opposée de la direction réelle)(Peduzzi, Concato et al. 1996). L'utilisation d'un trop grand nombre de variables dans le modèle statistique peut ainsi produire une estimation biaisée de l'effet, dans une direction imprévisible. Il est recommandé de ne pas inclure dans le modèle plus d'une variable pour 10 événements observés (Peduzzi, Concato et al. 1995, Peduzzi, Concato et al. 1996). Le choix des variables à inclure est donc une question primordiale. Il peut se faire sur des arguments théoriques et/ou en utilisant des algorithmes automatisés, mais doit veiller à ne pas inclure de variables fortement corrélées entre-elles. Le cas contraire conduirait à une multicolinéarité qui augmente la variance des coefficients de régression et les rend instables et difficiles à interpréter. Enfin, si les modèles statistiques sont de mise en œuvre apparente aisée, ils doivent être utilisés avec précaution. Leur validité repose sur un certain nombre d'hypothèses pas toujours éprouvées par les investigateurs. Des outils statistiques peuvent aider à vérifier l'adéquation du modèle (comme le test de Hosmer-Lemeshow pour la régression logistique) et sa qualité prédictive (aire sous la courbe ROC ou *c-statistic*).

A.5.2. Les méthodes novatrices

A.5.2.1. Les méthodes utilisant un score de propension

A.5.2.1.1. Le principe

La méthode des scores de propension a été développée par Rosenbaum et Rubin en 1983, mais il a fallu attendre la fin des années 2000 pour qu'elle devienne populaire (**Figure VIII**).

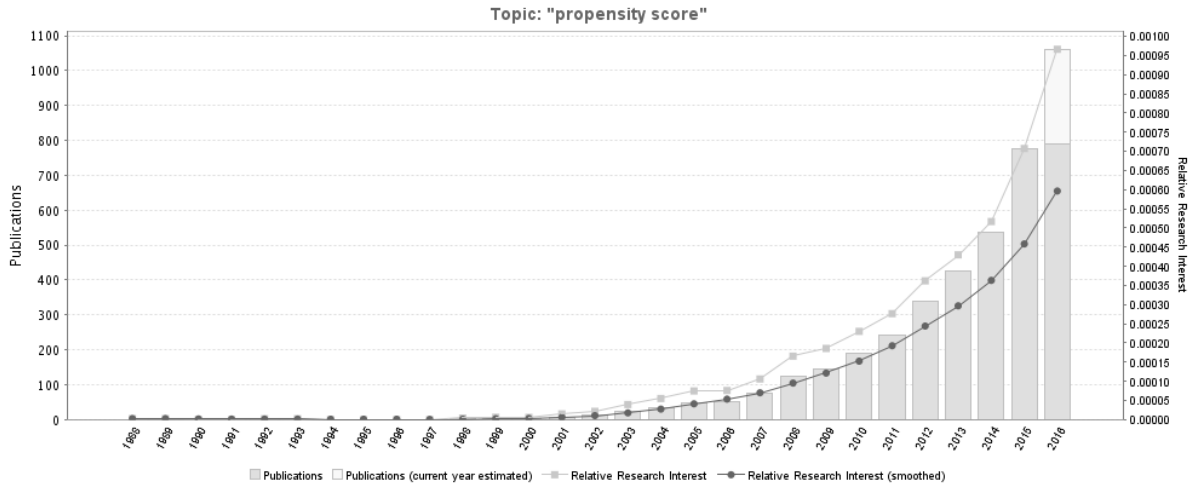


Figure VIII : Évolution annuelle du nombre de résultats à la requête '*propensity score*' dans PubMed entre 1988 et 2016. D'après le site GoPubMed.org accédé le 15/08/2016.

L'approche basée sur l'utilisation d'un score de propension (SP) cherche à rendre comparables les groupes exposés et témoins pour l'ensemble des facteurs de confusion potentiels mesurés. Le SP est défini comme la probabilité de recevoir l'intervention en fonction d'un ensemble de caractéristiques mesurées (Rosenbaum and Rubin 1983). Plus précisément, c'est la probabilité conditionnelle pour un individu i de recevoir l'intervention Z=1 selon ses caractéristiques x mesurées à son inclusion dans l'étude :

$$SP(x_i) = \Pr(Z_i = 1 | X_i = x_i)$$

Théoriquement, des individus ayant le même SP sont comparables pour l'ensemble des variables utilisées pour construire le score. L'avantage du SP par rapport à l'ajustement conventionnel est que le nombre de facteurs de confusion pris en compte n'est pas limité. Mais il ne permet pas de maîtriser le biais induit par des facteurs de confusion non mesurés (connus ou non), si bien qu'un biais de confusion résiduel demeure possible.

En pratique, ainsi que nous le détaillerons dans les paragraphes suivants, cette méthode nécessite d'abord de calculer le SP, puis d'utiliser ce score pour évaluer l'effet de l'intervention.

A.5.2.1.2. Le calcul du score de propension

Le SP est communément calculé par une régression logistique. Le modèle de régression logistique permet d'estimer la probabilité de recevoir l'intervention en fonction d'un certain nombre de caractéristiques x des individus mesurées à leur inclusion dans l'étude :

$$Logit(SP) = \alpha_1 + \alpha_2 x_1 + \ldots + \alpha_i x_i$$

Les chercheurs se sont d'abord intéressés aux modalités de choix des variables à inclure dans le modèle. S'il était clair que les variables mesurées après l'allocation de l'intervention devaient être exclues (D'Agostino 2007), la discussion portait sur le traitement des variables mesurées à l'inclusion dans l'étude, selon qu'elles sont associées à l'intervention et/ou au critère de jugement. Le consensus actuel recommande d'inclure les facteurs de confusion (liés à l'intervention et au critère de jugement) mais également les variables associées uniquement au critère de jugement, sans qu'elles soient obligatoirement associées à l'intervention (Rubin 1997, Brookhart, Schneeweiss et al. 2006, Austin, Grootendorst et al. 2007).

Dans un deuxième temps, les réflexions ont opéré un retour vers le principe même de l'utilisation d'un SP. L'objectif du score n'est pas de prédire parfaitement l'allocation de l'intervention, car dans ce cas il n'y aurait pas de superposition des distributions du SP chez les exposés et les témoins, et la méthode ne serait pas utilisable. L'objectif est bien de rendre comparable les groupes exposés et témoins. Ainsi, plutôt que d'adopter des consignes strictes concernant les variables à inclure ou non dans la construction du score, il est préférable de s'assurer que le modèle utilisé permet d'atteindre la comparabilité des groupes. Si ce

diagnostic met en évidence un déséquilibre de certaines variables entre les groupes après prise en compte du SP, c'est que la construction de ce score n'est pas satisfaisante et qu'elle doit donc être rediscutée.

Pour effectuer la comparaison des variables entres les groupes après utilisation du SP, les approches statistiques usuelles de comparaison des distributions, qui ont été initialement utilisées, ont été critiquées dans ce contexte (Austin 2009). En effet, la *p-value* n'est pas caractéristique des échantillons comparés mais fait référence à une population hypothétique, et est dépendante de la taille des échantillons (Ho, Imai et al. 2007). L'alternative recommandée est de calculer la différence standardisée (SDiff) de chaque variable entre les groupes, qui n'a pas les deux caractéristiques précitées de la *p-value* (Ali, Groenwold et al. 2015). La SDiff se calcule de la manière suivante :

- dans le cas d'une variable continue x et de variance s :

$$SDiff = \frac{100\ (\bar{x}_{intervention} - \bar{x}_{témoin})}{\sqrt{\frac{s^2_{intervention} - s^2_{témoin}}{2}}}$$

La SDiff est la différence des moyennes dans les deux groupes divisée par une estimation de l'écart-type de cette différence.

- dans le cas d'une variable qualitative dichotomique, p étant la proportion d'individus observant le caractère dans l'échantillon :

$$SDiff = \frac{100\ (p_{intervention} - p_{témoin})}{\sqrt{\frac{p_{intervention}(1 - p_{intervention}) + p_{témoin}(1 - p_{témoin})}{2}}}$$

Pour une variable donnée, une SDiff inférieure ou égale à 10% en valeur absolue permet communément de considérer la différence entre les deux groupes comme négligeable (Austin, Grootendorst et al. 2007). Comme nous l'avons noté, l'objectif de cette approche n'étant pas de prédire parfaitement l'exposition, il n'est pas non plus souhaitable de recourir à des tests d'adéquation ou de discrimination du modèle de calcul du SP (test de Hosmer-Lemeshow, aire sous la courbe ROC ou *c-statistic*)(Westreich, Cole et al. 2011). Ces tests permettent en effet de vérifier la performance de la prédiction de l'allocation de l'intervention, mais celle-ci ne garantit en rien la comparabilité des groupes en utilisant le score calculé.

Un développement particulier de la méthode basée sur le SP est l'utilisation d'un score dit à hautes dimensions (*hdPS*, *'high dimensional Propensity Score'*), qui consiste à mettre en œuvre un algorithme standardisé afin de sélectionner les covariables qui seront prises en compte (Schneeweiss, Rassen et al. 2009). Cette méthode, utilisée sur de larges bases de données, pourrait permettre d'améliorer encore la prise en compte du biais de confusion et ainsi de diminuer le biais résiduel. La théorie est de prendre en compte un très grand nombre de covariables mesurées et de parvenir à approcher les facteurs de confusion non mesurés. Cependant, l'intérêt supplémentaire apporté par une telle approche demeure hypothétique (Guertin, Rahme et al. 2016).

A.5.2.1.3. L'utilisation du score de propension

Une fois calculé, le score peut être utilisé de différentes manières : ajustement, stratification, appariement, pondération.

(i) Ajustement

C'est l'utilisation la plus simple du SP. Il n'y a pas d'étape intermédiaire entre le calcul du score et l'estimation de l'effet : le score est utilisé directement comme une variable d'ajustement dans le modèle de régression entre l'exposition et le critère de jugement. Derrière cette simplicité se cache des difficultés à vérifier la bonne mise en œuvre de cette approche. D'une part, il faut veiller à ce que le modèle de calcul du score soit correctement réalisé et que la relation entre le critère de jugement et le score soit correctement modélisée. D'autre part, la vérification de la comparabilité des groupes après prise en compte du score n'est pas évidente et nécessite la mise en œuvre de méthodes développées spécifiquement pour l'ajustement (Austin 2008).

(ii) Stratification

Les groupes sont distribués en plusieurs strates dont les bornes sont des valeurs du SP. Habituellement cette division est opérée sur la base des quintiles de la distribution du SP, car il a été montré que cette pratique permettait de réduire de 90% le biais généré par les facteurs de confusion mesurés si on estime un effet linéaire de l'intervention (Rosenbaum and Rubin 1984).

Une fois les strates constituées, l'effet de l'intervention est estimé au sein de chacune d'elles. Les estimations de chaque strate peuvent être poolées pour obtenir une estimation globale. Le score n'étant pas utilisé lors de l'analyse statistique, mais préalablement à celle-ci lors de la constitution des strates, cette méthode ne nécessite pas de modéliser la relation entre le critère de jugement et le score (Rubin 2004). Cependant, au sein d'une strate, les distributions du score chez les exposés et les témoins ne sont pas absolument superposables, si bien qu'il existe un déséquilibre résiduel des caractéristiques des exposés et des témoins.

(iii) Appariement

Des paires « individu exposé – individu témoin » sont constituées sur la base du SP. Dans l'idéal, les individus d'une même paire devraient avoir un score identique, mais cela est difficilement réalisable en pratique. Une différence est donc admise, dont la valeur maximale (‘*caliper’*) est au choix de l'investigateur. Par exemple, Austin et coll. recommandent de choisir un ‘*caliper’* égal à 0,2 fois l'écart-type du SP (Austin 2011). Cependant, plus l'investigateur sera exigeant sur la proximité des scores, plus il lui sera difficile de constituer des paires, et les individus non appariés seront alors exclus de l'analyse, ce qui restreint l'effectif. Comme évoqué concernant l'appariement direct sur les variables de confusion (§A.5.1.2, p.39), le risque est une perte de puissance statistique et éventuellement un biais de sélection, si les individus exclus ont des caractéristiques différentes des individus appariés. De nombreuses techniques d'appariement sont proposées, intéressant le choix des paires (‘*full matching*’, ‘*greedy matching*’, ‘*caliper*’, ‘*nearest neighbor*’, ‘*Mahalanobis distance*’,…)(Austin 2011, Lunt 2014, Austin and Stuart 2015), le nombre de témoin(s) par individu exposé (Austin 2010), l'existence ou non d'une remise (témoins utilisables dans des paires différentes)(Austin 2009). Comme à l'étape de construction du score, plutôt qu'avoir des certitudes théoriques sur les modalités d'appariement, il convient de faire des choix pragmatiques conciliant la proximité des scores, la taille de l'effectif et les considérations statistiques (des méthodes particulières doivent être utilisées lorsqu'un témoin peut être utilisé dans plusieurs paires).

Une fois les groupes « exposés » et « témoins » constitués, la méthode statistique utilisée pour évaluer l'effet de l'intervention devra prendre en compte l'appariement. Par exemple pour une analyse de survie, on utilisera un modèle marginal (Lin and Wei 1989) ou un modèle de Cox à fragilité (Cummings, McKnight et al. 2003). Dans cette approche comparant des paires « individu exposé – individu témoin » ayant le même score, l'effet évalué est l'effet moyen

chez les traités (‘*Average treatment among treated*’ *ATT*), comme lors d’un essai contrôlé randomisé. Le score n’étant pas utilisé lors de l’analyse statistique, mais préalablement lors de la constitution des groupes, cette méthode ne nécessite pas de modéliser la relation entre le critère de jugement et le score (Rubin 2004).

(iv) Pondération

Le SP est utilisé pour affecter à chaque individu un facteur de pondération. La pondération la plus communément réalisée est la pondération inverse (*IPTW*, ‘*Inverse ponderation treatment weighting*’). Dans cette méthode, on attribue un poids égal à $\frac{1}{SP}$ à un individu exposé et un poids égal à $\frac{1}{1-SP}$ à un témoin (Imbens 2000, Hirano and Imbens 2001). On crée ainsi un échantillon virtuel pondéré sur le score, dans lequel les caractéristiques des groupes exposés et témoins sont comparables pour les variables utilisées dans la construction du SP, et où l’effet de l’intervention est le même que dans la population cible. En d’autres termes, dans cette population pondérée, l’exposition est indépendante des facteurs de confusion mesurés et pris en compte dans le score. Cette méthode ne nécessite pas de modéliser la relation entre le critère de jugement et le score. Un inconvénient de cette approche est qu’elle est sensible aux individus ayant un poids extrême (Rubin 2004). La pondération par « standardisation », où les patients traités reçoivent un poids égal à 1 et les témoins un poids dépendant du SP, a été proposée mais est rarement employée (Sato and Matsuyama 2003).

(v) Restriction de la population par ‘*trimming*’

La restriction par ‘*trimming*’ n’est pas une méthode d’utilisation du SP à proprement parler, mais une option d’analyse rendue possible par l’existence du score. Cette option s’intéresse à

certains individus atypiques qui reçoivent un traitement contraire à celui pourtant prédit avec une importante probabilité d'après leurs caractéristiques mesurées initialement. Ces individus sont ceux situés aux extrêmes de la distribution du SP, dans les valeurs les plus élevées pour les individus non-traités et dans les valeurs les plus basses pour les individus traités. Ils présentent vraisemblablement d'autres caractéristiques non mesurées qui expliquent cette contradiction avec la prédiction, comme par exemple une fragilité singulière s'opposant à la réalisation de l'intervention, qui peuvent biaiser l'analyse de l'effet de l'intervention. Pour éviter ce biais, certains auteurs recommandent de restreindre la population analysée ('*trimming*'), en écartant les individus les plus extrêmes ('*outliers*') recevant une intervention contraire à la prédiction (Sturmer, Wyss et al. 2014). Il est suggéré de réaliser des analyses de sensibilité en comparant les résultats avec différentes valeurs de restriction (par exemple : 1%, 2,5%...) et sans restriction. Cette méthode pourrait permettre de réduire le biais de confusion généré par des variables non mesurées (Sturmer, Rothman et al. 2010).

A.5.2.1.4. Quelle méthode d'utilisation du score de propension privilégier ?

Parmi les quatre modalités de mise en œuvre du SP, l'investigateur s'interrogera légitimement sur celle qui lui permettra d'obtenir les résultats les moins biaisés, tout en conservant une puissance suffisante pour mettre en évidence une différence si elle existe. Si aucune d'elles ne possède une supériorité franche par rapport aux autres, la littérature (par des études de simulation) fournit des arguments de choix éclairants.

Avant de comparer les estimations relatives des méthodes, il faut toujours s'interroger sur le type d'effet que l'on veut estimer. En effet, on peut vouloir évaluer l'effet moyen chez les individus traités (*ATT*, '*Average treatment effect among treated*') ou l'effet moyen dans la population entière (*ATE*, '*Average treatment effect*'). L'*ATE* représente l'effet que l'on

mesurerait si on faisait passer l'ensemble des individus de la population du statut « non exposé » au statut « exposé ». Cet effet est donc propre à l'échantillon étudié. L'*ATT* est l'effet mesuré chez les individus qui ont réellement bénéficié de l'intervention. Selon la question posée et le point de vue de l'évaluateur, on peut vouloir estimer l'un ou l'autre de ces effets (Pirracchio, Carone et al. 2013). L'*ATE* est intéressant à évaluer lorsqu'il n'y a pas de barrière à la diffusion de l'intervention dans l'ensemble de la population : dans cette situation, il est pertinent de chercher à connaître l'effet *ATE* qui serait obtenu si on parvenait à faire bénéficier de l'intervention l'ensemble de la population. En revanche, dans les situations où il n'est pas réaliste d'envisager la délivrance de l'intervention à l'ensemble de la population, il est préférable d'évaluer l'effet *ATT* de l'intervention dans la sous-population des individus qui en ont réellement bénéficié. Il est donc important de savoir le type d'effet que les différentes méthodes d'utilisation du SP permettent d'évaluer (**Tableau 1**). *ATT* et *ATE* sont des effets dits marginaux, c'est-à-dire estimés au niveau d'une population, et non conditionnels (au niveau individuel). L'estimation d'un effet marginal est privilégiée dans une perspective de santé publique pour étayer la mise en œuvre d'une intervention dans une population, alors que l'estimation d'un effet conditionnel est plus appropriée dans une perspective clinique lorsqu'on s'interroge sur les effets potentiels de l'intervention chez un individu donné. L'ajustement sur le SP ne permet pas d'évaluer un effet marginal, mais un effet conditionnel (Austin 2014). Concernant la stratification, s'il est habituellement rapporté, comme dans le tableau 1, qu'elle permet d'estimer l'*ATT* ou l'*ATE* selon la méthode utilisée, Austin a récemment remis en cause cette position, jugeant qu'elle permettait plutôt d'évaluer un effet conditionnel (Austin 2014).

Tableau 1: Type d'effet évalué selon la méthode de score de propension utilisée (Adapté de (Deb, Austin et al. 2016)).

Méthode	**Effet moyen chez les traités (*ATT*)**	**Effet moyen dans la population (*ATE*)**
Appariement	Oui	Non (sauf *full matching*)
Stratification	Oui (poids ajustés)	Oui (poids égaux)
Ajustement	Non	Non
Pondération inverse	Oui (poids modifiés)	Oui

Parmi les quatre méthodes, l'ajustement est celle qui apparaît sans ambiguïté comme celle à ne pas retenir : d'une part du fait des difficultés de mise en œuvre évoquées plus haut, ensuite parce qu'elle se révèle moins performante que les autres méthodes (Wan and Mitra 2016).

Pour départager les autres méthodes, on dispose des informations suivantes :

- L'appariement permet un meilleur équilibre des variables entre les groupes que la stratification (Austin, Grootendorst et al. 2007).
- L'appariement engendre moins de biais que la stratification (Austin and Mamdani 2006, Austin, Grootendorst et al. 2007).
- Dans deux études de simulation, Austin a comparé la performance des différentes méthodes d'utilisation du SP pour évaluer l'effet d'une intervention sur un critère dépendant du temps. La première étude s'est intéressée au calcul de *hazard ratio* : l'appariement et la pondération IPTW se sont révélés plus performants que la stratification et l'ajustement (Austin 2013). La seconde étude était consacrée à l'estimation d'un effet absolu : là encore, l'appariement et la pondération IPTW se sont révélés plus performants que la stratification, avec un avantage pour l'appariement lorsque l'intervention est peu fréquente (moins de 10%)(Austin and Schuster 2014). À l'inverse, dans le contexte d'événements rares, Hajage et coll. ont conclu à une meilleure estimation en pondérant plutôt qu'en appariant (Hajage, Tubach et al. 2016).

- Dans les analyses de survie toujours, Austin a poursuivi ses travaux et conclu que l'estimation par pondération, comme avec appariement, est faiblement biaisée si la prédiction du traitement est faible à modérée, même lorsque le modèle du SP est mal spécifié, mais est moins fiable quand la prédiction est plus forte, même quand le modèle du SP est bien spécifié (Austin and Stuart 2015). L'estimation est alors particulièrement sensible aux extrêmes, et peut-être améliorée en restreignant l'échantillon analysé (par exemple par '*trimming*') ou en exigeant que les paires soient plus proches (Austin and Stuart 2015). Ces résultats rejoignent les conclusions faites par ailleurs (Kurth, Walker et al. 2006).
- La pondération inverse IPTW fournit des résultats peu biaisés pour les grands échantillons, mais est sensible aux extrêmes sur des échantillons plus petits (Sturmer, Schneeweiss et al. 2005).

Au total, l'appariement et la pondération inverse sont les méthodes d'utilisation du SP les plus performantes. On privilégiera l'appariement si on souhaite évaluer l'effet *ATT* et la pondération inverse si on souhaite évaluer l'effet *ATE*. Si le choix se porte sur l'appariement, il faut veiller à avoir un recouvrement suffisant des distributions du SP des deux groupes, sans quoi le nombre de paires constituées sera limité, entraînant une perte de puissance et éventuellement un biais de sélection. Globalement, les performances des méthodes utilisant un SP paraissent diminuer quand le biais d'indication augmente et/ou que l'effectif analysé diminue.

A.5.2.1.5. Quels sont les avantages théoriques des méthodes utilisant un score de propension, par rapport à l'ajustement conventionnel ?

L'avantage le plus souvent mis en avant des méthodes basées sur le SP, par rapport à l'ajustement conventionnel, est l'absence de limitation du nombre de facteurs de confusion mesurés utilisables dans l'analyse. Cepeda et coll. ont effectivement montré que l'analyse avec SP était moins biaisée que l'analyse conventionnelle en cas d'événements rares (moins de 8 événements par facteurs de confusion)(Cepeda, Boston et al. 2003). Cependant, lorsque les événements sont plus nombreux, l'analyse conventionnelle peut se révéler plus performante. Ces conclusions sont cohérentes avec les connaissances antérieures sur les méthodes conventionnelles, dans lesquelles il est recommandé de ne pas descendre sous le ratio de 10 événements par covariable incluse dans le modèle (Peduzzi, Concato et al. 1995, Peduzzi, Concato et al. 1996). Elles ont été confirmées par les travaux de Martens et coll., qui ont conclu que les méthodes basées sur le SP donnaient une estimation de l'effet plus proche de l'effet marginal réel qu'un modèle de régression logistique, et ce d'autant plus que les facteurs confondants étaient nombreux et que l'effet était important (Martens, Pestman et al. 2008).

Cependant, comme nous l'évoquions lors de la comparaison des différentes modalités d'utilisation du SP, il apparaît que l'analyse avec SP perd en performance lorsque l'effectif analysé diminue, car l'équilibre des variables entre les groupes comparés devient alors nécessairement moins bon (Rubin 1997).

On peut également citer les travaux de Cook et Goldman en faveur de l'analyse par stratification sur le SP par rapport à l'analyse conventionnelle, en cas de forte association entre l'exposition et les facteurs de confusion (Cook and Goldman 1989).

Il ne faut pas perdre de vue que l'analyse conventionnelle et l'analyse basée sur un SP partagent la même limite, celle de ne pas agir sur le biais résiduel généré par des facteurs de confusion non mesurés et/ou non connus. L'investigateur doit se garder de céder à la tentation de penser que le SP pourrait être indirectement lié à des facteurs non mesurés. En effet, Drake a montré que la direction et l'importance du biais induit par l'omission dans l'analyse d'un facteur de confusion étaient similaires avec une méthode conventionnelle multivariée et une méthode utilisant un SP (Drake 1993). Par conséquent, le SP ne diminue pas le biais induit par les variables non prises en compte. De manière analogue, Austin et coll. ont étudié la comparabilité des variables entre les groupes dans les quintiles du SP et montré que la construction du SP avec 18 variables issues d'une base de données médico-administrative ne permettait pas d'obtenir l'équilibre de variables cliniques non utilisées pour calculer le SP (Austin, Mamdani et al. 2005). D'autres travaux ont abouti à la même conclusion (Austin, Grootendorst et al. 2007).

Globalement, les méthodes basées sur un SP ont quelques avantages sur les méthodes conventionnelles : performance théorique meilleure, notamment en cas d'événements rares, facilitation de l'appariement, possibilité de restreindre l'échantillon par '*trimming*', possibilité d'analyser les interactions/effets selon la probabilité de bénéficier de l'intervention (Glynn, Schneeweiss et al. 2006). De plus, le SP autorise certains développements qui doivent encore être approfondis, comme les SP à hautes dimensions évoqués plus haut (§A.5.2.1.2, p.43), ou encore la calibration par le SP pour corriger des erreurs de mesure (Sturmer, Schneeweiss et al. 2005). Au contraire de l'ajustement conventionnel, les méthodes basées sur un SP ne permettent toutefois pas d'estimer un effet conditionnel (patient-spécifique), pourtant utile dans une approche clinique individuelle.

A.5.2.1.6. Mise en œuvre des méthodes utilisant un score de propension dans la littérature

Plusieurs revues de la littérature ont étudié l'utilisation faite des méthodes basées sur un SP, des modalités de leur mise en œuvre à leurs résultats, en comparant ces derniers à ceux issus d'une analyse par méthode conventionnelle. Ces revues aboutissent à la même conclusion, que les résultats sont peu différents entre les méthodes, avec toutefois une réserve majeure : le SP est souvent utilisé de manière inappropriée.

Les revues de la littérature ont effectivement dressé le constat de l'insuffisance d'informations publiées dans les articles concernant le développement et la validation des méthodes utilisant le SP. La première revue est celle de Weitzen et coll. en 2004, qui s'est intéressée à la qualité de 47 études publiées en 2001 (Weitzen, Lapane et al. 2004). Vingt-quatre (51%) études n'explicitaient pas les modalités de choix des variables utilisées pour le calcul du SP, 22 (47%) ne rapportaient pas la comparaison des variables dans les groupes après prise en compte du SP, alors que 18 (38%) évaluaient le pouvoir discriminant du modèle de construction du SP en indiquant une *c-statistic*, qui n'est pourtant pas pertinente (§A.5.2.1.2, p.44). Ensuite, dans un premier travail, Austin a critiqué l'utilisation de l'appariement sur le SP dans 47 études publiées entre 1996 et 2003 (Austin 2008). Seules deux (4%) études comparaient les caractéristiques des groupes exposés et témoins de façon appropriée et seules 13 (28%) études utilisaient des méthodes d'analyse statistique prenant en compte l'appariement. Austin a effectué le même travail d'analyse critique pour des articles publiés entre 2004 et 2006 dans le domaine cardiovasculaire, révélant que la méthodologie employée demeurait largement inadéquate. Il a ainsi recensé 44 articles publiés dans des revues de cardiologie entre 2004 et 2006 ayant eu recours à l'appariement sur le SP (Austin 2008). Moins de la moitié de ces études (n=20, 45%) rendaient compte des informations suffisantes concernant la réalisation de l'appariement, seules 4 (9%) avaient comparé avec une méthode

adaptée les caractéristiques initiales des deux groupes, et dans seulement un quart des études (n=11) la méthode d'analyse statistique de l'effet avait tenu compte de l'appariement. Austin a également fait un constat assez proche dans les revues de chirurgie cardiaque sur la même période 2004-2006 (Austin 2007). Quelques années plus tard, la situation restait largement perfectible : Ali et coll. ont passé en revue 296 articles publiés entre décembre 2011 et mai 2012 ayant utilisé un SP (Ali, Groenwold et al. 2015). La méthodologie de sélection des variables pour la construction du score n'a été explicitée que dans 102 (34%) articles, l'équilibre des variables a été vérifié et rapporté dans 177 (60%) études (cependant avec calcul de *p-values* dans 125 études). Parmi les 204 (69%) études avec appariement, les modalités d'appariement n'étaient rapportées que dans 67 (32,8%).

Malgré la limite précédente, les résultats obtenus par méthode conventionnelle et par méthode basée sur un SP ont été comparés dans plusieurs revues de la littérature. La première a retenu 43 études rapportant 78 estimations d'un effet d'une intervention au moyen de deux méthodes, une méthode de régression conventionnelle et une méthode utilisant un SP (Shah, Laupacis et al. 2005). Une différence significative entre les deux méthodes était observée dans seulement 8 (10%) cas. La concordance élevée des résultats entre les deux méthodes était illustrée par un score de kappa égal à 0,79. Dans les 8 cas non concordants, un effet significatif de l'intervention mis en évidence par la méthode conventionnelle n'était pas retrouvé par la méthode avec SP. Cela pourrait s'expliquer au moins partiellement par une perte de puissance liée à la réduction de l'effectif lors de l'appariement. L'utilisation du SP aboutissait à des estimations un peu plus proches de l'unité (de 6,4%). Dans la revue de Stürmer et coll., qui a inclus 69 études, seules 9 (13%) estimations différaient de plus de 20% entre les deux méthodes (Sturmer, Joshi et al. 2006). Enfin, dans une revue systématique d'études observationnelles consacrées au syndrome coronarien aigu, Dahabreh et coll. ont recensé 11 études rapportant une évaluation réalisée à la fois par régression conventionnelle et

par l'utilisation d'un SP (Dahabreh, Sheldrick et al. 2012). Les estimations étaient généralement très proches avec les deux méthodes.

Néanmoins, il faut garder à l'esprit que ces conclusions peuvent être faussées par un biais de publication, car on peut imaginer que les auteurs des études ont eu tendance à ne rapporter les résultats des différentes méthodes que dans le cas où elles aboutissaient à la même conclusion.

A.5.2.2. Les méthodes utilisant une variable instrumentale

A.5.2.2.1. De la théorie…

Les premières méthodes basées sur l'utilisation d'une variable instrumentale (VI) ont été développées dans les années 1920 en économétrie, où elles sont devenues d'utilisation courante (Greenland 2000). Elles ont commencé à être utilisées pour l'évaluation des interventions en santé dans les années 1990, le premier article didactique sur la VI à l'intention des épidémiologistes ayant été publié en 2000 (Greenland 2000), et d'autres ont suivi à partir de 2006 (Jamal Uddin and Klungel 2015)(**Figure IX**).

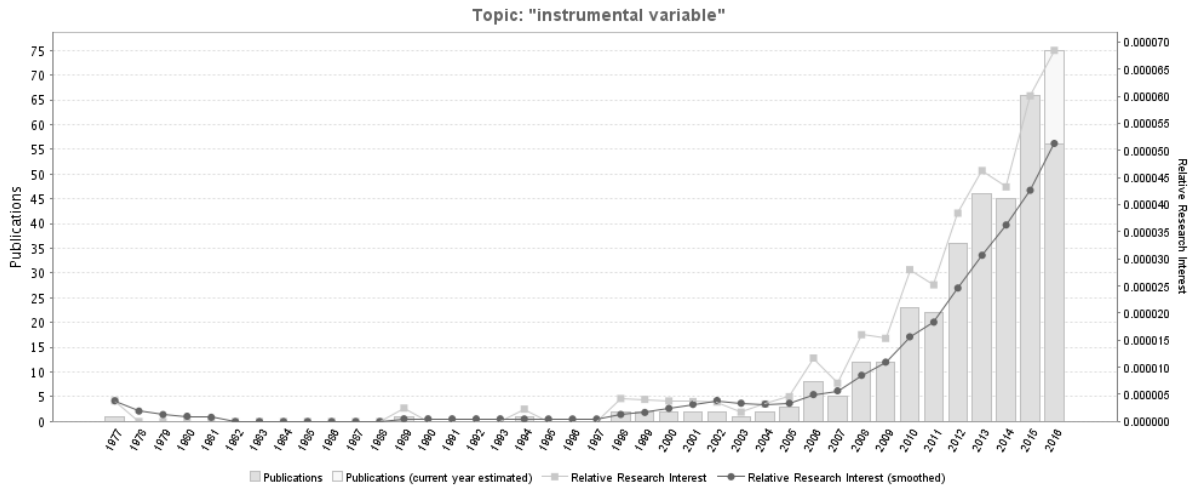


Figure IX : Évolution annuelle du nombre de résultats à la requête "instrumental variable" dans PubMed entre 1977 et 2016. D'après le site GoPubMed.org accédé le 15/08/2016.

Dans ces méthodes, une VI, encore appelée « instrument », est utilisée pour imiter le processus d'assignation de l'intervention d'une randomisation, permettant de limiter le biais généré par l'ensemble des facteurs de confusion, qu'ils aient été mesurés ou non.

L'approche basée sur une VI est réalisée en deux temps. Dans un premier temps, l'effet de l'instrument sur l'exposition est modélisé. Dans un second temps, le critère de jugement est comparé selon l'exposition prédite à la première étape, et non selon l'exposition réelle. Soit X l'exposition à l'intervention que l'on souhaite évaluer, Y le critère de jugement mesuré et Z la variable instrumentale. L'estimation de l'effet de X sur Y par Z fait appel à deux équations (E et F étant les termes d'erreur) :

$$(1) \quad Y = \alpha + \beta X + E$$

$$(2) \quad X = \gamma + \delta Z + F$$

La validité de la méthode repose sur le respect des trois conditions relatives à l'instrument Z choisi :

(i) Z a un effet causal sur X ;

(ii) Z et Y ne doivent pas avoir de cause commune (condition d'indépendance) ;

(iii) Z ne doit pas avoir d'effet sur Y autrement qu'indirectement par l'intermédiaire de X (condition d'exclusion).

Il faut souligner que ces trois conditions ne font pas explicitement référence aux facteurs de confusion U de la relation entre X et Y. L'association de Z avec U n'est qu'un cas particulier de violation des conditions (ii) et (iii). Il n'est pourtant pas rare de faire figurer l'absence d'association entre Z et U comme une condition de validité de l'instrument, comme illustré ci-après (**Figure X**). Concernant la condition (i), on adopte généralement une condition modifiée plus souple selon laquelle Z et X sont associés, que Z ait un effet causal sur X ou que Z et X aient une cause commune.

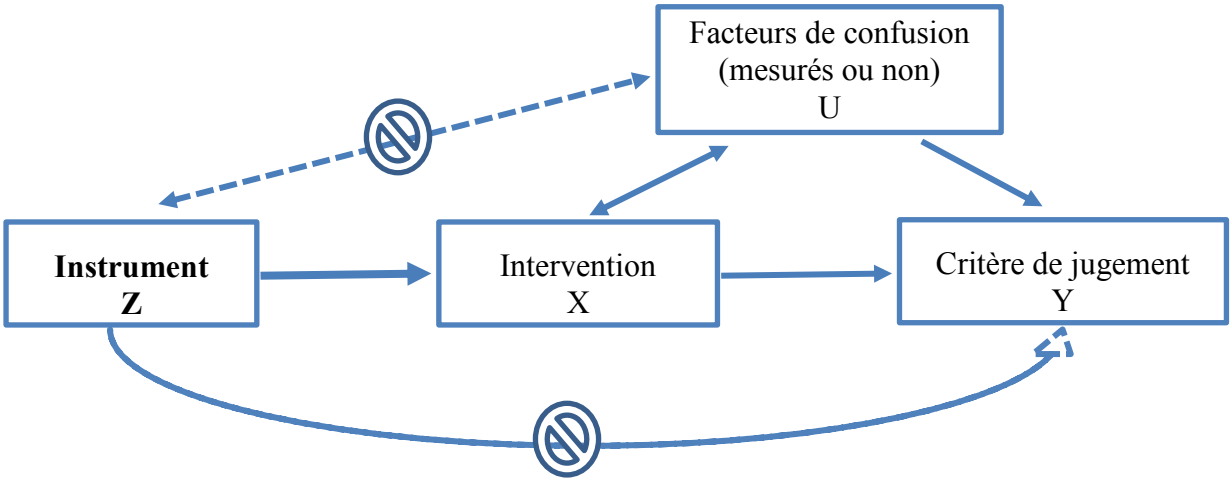


Figure X : Diagramme de causalité entre l'instrument, l'intervention, le critère de jugement et les facteurs de confusion. Adapté de (Chen and Briesacher 2011).

Pour illustrer cette méthode, il est classique de présenter l'étude de Permutt et Hebel, qui est considérée comme la première utilisation d'une VI en recherche médicale (**Figure XI**)(Permutt and Hebel 1989). Cette étude permet d'appréhender l'articulation entre l'instrument, l'exposition et le critère de jugement. L'objectif était d'étudier l'effet du tabagisme maternel pendant la grossesse (X) sur le poids de naissance de l'enfant (Y). L'instrument (Z) était un programme d'encouragement au sevrage tabagique qui était réalisé de façon aléatoire dans une population de parturientes. Z respecte les trois conditions de validité d'un instrument : (i) Z est par définition associé à X ; (ii) Z et Y n'ont pas de cause commune ; (iii) l'effet de Z sur Y passe par X.

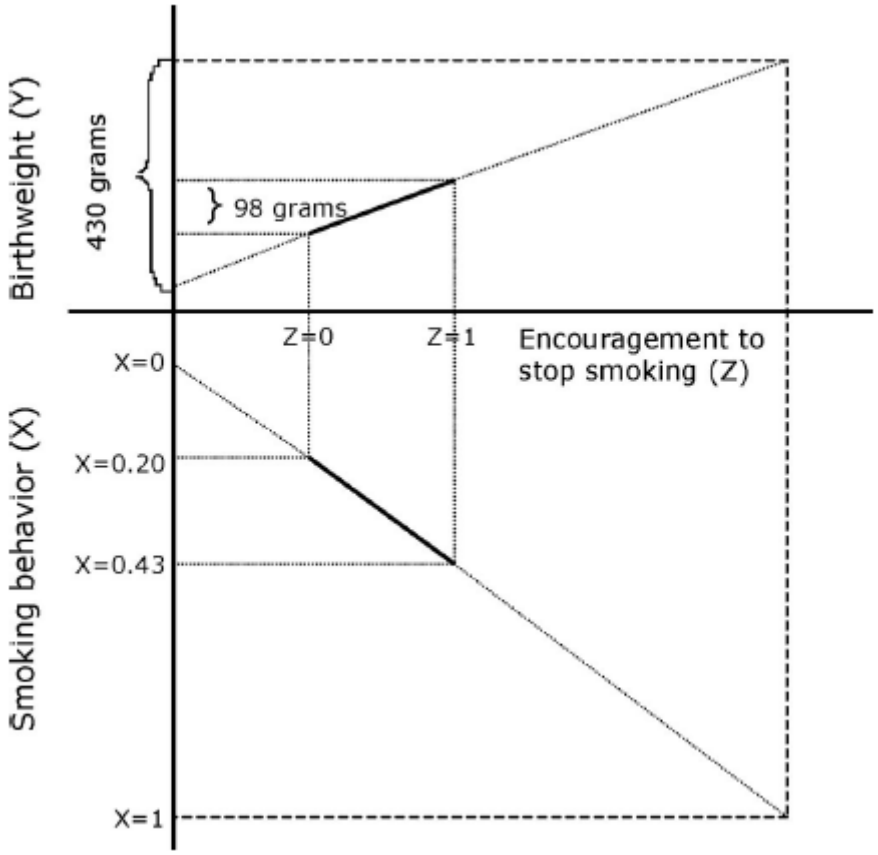


Figure XI : Estimation de l'effet du tabagisme X sur le poids de naissance Y par la variable instrumentale Z (programme d'encouragement au sevrage tabagique) dans l'étude de Permut et Hebel. Tiré de (Martens, Pestman et al. 2006).

On peut montrer que l'estimateur de l'effet de X sur Y en utilisant Z est le rapport de deux estimateurs des moindres carrés suivant (Martens, Pestman et al. 2006) :

$$(3) \quad \beta_{iv} = \frac{\widehat{\beta}_{ols(Z \to Y)}}{\widehat{\beta}_{ols(Z \to X)}}$$

Lorsque l'instrument et le critère de jugement sont dichotomiques, l'équation (3) devient :

$$(4) \quad \beta_{iv} = \frac{P(Y=1 \mid Z=1) - P(Y=1 \mid Z=0)}{P(X=1 \mid Z=1) - P(X=1 \mid Z=0)}$$

Puisque l'instrument est dichotomique, l'estimateur en intention de traiter $\widehat{\beta}_{\mathrm{ols}(\mathrm{Z} \to \mathrm{Y})}$ est simplement égal à la différence de poids moyen à la naissance entre les femmes qui ont

bénéficié de l'encouragement et celles qui n'en ont pas bénéficié, soit 98 grammes. Ensuite, l'estimateur $\hat{\beta}_{ols(Z \to X)}$ est égal à la différence entre les proportions de femmes tabagiques dans le groupe n'ayant pas bénéficié de l'encouragement et celui en ayant bénéficié, soit 0,23. Enfin, on calcule l'estimateur de l'effet β_{iv} = 98/0,23 = 430 grammes. Ainsi, on a estimé que le tabagisme maternel entraînait une réduction de 430 grammes du poids de naissance. Attention, dans cet exemple particulier l'intervention – le programme d'encouragement au sevrage tabagique – n'était pas l'objet de l'évaluation mais a été utilisé comme un instrument pour évaluer l'effet de l'exposition au tabagisme maternel. Dans le cadre habituel de l'évaluation d'une intervention en santé, X désignera l'exposition à l'intervention évaluée.

En raison de l'origine économétrique de la VI, la méthode statistique traditionnellement utilisée est, comme dans l'exemple précédent, celle des doubles moindres carrés ou méthode des moindres carrés en deux étapes (en anglais 2SLS pour '*two-stage least squares*')(Rassen, Schneeweiss et al. 2009). Progressivement, de nombreuses autres méthodes ont été développées pour s'adapter à tous les types de variables d'exposition et critères de jugement (par exemple, l'approche par inclusion des résidus en deux étapes pour les analyses de survie (Terza, Basu et al. 2008))(Davies, Smith et al. 2013, Klungel, Jamal Uddin et al. 2015).

Dans une méthode basée sur une VI, il faut comprendre que l'évaluation qui est réalisée est celle de l'effet moyen chez les patients dits « compliants » ('*compliers*') ou « marginaux » (Tableau 2), c'est-à-dire ceux dont l'allocation de l'intervention est affectée par l'instrument (cet effet est appelé '*Local Average Treatment Effet*' ou *LATE*)(Imbens and Angrist 1994). Prenons l'exemple de l'évaluation d'un nouveau traitement A comparativement au traitement

courant B, en utilisant un instrument dichotomique basé sur la préférence du prescripteur pour le traitement A ou B. Les patients '*compliers*' sont ceux qui bénéficieraient du traitement A dans le groupe « médecin préférant le traitement A », alors qu'ils n'en bénéficieraient pas dans le groupe « médecin préférant le traitement B ». Au contraire, les patients qui reçoivent toujours le traitement A quelle que soit la préférence du médecin ('*always-takers*'), et ceux qui ne le reçoivent jamais ('*never-takers*')(**Tableau 2**), ne sont pas pris en compte dans l'analyse. Une hypothèse de la méthode est qu'il n'existe pas d'individu '*defier*' qui recevrait le traitement A dans le groupe « médecin préférant le traitement B » et le traitement B dans le groupe « médecin préférant le traitement A ».

Tableau 2: Illustration des quatre types d'individus selon la compliance au traitement dans une analyse avec variable instrumentale. Exemple d'un instrument dichotomique basé sur la préférence de prescription du médecin et de deux traitements possibles A et B. Traduit de (Swanson, Miller et al. 2015).

		Individu pris en charge par un médecin préférant le traitement B	
		Traitement A prescrit	Traitement B prescrit
Individu pris en charge par un médecin préférant le traitement A	Traitement A prescrit	*Always-takers*	*Compliers*
	Traitement B prescrit	*Defiers*	*Never-takers*

Selon le caractère homogène ou hétérogène de l'effet de l'intervention dans la population, les différents types d'effet peuvent ne pas être équivalents. Considérons l'évaluation d'un réseau de soins spécialisé dans la prise en charge de patients atteints d'insuffisance cardiaque, en utilisant un instrument basé sur la préférence du médecin. Si ce réseau avait un effet homogène dans la population, il aurait un effet identique chez chaque patient atteint d'insuffisance cardiaque, quelles que soient ses caractéristiques. Dans cette situation très hypothétique, les effets *ATT*, *ATE* et *LATE* seraient équivalents (Fang, Brooks et al. 2012). Il en serait de même si l'effet était hétérogène, mais que cette hétérogénéité n'était pas liée à l'instrument, par exemple si l'effet du réseau était plus important chez les patients plus âgés,

mais que l'âge n'entrait pas en considération dans la préférence du médecin. En revanche, si l'âge était un élément de la préférence du médecin, l'effet moyen chez les traités *ATT* serait modifié selon la valeur de l'instrument : dans ce cas de figure où l'hétérogénéité de l'effet est liée à l'instrument, les effets *ATT*, *ATE* et *LATE* ne représentent pas la même chose. Ces considérations sont à réfléchir lorsqu'on compare les résultats d'analyse avec VI à ceux d'autres méthodes. En outre, certains auteurs ont mis en débat la pertinence d'évaluer l'effet *LATE*, notamment parce que les individus '*compliers*' ne sont pas identifiés et que conséquemment l'effet causal est mesuré dans une sous-population inconnue (Robins and Greenland 1996, Swanson and Hernan 2013, Swanson and Hernan 2014).

A.5.2.2.2. … à la pratique

Si les méthodes utilisant une VI apparaissent séduisantes, en ce qu'elles permettent théoriquement de prendre en compte à la fois les facteurs de confusion mesurés et non mesurés, elles reposent sur des hypothèses qui sont souvent peu familières aux chercheurs et qui ne sont pas complètement vérifiables, si bien que ces méthodes peuvent produire des estimations biaisées (Brookhart and Schneeweiss 2007, Grootendorst 2007, Baser 2009, Crown, Henk et al. 2011).

En premier lieu, il est bien souvent difficile de trouver un instrument approprié c'est-à-dire qui vérifie les trois conditions énoncées plus tôt (p.58). Les catégories d'instrument les plus fréquemment utilisées dans les études reposent sur des variations géographiques (exemple : proportion des individus qui bénéficient de l'intervention dans une zone géographique), des variations relatives à la structure où est délivrée ou proposée l'intervention (exemple : proportion des individus qui bénéficient de l'intervention dans un établissement de soins), des variations relatives au prescripteur ou à l'effecteur de l'intervention (exemple : taux de

prescription de l'intervention dans la patientèle d'un médecin, qui est une variable proxy de la préférence du médecin), la distance entre le domicile du patient et une structure d'intérêt (exemple : hôpital) ou enfin un paramètre temporel (exemple : la date ou l'heure d'admission à l'hôpital)(Davies, Smith et al. 2013, Garabedian, Chu et al. 2014). Seule l'hypothèse (i), c'est-à-dire l'association entre l'intervention et la VI, peut être vérifiée empiriquement, par le calcul d'odds-ratio, du coefficient R^2 ou de la statistique F. Lorsque cette hypothèse n'est pas respectée, on dit que l'instrument est faible ('*weak instrument*'). La précision de l'estimation diminue (et donc la taille de l'intervalle de confiance augmente) d'autant plus que la force de l'association entre l'instrument et l'intervention diminue (Martens, Pestman et al. 2006).

En revanche, les deux autres hypothèses ne peuvent être que justifiées théoriquement. Si certains tests peuvent permettre de conclure qu'elles ne sont pas respectées, aucun ne peut démontrer qu'elles le sont. Pour tester la condition d'indépendance (ii), on ne peut pas faire davantage que constater la non mise en évidence d'une association entre l'instrument et les facteurs de confusion mesurés, par exemple en décrivant les covariables selon les différentes valeurs de l'instrument et les groupes d'exposition. Cette vérification est insuffisante, puisque les facteurs de confusion non mesurés ne peuvent évidemment pas être étudiés ainsi. Ainsi, l'utilisation de la différence standardisée (déjà évoquée pour la comparabilité des groupes avec le SP, §A.5.2.1.2 p.44) pour effectuer cette vérification de la validité de l'instrument a été critiquée : il a été montré que la constatation de différences standardisées proches de zéro pour les variables mesurées ne garantissait pas l'absence d'association entre l'instrument et les facteurs non mesurés, et donc que l'utilisation de cette méthode diagnostique du respect de la condition (ii) pouvait être faussement rassurante (Ali, Uddin et al. 2014). La description des covariables selon les différentes valeurs de l'instrument et les groupes d'exposition est malgré tout classiquement préconisée dans les recommandations, à défaut d'autre méthode disponible (Martens, Pestman et al. 2006, Brookhart, Rassen et al. 2010, Boef, Dekkers et al. 2013,

Davies, Smith et al. 2013, Baiocchi, Cheng et al. 2014). Pourtant, cette description peut être trompeuse : la présentation de faibles déséquilibres des covariables ne préjuge pas de l'importance du biais dans l'analyse (Jackson and Swanson 2015), dans la mesure où le biais de confusion généré par le non-respect de la condition (ii) est par ailleurs amplifié par la faiblesse de l'instrument (condition (i))(Hernan and Robins 2006). L'équation (3) (p.60) montre que la faiblesse de l'association entre l'instrument et l'exposition (le dénominateur) rend l'estimateur particulièrement sensible à des petites variations du numérateur. Ainsi, une analyse avec VI peut être en fait plus biaisée qu'une analyse sans VI, alors que les facteurs de confusion paraissent plus équilibrés en tenant compte de l'instrument, parce que la condition (ii) n'est pas vérifiée et que, en plus, l'instrument est insuffisamment associé à l'intervention. Pour remédier à cela, Jackson et Swanson ont récemment proposé une méthode diagnostique graphique qui permet de comparer le biais résiduel de l'analyse avec ou sans VI, et ainsi de discuter la pertinence de recourir à une telle analyse, et de comparer plusieurs instruments entre eux (Jackson and Swanson 2015). De la même manière, la condition d'exclusion (iii) n'est pas vérifiable empiriquement, et son non-respect introduit un biais qui est amplifié par la faiblesse de l'instrument (Hernan and Robins 2006).

Pourtant, en pratique, malgré ce risque de biais en utilisant une VI, la plupart des études publiées ne rapportent pas les informations permettant d'étayer le respect des conditions de validité de l'instrument (Chen and Briesacher 2011, Davies, Smith et al. 2013). Eu égard à la condition d'indépendance (iii) qui n'est pas vérifiable, la plupart des instruments choisis seraient même vraisemblablement mis en défaut (Garabedian, Chu et al. 2014).

D'autres travaux ont permis de préciser davantage les limites des méthodes basées sur une VI.

Un certain nombre d'entre eux se sont intéressés à la taille de l'échantillon analysé et ont mis en avant l'importance de ce paramètre. Comme nous l'avons déjà noté, l'analyse avec VI est

sensible au biais, si les conditions (ii) et/ou (iii) ne sont pas vérifiées, tout particulièrement si l'instrument n'est pas fortement associé à l'intervention ('*weak instrument*'). Cela est d'autant plus vrai que l'effectif est petit (Bound, Jaeger et al. 1995, Small and Rosenbaum 2008). En d'autres termes, plus l'effectif est petit, plus l'instrument doit être fortement associé à l'intervention pour parvenir à un biais de même taille (Martens, Pestman et al. 2006). Même lorsque les hypothèses relatives à l'instrument sont vérifiées, la performance de l'analyse avec VI par rapport à l'analyse conventionnelle dépend fortement de la taille de l'échantillon (Boef, Dekkers et al. 2014). Si l'échantillon est petit, la faiblesse de l'instrument suffit à biaiser l'estimation, même si les autres conditions sont respectées (Martens, Pestman et al. 2006). Des échantillons de plusieurs milliers d'individus sont habituellement nécessaires pour que l'analyse avec VI fournisse une estimation plus fiable de l'effet que l'analyse sans VI (Boef, Dekkers et al. 2014). Toutefois, travailler sur un échantillon de grande taille n'est pas la garantie d'une analyse non-biaisée (Bound, Jaeger et al. 1995).

Les deux réserves suivantes sont habituellement moins connues. La première est la limite théorique des méthodes basées sur une VI dans les situations où le biais de confusion est important. En effet, il est montré que plus la confusion est importante, plus l'association théorique maximale entre l'instrument et l'intervention est faible (Cepeda, Boston et al. 2003). Dans un tel cas, une augmentation de la force de l'instrument est possible mais elle s'opère au prix de l'introduction d'une association de plus en plus forte entre l'intervention et l'erreur E dans l'équation (1), c'est-à-dire au prix d'une violation de l'hypothèse (iii)(Martens, Pestman et al. 2006). Ainsi, dans les situations où il existe une confusion importante, l'utilisation d'une VI aboutira nécessairement à une estimation biaisée par le non-respect des conditions (i) et/ou (iii).

La seconde réserve est l'existence d'une hypothèse additionnelle de la méthode souvent négligée. Si le respect des trois conditions relatives à l'instrument évoquées plus haut autorise à estimer les bornes inférieure et supérieure de l'effet, une hypothèse additionnelle doit être vérifiée pour permettre le calcul d'une estimation ponctuelle (Hernan and Robins 2006, Swanson and Hernan 2014). On pourra estimer l'effet *LATE* équivalent aux effets *ATE* et *ATT* (*cf.* §A.5.2.1.4, p.49), si l'effet de l'intervention est homogène dans la population étudiée ou s'il est hétérogène mais que l'hétérogénéité n'est pas liée à l'instrument (Hernan and Robins 2006, Swanson and Hernan 2014). Autrement, l'effet *LATE* ne sera pas équivalent aux effets *ATE* et *ATT*, et son évaluation nécessitera d'être en configuration de monotonicité (Hernan and Robins 2006, Swanson and Hernan 2013). La monotonicité signifie que l'instrument ne peut influencer l'allocation de l'intervention que dans une direction, et donc qu'on ne rencontre aucun individu *defiers* (**Tableau 2**). C'est-à-dire, dans la situation du tableau 2 illustrant le cas d'un instrument dichotomique basé sur la préférence du médecin pour l'un ou l'autre de deux traitements, qu'aucun patient ne recevrait le traitement B s'il était vu par un médecin préférant le traitement A, alors qu'il recevrait le traitement A s'il était vu par un médecin préférant le traitement B. Cependant, le respect de la monotonicité, qui ne peut pas être vérifié, est considéré comme improbable dans une étude observationnelle en utilisant un instrument basé sur la préférence (Swanson, Miller et al. 2015, Boef, le Cessie et al. 2016). De plus, Swanson et coll. ont montré qu'il n'était pas possible que la population étudiée soit distribuée de façon stricte entre les trois types '*always-takers*', '*never-takers*' et '*compliers*', et que la monotonicité n'était habituellement pas respectée, induisant un biais de l'estimation dont la direction n'était pas prévisible (Swanson, Miller et al. 2015).

Enfin, il a été souligné qu'une analyse avec VI ne prenant en compte qu'une partie des interventions possibles (par exemple, intervention A *vs* intervention B, sans retenir les

individus ne recevant aucune intervention ou recevant une intervention C) était nécessairement biaisée (Swanson, Robins et al. 2015), et que les méthodes basées sur une VI n'étaient pas adaptées aux exposition dépendantes du temps (Hernan and Robins 2006).

Pour mettre en œuvre de la manière la plus appropriée les méthodes basées sur une VI et se prémunir au mieux des risques de biais fortuits, des recommandations ont été proposées (Brookhart, Rassen et al. 2010, Davies, Smith et al. 2013, Swanson and Hernan 2013, Garabedian, Chu et al. 2014).

Pour conclure, les méthodes basées sur une VI ont l'avantage théorique indéniable sur les méthodes conventionnelles et celles utilisant un SP, de prendre en compte les facteurs de confusion non mesurés et/ou non connus. Cependant, en pratique, leur mise en œuvre soulève des contraintes majeures, en particulier les difficultés à identifier un instrument valide, l'impossibilité de vérifier par des tests la validité de l'instrument, la grande sensibilité aux biais et la nécessité d'effectifs importants. Des travaux de recherche doivent être menés pour chercher à résoudre ces contraintes. Actuellement, ces méthodes paraissent devoir être réservées aux études sur de larges échantillons, lorsqu'on dispose d'un instrument fortement associé à l'intervention.

A.5.2.3. Les résultats produits par une méthode utilisant un score de propension et ceux produits par une méthode utilisant une variable instrumentale sont-ils comparables ? Revue de la littérature (article 1).

A.5.2.3.1. Contexte

Comme nous l'avons vu précédemment, la recherche bibliographique sur les méthodes d'analyse permettant le limiter le biais d'indication fait apparaître l'émergence récente et la diffusion rapide des méthodes utilisant un SP ou une VI. Au cours de notre travail, nous avons même observé qu'un nombre croissant d'études publiées utilisaient de façon conjointe les deux méthodes pour évaluer une même intervention. Si les méthodes utilisant un SP et celles utilisant une VI ont fait l'objet, respectivement, de plusieurs revues critiques, aucun travail de recensement et d'analyse des études les utilisant conjointement n'avait encore été effectué. Dans le cadre de cette thèse, nous avons réalisé ce travail en menant une revue systématique de la littérature pour évaluer la concordance de ces deux méthodes. Nous voulions notamment savoir si l'utilisation des deux méthodes avait permis aux auteurs des études de parvenir aux mêmes conclusions.

Ce travail a fait l'objet d'une publication dans la revue *Journal of Clinical Epidemiology* (Laborde-Casterot, Agrinier et al. 2015), reproduit dans les pages suivantes.

ELSEVIER

CrossMark

Journal of Clinical Epidemiology 68 (2015) 1232−1240

Journal of Clinical Epidemiology

Performing both propensity score and instrumental variable analyses in observational studies often leads to discrepant results: a systematic review

Hervé Laborde-Castérot[a,b], Nelly Agrinier[a,c], Nathalie Thilly[a,c,*]

[a]*Lorraine University, Paris-Descartes University, EA 4360 Apemac, Avenue de la forêt de Haye, 54500 Vandoeuvre-lès-Nancy, France*
[b]*Université Paris 13, Sorbonne Paris Cité, UFR SMBH, 1 rue de Chablis, 93017, Bobigny, France*
[c]*Clinical Epidemiology and Evaluation, CIC-EC CIE6 Inserm, University Hospital of Nancy, Allée du Morvan, 54500 Vandoeuvre-lès-Nancy, France*

Accepted 2 April 2015; Published online 8 April 2015

Abstract

Objectives: Propensity score (PS) and instrumental variable (IV) are analytical techniques used to adjust for confounding in observational research. More and more, they seem to be used simultaneously in studies evaluating health interventions. The present review aimed to analyze the agreement between PS and IV results in medical research published to date.

Study Design and Setting: Review of all published observational studies that evaluated a clinical intervention using simultaneously PS and IV analyses, as identified in MEDLINE and Web of Science.

Results: Thirty-seven studies, most of them published during the previous 5 years, reported 55 comparisons between results from PS and IV analyses. There was a slight/fair agreement between the methods [Cohen's kappa coefficient = 0.21 (95% confidence interval: 0.00, 0.41)]. In 23 cases (42%), results were nonsignificant for one method and significant for the other, and IV analysis results were nonsignificant in most situations (87%).

Conclusion: Discrepancies are frequent between PS and IV analyses and can be interpreted in various ways. This suggests that researchers should carefully consider their analytical choices, and readers should be cautious when interpreting results, until further studies clarify the respective roles of the two methods in observational comparative effectiveness research.

Keywords: Instrumental variable; Propensity score; Confounding by indication; Observational studies; Comparative effectiveness research; Statistical methods

1. Introduction

Evidence-based medicine has conferred on randomized controlled trials (RCTs) a high level of evidence concerning the results on efficacy of clinical interventions. RCTs minimize bias and control confounding and are therefore considered the gold standard of design validity [1]. However, efficacy does not necessarily mean effectiveness. RCTs are generally conducted under ideal conditions, among highly selected patients followed by hyperspecialized physicians, and often fail to demonstrate the generalizability of their results in a real-world setting [2]. Moreover, it is not always possible, for practical or ethical reasons, to carry out an RCT [2]. Thus, complementary, or alternative approaches when RCTs are not possible, is needed to evaluate the effectiveness of clinical interventions. Well-designed observational studies may be useful to evaluate real-world usage patterns and the effects of clinical interventions [2,3]. However, they are prone to bias, particularly to confounding by indication: assignment to intervention does not occur by chance but depends on patient characteristics that can influence the effect of the intervention on the outcome [4]. In other words, the apparent effectiveness of an intervention may be explained by preintervention differences in risk factors between patients who received the

Conflict of interest: None.
Funding: None.

* Corresponding author. Service d'Epidémiologie et Evaluation cliniques—CHU de Nancy, Hôpitaux de Brabois. Allée du Morvan - 54500 Vandœuvre-lès-Nancy. Tél.: (33) 3-83-85-21-63; fax: (33) 3-83-85-12-05.
E-mail address: n.thilly@chu-nancy.fr (N. Thilly).

http://dx.doi.org/10.1016/j.jclinepi.2015.04.003
0895-4356/

What is new?

Key findings

- More and more observational studies simultaneously use propensity score (PS) and instrumental variable (IV) approaches to evaluate the same intervention, often leading to nonconcordant results that may be difficult to interpret.
- Discrepancies between results of the two methods can be explained by an expected difference in the control of confounding by indication, but also by theoretical differences between methods or by unintended consequences of inappropriate use.

What is the implication and what should change now?

- Researchers should be aware of the impact on results of the analytical technique chosen (PS or IV) and its appropriate use in observational studies, and readers should be cautious in the interpretation of results.
- Further studies are needed to investigate the agreement between PS and IV analyses in particular settings and precise the indications for each method.

intervention and those who did not. Consequently, analytical techniques are needed to address the problem of confounding in observational data [5]. Because of its ability to control for numerous potentially confounding factors, propensity score (PS) analysis, proposed by Rubin and Rosenbaum [6,7], has been increasingly used in this context for 15 years. PS represents the likelihood of a patient being assigned to an intervention on the basis of his or her preintervention characteristics. It may be applied in different ways: matching, stratification, adjustment, and weighting [8]. As compared with traditional regression models, the number of potential confounders considered in the analysis is not limited. Thus, PS methods theoretically increase comparability between groups by creating pseudorandomization of all possible measured confounders. However, even if PS method is able to reduce bias due to all measured confounders, it fails to limit bias due to unmeasured or unknown confounders [7].

To address this limit, instrumental variable (IV) analysis [9,10], widely used in economy during the last decades [11], has recently emerged in the field of clinical research and is increasingly used. This technique compares patient groups according to an IV—also named instrument—which is randomly distributed, rather than comparing patients with respect to the actual intervention received. A critical step is to find an appropriate instrument that should meet three requirements: (1) to be associated with the intervention (relevancy assumption); (2) not to directly affect the outcome of interest, but only indirectly affect it through the intervention assignment (exclusion restriction); and (3) to be independent of confounders [9,10]. Theoretically, if properly implemented, IV analysis differs from PS analysis in that it also aims to control for unmeasured or unknown confounders.

To date, the optimal approach (PS or IV) to adjust for confounders in observational studies has remained unclear, and researchers tend increasingly to use both methods—and compare results—to evaluate the effectiveness of an intervention. The way to use PS on one hand and IV on the other in medical research has been the focus of several reviews [8,12–17], but no publication has yet compared results obtained with both methods for the same analyses. These comparisons would help understand whether both methods lead to concordant results or not and interpret their respective findings. The purpose of this article was to systematically review the current medical literature in which the effects of interventions are estimated by both PS and IV analyses and to discuss the agreement between these two analytical methods.

2. Methods

2.1. Search strategy

A comprehensive search of the literature in MEDLINE and in the medical research part of Web of Science was performed to identify all published observational studies that evaluated a clinical intervention using both PS and IV analyses. We used the combination of terms "propensity" and "instrumental variable(s)," and limited the research to studies published up to March 31, 2014 (see Appendix at www.jclinepi.com, for search strategy). Additional publications have been identified by screening the references of the full-text articles selected.

2.2. Study selection

Records identified using the previously mentioned databases were independently screened by two authors on their titles and abstracts. Studies were eligible for the analysis if they satisfied the following selection criteria: (1) evaluation of an intervention using both PS and IV methods for the same analyses; (2) reporting of quantitative results for both methods, even if they were expressed differently, with a confidence interval (95% CI) or *P*-value for significance testing; (3) use of morbidity or mortality criteria as outcomes; and (4) publication in English. Records with insufficient description of the methodology, such as brief reports or congress abstracts, were not considered for review. If insufficient information on selection criteria was available in the abstract, the full-text article was considered. Finally, studies that met the criteria after a full-text assessment were included in the review.

2.3. Data abstraction

Two authors independently extracted the following information from included studies: year of publication, population studied, type of intervention, outcome, PS applications (adjustment, matching, stratification, or weighting) and results, type of IV and IV analysis results, and statistical significance obtained with both methods. When several PS applications were used for the same association, all were considered. If an outcome was assessed at different times (e.g., mortality at 1, 2, and 3 years), only the main endpoint was considered. Similarly, results from sensitivity analyses were not considered. Estimations of intervention effects were occasionally expressed using different target parameters (hazard ratio, odds ratio (OR), relative risk, risk difference, and relative reduction in risk). In such situations, to facilitate comparisons, PS and IV results were presented using the same target parameter, if provided by authors.

2.4. Agreement between PS and IV analyses

PS and IV analyses were considered as concordant when they both showed no statistically significant effect or a significant effect in the same direction. Agreement on the statistical significance of results between the two methods was measured using Cohen's kappa coefficient. When both methods reported a significant effect in the same direction for the same target parameter, a quantitative comparison was made by calculating the ratio of effect size using IV analysis to effect size using PS analysis. If several PS applications were used for the same outcome, an average PS effect size was defined as the arithmetic mean of the different PS estimations and used in the calculation.

As a sensitivity analysis, PS and IV effect estimations were reclassified and considered as concordant when both estimations pointed toward the same direction, irrespective of statistical significance.

3. Results

3.1. Study characteristics

The Preferred Reporting Items of Systematic reviews and Meta-Analyses (PRISMA) flowchart summarizing the data collection process is presented in Fig. 1. Thirty-seven articles met the selection criteria and were included in the review (Appendix/eTable 1 at www.jclinepi.com)

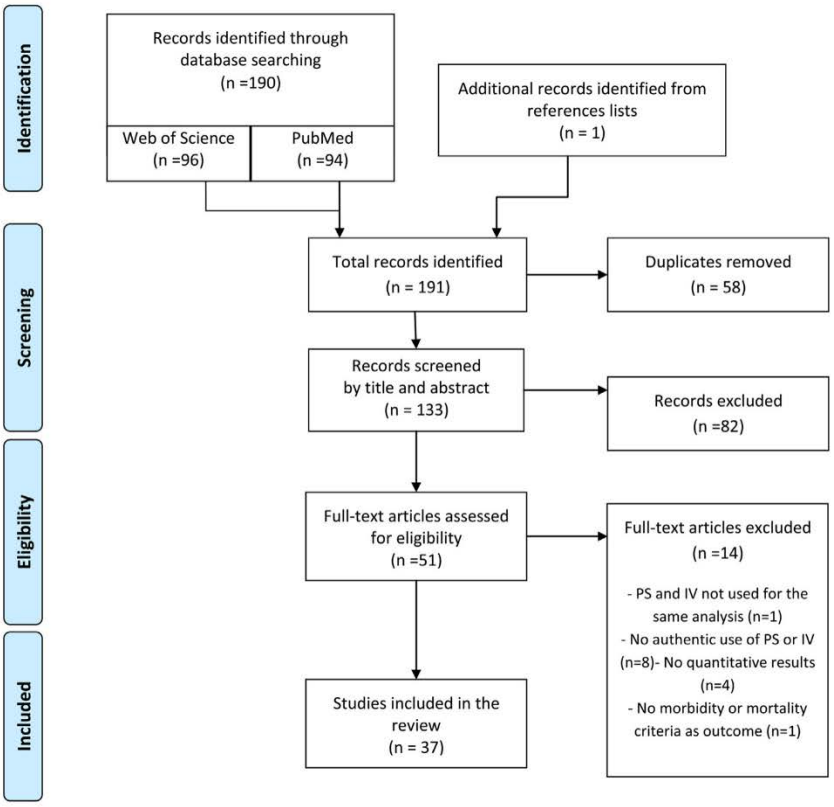


Fig. 1. PRISMA flowchart of article selection. PS, propensity score; IV, instrumental variable.

[18–54]. Most were published after 2009 (n = 31). Comparisons concerned intervention effectiveness in 27 studies, intervention safety in 9, and effectiveness and safety in 1. Oncology (n = 16) was the most frequent setting, followed by cardiology (n = 8) and psychiatry (n = 6). The intervention assessed was mainly drug exposure (n = 17), well ahead of radiotherapy (n = 6), surgery (n = 4), and interventional cardiology (n = 4).

All applications of PS were found: matching (n = 20), adjustment (n = 18), stratification (n = 9), and weighting (n = 8). Fourteen studies used several PS applications to estimate the same parameter (2 applications in 11 studies, 3 in 2 studies, and 4 in 1 study). For IV analyses, the most commonly used instrument types were geographic (n = 12, such as intervention rate in the area or distance between patient's residence and health center), related to physicians practice (n = 11, physician's prescription patterns, such as physician preference) and related to facilities (n = 7, intervention patterns in the facility where the patient is treated, such as facility intervention rate).

3.2. Comparison between PS and IV results

Our data collection abstracted a unique main result obtained by PS and IV analyses in 30 studies (81%) (Appendix/eTable 1 at www.jclinepi.com). The seven other studies reported several results for at least one of the two methods. Overall, 55 results allowing us to compare PS and IV methods were considered. The agreement between results from PS and IV analyses is presented in Table 1. Both methods gave concordant results in 31 cases (56%). Results were nonsignificant for one method and significant for the other in 23 cases (42%), and nonsignificant results were obtained by the IV analysis in most situations (87%). Cohen's kappa coefficient for agreement was 0.21 (95% CI: 0.00, 0.40). Overall, nineteen studies (51%) stated that PS and IV results were concordant, 14 (38%) that results were nonconcordant, and four (11%) reported several results, some concordant and some not.

Table 1. Concordance between results from propensity score and instrumental variable analyses

	Instrumental variable analysis results		
Effect result	**Significant effect in the same direction as PS**	**Nonsignificant effect**	**Significant effect in the opposite direction of PS[a]**
Propensity score analysis results			
Significant effect	18	20	1
Nonsignificant effect	3	13	0

Abbreviation: PS, propensity score.

Cohen's kappa coefficient = 0.21 (95% confidence interval: 0.00, 0.40).

[a] To calculate Cohen's kappa coefficient and be in the configuration of two judges and two modalities, the only IV significant result in the opposite of a significant PS result was considered as nonsignificant.

In the sensitivity analysis, 16 results considered as nonconcordant in the primary analysis were reclassified as concordant, five results considered as concordant were reclassified as nonconcordant, and one result considered as nonconcordant could not be reclassified because the hazard ratio from IV analysis equaled 1. Finally, 42 results (76%) were considered as concordant, 12 (22%) as nonconcordant, and 1 (2%) was unclassifiable. Cohen's kappa coefficient was 0.53 (95% CI: 0.26, 0.79) (moderate agreement). The mean ratio of effect size by IV analysis to effect size by PS analysis was 3.39 (95% CI: −0.25, 7.03).

Eleven concordant significant results obtained in the primary analysis were expressed with the same target parameter, which allowed us to compare results quantitatively (Table 2). In three of them, the IV estimation did not deviate by more than 15% from the PS estimation; the IV analysis estimations were further from the null in six cases, and closer to the null in two, compared with the PS analysis findings. The mean ratio of effect size by IV analysis to effect size by PS analysis was 1.32 (95% CI: 0.89, 1.75).

Discrepant results have been interpreted in several ways. In eight studies, authors considered IV analysis provided the true estimations [30,35,36,38–40,52,54]. In five studies, results have been interpreted as concordant, despite the nonsignificance of IV results and the significance of PS results [25–27,50,51]. Four studies discussed the advantages and limits of both methods without concluding that one method was superior to the other [34,37,43,44]. In the last study reporting discrepant results, authors did not discussed these results [20].

4. Discussion

Most of the studies considered in this review were published in the previous 5 years, confirming the trend toward simultaneous use of both analytical methods when evaluating effectiveness to better account for bias due to observational design. A large proportion of results using these two methods were nonconcordant [Cohen's kappa coefficient 0.21 (95% CI: 0.00, 0.40)], raising the question of what is the best analytical approach to controlling confounding by indication in observational studies evaluating the effectiveness or safety of interventions.

Although most of the studies also reported results of traditional risk-adjustment analyses, these results were not considered here as previous reviews showed that they have similar results to PS analysis [8,14]. Almost half of the studies used several PS applications that often yielded similar results. Although this suggests the equivalence of PS applications, it may be due to publication bias whereby authors only report results from different applications when they are concordant.

PS and IV methods are analytical approaches used in observational studies to consistently estimate the effects of intervention despite confounding by indication. When both methods give similar results, it tends to strengthen the credibility of the findings. It may be argued that bias

Table 2. Studies reporting concordant significant results expressed with the same target parameter

Authors	Patients studied	Type of intervention	Outcome	Propensity score analysis, target parameter PS application (A, M, S, W) estimation (95% CI)	Instrumental variable analysis, type of IV	Target parameter estimation (95% CI)	Ratio of effect size IV/PS
Suaya [28]	Coronary conditions	Cardiac rehabilitation	Mortality	RRR M 33.70% ($P < 0.001$)	Distance patient's residence—center and density of cardiac rehabilitation facilities	RRR 21.20% ($P < 0,001$)	0.63
Pratt [31]	Patients receiving antipsychotic medication	Conventional vs. atypical antipsychotic medication	Mortality	RD% A 11.5 (10.0, 13.0)	Physician's preference	RD% 23.8 (17.6, 30.0)	2.07
Chuang [21]	Recurrent epithelial ovarian, tubal, and peritoneal cancers	Secondary cytoreductive surgery	Mortality	HR M 0.75 (0.64, 0.86) HR S 0.73 (0.65, 0.82) HR W 0.71 (0.65, 0.77) HR A 0.72 (0.63, 0.83)	Oncologist's preference	HR 0.75 (0.65, 0.86)	0.92
Federspiel [35]	Acute coronary syndrome	Drug-eluting stents vs. bare metal stents	Repeat revascularization	HR M 0.9 (0.87, 0.93)	Month of treatment	HR 0.76 (0.63, 0.89)	2.40
Newman [38]	Newborns with hyperbilirubinemia	Phototherapy	Bilirubinemia	OR A 0.2 (0.1, 0.3)	Proportion of infants who had a particular bilirubinemia at hospital	OR 0.05 (0.015, 0.15)	1.19
Sheets [39]	Localized prostate cancer	Intensity-modulated radiation therapy vs. proton therapy	Gastrointestinal events	RR M 0.66 (0.55, 0.79)	Radiation therapy oncology group affiliation	RR 0.66 (0.49, 0.88)	1.00
Jacob [44]	Patients eligible for colorectal cancer (CRC) screening	Colonoscopy screening	CRC incidence	RD W −0.54 (−0.69, −0.39) RD M −0.56 (−0.99, −0.12)	Physician's rate of discretionary colonoscopy	RD −0.6 (−0.78, −0.31)	1.09
		Colonoscopy screening	Mortality	RD W −0.08 (−0.15, −0.02) RD M −0.10 (−0.24, −0.04)	Physician's rate of discretionary colonoscopy	RD -0.17 (−0.21, −0.14)	1.89
Margolis [45]	Diabetic foot ulcer	Hyperbaric oxygen therapy	Healing	HR A 0.68 (0.63, 0.73) HR S 0.61 (0.58, 0.65) HR M 0.68 (0.63, 0.73)	Center's treatment rate	HR 0.43 (0.35, 0.52)	1.66
Pöttgen [46]	Stage IIIA/B non–small cell lung cancer	Accelerated hyperfractionated radiotherapy vs. conventional fractionation	Complete response	OR W 3.04 (1.93, 4.76) OR W 3.73 (2.09, 6.67)	Year of treatment and trial participation	OR 1.67 (1.22, 2.17)	0.28
Beadle [50]	Head and neck cancer	Intensity-modulated radiation therapy	Mortality	HR M 0.72 (0.59, 0.90)	Provider experience	HR 0.60 (0.41, 0.88) Overall mean ratio	1.43 1.32 (0.89, 1.75)

Abbreviations: target parameter: PS, propensity score; CI, confidence interval; IV, instrumental variable; RRR, relative reduction in risks; RD, risk difference; HR, hazard ratio; OR, odds ratio; RR, relative risk.

Studies are sorted by year of publication.

Propensity score applications: A, adjustment; M, matching; S, stratification; W, weighting.

is fully controlled by taking into account measured potential confounders. However, residual confounding remains possible and does not necessarily alter the findings in a way that makes both results conflicting, especially when both methods were unable to fully correct for confounding. However, when the two methods give nonconcordant results, what result should be considered? IV results are often considered by authors to be the "true" estimations because of the ability of IV analysis to control bias due to unmeasured or unknown confounders in addition to measured ones [30,35,36,38–40,52,54]. Nevertheless, it is important to be aware of the limits of IV and not to unquestioningly consider IV analysis to be an "epidemiologist's dream which comes true" [55,56].

The primary challenge to implementing a valid IV analysis is to find a reliable instrument. Meeting that challenge is complicated and not even always possible. As mentioned in introduction, a good instrument should be strongly associated with the intervention and indirectly affects the outcome of interest through the intervention assignment. Moreover, the instrument should be independent of confounders. Davies et al. [15] reviewed 90 studies using IV analysis and revealed frequent use of instruments that were inappropriate because they did not fulfill the above conditions. Consequently, studies using IV analysis may report results that are in fact biased, although they are presented as being exempt from residual confounding. For instance, using a weak instrument (i.e., an instrument not strongly associated with the intervention) leads to various consequences [57−60], such as biased or imprecise (large standard error) estimations. In our review, when PS and IV results were not concordant, PS analysis reported a significant effect and IV analysis did not, in most cases (87%). The nonsignificant effect more often found with IV analysis may be the consequence of a lack of power due to the imprecise estimations. Paradoxically, when PS and IV results were concordant and statistically significant, the effect size was on average 32% larger in magnitude in the IV analysis than the PS analysis. However, the use of weak instruments in IV analysis can either increase or decrease the average intervention effect depending on the sign of the covariance between the instrument and the intervention as well as its relationship with the outcome directly [57,60]. This result yet requires confirmation because comparison of the effect size was limited to 11 of the studies included and did not reach statistical significance. Such a trend was, for example, reported by Newman et al. [41] in their study of the efficacy of phototherapy for newborns with hyperbilirubinemia. They showed a significant effect of phototherapy with an OR of 0.2 by PS analysis and an OR of 0.05 by IV analysis. The authors judged the IV estimation implausible, and they explained the result in terms of the choice of an instrument highly associated with one of the confounders. Indeed, weak instruments tend to lead to residual correlation of the instrument with the error term in the outcome equation [57]. Under these circumstances, IV can no longer be assumed to be an unbiased estimator.

Some might be inclined to compare results obtained with RCTs and observational studies assessing the same intervention when both designs are available. Reviews focusing on the agreement between results from PS approach and RCT report conflicting conclusions, with either less [61] or more [62] effect of the intervention with PS than with RCT. Concerning IV method, the review from Chen and Briesacher [16] reported that 15 studies of 16 considered showed concordant results between IV and RCT. Similar comparisons with RCT results, considered as the gold standard, were made in several studies in the present review [18,28,30,35,38,40,54]. In most studies, similar findings between IV and RCT were considered as guaranteeing unbiased results. However, even if the IV method fully controls bias due to unbalanced confounders between both groups, the characteristics of patients considered in observational studies generally differ enough from those included in RCT to obtain not surprisingly different results. Thus, considering IV method superior to PS on the basis of results similar to those achieved with RCT may not be appropriate.

When comparing PS and IV results, it is important to keep in mind the conceptual difference between the two methods [63]. In current practice, allocation of treatments or interventions of interest often results from professional decision or choice. According to the PS method used (adjustment, matching, stratification, and weighting) [64], PS analysis produces estimations of either the average treatment effect (ATE) at the population level or the average treatment/intervention effect in the treated patients (ATT), whereas IV analysis produces estimations of the local average treatment/intervention effect (LATE) for marginal patients. Marginal patients are the subset of patients whose treatment/intervention choices are affected by variation in specific factors named ''instruments'' in IV analysis, for example, physician preference [65]. When intervention effects are homogeneous within the study sample or when they are heterogeneous, but the intervention decision is not related to this intervention effects heterogeneity (e.g., when there is no available evidence suggesting who will benefit more from the intervention, although there is true underlying heterogeneity of intervention effects across patients), ATE/ATT and LATE are equivalent. In this case, PS estimations equal to IV estimations. When intervention effects are heterogeneous and the intervention decision is related to this heterogeneity (e.g., when there is clinical evidence suggesting that certain subgroups of patients are more prone to benefit from a particular intervention and physicians sort patients with a greater intervention benefit toward intervention), ATE/ATT and LATE are different estimations. In this case, PS and IV estimations may differ despite both being completely correct.

Finally, Fig. 2 provides a flowchart for readers whenever confronting with nonconcordant PS and IV analyses results.

This analysis of agreement between PS and IV results reflects the use of these methods in effectiveness and/or safety evaluation studies to date. Results should be interpreted in the light of three potential limitations. First, a publication bias is always plausible, particularly if authors were more likely to report results from the two methods when they are concordant. Second, the use of different target parameters to express results was a true obstacle to compare the effect size between the two methods. Last, our definition of concordance based on the statistical significance raises concerns about results interpretation. For example, some effect estimations appeared to be similar but were classified as nonconcordant results because one reached the statistical significance and the other did not. In the sensitivity analysis defining concordant results as

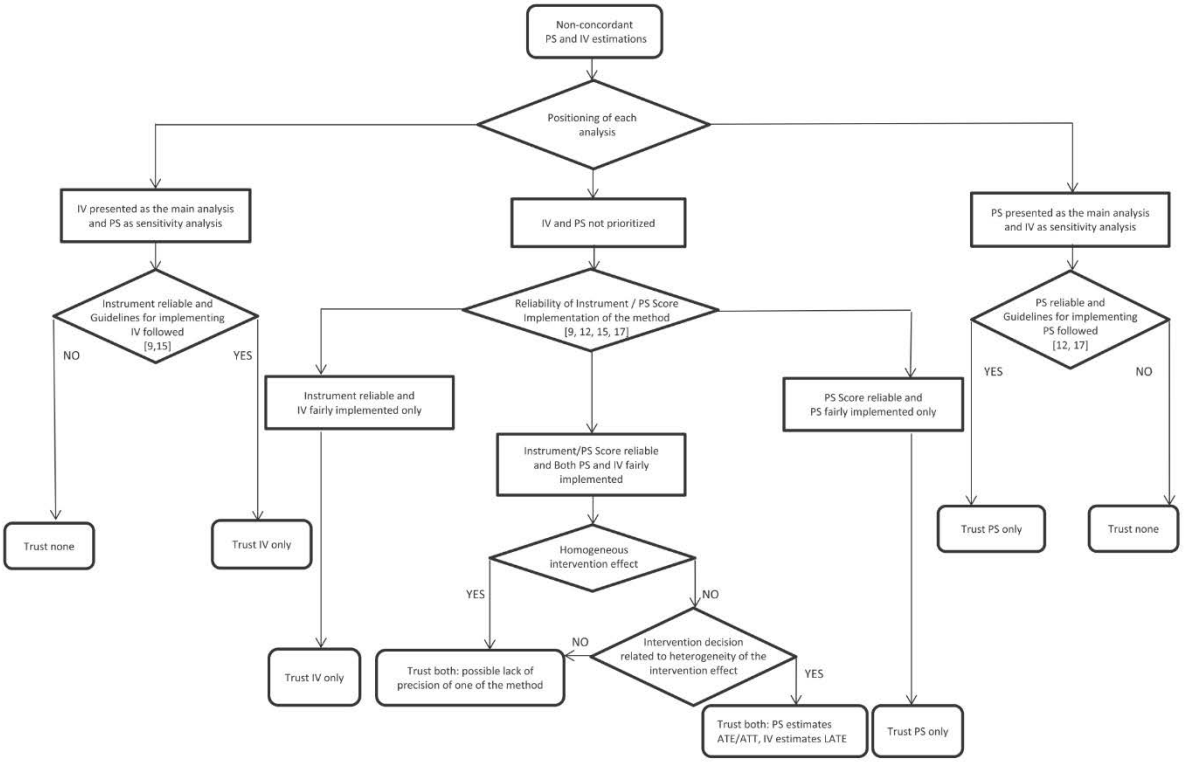


Fig. 2. Flowchart for interpretation of nonconcordant propensity score (PS) and instrumental variable (IV) analyses results both used in observational studies. ATE, average treatment effect; ATT, average treatment/intervention effect in the treated patients; LATE, local average treatment/intervention effect.

estimations pointing toward the same direction whatever the *P*-value, 76% of the results were then considered as being concordant, raising the agreement between PS and IV analysis from fair to moderate. Nevertheless, such a broad definition of concordance does not consider effect size and leads to classify both estimations that are very distant in magnitude as concordant, resulting in a likely overestimation of the ratio of effect size.

In conclusion, nonconcordance between PS and IV results found in this review may reflect different situations not mutually exclusive: (1) the superiority of one method vs. the other by a better control of confounders; (2) the lack of statistical power for one of the two methods; (3) the inappropriate use of one or both method(s), particularly for the IV method because of the difficulty to find a strong instrument that is actually uncorrelated with the outcome of interest; (4) the correct estimation by each method of intervention effects, which are heterogeneous across the study sample with a treatment decision related to treatment-effects heterogeneity (ATE or ATT/LATE). Further methodological research is required to precise the respective convenience of each method, PS and IV. For the moment, researchers should be aware of the impact on results of the analytical technique chosen and its appropriate use in observational studies. Likewise, readers should be cautious in the interpretation of results, especially when underlying assumptions of the technique used are violated, for example, unmeasured confounders are not ignorable when measured confounders are controlled in case of PS use, or instruments are related to the outcome directly or to unmeasured confounders in case of IV use. Under those circumstances, the technique used can no longer be assumed to result in an unbiased estimator. PS and IV methods should be theoretically justified beforehand rather than being blindly applied together. The current trend to avoid is increasing statistical tests using simultaneously PS and IV in multiple analyses, without prioritize them, and choosing arbitrarily from several discrepant results.

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jclinepi.2015.04.003.

References

[1] Abel U, Koch A. The role of randomization in clinical studies: myths and beliefs. J Clin Epidemiol 1999;52:487–97.
[2] Black N. Why we need observational studies to evaluate the effectiveness of health care. BMJ 1996;312:1215–8.

[3] Concato J, Lawler EV, Lew RA, Gaziano JM, Aslan M, Huang GD. Observational methods in comparative effectiveness research. Am J Med 2010;123:e16−23.
[4] Grimes DA, Schulz KF. Bias and causal associations in observational research. Lancet 2002;359:248−52.
[5] Klungel OH, Martens EP, Psaty BM, Grobbee DE, Sullivan SD, Stricker BH, et al. Methods to assess intended effects of drug treatment in observational studies are reviewed. J Clin Epidemiol 2004; 57:1223−31.
[6] Rosenbaum P, Rubin D. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70:41−55.
[7] Joffe MM, Rosenbaum PR. Invited commentary: propensity scores. Am J Epidemiol 1999;150:327−33.
[8] Sturmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. 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.
[9] Brookhart MA, Rassen JA, Schneeweiss S. Instrumental variable methods in comparative safety and effectiveness research. Pharmacoepidemiol Drug Saf 2010;19:537−54.
[10] Greenland S. An introduction to instrumental variables for epidemiologists. Int J Epidemiol 2000;29:1102.
[11] Stock JH, Trebbi F. Retrospectives: who invented instrumental variable regression? J Econ Perspect 2003;17:177−94.
[12] Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. Stat Med 2008;27: 2037−49.
[13] Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. Principles for modeling propensity scores in medical research: a systematic literature review. Pharmacoepidemiol Drug Saf 2004;13:841−53.
[14] Shah BR, Laupacis A, Hux JE, Austin PC. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. J Clin Epidemiol 2005;58: 550−9.
[15] Davies NM, Smith GD, Windmeijer F, Martin RM. Issues in the reporting and conduct of instrumental variable studies: a systematic review. Epidemiology 2013;24:363−9.
[16] Chen Y, Briesacher BA. Use of instrumental variable in prescription drug research with observational data: a systematic review. J Clin Epidemiol 2011;64:687−700.
[17] Ali MS, Groenwold RH, Belitser SV, Pestman WR, Hoes AW, Roes KC, 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.
[18] Earle CC, Tsai JS, Gelber RD, Weinstein MC, Neumann PJ, Weeks JC. Effectiveness of chemotherapy for advanced lung cancer in the elderly: instrumental variable and propensity analysis. J Clin Oncol 2001;19:1064−70.
[19] Punglia RS, Saito AM, Neville BA, Earle CC, Weeks JC. Impact of interval from breast conserving surgery to radiotherapy on local recurrence in older women with breast cancer: retrospective cohort analysis. BMJ 2010;340:c845.
[20] Saito AM, Landrum MB, Neville BA, Ayanian JZ, Earle CC. The effect on survival of continuing chemotherapy to near death. BMC Palliat Care 2011;10:14.
[21] Chuang C-M, Chou Y-J, Yen M-S, Chao K-C, Twu N-F, Wu H-H, et al. The role of secondary cytoreductive surgery in patients with recurrent epithelial ovarian, tubal, and peritoneal cancers: a comparative effectiveness analysis. Oncologist 2012;17:847−55.
[22] Schneeweiss S, Setoguchi S, Brookhart A, Dormuth C, Wang PS. Risk of death associated with the use of conventional versus atypical antipsychotic drugs among elderly patients. Can Med Assoc J 2007; 176:627−32.
[23] Wang PS, Schneeweiss S, Avorn J, Fischer MA, Mogun H, Solomon DH, et al. Risk of death in elderly users of conventional vs. atypical antipsychotic medications. N Engl J Med 2005;353: 2335−41.
[24] Stukel TA, Fisher ES, Wennberg DE, Alter DA, Gottlieb DJ, Vermeulen MJ. 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;297:278−85.
[25] Bateman BT, Bykov K, Choudhry NK, Schneeweiss S, Gagne JJ, Polinski JM, et al. Type of stress ulcer prophylaxis and risk of nosocomial pneumonia in cardiac surgical patients: cohort study. BMJ 2013;347:f5416.
[26] Huybrechts KF, Brookhart MA, Rothman KJ, Silliman RA, Gerhard T, Crystal S, et al. Comparison of different approaches to confounding adjustment in a study on the association of antipsychotic medication with mortality in older nursing home patients. Am J Epidemiol 2011;174:1089−99.
[27] Schneeweiss S, Seeger JD, Landon J, Walker AM. Aprotinin during coronary-artery bypass grafting and risk of death. N Engl J Med 2008;358:771−83.
[28] Suaya JA, Stason WB, Ades PA, Normand S-LT, Shepard DS. Cardiac rehabilitation and survival in older coronary patients. J Am Coll Cardiol 2009;54:25−33.
[29] Bosco JL, Silliman RA, Thwin SS, Geiger AM, Buist DS, Prout MN, et al. A most stubborn bias: no adjustment method fully resolves confounding by indication in observational studies. J Clin Epidemiol 2010;63:64−74.
[30] Hadley J, Yabroff KR, Barrett MJ, Penson DF, Saigal CS, Potosky AL. Comparative effectiveness of prostate cancer treatments: evaluating statistical adjustments for confounding in observational data. J Natl Cancer Inst 2010;102:1780−93.
[31] Pratt N, Roughead EE, Ryan P, Salter A. Antipsychotics and the risk of death in the elderly: an instrumental variable analysis using two preference based instruments. Pharmacoepidemiol Drug Saf 2010; 19:699−707.
[32] Wisnivesky JP, Halm E, Bonomi M, Powell C, Bagiella E. Effectiveness of radiation therapy for elderly patients with unresected stage I and II non-small cell lung cancer. Am J Respir Crit Care Med 2010; 181:264−9.
[33] Wisnivesky JP, Halm EA, Bonomi M, Smith C, Mhango G, Bagiella E. Postoperative radiotherapy for elderly patients with stage III lung cancer. Cancer 2012;118:4478−85.
[34] Pirracchio R, Sprung C, Payen D, Chevret S. Benefits of ICU admission in critically ill patients: whether instrumental variable methods or propensity scores should be used. BMC Med Res Methodol 2011; 11:132.
[35] Venkitachalam L, Lei Y, Magnuson EA, Chan PS, Stolker JM, Kennedy KF, et al. Survival benefit with drug-eluting stents in observational studies fact or artifact? Circ Cardiovasc Qual Outcomes 2011;4:587−94.
[36] Fang G, Brooks JM, Chrischilles EA. Comparison of instrumental variable analysis using a new instrument with risk adjustment methods to reduce confounding by indication. Am J Epidemiol 2012;175:1142−51.
[37] Parmar AD, Sheffield KM, Han Y, Vargas GM, Guturu P, Kuo Y-F, et al. Evaluating comparative effectiveness with observational data endoscopic ultrasound and survival in pancreatic cancer. Cancer 2013;119:3861−9.
[38] Federspiel JJ, Stearns SC, Sheridan BC, Kuritzky JJ, D'Arcy LP, Crespin DJ, et al. Evaluating the effectiveness of a rapidly adopted cardiovascular technology with administrative data: the case of drug-eluting stents for acute coronary syndromes. Am Heart J 2012;164:207−14.
[39] Lee C-C, Ho H-C, Hsiao S-H, Huang T-T, Lin H-Y, Li S-C, et al. Infectious complications in head and neck cancer patients treated with cetuximab: propensity score and instrumental variable analysis. PLoS One 2012;7:e50163.

[40] Suh HS, Hay JW, Johnson KA, Doctor JN. Comparative effectiveness of statin plus fibrate combination therapy and statin monotherapy in patients with type 2 diabetes: use of propensity-score and instrumental variable methods to adjust for treatment-selection bias. Pharmacoepidemiol Drug Saf 2012;21:470−84.
[41] Newman TB, Vittinghoff E, McCulloch CE. Efficacy of phototherapy for newborns with hyperbilirubinemia: a cautionary example of an instrumental variable analysis. Med Decis Making 2012;32:83−92.
[42] Sheets NC, Goldin GH, Meyer AM, Wu Y, Chang Y, Sturmer T, et al. Intensity-modulated radiation therapy, proton therapy, or conformal radiation therapy and morbidity and disease control in localized prostate cancer. JAMA 2012;307:1611−20.
[43] Valenstein M, Kim HM, Ganoczy D, Eisenberg D, Pfeiffer PN, Downing K, et al. Antidepressant agents and suicide death among US Department of Veterans Affairs patients in depression treatment. J Clin Psychopharmacol 2012;32:346−53.
[44] Bekelman JE, Handorf EA, Guzzo T, Evan Pollack C, Christodouleas J, Resnick MJ, et al. Radical cystectomy versus bladder-preserving therapy for muscle-invasive urothelial carcinoma: examining confounding and misclassification biasin cancer observational comparative effectiveness research. Value Health 2013;16:610−8.
[45] Hebert PL, McBean AM, O'Connor H, Frank B, Good C, Maciejewski ML. Time until incident dementia among medicare beneficiaries using centrally acting or non-centrally acting ACE inhibitors. Pharmacoepidemiol Drug Saf 2013;22:641−8.
[46] Huesch MD. External adjustment sensitivity analysis for unmeasured confounding: an application to coronary stent outcomes, Pennsylvania 2004-2008. Health Serv Res 2013;48:1191−214.
[47] Jacob BJ, Sutradhar R, Moineddin R, Baxter NN, Urbach DR. Methodological approaches to population based research of screening procedures in the presence of selection bias and exposure measurement error: colonoscopy and colorectal cancer outcomes in Ontario. BMC Med Res Methodol 2013;13:59.
[48] Margolis DJ, Gupta J, Hoffstad O, Papdopoulos M, Glick HA, Thom SR, et al. Lack of effectiveness of hyperbaric oxygen therapy for the treatment of diabetic foot ulcer and the prevention of amputation a cohort study. Diabetes Care 2013;36:1961−6.
[49] Poettgen C, Eberhardt W, Graupner B, Theegarten D, Gauler T, Freitag L, et al. Accelerated hyperfractionated radiotherapy within trimodality therapy concepts for stage IIIA/B non-small cell lung cancer: markedly higher rate of pathologic complete remissions than with conventional fractionation. Eur J Cancer 2013;49:2107−15.
[50] Steingrub JS, Lagu T, Rothberg MB, Nathanson BH, Raghunathan K, Lindenauer PK. Treatment with neuromuscular blocking agents and the risk of in-hospital mortality among mechanically ventilated patients with severe sepsis. Crit Care Med 2014;42:90−6.
[51] Thomas KH, Martin RM, Davies NM, Metcalfe C, Windmeijer F, Gunnell D. Smoking cessation treatment and risk of depression, suicide, and self harm in the Clinical Practice Research Datalink: prospective cohort study. BMJ 2013;347:f5704.
[52] Wright JD, Ananth CV, Tsui J, Glied SA, Burke WM, Lu Y-S, et al. Comparative effectiveness of upfront treatment strategies in elderly women with ovarian cancer. Cancer 2014;120:1246−54.
[53] Beadle BM, Liao K-P, Elting LS, Buchholz TA, Ang KK, Garden AS, et al. Improved survival using intensity-modulated radiation therapy in head and neck cancers: a SEER-Medicare analysis. Cancer 2014;120:702−10.
[54] VanDyke RD, McPhail GL, Huang B, Fenchel MC, Amin RS, Carle AC, et al. Inhaled tobramycin effectively reduces FEV1 decline in cystic fibrosis. An instrumental variables analysis. Ann Am Thorac Soc 2013;10:205−12.
[55] Hernan MA, Robins JM. Instruments for causal inference: an epidemiologist's dream? Epidemiology 2006;17:360−72.
[56] Baser O. Too much ado about instrumental variable approach: is the cure worse than the disease? Value Health 2009;12:1201−9.
[57] Bound J, Jaeger DA, Baker RM. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. J Am Stat Assoc 1995;90:443−50.
[58] Martens EP, Pestman WR, de Boer A, Belitser SV, Klungel OH. Instrumental variables: application and limitations. Epidemiology 2006;17:260−7.
[59] Boef AG, Dekkers OM, Vandenbroucke JP, le Cessie S. Sample size importantly limits the usefulness of instrumental variable methods, depending on instrument strength and level of confounding. J Clin Epidemiol 2014;67:1258−64.
[60] Crown WH, Henk HJ, Vanness DJ. Some cautions on the use of instrumental variables estimators in outcomes research: how bias in instrumental variables estimators is affected by instrument strength, instrument contamination, and sample size. Value Health 2011;14:1078−84.
[61] Zhang Z, Ni H, Xu X. Observational studies using propensity score analysis underestimated the effect sizes in critical care medicine. J Clin Epidemiol 2014;67:932−9.
[62] Dahabreh IJ, Sheldrick RC, Paulus JK, Chung M, Varvarigou V, Jafri H, 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.
[63] Fang G, Brooks JM, Chrischilles EA. Apples and oranges? Interpretations of risk adjustment and instrumental variable estimates of intended treatment effects using observational data. Am J Epidemiol 2012;175:60−5.
[64] Austin PC. A tutorial and case study in propensity score analysis: an application to estimating the effect of in-hospital smoking cessation counseling on mortality. Multivariate Behav Res 2011;46:119−51.
[65] Harris KM, Remler DK. Who is the marginal patient? Understanding instrumental variables estimates of treatment effects. Health Serv Res 1998;33:1337−60.

Appendix: Search strategy

Search Strategy for Medline

The database Medline was searched via Pubmed using the following search queries:
propensity AND (instrumental AND variable*) AND ("0001/01/01"[PDAT] : "2014/03/31"[PDAT]) AND English[lang]
This search retrieved 94 results.

Search strategy for Web of Science

At March 31, 2014 Web of Science was searched for:
TOPIC: (propensity) AND TOPIC: (instrumental variable*)
Refined by:
RESEARCH DOMAINS: (SCIENCE TECHNOLOGY)
AND RESEARCH AREAS: (ALLERGY AND ANESTHESIOLOGY AND CARDIOVASCULAR SYSTEM CARDIOLOGY AND CRITICAL CARE MEDICINE AND ENDOCRINOLOGY METABOLISM AND GENERAL INTERNAL MEDICINE AND GERIATRICS GERONTOLOGY AND INFECTIOUS DISEASES AND MICROBIOLOGY AND OBSTETRICS GYNECOLOGY AND ONCOLOGY AND ORTHOPEDICS AND PEDIATRICS AND PHARMACOLOGY PHARMACY AND PSYCHIATRY AND PSYCHOLOGY AND PUBLIC ENVIRONMENTAL OCCUPATIONAL HEALTH AND RADIOLOGY NUCLEAR MEDICINE MEDICAL IMAGING AND REHABILITATION AND REPRODUCTIVE BIOLOGY AND RESPIRATORY SYSTEM AND RHEUMATOLOGY AND SUBSTANCE ABUSE AND SURGERY AND TOXICOLOGY AND TRANSPLANTATION)
AND DOCUMENT TYPES: (ARTICLE)
AND LANGUAGES: (ENGLISH)
This search retrieved 96 results.

eTable 1. Studies characteristics and results.

Authors	Patients studied	Type of intervention	Outcome	Propensity score analysis Target parameter PS application (A, M, S, W) Estimation (95% CI)	Instrumental variable analysis Type of IV	 Target parameter Estimation (95% CI)	Statistical significance PS/IV
Earle [15]	Stage IV non-small-cell lung cancer	Chemotherapy	Mortality	HR S 0.78 to 0.85 across quintiles (all are significantly less than 1)	Chemotherapy utilization in health area	RD% -9 (-23, -4)	+/+
Wang [20]	Patients receiving antipsychotic medication	Conventional vs atypical antipsychotic medication	Mortality	HR A 1.37 (1.27, 1.49) HR S 1.17 to 1.58 across deciles	Physician's preference	RD% 0.073 (0.020, 0.126)	+/+
Schneeweiss [19]	Patients receiving antipsychotic medication	Conventional vs atypical antipsychotic medication	Mortality	HR A 1.39 (1.30, 1,49)	Physician's preference	RD% 4.2 (1.2, 7.3)	+/+
Stukel [21]	Acute myocardial infarction	Cardiac catheterization	Mortality	RR A 0.538 (0.529, 0.547) RR M 0.538 (0.518, 0.558)	Regional cardiac catheterization rate	RR 0.84 (0.79, 0.90)	+/+
Schneeweiss [24]	Patients receiving coronary-artery bypass grafting	Aprotinin vs aminocaproic acid	Mortality	RR M 1.32 (1.08, 1.63)	Surgeons' preference	RD% 0.6 (0.00, 1.21)	+/NS
Suaya [25]	Coronary conditions	Cardiac rehabilitation	Mortality	RRR M 33.70% (p<0.001)	Distance patient's residence - center and density of cardiac rehabilitation facilities	RRR 21.20% (p<0,001)	+/+
Bosco [26]	Stage I-IIB breast cancer	Adjuvant chemotherapy	Cancer recurrence	HR S 1.3 (0.8, 2.0) HR A 1.1 (0.7, 1.7)	Surgeon's preference	HR 0.9 (0.2, 4.3)	NS/NS
Hadley [27]	Stage T1-T2 prostate cancer	Conservative management vs radical prostatectomy	Mortality	HR W 1.6 (1.40, 1.83) HR W 1.39 (1.10, 1.76)	Local area treatment pattern	HR 0.73 (0.08, 6.73)	+/NS
Pratt [28]	Patients receiving antipsychotic medication	Conventional vs atypical antipsychotic medication	Mortality	RD% A 11.5 (10.0, 13.0)	Physician’s preference	RD% 23.8 (17.6, 30.0)	+/+
Punglia [16]	Stage 0-II breast cancer	Time from surgery to radiotherapy	Recurrence	HR A 1.19 (1.01, 1.39)	Distance to radiotherapy facility	RD% 0.96 (p=0,026)	+/+
Wisnivesky [29]	Unresected stage I and II non-small cell lung cancer	Radiotherapy	Mortality	HR A 0.74 (0.70, 0.78) HR S 0.70 to 0.80 across quintiles (all are significantly less than 1)	Intensity of radiotherapy use in the regions	RD% -18.1 (-38.0, -2.2)	+/+
Huybrechts [23]	Patients receiving antipsychotic medication after admission to a nursing home	Conventional vs atypical antipsychotic medication	Mortality	RD% A 7.79 (6.61, 8.96)	Nursing home prescribing preference	RD% 8.84 (-1.28, 18.95)	+/NS
Pirrachio [31]	Patients evaluated for intensive care unit admission	Intensive care unit admission	Mortality	OR M 1.23 (1.04, 1.45)	Physician's specialization	OR 0.73 (0.24, 2.45)	+/NS
Saito [17]	Advanced non-small cell lung cancer	Aggressive approach chemotherapy vs standard chemotherapy	Mortality	HR A 1.21 (1.00, 1.48) HR S 0.85 to 1.32 across quintiles (all are non-significant)	Treatment rate in health area	RD% 53.5 (17.2, 124.1)	NS/NS
		Chemotherapy vs no chemotherapy	Mortality	HR A 0.76 (0.71, 0.82)	Treatment rate in health area	RD% -6.7 (-19.9, -6.6)	+/NS

Authors	Patients studied	Type of intervention	Outcome	Propensity score analysis Target parameter PS application (A, M, S, W) Estimation (95% CI)	Instrumental variable analysis Type of IV	 Target parameter Estimation (95% CI)	Statistical significance PS/IV
				HR S 0.73 to 0.85 across quintiles (all are significantly less than 1)			
Venkitachalam [32]	Patients receiving percutaneous coronary intervention with stent replacement	Drug-eluting stents vs bare metal stents	Mortality	RD% M -1.8 (-3.3, -0.3)	Enrollment year	RD% 2.0 (-1.8, 5.7)	+/NS
		Drug-eluting stents vs bare metal stents	Revascularization	RD% M -3.0 (-4.5, -1.4)	Enrollment year	RD% -4.2 (-8.9, 0.4)	+/NS
Chuang [18]	Recurrent epithelial ovarian, tubal and peritoneal cancers	Secondary cytoreductive surgery	Mortality	HR M 0.75 (0.64, 0.86)	Oncologist's preference	HR 0.75 (0.65, 0.86)	+/+
				HR S 0.73 (0.65, 0.82)			
				HR W 0.71 (0.65, 0.77)			
				HR A 0.72 (0.63, 0.83)			
Fang [33]	New-onset and uncontrolled persistant asthma	Long-term control therapy	Acute exacerbation	RD A 0.01 (-0.014, 0.033)	Area treatment rate	RD -0.161 (-0.320, -0.001)	NS/+
				RD M 0.016 (-0.022, 0.053)			
Federspiel [35]	Acute coronary syndrome	Drug-eluting stents vs bare metal stents	Mortality	HR M 0.8 (0.77, 0.83)	Month of treatment	HR 0.99 (0.88, 1.11)	+/NS
		Drug-eluting stents vs bare metal stents	Repeat revascularization	HR M 0.9 (0.87, 0.93)	Month of treatment	HR 0.76 (0.63, 0.89)	+/+
Lee [36]	Head and neck cancer	Cetuximab	Infectious complications	OR S 2.27 (1.46, 3.54)	Treatment hospital rate	OR 0.87 (0.61, 1.14)	+/NS
Newman [38]	Newborns with hyperbilirubinemia	Phototherapy	Bilirubinemia	OR A 0.2 (0.1, 0.3)	Proportion of infants who had a particular bilirubinemia at hospital	OR 0.05 (0.015, 0.15)	+/+
Sheets [39]	Localized prostate cancer	Intensity-modulated radiation therapy vs proton therapy	Gastrointestinal events	RR M 0.66 (0.55, 0.79)	Radiation Therapy Oncology Group affiliation	RR 0.66 (0.49, 0.88)	+/+
		Intensity-modulated radiation therapy vs proton therapy	Urinary nonincontinence events	RR M 1.25 (0.99, 1.58)	Radiation Therapy Oncology Group affiliation	RR 1.1 (0.78, 1.58)	NS/NS
		Intensity-modulated radiation therapy vs proton therapy	Urinary incontinence events	RR M 0.96 (0.70, 1.32)	Radiation Therapy Oncology Group affiliation	RR 1.03 (0.63, 1.71)	NS/NS
		Intensity-modulated radiation therapy vs proton therapy	Erectile dysfunction events	RR M 0.89 (0.70, 1.12)	Radiation Therapy Oncology Group affiliation	RR 0.78 (0.54, 1.13)	NS/NS
		Intensity-modulated radiation therapy vs proton therapy	Additional cancer therapy	RR M 1.26 (0.86, 1.84)	Radiation Therapy Oncology Group affiliation	RR 1.6 (0.85, 3.00)	NS/NS
Suh [37]	Type 2 diabetes	Statin plus fibrate vs fibrate	Cardiovascular events	OR M 0.53 (0.36, 0.79)	Physician's preference	OR 1.83 (0.85, 3.95)	+/NS
				OR A 0.62 (0.46, 0.84)			
Valenstein [40]	Depression	Bupropion vs Citalopram	Mortality	RD‰ W -35.58 (-58.50, -12.65)	Prescription rate in the facility	RD‰ -78.31 (-360.53, 203.91)	+/NS

Authors	Patients studied	Type of intervention	Outcome	Propensity score analysis Target parameter PS application (A, M, S, W) Estimation (95% CI)	Instrumental variable analysis Type of IV	 Target parameter Estimation (95% CI)	Statistical significance PS/IV
		Sertraline vs Citalopram	Mortality	RD‱ W -19.89 (-37.87, -1.91)	Prescription rate in the facility	RD‱ -8.04 (-109.53, 93.45)	+/NS
		Fluoxetine vs Citalopram	Mortality	RD‱ W -18.98 (-40.56, 2.59)	Prescription rate in the facility	RD‱ -42.53 (-138.73, 53.67)	NS/NS
		Venlafaxine vs Citalopram	Mortality	RD‱ W -1.02 (-38.42, 36.38)	Prescription rate in the facility	RD‱ -45.36 (-556.58, 465.86)	NS/NS
		Paroxetine vs Citalopram	Mortality	RD‱ W 1.11 (-23.00, 25.21)	Prescription rate in the facility	RD‱ -60.95 (-193.77, 71.87)	NS/NS
		Mirtazapine vs Citalopram	Mortality	RD‱ W -9.99 (-48.66, 28.67)	Prescription rate in the facility	RD‱ -149.17 (-406.23, 107.89)	NS/NS
Wisnivesky [30]	Stage III lung cancer	Postoperative radiotherapy	Mortality	HR A 1.11 (0.97, 1.27) HR M 1.10 (0.95, 1.27) HR S 1.12 (0.98, 1.28)	Geographic variability in the radiotherapy use	RD 0.08 (-0.15, 0.24)	NS/NS
Bateman [22]	Coronary artery bypass graft surgery	Proton pump inhibitors vs H2 receptor antagonists	Nosocomial pneumonia	RD‰ M 6.1 (-0.6, 12.8)	Hospital's preference	RD‰ 8.2 (1.1, 15.4)	NS/+
Bekelman [41]	Muscle-invasive urothelial carcinoma	Radical cystectomy vs bladder-preserving therapy	Mortality	HR A 1.26 (1.05, 1.53) HR W 1.27 (1.06, 1.53)	Local area cystectomy rate	HR 1.06 (0.78, 1.31)	+/NS
Hebert [42]	Treated by angiotensin-converting enzyme inhibitors (ACEI)	Centrally acting ACEIs vs non-Centrally acting ACEIs	Alzheimer's disease and related dementias	HR M 1.02 (0.98, 1.07)	Physician's preference and geographical rate of prescription	HR 0.99 (0.96, 1.04)	NS/NS
Huesch [43]	Patients receiving percutaneous coronary intervention with stent replacement	Drug-eluting stents vs bare metal stents	Mortality	HR A 0.78 (0.73, 0.83)	Physician's preference	RD% -7.5 (-13.9, -1.2)	+/+
Jacob [44]	Patients eligible for colorectal cancer (CRC) screening	Colonoscopy screening	CRC incidence	RD W -0.54 (-0.69, -0.39) RD M -0.56 (-0.99, -0.12)	Physician's rate of discretionary colonoscopy	RD -0.6 (-0.78, -0.31)	+/+
		Colonoscopy screening	Mortality	RD W -0.08 (-0.15, -0.02) RD M -0.10 (-0.24, -0.04)	Physician's rate of discretionary colonoscopy	RD -0.17 (-0.21, -0.14)	+/+
Margolis [45]	Diabetic foot ulcer	Hyperbaric oxygen therapy	Healing	HR A 0.68 (0.63, 0.73) HR S 0.61 (0.58, 0.65) HR M 0.68 (0.63, 0.73)	Center's treatment rate	HR 0.43 (0.35, 0.52)	+/+
Parmar [34]	Pancreatic cancer	Endoscopic ultrasound evaluation	Mortality	HR M 0.77 (0.70, 0.84) HR S 0.70 to 1.01 across quintiles (only Q1 was not significantly less than 1) HR A 0.79 (0.74, 0.85)	Treatment rate in the area	HR 1.00 (0.73, 1.36)	+/NS

Authors	Patients studied	Type of intervention	Outcome	Propensity score analysis Target parameter PS application (A, M, S, W) Estimation (95% CI)	Instrumental variable analysis Type of IV	 Target parameter Estimation (95% CI)	Statistical significance PS/IV
Pöttgen [46]	Stage IIIA/B non-small cell lung cancer	Accelerated hyperfractionated radiotherapy vs conventional fractionation	Complete response	OR W 3.04 (1.93, 4.76)	Year of treatment and trial participation	OR 1.67 (1.22, 2.17)	+/+
				OR W 3.73 (2.09, 6.67)			
Steingrub [47]	Mechanically ventilated patients with severe sepsis	Neuromuscular blocking agent	Mortality	RD% M -4.4 (-7.6, -1.3)	Hospital treatment rate	RD% -4.3 (-11.5, 1.5)	+/NS
Thomas [48]	Smoking cessation product users	Bupoprion vs nicotine	Fatal and non-fatal self-harm	HR M 0.87 (0.31, 2.40)	Physicians' preference	RD‰ -3.9 (-7.0, -0.9)	NS/+
		Varenicline vs nicotine	Fatal and non-fatal self-harm	HR M 0.87 (0.51, 1.48)	Physicians' preference	RD‰ 0.4 (-0.8, 1.5)	NS/NS
		Bupoprion vs nicotine	Depression	HR M 0.66 (0.48, 0.90)	Physicians' preference	RD‰ -6.6 (-28.0, 14.8)	+/NS
		Varenicline vs nicotine	Depression	HR M 0.78 (0.67, 0.90)	Physicians' preference	RD‰ -5.9 (-12.6, 0.8)	+/NS
		Bupoprion vs nicotine	Mortality	HR M 0.34 (0.14, 0.82)	Physicians' preference	RD‰ -4.2 (-10.5, 2.1)	+/NS
		Varenicline vs nicotine	Mortality	HR M 0.37 (0.26, 0.53)	Physicians' preference	RD‰ -0.8 (-2.8, 1.1)	+/NS
VanDyke [51]	Cystic fibrosis with chronic *P. aeruginosa* infections	Inhaled tobramycin	Decline in forced expiratory volume in one second	RD% W -1.71 (-2.31, -1.12)	Center-level prescribing rates	RD% 2.55 (0.16, 4.94)	+/+ in opposite direction
Beadle [50]	Head and neck cancer	Intensity-modulated radiation therapy	Mortality	HR M 0.72 (0.59, 0.90)	Provider experience	HR 0.60 (0.41, 0.88)	+/+
Wright [49]	Stage II-IV ovarian cancer	Neoadjuvant chemotherapy vs primary surgery	Mortality	HR M 1.24 (1.15, 1.34)	Local area treatment pattern	HR 1.04 (0.67, 1.60)	+/NS
				HR W 1.30 (1.25, 1.35)			

Studies are sorted by year of publication.

Target parameter: HR, hazard ratio; OR, odds ratio; RR, relative risk; RD, risk difference; RRR, relative reduction in risks.

PS: propensity score. IV: instrumental variable.

Propensity score applications: A, adjustment; M, matching; S, stratification; W, weighting.

Statistical significance: NS, non-significant effect; +, significant effect.

A.5.2.3.2. Discussion

Nous avons identifié 37 études observationnelles dans lesquelles a été évalué l'effet d'une intervention clinique en utilisant à la fois une méthode basée sur un SP, et une autre basée sur une VI. La plupart des études ont été publiées au cours des 5 années précédant ce travail. Certaines études rapportaient plusieurs évaluations (plusieurs interventions ou critères de jugement) et ce sont au total 55 couples de résultats « résultat avec SP – résultat avec VI » qui ont été comparés. Globalement, l'agrément entre les deux méthodes a été jugé faible (coefficient kappa de Cohen 0,21 [0,00 ; 0,41]). Dans 23 cas (42%), les résultats n'étaient pas statistiquement significatifs avec une méthode mais significatifs avec l'autre, l'analyse basée sur une VI étant alors le plus souvent celle qui donnait des résultats non significatifs (87%). Un autre résultat intéressant était l'interprétation de la non-concordance faite par les auteurs des 18 études concernées. Dans 8 (44%) études, les auteurs ont considéré que l'analyse basée sur une VI fournissait l'estimation la plus valide ; dans 5 (28%) études, les résultats ont été considérés comme concordants, malgré l'absence de significativité statistique de l'estimation fournie par l'analyse basée sur une VI ; dans 4 (22%) études, les avantages et limites des deux méthodes ont été discutés, sans privilégier l'une ou l'autre ; enfin, dans la dernière étude (6%), aucune discussion de la non-concordance n'a été faite.

L'absence de concordance fréquemment constatée entre les conclusions tirées des deux méthodes a été discutée dans l'article à la lumière des forces et faiblesses de chaque méthode, telles qu'envisagées dans la première partie de ce manuscrit. Un algorithme d'interprétation de résultats non-concordants a été proposé.

Si l'utilisation des méthodes basées sur un SP ou une VI avait fait l'objet antérieurement d'analyses critiques révélant le caractère souvent inapproprié de leur mise en œuvre, ce travail original a mis en avant la nécessité de porter une plus grande attention aux caractéristiques respectives des méthodes d'analyse disponibles pour prendre en compte le biais d'indication

dans les études observationnelles. À la phase pré-analytique, les choix méthodologiques doivent permettre d'identifier la méthode d'analyse la plus adaptée, en fonction notamment de la question posée (type d'effet que l'on souhaite évaluer ?), de la taille de l'échantillon et de l'importance du biais d'indication attendu. Pour le lecteur, la comparaison de résultats publiés doit être critique, en questionnant les méthodes d'analyse utilisées, leurs forces et faiblesses, et ne pas se limiter à la simple confrontation numérique des estimations de l'effet.

B. DEUXIÈME PARTIE : APPLICATIONS

La première partie de la thèse a permis d'étudier et de comparer les différentes méthodes d'analyse qui permettent de minimiser le biais d'indication dans les études observationnelles. Le prolongement de ce travail a été de mettre en application ces méthodes pour l'évaluation de trois interventions de santé dans le cadre de deux études observationnelles de cohorte.

Chacune de ces évaluations a fait l'objet d'un article présenté dans cette deuxième partie :

- évaluation de l'effet d'un réseau de soins spécialisé dans l'insuffisance cardiaque sur la mortalité (article 2) ;
- évaluation de l'effet des stratégies médicamenteuses appropriées dans l'insuffisance cardiaque sur la mortalité (article 3) ;
- évaluation de l'effet des stratégies antithrombotiques chez les patients hémodialysés sur le risque hémorragique (article 4).

B.1. Évaluation de l'effet d'un réseau de soins spécialisé dans l'insuffisance cardiaque sur la mortalité (article 2)

B.1.1 Contexte

Cette première application, comme la suivante, est l'évaluation de l'efficacité d'une intervention complexe dans la prise en charge de l'insuffisance cardiaque, en exploitant les données issues de l'étude de cohorte EPICAL 2 (Epidémiologie et Pronostic de l'Insuffisance Cardiaque Aiguë en Lorraine). Cette cohorte comprend 2254 patients hospitalisés pour insuffisance cardiaque aiguë en Lorraine entre janvier 2011 et octobre 2012. Un recueil de données initiales (sociodémographiques, cliniques, biologiques, thérapeutiques) a été réalisé chez les patients inclus, qui ont ensuite été suivis à 6 mois, 1 an, 2 ans et 3 ans, avec un enregistrement des interventions thérapeutiques réalisées (interventions médicamenteuses et non médicamenteuses), des hospitalisations et du statut vital. L'étude EPICAL 2 offre un cadre de recherche pour évaluer l'efficacité de différents aspects, intra- et extra-hospitaliers, de la prise en charge de l'insuffisance cardiaque.

Le premier travail présenté dans ce paragraphe porte sur l'évaluation de l'efficacité d'une prise en charge par un réseau de soins spécialisé dans l'insuffisance cardiaque. La prise en charge thérapeutique des maladies chroniques telle que l'insuffisance cardiaque fait appel à un ensemble de moyens médicamenteux et non médicamenteux, dispensés par différents acteurs médicaux (cardiologue, médecin généraliste, pharmacien) et para-médicaux (infirmière, diététicien, kinésithérapeute…). Afin d'améliorer la continuité et la coordination des soins des patients, des réseaux de soins se sont développés au début des années 2000. L'inclusion des patients atteints d'insuffisance cardiaque dans ces réseaux fait aujourd'hui partie des recommandations de bonne pratique, leur efficacité ayant été montrée par de nombreux ECR (Ponikowski, Voors et al. 2016). Par exemple, sur la base des résultats de 11 ECR, une revue systématique Cochrane a conclu que ces réseaux de soins dans l'insuffisance

cardiaque étaient associés à une diminution de la mortalité toutes-causes à 1 an (OR 0,66 [IC95% 0,47 – 0,91])(Takeda, Taylor et al. 2012).

Cependant, de telles études expérimentales randomisées ont été menées dans des conditions idéales, sur des patients sélectionnés et pris en charge par des équipes spécialisées et entraînées, ce qui limite la généralisation de leurs résultats à la pratique courante. Aussi était-il souhaitable de disposer en complément de données d'évaluation issues d'études observationnelles. À notre connaissance, aucune étude observationnelle prospective comparative, mettant en œuvre des méthodes appropriées pour limiter le biais d'indication, et disposant d'un échantillon de patients représentatif de la pratique courante, n'avait été réalisée pour évaluer l'efficacité d'un réseau de soins spécialisé dans l'insuffisance cardiaque en termes de mortalité toutes-causes. L'objectif de ce travail était donc de mener une telle évaluation dans le cadre de l'étude EPICAL 2.

En Lorraine, le réseau de soins ICALOR (Insuffisance CArdiaque en LORraine) propose une prise en charge multidisciplinaire coordonnée pour les patients souffrant d'insuffisance cardiaque chronique. Cette prise en charge comprend un parcours de soins adapté aux patients, un suivi rapproché avec des visites infirmières au domicile, un partage des informations entre les différents acteurs de soins au moyen d'un dossier médical informatisé et un programme d'éducation thérapeutique (Agrinier, Altieri et al. 2013).

L'objectif de ce travail a été d'évaluer l'effet de la prise en charge des patients par le réseau de soins ICALOR sur la mortalité sur une période d'un an. Les différentes méthodes d'analyse permettant de minimiser le biais d'indication, identifiées et étudiées en première partir de thèse, ont été envisagées (méthodes conventionnelles, utilisation d'un score de propension, utilisation d'une variable instrumentale), en évitant l'écueil d'une utilisation

inappropriée des méthodes novatrices (score de propension et variable instrumentale), pointé du doigt dans la littérature.

L'article est reproduit dans les pages suivantes, sous sa forme publiée dans la revue *Medicine* (Laborde-Casterot, Agrinier et al. 2016).

Observational Study

Medicine®

OPEN

Effectiveness of a multidisciplinary heart failure disease management programme on 1-year mortality

Prospective cohort study

Hervé Laborde-Castérot, MD, MPH[a], Nelly Agrinier, MD, PhD[a,b], Faiez Zannad, MD, PhD[b,c], Alexandre Mebazaa, MD, PhD[d,e,f], Patrick Rossignol, MD, PhD[b,c], Nicolas Girerd, MD, PhD[b,c], François Alla, MD, PhD[a], Nathalie Thilly, PharmD, PhD[a,g,*]

Abstract

We performed a multicenter prospective observational cohort study (Epidémiologie et Pronostic de l'Insuffisance Cardiaque Aiguë en Lorraine, Epidemiology and Prognosis of Acute Heart Failure in Lorraine [EPICAL2]) to evaluate the effectiveness on mortality of a community-based multidisciplinary disease management programme (DMP) for heart failure (HF) patients.

Between October 2011 and October 2012, 1816 patients, who were hospitalized for acute HF or who developed acute HF during a hospitalization, were included from 21 hospitals in a northeast region of France. At hospital admission, their mean age was 77.3 (standard deviation [SD] 11.6) years and mean left ventricular ejection fraction was 45.0 (SD 16.0)%. A subset of patients were enrolled in a multidimensional DMP for HF (n = 312, 17.2%), based on structured patient education, home monitoring visits by HF-trained nurses, and automatic alerts triggered by significant clinical and biological changes to the patient. The DMP involved general practitioners, nurses, and cardiologists collaborating via an individual web-based medical electronic record. The outcome was all-cause mortality from the 3rd to the 12th month after discharge. During the follow-up, a total of 377 (20.8%) patients died: 321 (21.3%) in the control group and 56 (17.9%) in the DMP group. In a propensity score analysis, DMP was associated with lower 1-year all-cause mortality (hazard ratio 0.65, 95% CI 0.46–0.92). Instrumental variable analysis gave similar results (hazard ratio 0.56, 0.27–1.16).

In a real world setting, a multidimensional DMP for HF with structured patient education, home nurse monitoring, and appropriate physician alerts may improve survival when implemented after discharge from hospitalization due to worsening HF.

Abbreviations: AHF = acute heart failure, BMI = body mass index, BNP = B-type natriuretic peptide, BP = blood pressure, CI = confidence interval, eGFR = estimated glomerular filtration rate, EPICAL2 = Epidémiologie et Pronostic de l'Insuffisance Cardiaque Aiguë en Lorraine, Epidemiology and Prognosis of Acute Heart Failure in Lorraine, GP = general practitioner, HF = heart failure, HF-DMP = heart failure disease management programme, HR = hazard ratio, ICALOR = Insuffisance CArdiaque en LORraine, Heart Failure in Lorraine, IPTW = inverse probability of treatment weighting, IV = instrumental variable, LVEF = left ventricular ejection fraction, PS = propensity score, SD = standard deviation, SDiff = standardized difference.

Keywords: disease management programme, heart failure, instrumental variable, observational study, propensity score

1. Introduction

Heart failure (HF) is a major public health problem, affecting approximately 1% to 2% of the adult population in developed countries, with the prevalence rising to ≥10% among people aged 70 years or more.[1] Chronic HF is characterized by repeated hospitalizations and high mortality.[2] A decrease in HF mortality

Editor: Perbinder Grewal.

Funding/support: The EPICAL2 cohort study received a grant from the National Hospital Programme of Clinical Research (PHRC 2009) of the French Ministry of Health.

All researchers were independent from funders.

The authors have no conflict of interest to disclose.

Supplemental Digital Content is available for this article.

[a] University of Lorraine, APEMAC EA4360, [b] Inserm U1116, CIC-P 1433, University Hospital of Nancy, Nancy, [c] F-CRIN INI-CRCT network, [d] Inserm U942, [e] University Paris Diderot, Sorbonne Paris Cité, [f] Department of Anesthesia and Critical Care, Hôpitaux Universitaires Saint-Louis Lariboisière, APHP, Paris, [g] Inserm CIC-EC 6, Clinical Epidemiology and Evaluation, University Hospital of Nancy, Nancy, France.

[*] Correspondence: Nathalie Thilly, Inserm CIC-EC 6, Clinical Epidemiology and Evaluation, University Hospital of Nancy, Allée du Morvan, 54500 Vandoeuvre-lès-Nancy, France (e-mail: n.thilly@chru-nancy.fr).

Medicine (2016) 95:37(e4399)

Received: 19 April 2016 / Received in final form: 1 July 2016 / Accepted: 1 July 2016

http://dx.doi.org/10.1097/MD.0000000000004399

has been observed in western countries over recent decades[2,3] and probably relates to improvements in the management of HF, with the development of evidence-based therapies and clinical guidelines.[4]

Heart failure disease management programmes (HF-DMPs) are designed to improve outcomes through structured follow-up based on patient education, optimization of medical treatment, psychosocial support, and improved access to care.[2] They are strongly recommended in HF guidelines to reduce the risk of HF hospitalization, based on the highest level of evidence.[2] A Cochrane review of 11 randomized controlled trials concluded that case management interventions for HF were associated with a significant reduction in all-cause mortality at 12 months follow-up (odds ratio 0.66, 95% confidence interval [CI] 0.47–0.91).[5] However, randomized controlled trials are generally conducted under ideal conditions, among selected patients being cared for by hyper-specialized physicians, none of which reflect real-world conditions. Accordingly, the generalizability of results from randomized controlled trials is open to question, particularly when the trials involve complex interventions such as HF-DMP, which are greatly context dependent.[6] In addition, the magnitude of an intervention's effect under real-world conditions may be lower than in clinical trials. Thus, as a complement to trials, well-designed observational studies are useful to ascertain and quantify the effectiveness of HF-DMP in real-world settings.[7]

In this context, we used data from the Epidémiologie et Pronostic de l'Insuffisance Cardiaque Aiguë en Lorraine, Epidemiology and Prognosis of Acute Heart Failure in Lorraine (EPICAL2) cohort study to assess the effectiveness on all-cause 1-year mortality after hospitalization for acute heart failure (AHF), of a multidisciplinary community-based HF-DMP, implemented over several years in a large area of France. Our research hypothesis is that HF-DMP is an effective way to reduce mortality in a real-world setting, as demonstrated in randomized controlled trials.

2. Methods

2.1. Setting, design, and population

The EPICAL2 study was a prospective, observational community-based cohort study involving 21 volunteer hospitals spread over the Lorraine region of northeast France (population of 2,350,000, according to the 2012 census). The cohort enrolled comprised 2254 consecutive adult HF patients hospitalized between October 2011 and October 2012 in cardiology intensive care units, cardiology departments, or emergency departments at the hospitals concerned. Patients living in Lorraine and hospitalized for AHF were included, as were those who developed AHF during hospitalization. Eligible patients were identified either by physicians from the participating departments or by trained clinical research assistants who regularly visited the departments. Included patients were then followed for 3 years after discharge from the index hospitalization or until death if it happened first. The objectives of this cohort study were: to describe morbidity and mortality in the short-term (0–6 months) and mid-term (up to 3 years) and to identify the main prognostic factors; to assess the effectiveness of various aspects of care, in or out of hospital.

In the present investigation, patients who died during the index hospitalization were excluded, leaving 2070 who were alive at hospital discharge (Fig. 1). Independently of EPICAL2, some of them may have been enrolled in routine care in an HF-DMP named Insuffisance CArdiaque en LORraine, Heart Failure in Lorraine (ICALOR), which was the only specialized DMP for HF patients implemented in Lorraine in 2011 to 2012. This HF-DMP was accessible to all HF patients living in the region, whatever the severity of their HF and ejection fraction. No specific inclusion (other than the HF diagnosis and area of residence) and exclusion criteria were established to be enrolled in ICALOR. The proposal to enroll a patient was left to physician discretion during hospitalization or outpatient care, and the patient was free to refuse it or to formally accept by signing a written consent. As an observational study, EPICAL2 did not affect the participation of patients in HF-DMP. The list of all HF-DMP enrolled patients and dates of consent were obtained from the ICALOR administrative database. A patient was considered as exposed to HF-DMP if he (she) signed the consent for HF-DMP enrollment before or during the index hospitalization, or during the 1st month after discharge, and had benefited from at least 1 part of the programme. For patients enrolled in an HF-DMP after discharge from index hospitalization, the time period between the cohort enrollment and the HF-DMP consent is "immortal," as patients must survive this period in order to be exposed to HF-DMP.[8] As recommended by Suissa,[8] an adapted strategy was then implemented to control the immortal time bias: patients who died during the 1st month after hospital discharge and those who signed a consent to HF-DMP after the 1st month were excluded from this analysis. A total of 1816 patients were therefore considered in the present investigation (Fig. 1).

The EPICAL2 cohort study was conducted according to the principles of the Declaration of Helsinki and approved by national ethics committees (Comité Consultatif sur le Traitement de l'Information en Matière de Recherche, Comission Nationale de l'Informatique et des Libertés). All eligible patients were informed about the study protocol and were free to refuse to be included in the cohort.

2.2. The heart-failure disease management programme

ICALOR is a community-based HF-DMP 1st implemented in 2006 and independent of the EPICAL2 study described in detail elsewhere.[9] Its main objectives were to reduce mortality and cardiology hospitalizations in HF patients. Briefly, in 2012, the programme involved 101 cardiologists, 926 general practitioners (GPs), and 1603 volunteer nurses collaborating in the home care of HF patients living in Lorraine and receiving the HF-DMP. After providing written consent, each patient's GP set up an individual web-based medical electronic record, accessible to all health professionals involved in ICALOR and the patient's care. Patients enrolled in ICALOR benefited first from a structured educational programme delivered by a dedicated team of trained nurses, and then received regular home visits from HF trained nurses who continued patient and family education and delivered information and counseling aimed at improving adherence to diet, medications, and patient self-care. The nurse intervention also focused on detection of worsening HF and managing comorbidities. At each home visit, nurses monitored blood pressure (BP), heart rate, and weight, and updated the patient's electronic records with routine laboratory results. The frequency of home-nurse visits were adapted to the clinical severity and evolution of the patient's condition. The ICALOR coordinating center received automatic alerts from the web-based electronic record system according to preset thresholds of variations in the following indicators: dyspnea, weight gain or loss, heart rate,

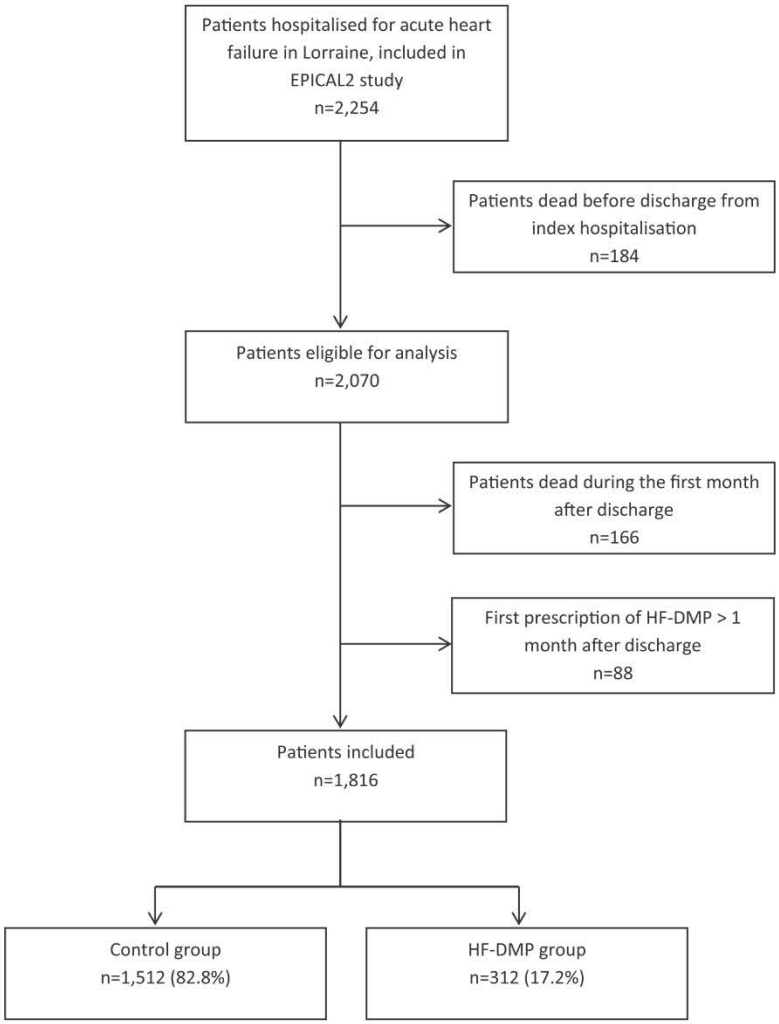


Figure 1. Flow chart for the selection of patients included in the evaluation of the heart failure disease management programme (HF-DMP).

edema, BP as monitored by the home-nurse, and B-type natriuretic peptide (BNP) or NT-proBNP, hemoglobin, potassium, sodium, and creatinine serum levels from laboratory assessments prescribed by the GP as a part of routine care. The coordinating center checked the validity of the alert and called the GP and/or the cardiologist to determine the appropriate action. The ICALOR HF-DMP included 304 (standard deviation [SD] 138) patients/year and triggered >500 alerts/year, more than half of them for weight gain.

2.3. Data collection and main outcome

A standardized form was used to interrogate medical records for socio-demographic and clinical data at inclusion, as well as biological and therapeutic data at inclusion, during the index hospitalization and discharge. In the present investigation, the main variables previously identified as impacting mortality in HF patients were considered as potential confounding factors: socio-demographic – age (<65, 65–79, ≥80 years), sex, living alone, type of hospital (local, regional, and teaching); clinical – HF etiology (ischemic or not), history of diabetes, chronic kidney disease, stroke or transient ischemic attack, chronic obstructive pulmonary disease or asthma, cancer, previous hospitalization for AHF, body mass index (BMI: underweight or normal weight < 25, overweight or obese ≥25 kg/m^2), left ventricular ejection fraction (LVEF: <30, 30–44, ≥45%), low systolic BP (<115 mm Hg), and prolonged QRS duration on electrocardiogram (>120 ms); biological at admission – low glomerular filtration rate estimated by the MDRD equation[10] (estimated glomerular filtration rate [eGFR]) (<60 mL/min/1.73 m^2), hyponatremia (<135 mmol/L), hypokalemia (<3.8 mmol/L), anemia (hemoglobin < 10 g/dL), and BNP or NT-proBNP (BNP > 400 pg/mL or NT-proBNP > 450 pg/mL in patients <50 years old, NT-proBNP > 900 pg/mL in patients between 50 and 75 years old, NT-proBNP > 1800 pg/mL in patients >75 years old – according to the state of the art at the beginning of the cohort[11]).

All data were collected and checked for completeness, according to French Good Practices in Epidemiology,[12] by 5

Table 1

Baseline characteristics of included patients, overall, and according to heart failure disease management programme (HF-DMP) enrollment.

	Overall (n = 1816)		Control group (n = 1504)		HF-DMP group (n = 312)		SDiff, %
Socio-demographic characteristics							
Male sex	892	(49.1%)	706	(46.9%)	186	(59.6%)	**25.61**
Age < 65 years	289	(15.9%)	230	(15.3%)	59	(18.9%)	**22.89**
65–79 years	582	(32.1%)	462	(30.7%)	120	(38.5%)	
≥80 years	945	(52.0%)	812	(54.0%)	133	(42.6%)	
Living alone	612	(33.7%)	521	(34.6%)	91	(29.1%)	**11.91**
Type of hospital: local	782	(43.1%)	590	(39.2%)	192	(61.5%)	**59.71**
Regional	461	(25.4%)	435	(28.9%)	26	(8.3%)	
Teaching	573	(31.5%)	479	(31.9%)	94	(30.1%)	
Medical history							
Ischemic HF etiology	431	(23.7%)	330	(21.9%)	101	(32.4%)	**23.61**
Diabetes	654	(36.0%)	510	(33.9%)	144	(46.2%)	**25.19**
Chronic kidney disease	415	(22.9%)	321	(21.3%)	94	(30.1%)	**20.20**
Stroke/TIA	231	(12.7%)	191	(12.7%)	40	(12.8%)	0.36
COPD/asthma	410	(22.6%)	326	(21.7%)	84	(26.9%)	**12.26**
Cancer	260	(14.3%)	225	(15.0%)	35	(11.2%)	**11.11**
Previous hospitalization for acute HF	641	(35.3%)	474	(31.5%)	167	(53.5%)	**45.67**
Clinical and biological status at admission							
Overweight or obese	1233	(67.9%)	993	(66.1%)	240	(77.0%)	**30.82**
LVEF < 30%	322	(17.7%)	238	(15.8%)	84	(26.9%)	**31.60**
30–44%	573	(31.6%)	469	(31.2%)	104	(33.2%)	
≥45%	922	(50.8%)	798	(53.1%)	124	(39.9%)	
Low systolic BP[1]	335	(18.4%)	267	(17.8%)	68	(21.7%)	9.86
Low eGFR[2]	1172	(65.3%)	953	(64.0%)	219	(71.6%)	**16.33**
Prolonged QRS duration[3]	246	(13.5%)	189	(12.3%)	57	(18.2%)	**15.67**
Hyponatremia[4]	322	(18.0%)	267	(18.0%)	55	(17.9%)	0.26
Hypokalemia[5]	439	(24.6%)	361	(24.4%)	78	(25.4%)	2.31
Anemia[6]	188	(10.5%)	151	(10.2%)	37	(12.0%)	5.60
Increased BNP/NT-proBNP[7]	1318	(79.9%)	1076	(79.9%)	246	(79.9%)	0.18

Figures are numbers (percentage %) unless stated otherwise. 1 = systolic BP < 115 mm Hg; 2 = eGFR < 60 mL/min/1.73 m^2; 3 = QRS > 120 ms; 4 = natremia < 135 mmol/L; 5 = kalemia < 3.8 mmol/L; 6 = hemoglobin < 10 g/dL; 7 = BNP > 400 pg/mL or (NT-proBNP > 450 pg/mL in patients < 50 years old, NT-proBNP > 900 pg/mL in patients between 50 and 75 years old, NT-proBNP > 1800 pg/mL in patients > 75 years old). BNP = B-type natriuretic peptide, BP = blood pressure, COPD = chronic obstructive pulmonary disease, eGFR = glomerular filtration rate, HF = heart failure, LVEF = left ventricular ejection fraction, NT-proBNP = N-terminal pro-B-type natriuretic peptide, SDiff = absolute standardized difference, TIA = transient ischemic attack.

trained clinical assistants. Patient enrollment and quality of data collection were regularly controlled by a steering committee of 4 epidemiologists, 1 cardiologist, 1 nephrologist, and 1 cardiology intensive care physician. Ten percent of completed standardized forms were audited by an independent clinical research assistant who compared, for each form, data collected and the patient medical record.

2.4. Outcome of interest

The outcome of interest was all-cause mortality; the 1-year vital status of each patient and the date of death, if appropriate, were collected through civil registries. Survival time was calculated from the beginning of the 3rd month after discharge from the index hospitalization. This time zero for survival analysis was chosen because the 1st interventions of the HF-DMP might have been delayed by up to 1 month after patients gave their written consent. Surviving patients were censored at the end of the 1-year follow-up.

2.5. Statistical analysis

We first compared baseline characteristics of patients who received the HF-DMP with controls by calculating standardized differences (SDiffs), which indicate the degree of systematic differences in covariates between groups. Empirically, an absolute SDiff of <10% indicates a negligible difference in mean or percentage of the covariates between groups.[13] Cox proportional hazards models were then used to assess the effect on 1-year mortality of HF-DMP. To minimize potential bias and confounding effects, a propensity score (PS) was estimated by using a multivariate logistic regression model including all patients' baseline characteristics as shown in Table 1. The PS represents the likelihood of having received the HF-DMP depending on baseline characteristics of patients. The inverse probability of treatment weighting (IPTW) method was then applied by using the PS to assign individual weights to all observations, which allows some of characteristics of randomized controlled trials to be mimicked in observational study.[14] Weights were stabilized by the marginal prevalence of the HF-DMP exposure.[15] To check the accuracy of the PS model, we assessed covariate balance between groups after weighting by calculating SDiffs.[16] Four Cox models were applied: model 1, PS analysis using stabilized IPTW; model 2, model 1 + adjustment for covariates for which SDiffs were >10% after IPTW; model 3, model 1 with trimming of 2.5% of patients (un)treated contrary to PS prediction on both sides;[15] and model 4, model 2 with trimming of 2.5% of patients (un)treated contrary to PS prediction on both sides. Kaplan–Meier survival curves stratified by HF-DMP and control groups were constructed.

For 6 covariates with missing values (living alone, BMI, history of hospitalization for HF, systolic BP, LVEF, and prolonged QRS duration), we obtained values by multiple imputations as recommended for the Cox model analysis.[17] This was achieved through regression switching imputation using linear or logistic regression models respectively for quantitative or qualitative incomplete covariate fitted. This procedure was repeated 5 times to obtain 5 draws for each missing value in 5 distinct datasets. Baseline characteristics of patients were described and compared in each dataset (see Supplemental content, Tables S1–S6, http://links.lww.com/MD/B268). However, to be concise and summarize the results, Rubin approach was adopted, whereby the average percentage of each imputed covariate from the dataset of 5 is reported, the corresponding SDiff is calculated from these averages.[18] To check the influence of multiple imputations on parameter estimates, a sensitivity analysis was conducted with the original nonimputed dataset.

The fit of the proportional hazards model was checked visually by plotting log (−log [survival]) versus log (time). Results are reported as pooled hazard ratios (HRs) with a 95% CI. All analyses were performed with SAS version 9.4 software (SAS Institute, Inc., Cary, NC).

2.6. Confirmatory analysis

To further assess the validity of our results, because of possible residual bias due to unmeasured confounders with the PS approach, we performed an additional instrumental variable (IV) analysis. This method requires a valid instrument which is strongly associated with the intervention (HF-DMP) (1st condition), does not directly affect mortality (2nd condition), and is not associated with confounding factors (3rd condition).[19] A hospital's prescribing preference, defined as the prevalence of patients included in the HF-DMP in each hospital, appeared to be a good candidate. To test the association between the instrument and measured covariates, we classified hospitals into 2 groups according to their HF-DMP prescription rates ("DMP preference" group for hospitals above the median and "no-DMP preference" group for hospitals below the median) and compared characteristics of patients in both groups with absolute SDiffs. For the analysis, in the presence of a binary outcome and a binary exposure, we used 2-stage residual inclusion.[20] The 1st-stage model used the IV and all observed covariates to predict the probability of receiving the HF-DMP. In the 2nd-stage, residuals from the 1st-stage were included as an additional variable along with the exposure and all observed covariates to model mortality in a Cox regression.

3. Results

3.1. Crude baseline characteristics

Overall, a total of 1816 patients were included in the present investigation (Table 1). The mean age was 77.3 (SD 11.6) years and 49.1% were men. The etiology of HF was ischemic for 23.7% of patients and the mean LVEF at hospital admission was 45.0 (SD 16.0)%. A total of 312 patients (17.2%) were enrolled in HF-DMP: 112 (35.9%) prior to index hospitalization, 90 (28.9%) during the index hospitalization, and 110 (35.3%) during the 1st month after discharge. In the HF-DMP group, patients received a median of 9 nurse home visits per year [Q1-Q3: 5–14] and triggered a median of 2.5 automatic alerts per year [Q1–Q3: 1–4]. Compared to controls, patients receiving HF-DMP were younger (mean age 74.7 [SD 11.2] years vs 77.8 [SD 11.6]), more likely to be men and hospitalized in a local hospital, and less likely to live alone. They also had more comorbidities (diabetes, chronic kidney disease, chronic obstructive pulmonary disease, overweight, or obesity), and were more likely to have ischemic HF. In addition, they had markers of high-risk HF as demonstrated by a higher frequency of a history of hospitalization for AHF and QRS enlargement along with lower LVEF, systolic BP, and eGFR.

The mean percentage of missing data among all 20 covariates considered was 4.5%; the highest proportions were for QRS duration (22.9%), BMI (19.7%), and LVEF (18.9%), whereas previous hospitalization for AHF, systolic BP, eGFR, natremia, kalemia, and anemia were missing less than 2% of data. Original nonimputed baseline characteristics of patients are available in Supplemental content (Table S1, http://links.lww.com/MD/B268).

3.2. Baseline characteristics using IPTW and IV methods

Table 2 presents characteristics of patients in the 2 HF-DMP groups after using IPTW and IV methods. For both methods, covariates were well balanced between groups except for the type of hospital (SDiff = 15.05% for IPTW, SDiff = 137.64% for IV). In addition, SDiffs after IPTW were borderline for chronic kidney disease in average results (SDiff = 9.77%) and was slightly above 10% in 2 datasets (Supplemental content, Tables S5–S6, http://links.lww.com/MD/B268). In the IV analysis, the proportion of patients enrolled in the HF-DMP was 32.9% (n = 222) in the "DMP preference" group and 9.2% (n = 105) in the "no-DMP preference" group (odds ratio 4.83 [3.71–6.31]).

3.3. Survival analysis

Between the 3rd and the 12th months after the index hospitalization discharge, a total of 377 (20.8%) died, 321 (21.3%) controls, and 56 (17.9%) in the HF-DMP group (unadjusted Cox model HR 0.82, 95% CI 0.61–1.09). No patient was lost to follow-up. Kaplan–Meier survival curves of being dead according to the group (HF-DMP or control) show that mortality was higher in the control than the HF-DMP group, although the difference did not reach the statistical significance (Fig. 2).

Table 3 shows the estimations of the HF-DMP effect using the 4 Cox models. In the IPTW analysis performed in all included patients, HF-DMP was associated with significantly lower all-cause mortality (model 1 HR 0.65, 95% CI 0.46–0.92). Additional adjustment on the 2 covariates insufficiently balanced after IPTW (type of hospital and chronic kidney disease), provided similar intervention effect estimates (model 2 HR 0.63, 0.45–0.89). Slightly lower HRs were found when trimming the sample analyzed (model 3 HR 0.58, 0.40–0.86, and model 4 HR 0.58, 0.39–0.85). Multiple imputation provided similar point estimate as compared with analyses in nonimputed data and reduced estimation uncertainty (model 1 in nonimputed dataset HR 0.66, 0.38–1.12). IV analysis provided associations of similar strength to those observed in models 3 and 4 but did not reach statistical significance (HR 0.56, 0.27–1.16). The results of the survival analysis in the 5 datasets are available as Supplemental content (Table S7, http://links.lww.com/MD/B268).

Table 2

Characteristics of patients according to HF-DMP enrollment, after inverse probability of treatment weighting and in preference-based instrumental variable method.

	Weighted population			Instrumental variable analysis		
	Control group (n = 1323)	HF-DMP group (n = 294)	SDiff, %	No-DMP preference (n = 1141)	DMP preference (n = 675)	SDiff, %
Socio-demographic characteristics						
Male sex	48.4%	49.1%	1.48	48.4%	50.4%	4.00
Age < 65 years	14.2%	13.9%	3.36	15.2%	17.2%	7.93
65–79 years	31.6%	30.2%		31.4%	33.2%	
≥80 years	54.1%	55.8%		53.5%	49.6%	
Living alone	34.0%	36.1%	4.32	33.8%	33.9%	0.21
Type of hospital: local	45.8%	47.6%	**15.05**	43.7%	42.1%	**137.64**
Regional	22.6%	16.8%		40.4%	0.0%	
Teaching	31.5%	35.5%		16.0%	57.9%	
Medical history						
Ischemic HF etiology	24.5%	28.2%	8.59	24.4%	22.7%	4.01
Diabetes	36.6%	38.1%	2.94	34.7%	38.2%	7.28
Chronic kidney disease	23.5%	27.8%	9.77	22.7%	23.1%	0.95
Stroke/TIA	12.6%	9.7%	9.21	13.1%	12.2%	2.71
COPD/asthma	22.7%	21.4%	3.23	23.1%	21.6%	3.60
Cancer	14.0%	12.9%	3.22	14.6%	13.9%	2.00
Previous hospitalization for acute HF	36.7%	37.2%	1.08	35.0%	35.7%	1.38
Clinical and biological status at admission						
Overweight or obese	67.6%	66%	6.20	67.0%	69.6%	5.61
LVEF < 30%	17.9%	17.9%	0.46	17.4%	18.3%	2.21
30–44%	30.9%	31.1%		31.6%	31.5%	
≥45%	51.3%	51.0%		51.1%	50.3%	
Low systolic BP[1]	17.6%	16.7%	2.28	17.4%	20.1%	7.02
Low eGFR[2]	66.5%	67.9%	2.94	65.9%	64.2%	3.57
Prolonged QRS duration[3]	13.7%	14.7%	2.75	12.5%	15.4%	8.61
Hyponatremia[4]	18.2%	17.5%	1.62	17.7%	18.4%	1.82
Hypokemia[5]	24.0%	26.7%	6.30	25.8%	22.5%	7.72
Anemia[6]	10.8%	11.1%	0.96	10.4%	10.8%	1.30
Increased BNP/NT-proBNP[7]	80.0%	78.8%	2.97	79.9%	79.9%	0.00

1 = systolic BP < 115 mm Hg; 2 = eGFR < 60 mL/min/1.73 m^2; 3 = QRS > 120 ms; 4 = natremia < 135 mmol/L; 5 = kalemia < 3.8 mmol/L; 6 = hemoglobin < 10 g/dL; 7 = BNP > 400 pg/mL or (NT-proBNP > 450 pg/mL in patients < 50 years old, NT-proBNP > 900 pg/mL in patients between 50 and 75 years old, NT-proBNP > 1800 pg/mL in patients > 75 years old). BNP = B-type natriuretic peptide, BP = blood pressure, COPD = chronic obstructive pulmonary disease, eGFR = glomerular filtration rate, HF = heart failure, HF-DMP = heart failure disease management programme, LVEF = left ventricular ejection fraction, NT-proBNP = N-terminal pro-B-type natriuretic peptide, SDiff = absolute standardized difference, TIA = transient ischemic attack.

4. Discussion

4.1. Principal results and possible explanations

Using a prospective cohort of 1816 patients with AHF hospitalized in Lorraine, we found evidence of real-world effectiveness of a multidisciplinary community-based HF-DMP in terms of all-cause 1-year mortality after hospitalization for HF. Several analytical methods (PS weighting, IV analysis) and models (with or without adjustments and/or trimming) were used to account for potentially confounding characteristics of patients that might affect their enrollment in HF-DMP. They all showed that patients enrolled in HF-DMP had a 35% to 44% reduced risk of death during the year after hospitalization, as compared to patients not enrolled. Our findings suggest that the efficacy of HF-DMP demonstrated in clinical trials may be translated into real-world clinical practice and further support the current guidelines recommending DMP a way of organizing care to improve outcome for HF patients.[2] A DMP may include various health interventions with diverse design characteristics,[21] and evaluations of such programmes should be interpreted in the light of type and characteristics of interventions considered. First, ICALOR interventions include patient education and training for caregivers aimed at enhancing patients' adherence to

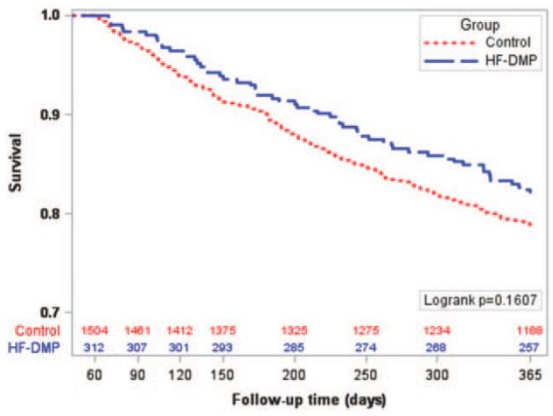


Figure 2. Kaplan–Meier survival curves according to the group (heart failure disease management programme [HF-DMP] and control). The numbers of patients still at risk of death are indicated.

Table 3

Estimations of the heart failure disease management programme effect.

Model	HR	95% CI
Model 1: propensity score analysis using IPTW	0.65	0.46–0.92
Model 2: model 1 + adjustment for type of hospital and chronic kidney disease	0.63	0.45–0.89
Model 3: model 1 + trimming of 2.5% of patients on both sides	0.58	0.40–0.86
Model 4: model 2 + trimming of 2.5% of patients on both sides	0.58	0.39–0.85
Instrumental variable analysis	0.56	0.27–1.16

95% CI = 95% confidence interval, HR = hazard ratio, IPTW = inverse probability of treatment weighting.

medications, lifestyle, and dietary habits as well as medication prescription (i.e., according to guidelines), the final outcome being a reduced risk of decompensation of HF and then of HF hospitalization and death. Second, ICALOR interventions also include close home-monitoring for worsening signs and decompensation symptoms, which allows for prompt therapeutic adjustment at the beginning of HF aggravation. Here again, the objective is to avoid rehospitalization and, ultimately, death. In a previous ecological investigation, we found ICALOR to be associated with a reduction in HF hospitalizations in Lorraine, as compared to other regions of France.[9] It could be hypothesized that the HF-DMP decreases the risk of HF hospitalizations, which leads to a lower likelihood of 1-year mortality. Last, ICALOR HF-DMP was implemented in Lorraine in 2006, that is, several years before the present investigation.[9] It could therefore be assumed that years of experience of ICALOR resulted in a trained multidisciplinary network and contributed to the HF-DMPs effectiveness.

4.2. Comparisons with other studies

It has been shown that patients included in most clinical trials involving HF are not representative of the whole spectrum of HF patients actually managed in clinical practice,[22] although clinical guidelines are based on results from such trials. This observation was confirmed in the present investigation, as patients enrolled in the ICALOR HF-DMP appeared to have more severe conditions than their peers included in clinical trials. For instance, the prevalence of eGFR below 60 mL/min/1.73 m^2 observed in the present investigation was double that reported in a large (>1 million patients) meta-analysis assessing the impact of impaired kidney function in HF patients.[23] Moreover, all 11 randomized controlled trials considered in the Cochrane meta-analysis reviewing the efficacy of case management interventions for HF on mortality excluded patients with comorbidities such as chronic kidney disease or those with a limited life expectancy, raising concern about the generalizability of their results.[5] In addition, trials are generally conducted in tightly controlled conditions by research teams and involve strict protocols that maximize both intervention implementation and patient adherence.[24] Consequently, the valuable contribution of the EPICAL2 cohort to existing evidence is related to its high external validity. To date, many observational studies have been conducted to assess the effectiveness of HF-DMP in relation to various outcomes, but those evaluating the HF-DMP effect on mortality in a prospective cohort design are scarce. To our knowledge only 2 observational prospective studies have published results concerning the efficacy of an HF-DMP including home nurse visits on all-cause mortality. The 1st, by Lowery et al,[25] included 969 patients (458 [47%] in the HF-DMP group) and showed a significant 1-year all-cause mortality reduction in the HF-DMP group (HR 0.43, $P < 0.001$). However, this study was conducted in a very particular setting and applied specific exclusion criteria limiting the generalizability of the results to the whole spectrum of patients managed in clinical practice (patients were included in medical centers for veterans and had no comorbidity associated with a predicted life expectancy ≤6 months). The 2nd published study, by Bonarek-Hessamfar et al,[26] included 362 patients (129 [36%] in HF-DMP group) with severe HF, with no indication for surgical or interventional treatment and no major disease reducing the short-term vital prognosis. After a 2-year follow-up period, the all-cause mortality was significantly lower in the HF-DMP group (HR 0.37, 0.16–0.89). However, analyses were adjusted for sex, age, and NYHA stage; the lack of control for many potential confounding factors foreshadowed a probable residual bias.

4.3. Strengths and limitations of our study

The main strengths of our study are the large cohort design, including an unselected sample of HF patients as encountered in current practice, the quality of individual data collection and the completeness of the 1-year follow-up, but also the use of analytical methods considered to be effective to control for confounders, in such an extensive manner that they are referred to as pseudo-randomized methods.[27] The proposal of HF-DMP was left to the discretion of physicians and depended on their perception of the benefits to each patient of being enrolled, according to his (her) baseline characteristics. Thus, patients enrolled in ICALOR HF-DMP differed significantly in several ways from those not enrolled. These differences constituted a confounding by indication that usually threatens the internal validity of results from all observational studies.[28] In our investigation, the risk was of underestimating the effect of HF-DMP, as patients enrolled in the HF-DMP had more severe HF than those who did not. To address the indication bias due the lack of random allocation of the HF-DMP, we used several analytical methods and models and obtained similar results, which tend to strengthen the credibility of our findings. At first we conducted a PS analysis using IPTW, which is considered by Austin[29] to be the most effective PS method with which to reduce bias in observational studies dealing with time-to-event outcomes. After weighting on PS, patient baseline characteristics were well balanced between the groups (SDiff < 10%), except for the type of hospital and, marginally, for chronic kidney disease, but additional adjustment on these covariates did not impact on the result. IPTW analysis estimates the average effect of HF-DMP in the entire population and showed a significant 1-year mortality reduction in the HF-DMP group (model 1 HR 0.65, 0.46–0.92, and model 2 HR 0.63, 0.45–0.89). Slightly lower HRs were observed when excluding patients treated contrary to the prediction (model 3 HR 0.58, 0.40–0.86, and

model 4 HR 0.58, 0.39–0.85), as recommended by Stürmer et al.[15] PS analysis is recognized to be able to reduce bias due to all measured confounders, but fails to limit bias due to unmeasured or unknown confounders.[15] To address this limit, IV analysis is increasingly used in clinical research, as it is able to control for measured confounders as well as unmeasured or unknown ones.[19,27] In our investigation, hospital HF-DMP prescription preference appeared to be a good instrument for the analysis. By definition, hospital HF-DMP prescription rates are associated with HF-DMP (1st condition for IV use). Concerning the 2nd condition for IV use, hospital HF-DMP prescription preference can be assumed not to be associated with patient outcome (differences in the use of HF-DMP across hospitals are much more likely to be associated with nonmedical factors, such as local habits, rather than factors related to the HF itself). The 3rd condition for IV use, stipulating that the instrument must not be associated with confounding factors, was checked as patient characteristics were well balanced between "DMP preference" and "no-DMP preference" groups, except for 1 covariate, the type of hospital. This imbalance was explained by the presence of the only 2 regional hospitals in the "no-DMP preference" group. If regional and teaching hospitals were considered as one (i.e., large versus small care centers), the imbalance disappeared (data not shown). Finally, the IV analysis tends to confirm the 1-year mortality reduction associated with HF-DMP found in the PS analysis, even if the result did not reach statistical significance (HR 0.56, 0.27–1.16). As expected, IV analysis led to larger CIs than PS analysis because of a relatively limited sample size for the use of such a method but point estimate should be considered.[30,31]

Our results should be interpreted in the light of some limitations. First, eligible patients hospitalized in Lorraine during the study period were probably not all identified, and included in the EPICAL2 cohort. Identification of eligible patients based on declaration by physicians of the participating departments and/or by clinical research assistants was certainly not comprehensive. However, this type of recruitment also makes it unlikely that patients included in EPICAL2 were not representative of all eligible patients. In a previous study, the ICALOR HF-DMP coverage ratio, that is, the number of patients included in ICALOR over the number of HF patients requiring hospitalization in Lorraine was estimated at 18%, similar to the ratio that we found in the present investigation.[9] The 2nd limitation was related to the rate of missing data in medical records for some covariates (i.e., QRS duration, BMI, and LVEF) which led us to use multiple imputation. However, a sensitivity analysis conducted with the original nonimputed dataset gave an HR similar to the analysis with imputed data, attesting to the lack of bias due to multiple imputation. The only difference is the nonsignificant result obtained with nonimputed data because of a loss of statistical power caused by the exclusion of patients with missing data. The 3rd limitation is related to some interesting questions that could not have been explored in the present investigation.

(1) The limited number of events (377 deaths) in our study sample did not allow us to perform subgroups analyses such as analyses according to the LVEF level. Indeed, the question of the homogeneity of the HF-DMP effectiveness in reduced and preserved LVEF patients is particularly relevant as a large part of the HF-DMP activity is focused on optimizing and enhancing patient adherence to medications.

(2) It would have been interesting to evaluate the effectiveness of the HF-DMP in terms of cardiac-specific mortality. However, cause of death was not collected and the choice of all-cause mortality as outcome is justified by the main objective of ICALOR to reduce overall mortality, whatever the cause.

The last limitation is the lack of some data that should have been considered in the PS because of their potential impact on prognosis, such as:

(1) Biological results (especially BNP) and NYHA class collected at admission but unfortunately not discharge because these measurements are not systematically performed at discharge within routine clinical care in the hospitals involved in the EPICAL2 study and were not then available in medical records.
(2) Remote monitoring, not collected in EPICAL2, is possible with some implantable cardiac devices and is likely to improve quality of care and hence survival. However, the small number of patients with such devices in our sample (n = 91, 5.0%) makes a noteworthy bias unlikely.

4.4. Conclusions and policy implications

This investigation shows that HF-DMP is likely to be effective in a real-world setting, with a reduction of 40% in 1-year mortality after hospitalization for HF patients enrolled in a community-based HF-DMP. Our results and those from previous randomized controlled and observational studies on HF-DMP provide strong evidence for a large decrease in HF mortality associated with HF-DMP in clinical practice. This finding is of critical importance given the low proportion of western populations that have access to HF-DMP.[32,33] Health care systems in several countries do not provide HF-DMP at a national level or enforce the implementation of HF-DMP in all health care areas, resulting in moderate to poor coverage of HF-DMP across Europe.[32] Given the high prevalence of HF and its poor prognosis, further promotion and development of community-based case management HF-DMP should significantly reduce the public health burden of HF.

Acknowledgments

The authors thank all physicians from the 21 centers participating in EPICAL2 (Hôpitaux de Brabois et Hôpital Central, CHU Nancy; CH Luneville; Espace chirurgical Ambroise Paré Nancy; CH Alpha Santé Mont-Saint-Martin; CH Pont-à-Mousson; CH Saint-Nicolas Verdun; Hôpital Bon-Secours CHR Metz; CH Freyming Merlebach; Hôpital Sainte-Blandine Metz; Hôpital Bel Air CHR Thionville; CH Marie-Madeleine Forbach; Hôpital Alpha Santé Hayange; CH Saint-Nicolas Sarrebourg; Hôpital Lemire Saint-Avold; Hôpital des Armées Legouest Metz; Clinique Claude Bernard Metz; CH Saint-Charles Saint-Dié; CH Jean Monnet Epinal; CH Neufchateau; CH Vittel).

References

[1] Santulli G. Epidemiology of cardiovascular disease in the 21st century: updated numbers and updated facts. J Cardiovasc Dis 2013;1:1–2.
[2] McMurray JJ, Adamopoulos S, Anker SD, et al. ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure 2012: the task force for the diagnosis and treatment of acute and chronic heart failure 2012 of the European Society of Cardiology. Developed in collaboration with the Heart Failure Association (HFA) of the ESC. Eur Heart J 2012;33:1787–847.

[3] Laribi S, Aouba A, Nikolaou M, et al. Trends in death attributed to heart failure over the past two decades in Europe. Eur J Heart Fail 2012;14:234–9.
[4] Gabet A, Juilliere Y, Lamarche-Vadel A, et al. National trends in rate of patients hospitalized for heart failure and heart failure mortality in France, 2000–2012. Eur J Heart Fail 2015;17:583–90.
[5] Takeda A, Taylor SJ, Taylor RS, et al. Clinical service organisation for heart failure. Cochrane Database Syst Rev 2012;9:CD002752.
[6] Victora CG, Habicht JP, Bryce J. Evidence-based public health: moving beyond randomized trials. Am J Public Health 2004;94:400–5.
[7] Black N. Why we need observational studies to evaluate the effectiveness of health care. BMJ 1996;312:1215–8.
[8] Suissa S. Immortal time bias in pharmaco-epidemiology. Am J Epidemiol 2008;167:492–9.
[9] Agrinier N, Altieri C, Alla F, et al. Effectiveness of a multidimensional home nurse led heart failure disease management program – a French nationwide time-series comparison. Int J Cardiol 2013;168:3652–8.
[10] Levey AS, Coresh J, Balk E, et al. National Kidney Foundation practice guidelines for chronic kidney disease: evaluation, classification, and stratification. Ann Intern Med 2003;139:137–47.
[11] Maisel A, Mueller C, Adams KJr, et al. State of the art: using natriuretic peptide levels in clinical practice. Eur J Heart Fail 2008;10:824–39.
[12] ADELF, ADEREST, AEEMA, EPITER. Recommendations for professional standards and good epidemiological practices (version France 2007). Rev Epidemiol Sante Publique 2008;56(Spec No 1):S121–75.
[13] 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.
[14] Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behav Res 2011;46:399–424.
[15] Stürmer T, Wyss R, Glynn RJ, et al. Propensity scores for confounder adjustment when assessing the effects of medical interventions using nonexperimental study designs. J Intern Med 2014;275:570–80.
[16] Ali MS, Groenwold RH, 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.
[17] Marshall A, Altman DG, Holder RL. Comparison of imputation methods for handling missing covariate data when fitting a Cox proportional hazards model: a resampling study. BMC Med Res Methodol 2010;10:112.
[18] Rubin DB. Multiple implutation after 18+ years. J Am Stat Assoc 1996;91:17.
[19] Brookhart MA, Rassen JA, Schneeweiss S. Instrumental variable methods in comparative safety and effectiveness research. Pharmacoepidemiol Drug Saf 2010;19:537–54.
[20] Terza JV, Basu A, Rathouz PJ. Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. J Health Econ 2008;27:531–43.
[21] Clark AM, Thompson DR. Heart failure disease management programmes: a new paradigm for research. Heart 2012;98:1476–7.
[22] Badano LP, Di Lenarda A, Bellotti P, et al. Patients with chronic heart failure encountered in daily clinical practice are different from the "typical" patient enrolled in therapeutic trials. Ital Heart J 2003;4:84–91.
[23] Damman K, Valente MA, Voors AA, et al. Renal impairment, worsening renal function, and outcome in patients with heart failure: an updated meta-analysis. Eur Heart J 2014;35:455–69.
[24] Silverman SL. From randomized controlled trials to observational studies. Am J Med 2009;122:114–20.
[25] Lowery J, Hopp F, Subramanian U, et al. Evaluation of a nurse practitioner disease management model for chronic heart failure: a multisite implementation study. Congest Heart Fail 2012;18:64–71.
[26] Bonarek-Hessamfar M, Benchimol D, Lauribe P, et al. Multidisciplinary network in heart failure management in a community-based population: results and benefits at 2 years. Int J Cardiol 2009;134:120–2.
[27] Klungel OH, Martens EP, Psaty BM, et al. Methods to assess intended effects of drug treatment in observational studies are reviewed. J Clin Epidemiol 2004;57:1223–31.
[28] Salas M, Hofman A, Stricker BH. Confounding by indication: an example of variation in the use of epidemiologic terminology. Am J Epidemiol 1999;149:981–3.
[29] Austin PC. The performance of different propensity score methods for estimating marginal hazard ratios. Stat Med 2013;32:2837–49.
[30] Boef AG, Dekkers OM, Vandenbroucke JP, et al. Sample size importantly limits the usefulness of instrumental variable methods, depending on instrument strength and level of confounding. J Clin Epidemiol 2014;67:1258–64.
[31] Laborde-Casterot H, Agrinier N, Thilly N. Performing both propensity score and instrumental variable analyses in observational studies often leads to discrepant results: a systematic review. J Clin Epidemiol 2015;68:1232–40.
[32] Jaarsma T, Stromberg A, De Geest S, et al. Heart failure management programmes in Europe. Eur J Cardiovasc Nurs 2006;5:197–205.
[33] Driscoll A, Worrall-Carter L, Hare DL, et al. Evidence-based chronic heart-failure management programmes: reality or myth? BMJ Qual Saf 2011;20:31–7.

Table S1. Baseline characteristics of the patients according to heart failure disease management programme (HF-DMP) enrolment. Crude data without multiple imputation for missing data.

	Overall (n=1816)		Control group (n=1504)		HF-DMP group (n=312)		SDiff %
Socio-demographic characteristics							
Male sex	892	49.1%	706	46.9%	186	59.6%	**25.61**
Age < 65 years	289	15.9%	230	15.3%	59	18.9%	**22.89**
65-79 years	582	32.1%	462	30.7%	120	38.5%	
≥ 80 years	945	52.0%	812	54.0%	133	42.6%	
Living alone	510	32.0%	428	33.1%	82	27.4%	**12.27**
Type of hospital: Local	782	43.1%	590	39.2%	192	61.5%	**59.71**
Regional	461	25.4%	435	28.9%	26	8.3%	
Teaching	573	31.5%	479	31.9%	94	30.1%	
Medical history							
Ischaemic HF aetiology	431	23.7%	330	21.9%	101	32.4%	**23.61**
Diabetes	654	36.0%	510	33.9%	144	46.2%	**25.19**
Chronic kidney disease	415	22.9%	321	21.3%	94	30.1%	**20.20**
Stroke / TIA	231	12.7%	191	12.7%	40	12.8%	0.36
COPD / asthma	410	22.6%	326	21.7%	84	26.9%	**12.26**
Cancer	260	14.3%	225	15.0%	35	11.2%	**11.11**
Previous hospitalisation for acute HF	635	35.1%	469	31.3%	166	53.4%	**45.87**
Clinical and biological status at admission							
Overweight or obese	994	68.2%	595	66.3%	202	78.3%	**34.01**
LVEF < 30%	262	17.8%	191	15.5%	71	30.0%	**40.83**
30-44%	465	31.6%	385	31.2%	80	33.8%	
≥ 45%	746	50.6%	660	53.4%	86	36.3%	
Low systolic BP[1]	333	18.5%	266	17.8%	67	21.9%	**10.27**
Low eGFR[2]	1172	65.3%	953	64.0%	219	71.6%	**16.33**
Prolonged QRS duration[3]	172	12.3%	127	11.0%	45	18.4%	**21.17**
Hyponatraemia[4]	322	18.0%	267	18.0%	55	17.9%	0.26
Hypokalaemia[5]	439	24.6%	361	24.4%	78	25.4%	2.31
Anaemia[6]	188	10.5%	151	10.2%	37	12.0%	5.60
Increased BNP/NT-proBNP[7]	1318	79.9%	1076	79.9%	246	79.9%	0.18

Abbreviations. BNP: B-type natriuretic peptide; BP: blood pressure; COPD: chronic obstructive pulmonary disease; eGFR: glomerular filtration rate; HF: heart failure; LVEF: left ventricular ejection fraction; NT-proBNP: N-terminal pro-B-type natriuretic peptide; SDiff: absolute standardized difference; TIA: transient ischaemic attack.
Notes. 1: systolic BP < 115 mmHg; 2: eGFR < 60mL/min/1.73m^2; 3: QRS > 120 ms; 4: natraemia < 135 mmol/L; 5: kalaemia < 3.8 mmol/L; 6: haemoglobin < 10 g/dL; 7: BNP > 400 pg/mL or (NT-proBNP > 450 pg/mL in patients < 50 years old, NT-proBNP > 900 pg/mL in patients between 50 and 75 years old, NT-proBNP > 1800 pg/mL in patients > 75 years old)

Table S2. Characteristics of patients before and after inverse probability of treatment weighting and in preference-based instrumental variable approach in dataset 1.

		Unweighted population			Weighted population			Instrumental variable analysis		
	Overall	Control group	HF-DMP group	SDiff %	Control group	HF-DMP group	SDiff %	No-DMP preference	DMP preference	SDiff %
N	1816	1504	312		1326	297		1141	675	
Socio-demographic characteristics										
Male sex		46.9%	59.6%	**25.61**	48.4%	49.3%	1.77	48.4%	50.4%	3.99
Age < 65 years		15.3%	18.9%	**22.89**	14.1%	14.0%	3.67	15.2%	17.2%	8.00
65-79 years		30.7%	38.5%		31.7%	30.2%		31.4%	33.2%	
≥ 80 years		54.0%	42.6%		54.1%	55.8%		53.5%	49.6%	
Living alone		35.7%	28.9%	**14.70**	34.9%	36.7%	3.95	34.5%	34.5%	0.03
Type of hospital: Local		39.2%	61.5%	**59.71**	45.8%	47.5%	**15.86**	43.7%	42.1%	**137.64**
Regional		28.9%	8.3%		22.6%	16.5%		40.4%	0.0%	
Teaching		31.9%	30.1%		31.6%	36.0%		16.0%	57.9%	
Medical history										
Ischaemic HF aetiology		21.9%	32.4%	**23.61**	24.5%	27.9%	7.69	24.4%	22.7%	4.00
Diabetes		33.9%	46.2%	**25.19**	36.7%	37.9%	2.58	34.7%	38.2%	7.31
Chronic kidney disease		21.3%	30.1%	**20.20**	23.5%	27.2%	8.41	22.7%	23.1%	0.98
Stroke / TIA		12.7%	12.8%	0.36	12.7%	9.6%	9.62	13.1%	12.2%	2.74
COPD / asthma		21.7%	26.9%	**12.26**	22.7%	21.3%	3.34	23.1%	21.6%	3.62
Cancer		15.0%	11.2%	**11.11**	14.0%	12.8%	3.46	14.6%	13.9%	1.78
Previous hospitalisation for acute HF		31.5%	53.5%	**45.67**	36.7%	37.1%	0.71	35.0%	35.7%	1.54
Clinical and biological status at admission										
Overweight or obese		65.8%	77.9%	**32.90**	67.5%	65.4%	7.51	66.6%	69.9%	7.34
LVEF < 30%		15.6%	26.3%	**29.33**	17.8%	18.4%	3.34	17.2%	17.8%	1.71
30-44%		31.1%	32.1%		31.0%	32.0%		31.2%	31.3%	
≥ 45%		53.4%	41.7%		51.2%	49.6%		51.6%	51.0%	
Low systolic BP[1]		17.7%	21.8%	**10.34**	17.5%	16.5%	2.64	17.4%	20.0%	6.56
Low eGFR[2]		64.0%	71.6%	**16.33**	66.5%	67.5%	2.18	65.9%	64.2%	3.48
Prolonged QRS duration[3]		13.0%	18.3%	**14.65**	13.8%	14.6%	2.20	13.2%	15.1%	5.64
Hyponatraemia[4]		18.0%	17.9%	0.26	18.1%	17.4%	1.76	17.7%	18.4%	1.83
Hypokalaemia[5]		24.4%	25.4%	2.31	24.0%	26.9%	6.64	25.8%	22.5%	7.67
Anaemia[6]		10.2%	12.0%	5.60	10.8%	11.2%	1.24	10.4%	10.8%	1.34
Increased BNP/NT-proBNP[7]		79.9%	79.9%	0.18	80.0%	79.1%	2.22	79.9%	79.9%	0.09

Abbreviations. BNP: B-type natriuretic peptide; BP: blood pressure; COPD: chronic obstructive pulmonary disease; eGFR: glomerular filtration rate; HF: heart failure; LVEF: left ventricular ejection fraction; NT-proBNP: N-terminal pro-B-type natriuretic peptide; SDiff: absolute standardized difference; TIA: transient ischaemic attack.
Notes. 1: systolic BP < 115 mmHg; 2: eGFR < 60mL/min/1.73m^2; 3: QRS > 120 ms; 4: natraemia < 135 mmol/L; 5: kalaemia < 3.8 mmol/L; 6: haemoglobin < 10 g/dL; 7: BNP > 400 pg/mL or (NT-proBNP > 450 pg/mL in patients < 50 years old, NT-proBNP > 900 pg/mL in patients between 50 and 75 years old, NT-proBNP > 1800 pg/mL in patients > 75 years old)

Table S3. Characteristics of patients before and after inverse probability of treatment weighting and in preference-based instrumental variable approach in dataset 2.

		Unweighted population			Weighted population			Instrumental variable analysis		
	Overall	Control group	HF-DMP group	SDiff %	Control group	HF-DMP group	SDiff %	No-DMP preference	DMP preference	SDiff %
N	1816	1504	312		1326	297		1141	675	
Socio-demographic characteristics										
Male sex		46.9%	59.6%	**25.61**	48.4%	49.5%	2.28	48.4%	50.4%	3.99
Age < 65 years		15.3%	18.9%	**22.89**	14.3%	14.0%	3.67	15.2%	17.2%	8.00
65-79 years		30.7%	38.5%		31.6%	30.1%		31.4%	33.2%	
≥ 80 years		54.0%	42.6%		54.1%	55.9%		53.5%	49.6%	
Living alone		34.5%	28.9%	**12.20**	33.9%	35.5%	3.36	33.9%	32.9%	2.18
Type of hospital: Local		39.2%	61.5%	**59.71**	45.8%	48.3%	**15.41**	43.7%	42.1%	**137.64**
Regional		28.9%	8.3%		22.6%	16.5%		40.4%	0.0%	
Teaching		31.9%	30.1%		31.6%	35.1%		16.0%	57.9%	
Medical history										
Ischaemic HF aetiology		21.9%	32.4%	**23.61**	24.5%	28.1%	8.19	24.4%	22.7%	4.00
Diabetes		33.9%	46.2%	**25.19**	36.6%	37.3%	1.39	34.7%	38.2%	7.31
Chronic kidney disease		21.3%	30.1%	**20.20**	23.5%	27.6%	9.40	22.7%	23.1%	0.98
Stroke / TIA		12.7%	12.8%	0.36	12.6%	9.8%	8.77	13.1%	12.2%	2.74
COPD / asthma		21.7%	26.9%	**12.26**	22.7%	21.6%	2.65	23.1%	21.6%	3.62
Cancer		15.0%	11.2%	**11.11**	14.1%	13.1%	2.71	14.6%	13.9%	1.78
Previous hospitalisation for acute HF		31.5%	53.5%	**45.67**	36.7%	37.0%	0.59	35.1%	35.7%	1.35
Clinical and biological status at admission										
Overweight or obese		65.7%	77.9%	**34.23**	67.6%	65.9%	7.56	66.6%	69.8%	7.04
LVEF < 30%		16.0%	26.6%	**32.53**	17.8%	18.1%	0.63	17.6%	18.1%	2.13
30-44%		30.5%	34.3%		30.4%	30.3%		31.5%	30.5%	
≥ 45%		53.6%	39.1%		51.8%	51.6%		50.9%	51.4%	
Low systolic BP[1]		17.8%	21.5%	9.21	17.5%	16.6%	2.37	17.4%	20.2%	6.93
Low eGFR[2]		64.0%	71.6%	**16.33**	66.5%	67.8%	2.95	65.9%	64.2%	3.48
Prolonged QRS duration[3]		13.0%	16.7%	**10.43**	13.6%	14.2%	1.65	12.3%	15.9%	**10.32**
Hyponatraemia[4]		18.0%	17.9%	0.26	18.2%	18.2%	0.14	17.7%	18.4%	1.83
Hypokalaemia[5]		24.4%	25.4%	2.31	23.9%	26.8%	6.69	25.8%	22.5%	7.67
Anaemia[6]		10.2%	12.0%	5.60	10.8%	11.0%	0.61	10.4%	10.8%	1.34
Increased BNP/NT-proBNP[7]		79.9%	79.9%	0.18	80.0%	79.4%	1.46	79.9%	79.9%	0.09

Abbreviations. BNP: B-type natriuretic peptide; BP: blood pressure; COPD: chronic obstructive pulmonary disease; eGFR: glomerular filtration rate; HF: heart failure; LVEF: left ventricular ejection fraction; NT-proBNP: N-terminal pro-B-type natriuretic peptide; SDiff: absolute standardized difference; TIA: transient ischaemic attack.
Notes. 1: systolic BP < 115 mmHg; 2: eGFR < 60mL/min/1.73m^2; 3: QRS > 120 ms; 4: natraemia < 135 mmol/L; 5: kalaemia < 3.8 mmol/L; 6: haemoglobin < 10 g/dL; 7: BNP > 400 pg/mL or (NT-proBNP > 450 pg/mL in patients < 50 years old, NT-proBNP > 900 pg/mL in patients between 50 and 75 years old, NT-proBNP > 1800 pg/mL in patients > 75 years old)

Table S4. Characteristics of patients before and after inverse probability of treatment weighting and in preference-based instrumental variable approach in dataset 3.

		Unweighted population			Weighted population			Instrumental variable analysis		
	Overall	Control group	HF-DMP group	SDiff %	Control group	HF-DMP group	SDiff %	No-DMP preference	DMP preference	SDiff %
N	1816	1504	312		1326	297		1141	675	
Socio-demographic characteristics										
Male sex		46.9%	59.6%	**25.61**	48.3%	48.8%	1.02	48.4%	50.4%	3.99
Age < 65 years		15.3%	18.9%	**22.89**	14.3%	13.9%	3.13	15.2%	17.2%	8.00
65-79 years		30.7%	38.5%		31.6%	30.5%		31.4%	33.2%	
≥ 80 years		54.0%	42.6%		54.1%	55.7%		53.5%	49.6%	
Living alone		34.2%	29.1%	9.51	33.8%	36.3%	5.35	33.3%	33.8%	1.00
Type of hospital: Local		39.2%	61.5%	**59.71**	45.8%	47.3%	**14.64**	43.7%	42.1%	**137.64**
Regional		28.9%	8.3%		22.6%	17.0%		40.4%	0.0%	
Teaching		31.9%	30.1%		31.6%	35.7%		16.0%	57.9%	
Medical history										
Ischaemic HF aetiology		21.9%	32.4%	**23.61**	24.4%	28.0%	8.28	24.4%	22.7%	4.00
Diabetes		33.9%	46.2%	**25.19**	36.7%	38.6%	4.04	34.7%	38.2%	7.31
Chronic kidney disease		21.3%	30.1%	**20.20**	23.5%	27.8%	9.85	22.7%	23.1%	0.98
Stroke / TIA		12.7%	12.8%	0.36	12.6%	9.6%	9.37	13.1%	12.2%	2.74
COPD / asthma		21.7%	26.9%	**12.26**	22.7%	21.2%	3.79	23.1%	21.6%	3.62
Cancer		15.0%	11.2%	**11.11**	14.0%	12.7%	3.76	14.6%	13.9%	1.78
Previous hospitalisation for acute HF		31.5%	53.5%	**45.67**	36.8%	37.0%	0.48	35.1%	35.7%	1.35
Clinical and biological status at admission										
Overweight or obese		65.6%	76.3%	**29.00**	67.1%	66.5%	2.89	66.5%	68.9%	5.37
LVEF < 30%		15.8%	27.2%	**30.10**	18.0%	17.6%	1.19	17.6%	17.9%	1.55
30-44%		31.6%	31.4%		30.8%	31.1%		31.8%	31.1%	
≥ 45%		52.7%	41.4%		51.2%	51.3%		50.6%	51.0%	
Low systolic BP[1]		17.7%	21.8%	**10.34**	17.5%	16.6%	2.40	17.4%	20.0%	6.56
Low eGFR[2]		64.0%	71.6%	**16.33**	66.5%	68.0%	3.21	65.9%	64.2%	3.48
Prolonged QRS duration[3]		12.1%	18.9%	**18.90**	13.6%	14.2%	1.52	12.3%	15.0%	7.86
Hyponatraemia[4]		18.0%	17.9%	0.26	18.2%	17.1%	2.92	17.7%	18.4%	1.83
Hypokalaemia[5]		24.4%	25.4%	2.31	23.9%	26.4%	5.71	25.8%	22.5%	7.67
Anaemia[6]		10.2%	12.0%	5.60	10.8%	11.2%	1.22	10.4%	10.8%	1.34
Increased BNP/NT-proBNP[7]		79.9%	79.9%	0.18	80.0%	77.9%	5.14	79.9%	79.9%	0.09

Abbreviations. BNP: B-type natriuretic peptide; BP: blood pressure; COPD: chronic obstructive pulmonary disease; eGFR: glomerular filtration rate; HF: heart failure; LVEF: left ventricular ejection fraction; NT-proBNP: N-terminal pro-B-type natriuretic peptide; SDiff: absolute standardized difference; TIA: transient ischaemic attack.
Notes. 1: systolic BP < 115 mmHg; 2: eGFR < 60mL/min/1.73m^2; 3: QRS > 120 ms; 4: natraemia < 135 mmol/L; 5: kalaemia < 3.8 mmol/L; 6: haemoglobin < 10 g/dL; 7: BNP > 400 pg/mL or (NT-proBNP > 450 pg/mL in patients < 50 years old, NT-proBNP > 900 pg/mL in patients between 50 and 75 years old, NT-proBNP > 1800 pg/mL in patients > 75 years old)

Table S5. Characteristics of patients before and after inverse probability of treatment weighting and in preference-based instrumental variable approach in dataset 4.

		Unweighted population			Weighted population			Instrumental variable analysis		
	Overall	Control group	HF-DMP group	SDiff %	Control group	HF-DMP group	SDiff %	No-DMP preference	DMP preference	SDiff %
N	1816	1504	312		1313	283		1141	675	
Socio-demographic characteristics										
Male sex		46.9%	59.6%	**25.61**	48.3%	48.5%	0.50	48.4%	50.4%	3.99
Age < 65 years		15.3%	18.9%	**22.89**	14.2%	14.0%	3.78	15.2%	17.2%	8.00
65-79 years		30.7%	38.5%		31.6%	30.1%		31.4%	33.2%	
≥ 80 years		54.0%	42.6%		54.1%	55.9%		53.5%	49.6%	
Living alone		35.2%	30.1%	**10.77**	34.6%	36.6%	4.22	34.2%	34.5%	0.71
Type of hospital: Local		39.2%	61.5%	**59.71**	45.9%	47.7%	**13.88**	43.7%	42.1%	**137.64**
Regional		28.9%	8.3%		22.6%	17.2%		40.4%	0.0%	
Teaching		31.9%	30.1%		31.3%	35.1%		16.0%	57.9%	
Medical history										
Ischaemic HF aetiology		21.9%	32.4%	**23.61**	24.4%	28.7%	9.70	24.4%	22.7%	4.00
Diabetes		33.9%	46.2%	**25.19**	36.6%	38.1 %	3.15	34.7%	38.2%	7.31
Chronic kidney disease		21.3%	30.1%	**20.20**	23.6%	28.3%	**10.85**	22.7%	23.1%	0.98
Stroke / TIA		12.7%	12.8%	0.36	12.7%	9.7%	9.55	13.1%	12.2%	2.74
COPD / asthma		21.7%	26.9%	**12.26**	22.8%	21.4%	3.30	23.1%	21.6%	3.62
Cancer		15.0%	11.2%	**11.11**	14.0%	12.8%	3.71	14.6%	13.9%	1.78
Previous hospitalisation for acute HF		31.5%	53.5%	**45.67**	36.6%	36.9%	0.47	35.0%	35.7%	1.54
Clinical and biological status at admission										
Overweight or obese		66.9%	76.9%	**28.59**	68.1%	65.4%	8.10	67.8%	70.1%	5.70
LVEF < 30%		15.4%	27.6%	**34.50**	17.6%	17.0%	0.97	16.6%	19.0%	6.99
30-44%		31.5%	33.7%		31.0%	30.7%		31.7%	32.5%	
≥ 45%		53.1%	38.8%		51.4%	51.9%		51.7%	48.9%	
Low systolic BP[1]		17.8%	21.5%	9.21	17.6%	16.7%	1.58	17.4%	20.2%	6.93
Low eGFR[2]		64.0%	71.6%	**16.33**	66.5%	68.5%	4.24	65.9%	64.2%	3.48
Prolonged QRS duration[3]		12.5%	19.2%	**18.50**	13.9%	16.0%	5.87	12.2%	16.2%	**11.39**
Hyponatraemia[4]		18.0%	17.9%	0.26	18.1%	17.2%	2.23	17.7%	18.4%	1.83
Hypokalaemia[5]		24.4%	25.4%	2.31	24.0%	26.7%	6.05	25.8%	22.5%	7.67
Anaemia[6]		10.2%	12.0%	5.60	10.8%	11.1%	1.20	10.4%	10.8%	1.34
Increased BNP/NT-proBNP[7]		79.9%	79.9%	0.18	79.9%	78.5%	3.62	79.9%	79.9%	0.09

Abbreviations. BNP: B-type natriuretic peptide; BP: blood pressure; COPD: chronic obstructive pulmonary disease; eGFR: glomerular filtration rate; HF: heart failure; LVEF: left ventricular ejection fraction; NT-proBNP: N-terminal pro-B-type natriuretic peptide; SDiff: absolute standardized difference; TIA: transient ischaemic attack.
Notes. 1: systolic BP < 115 mmHg; 2: eGFR < 60mL/min/1.73m^2; 3: QRS > 120 ms; 4: natraemia < 135 mmol/L; 5: kalaemia < 3.8 mmol/L; 6: haemoglobin < 10 g/dL; 7: BNP > 400 pg/mL or (NT-proBNP > 450 pg/mL in patients < 50 years old, NT-proBNP > 900 pg/mL in patients between 50 and 75 years old, NT-proBNP > 1800 pg/mL in patients > 75 years old).

Table S6. Characteristics of patients before and after inverse probability of treatment weighting and in preference-based instrumental variable approach in dataset 5.

		Unweighted population			Weighted population			Instrumental variable analysis		
	Overall	Control group	HF-DMP group	SDiff %	Control group	HF-DMP group	SDiff %	No-DMP preference	DMP preference	SDiff %
N	1816	1504	312		1326	297		1141	675	
Socio-demographic characteristics										
Male sex		46.9%	59.6%	**25.61**	48.4%	49.4%	2.06	48.4%	50.4%	3.99
Age < 65 years		15.3%	18.9%	**22.89**	14.2%	13.8%	3.35	15.2%	17.2%	8.00
65-79 years		30.7%	38.5%		31.7%	30.5%		31.4%	33.2%	
≥ 80 years		54.0%	42.6%		54.0%	55.7%		53.5%	49.6%	
Living alone		33.6%	28.5%	**11.08**	33.0%	35.4%	5.14	32.3%	33.6%	2.93
Type of hospital: Local		39.2%	61.5%	**59.71**	45.8%	47.3%	**14.95**	43.7%	42.1%	**137.64**
Regional		28.9%	8.3%		22.6%	16.9%		40.4%	0.0%	
Teaching		31.9%	30.1%		31.6%	35.8%		16.0%	57.9%	
Medical history										
Ischaemic HF aetiology		21.9%	32.4%	**23.61**	24.5%	28.5%	9.14	24.4%	22.7%	4.00
Diabetes		33.9%	46.2%	**25.19**	36.6%	38.4%	3.77	34.7%	38.2%	7.31
Chronic kidney disease		21.3%	30.1%	**20.20**	23.5%	28.0%	**10.18**	22.7%	23.1%	0.98
Stroke / TIA		12.7%	12.8%	0.36	12.6%	10.0%	8.27	13.1%	12.2%	2.74
COPD / asthma		21.7%	26.9%	**12.26**	22.8%	21.5%	2.98	23.1%	21.6%	3.62
Cancer		15.0%	11.2%	**11.11**	14.0%	13.2%	2.39	14.6%	13.9%	1.78
Previous hospitalisation for acute HF		31.5%	53.5%	**45.67**	36.7%	38.1%	2.84	35.0%	35.7%	1.54
Clinical and biological status at admission										
Overweight or obese		66.4%	76.3%	**30.21**	67.6%	66.6%	5.78	67.5%	69.0%	3.56
LVEF < 30%		16.2%	26.9%	**32.11**	18.2%	18.3%	0.71	17.8%	18.5%	2.75
30-44%		31.2%	34.6%		31.2%	31.3%		31.6%	32.2%	
≥ 45%		52.6%	38.5%		50.7%	50.4%		50.7%	49.3%	
Low systolic BP[1]		17.8%	21.8%	9.99	17.7%	17.1%	1.61	17.4%	20.3%	7.30
Low eGFR[2]		64.0%	71.6%	**16.33**	66.5%	67.6%	2.35	65.9%	64.2%	3.48
Prolonged QRS duration[3]		12.3%	18.0%	**15.81**	13.6%	14.3%	1.83	12.3%	15.0%	7.86
Hyponatraemia[4]		18.0%	17.9%	0.26	18.2%	17.8%	1.08	17.7%	18.4%	1.83
Hypokalaemia[5]		24.4%	25.4%	2.31	24.0%	26.7%	6.33	25.8%	22.5%	7.67
Anaemia[6]		10.2%	12.0%	5.60	10.8%	11.0%	0.56	10.4%	10.8%	1.34
Increased BNP/NT-proBNP[7]		79.9%	79.9%	0.18	79.9%	78.9%	2.59	79.9%	79.9%	0.09

Abbreviations. BNP: B-type natriuretic peptide; BP: blood pressure; COPD: chronic obstructive pulmonary disease; eGFR: glomerular filtration rate; HF: heart failure; LVEF: left ventricular ejection fraction; NT-proBNP: N-terminal pro-B-type natriuretic peptide; SDiff: absolute standardized difference; TIA: transient ischaemic attack.
Notes. 1: systolic BP < 115 mmHg; 2: eGFR < 60mL/min/1.73m^2; 3: QRS > 120 ms; 4: natraemia < 135 mmol/L; 5: kalaemia < 3.8 mmol/L; 6: haemoglobin < 10 g/dL; 7: BNP > 400 pg/mL or (NT-proBNP > 450 pg/mL in patients < 50 years old, NT-proBNP > 900 pg/mL in patients between 50 and 75 years old, NT-proBNP > 1800 pg/mL in patients > 75 years old).

Table S7. Estimations of the heart failure disease management programme effect in the 5 datasets

	Dataset 1		Dataset 2		Dataset 3		Dataset 4		Dataset 5		Pooled results	
	HR	95% IC	HR	95% IC	HR	95% IC	HR	95% IC	HR	95% IC	HR	95% IC
Model 1: propensity score analysis using IPTW	0.66	0.47 to 0.91	0.68	0.49 to 0.94	0.62	0.44 to 0.87	0.63	0.45 to 0.88	0.67	0.48 to 0.93	0.65	0.46 to 0.92
Model 2: model 1 + adjustment for type of hospital and history of chronic kidney disease	0.64	0.46 to 0.89	0.66	0.48 to 0.92	0.61	0.43 to 0.85	0.61	0.44 to 0.85	0.65	0.47 to 0.90	0.63	0.45 to 0.89
Model 3: model 1 + trimming of 2.5% of patients on both sides	0.56	0.38 to 0.82	0.56	0.38 to 0.81	0.58	0.40 to 0.85	0.60	0.42 to 0.86	0.62	0.43 to 0.89	0.58	0.40 to 0.86
Model 4: model 2 with trimming of 2.5% of patients on both sides	0.55	0.38 to .81	0.55	0.38 to 0.80	0.58	0.40 to 0.84	0.60	0.41 to 0.86	0.61	0.43 to 0.88	0.58	0.39 to 0.85
Instrumental variable	0.58	0.28 to 1.20	0.58	0.28 to 1.20	0.56	0.27 to 1.14	0.55	0.27 to 1.14	0.53	0.25 to 1.09	0.56	0.27 to 1.16

Abbreviations. 95% CI: 95% confidence interval; HR: hazard ratio; IPTW: Inverse probability of treatment weighting.

B.1.2. Discussion

L'étude EPICAL 2 est un cas d'école illustrant le risque de biais d'indication dans les études observationnelles. L'intervention évaluée dans notre travail, le réseau de soins ICALOR, était proposée aux patients à la libre appréciation des équipes médicales les prenant en charge. Tout patient habitant en Lorraine et présentant une IC était susceptible d'être inclus dans le réseau, si leur médecin le leur proposait et qu'il l'acceptait. L'étude, observationnelle, ne modifiait en rien le processus spontané d'allocation du traitement. Ainsi, 312 (17,2%) patients issus de la cohorte EPICAL2 étaient pris en charge par le réseau ICALOR. Ces patients n'étaient pas comparables à ceux du groupe témoin, puisque, par exemple, ils étaient davantage des hommes (59,6% vs 46,9%), étaient plus jeunes (42,6% avaient plus de 80 ans vs 54,0%), avaient davantage de comorbidités (46,2% étaient diabétiques vs 33,9%), ou encore avaient une fraction d'éjection du ventricule gauche plus basse (inférieure à 45% pour 61,1% des patients vs 46,9%). Les médecins avaient donc proposé une prise en charge par ICALOR à des patients présentant, dans l'ensemble, des caractéristiques particulières, vraisemblablement parce qu'ils attendaient un bénéfice plus important du réseau chez ces patients, alors que ces caractéristiques n'étaient pas des critères d'éligibilité au réseau. Ces caractéristiques étant des facteurs pronostiques de l'insuffisance cardiaque, on était confronté au risque de biais d'indication qui fausserait les résultats de l'évaluation et l'interprétation sur l'effet du réseau. Les facteurs de mauvais pronostic étant davantage prévalents dans le groupe ICALOR, le risque était de sous-estimer l'effet du réseau. Il fallait donc recourir à des méthodes d'analyse limitant le biais d'indication.

Notre analyse principale a été effectuée par pondération inverse sur un SP. Le score a été calculé en prenant en compte l'ensemble des caractéristiques initiales mesurées des individus. Cette méthode a permis de montrer une diminution significative de la mortalité dans le groupe réseau (HR 0,65 [IC95% 0,46 – 0,92]), qui n'apparaissait pas dans l'analyse brute sans prise

en compte des facteurs de confusion (HR 0,82 [IC95% 0,61 – 1,09]). Notre choix méthodologique ayant retenu l'utilisation d'un SP, nous n'avons pas rapporté dans l'article le résultat d'une analyse conventionnelle ajustée sur toutes les variables mesurées initialement. Ce résultat (HR 0,69 [IC95% 0,50 – 0,95]) est très proche de celui obtenu avec la méthode basée sur un SP, comme cela est habituellement rapporté dans la littérature (*cf.* §A.5.2.1.6, p.55).

Nous avons également envisagé l'utilisation d'une VI, dans l'hypothèse d'un biais résiduel avec la méthode basée sur un SP, du fait de l'existence probable de facteurs de confusion non mesurés. Toutefois, compte tenu des limites de la méthode basée sur une VI (*cf.* §A.5.2.2.2, p.63 : risque d'une plus grande incertitude de l'estimation en raison de l'effectif relativement limité, respect très incertain de la monotonicité), nous l'avons considérée comme une analyse de sensibilité. Nous avons identifié un instrument basé sur la préférence des médecins concernant le réseau au niveau de chaque établissement de santé. Il est apparu que le recours au réseau ICALOR était très variable entre les 21 hôpitaux participant à l'étude EPICAL 2. Cette différence dans les pratiques ne pouvait s'expliquer seulement par un recrutement de patients aux caractéristiques différentes, mais reflétait des habitudes différentes que l'on peut assimiler à une préférence. L'instrument retenu était donc le pourcentage de patients participant à ICALOR dans chaque établissement de santé. Le respect des conditions de validité de cet instrument a été discuté dans l'article.

L'analyse basée sur une VI a confirmé la réduction de la mortalité associée au réseau, même si le résultat n'était pas statistiquement significatif (HR 0,56 [IC95% 0,27 – 1,16]). Comme attendu, l'analyse basée sur une VI a produit une estimation avec un intervalle de confiance plus large que l'analyse basée sur un SP, mais la position de l'estimateur doit être considérée (Boef, Dekkers et al. 2014). Il est intéressant de noter que l'effet *LATE* chez les patients marginaux (HR 0,56) était très proche de l'effet moyen *ATE* dans l'analyse basée sur un SP

avec restriction de la population par '*trimming*' (HR 0,58) et sensiblement plus important que dans l'analyse avec SP réalisé sur l'échantillon entier (HR 0,65). La réduction par '*trimming*' dans l'analyse avec SP pourrait diminuer le biais produit par des facteurs de confusion non mesurés et augmenter la validité interne des résultats quand l'hétérogénéité de l'effet de l'intervention est liée à des facteurs de confusion non mesurés chez des individus traités contrairement à la prédiction (Sturmer, Rothman et al. 2010). Cela suggère que l'effet du réseau pourrait être hétérogène en raison de facteurs de confusion non mesurés liés à la fragilité des individus.

B.2. Évaluation de l'effet des stratégies médicamenteuses appropriées dans l'insuffisance cardiaque sur la mortalité (article 3)

B.2.1. Contexte

Ce travail est une deuxième exploitation des données issues de l'étude de cohorte EPICAL 2, et vise à évaluer l'effet de la conformité aux recommandations de bonne pratique des prescriptions médicamenteuses en sortie d'hospitalisation chez les patients atteints d'insuffisance cardiaque à fraction d'éjection réduite.

Les prescriptions étudiées concernaient les inhibiteurs de l'enzyme de conversion (IEC) (ou les antagonistes de l'angiotensine (ARA2) en cas d'intolérance et/ou contre-indication aux IEC) et les bêtabloquants (β-). Leur efficacité chez les patients présentant une insuffisance cardiaque à fraction d'éjection réduite, pour prévenir les réadmissions et les décès, a été largement démontrée dans des ECR au cours des 30 dernières années. Leur prescription est aujourd'hui recommandée en traitement de fond de l'insuffisance cardiaque à fraction d'éjection réduite (Ponikowski, Voors et al. 2016).

Cependant, il faut souligner que les patients inclus dans les ECR présentent des caractéristiques particulières (en termes d'âge, de co-morbidités, de co-prescriptions, …), et que le contexte particulier de l'expérimentation est éloigné de la pratique courante induisant des comportements particuliers aussi bien chez les patients (observance optimisée) que chez les soignants (suivi rapproché). Il est donc pertinent de réaliser des études observationnelles comparatives pour vérifier l'efficacité des prescriptions médicamenteuses recommandées en conditions de pratiques courantes de soins.

Dans ce contexte, l'objectif de ce travail était l'évaluer l'effet sur la mortalité sur une période de un an, du respect des recommandations de bonne pratique pour la prescription des

IEC/ARA2 et β- à la sortie d'hospitalisation chez des patients présentant une insuffisance cardiaque à fraction d'éjection réduite.

L'article issu de ce travail a été soumis à la revue *BMJ Quality & Safety* dans la version reproduite dans les pages suivantes.

Effectiveness of guideline-consistent heart failure drug prescriptions at hospital discharge on 1-year mortality: results from the EPICAL2 cohort study

Amandine BUSSON[1,2†], Hervé LABORDE-CASTEROT[2†,] , François ALLA[1,2], Ziyad MESSIKH[1], Isabelle CLERC-URMES[1], Alexandre Mebazaa[3,4,5], Nathalie THILLY [1,2], Nelly AGRINIER [1,2]

[1] Inserm, CIC-1433, Epidémiologie Clinique, CHRU Nancy, Nancy, France

[2] University of Lorraine, EA 4360 Apemac, Nancy, France

[3] Inserm U942, Paris, F-75000, France

[4] University Paris Diderot, Sorbonne Paris Cité, Paris, F-75000, France

[5] Department of Anesthesia and Critical Care, Hôpitaux Universitaires Saint-Louis Lariboisière, APHP, Paris, F-75000, France.

[†] These authors are first co-authors.

*Correspondence to: Nelly AGRINIER, CIC-EC, CHRU Nancy, Hôpitaux de Brabois, Allée du morvan, 54500 Vandoeuvre lès Nancy, France. Phone: +33 3 83 85 21 63; Fax: +33 3 83 85 12 05; e-mail: n.agrinier@chru-nancy.fr

Keywords (3): Clinical practice guidelines, Chronic disease management, Pharmacoepidemiology

Word Count: 3641 (excluding, title page, abstract, reference, figures and tables)

ABSTRACT

Background: We aimed to assess the effectiveness of recommended drug prescriptions at hospital discharge on 1-year mortality in patients with heart failure (HF) and reduced ejection fraction (HFREF).

Methods: We used data from the EPICAL2 cohort study. HF patients ≥ 18 years old with left ventricular ejection fraction (LVEF) <40% who were alive at discharge were included and followed up for mortality. Socio-demographic, clinical and therapeutic data were collected at admission. Therapeutic data were collected at discharge and at 6 month. Prescription of an angiotensin-converting enzyme (ACE) inhibitor (or an angiotensin II receptor blocker [ARB] in case of ACE inhibitor intolerance) and a β-blocker at discharge were considered "guideline-consistent discharge prescription" (GCDP). The effect of GCDP was assessed by a frailty Cox model after propensity score (PS) matching.

Results: Among 632 patients included, the mean (SD) age was 73.8 (12.9) years; 65% were male. A total of 414 (65.5%) patients received GCDP, and 81.8% still had guideline consistent prescription at 6 months. A total of 165 patients died during the follow-up, 78 in the GCDP group and 87 in the other group. Before PS matching, patients with GCDP were younger (|StDiff|=48.91%) and had higher body mass index (BMI) (|StDiff|=12.54%), lower LVEF (|StDiff|=21.96%) and lower Charlson index (|StDiff|=55.78%) than patients without GCDP. After PS matching and adjustment for residual unbalanced characteristics (BMI and history of hypertension), GCDP was associated with reduced mortality (pooled HR= 0.54, 95% CI [0.35-0.83]).

Conclusion: Prescription of ACE (or ARB) inhibitors and β-blockers for patients with HFREF may be low despite the evidence for morbidity and mortality improvement with these medications but effective in reducing 1-year mortality in a real-world setting.

INTRODUCTION

In France in 2002, heart failure (HF) prevalence reached 2.2% in the general population and 11.9% in patients 60 years and older [1]. In 2010, as the main cause of death, this syndrome caused almost 24,000 deaths [2], for a total of 539,000 deaths in France [3]. The number of hospital stays due to HF and the crude HF hospitalisation rate increased by 35% and 26%, respectively, from 1997 to 2008 in France [4]. In HF patients, hospitalisations due to HF accounted for 63% of global health expenses in France in 2007 (i.e., more than one billion Euros) [5]. Accordingly, reducing HF mortality and acute decompensation was included as an objective in the last public health act in France [6].

From results of randomised controlled trials [7- 9], the European Society of Cardiology (ESC) recommended the prescription of angiotensin-converting enzyme (ACE) inhibitors (or angiotensin II receptor antagonists [ARBs] in case of ACE inhibitor intolerance) and β-blockers for all patients with HF and reduced ejection fraction (HFREF) [10-11] to prevent HF hospitalisations and death. These drugs have shown efficacy under experimental conditions of clinical trials in highly selected patients. However, HF patients participating in trials are usually younger, have fewer comorbidities and more recent HF onset than HF patients in current medical practice [12]. In addition, the former patients undergo optimal regimens under close monitoring. These ideal experimental conditions, essential to establish causality, do not correspond to real-life patients and practices in HF. HF patients are usually older than trial patients, and they tend to have several comorbidities [13]. The syndrome is more severe, on average, in real-life HF patients than in trial HF patients [14]. In addition, real-life HF patients are less compliant with drug prescriptions and dietary habits than trial HF patients [15].

We hypothesised that recommended HF drugs are efficient in a real-world settings (i.e., in a population-based sample of HFREF patients). We used data from the Epidemiology and

Prognostic in Acute Heart Failure in Lorraine (EPICAL2) cohort to assess the effectiveness of recommended HF drug prescriptions at hospital discharge on 1-year mortality in HFREF patients in a real-world setting.

METHODS

Setting, design and sampling

The EPICAL2 study is an observational, prospective, population-based, and multicentre cohort study involving 21 volunteer hospitals spread over the Lorraine region of Northeast France (population of 2,350,000, according to the 2012 census). The cohort enrolled 2,254 consecutive adult HF patients hospitalised between October 2011 and October 2012 in cardiology intensive care units, cardiology departments or emergency departments at the hospitals. Patients living in Lorraine and hospitalised for acute heart failure (AHF) were included, as were those in whom AHF developed during hospitalisation. Eligible patients were identified by physicians from the participating departments or by trained clinical research assistants who regularly visited the departments. Included patients were then followed up for 3 years after discharge from the index hospitalisation or until death. The objectives of this cohort study were to 1) describe morbidity and mortality in the short term (0 to 6 months) and midterm (up to 3 years) and identify the main prognostic factors and 2) assess the effectiveness of various aspects of care, in or out of hospital. For the present investigation, we focused on the 632 HFREF patients defined by left ventricular ejection fraction (LVEF) <40% at admission who were alive at discharge (Figure 1).

Data collection

Socio-demographic, medical history and clinical characteristics were collected at hospital admission. Therapeutic characteristics were collected at admission, at hospital discharge and 6 months after discharge. Except for therapeutic characteristics collected by patients' general

practitioners by standardized interviews 6 months after discharge, all data were collected from medical records by using a standardised form. Unless otherwise specified, all variables were collected and treated as categorical variables.

Socio-demographic characteristics

Socio-demographic data collected were sex, age, area of residence, type of residence (living in a retirement or nursing home or not), and body mass index (BMI). Age and BMI were collected as continuous variables and then classified in 3 categories (age: ≤65, 66-80, and >80 years, and BMI: underweight or normal [<25 kg/m²], overweight [25-30 kg/m²], and obese [>30 kg/m²]).

Medical history

We collected ischemic factors precipitating the actual HF decompensation, defined as coronary syndrome with or without ST elevation identified as precipitating HF by a cardiologist; cardiovascular risk factors such as dyslipidaemia, hypertension, smoking, alcohol abuse, and family history of cardiovascular disease; previous HF hospitalisation(s), acute coronary syndrome with or without ST elevation, stroke or transient ischemic attack, peripheral arterial disease, or other cardiovascular conditions (valvular heart disease, pulmonary embolism and arrhythmia); previous cardiovascular interventions such cardiac resynchronisation therapy, and cardiac stimulation; and comorbidities such as diabetes mellitus, asthma or chronic obstructive pulmonary disease, severe chronic respiratory insufficiency, chronic kidney disease, depressive disorder, haematological malignancy, cancer, cirrhosis, peptic ulcer, and AIDS. To summarise some of the comorbidities, the Charlson index was calculated by using age, history of HF hospitalisation(s), and all the aforementioned comorbidities except severe chronic respiratory insufficiency [16]. The Charlson index was classified into 3 categories (≤ 5; 6-8; ≥ 9).

Clinical characteristics

We collected clinical characteristics such acute pulmonary oedema, peripheral oedema, orthopnoea, hepatojugular reflux, high blood pressure (HBP) (defined as systolic BP ≥140 mmHg or diastolic BP ≥90 mmHg), wide QRS complex (>0.12 s), left bundle branch block, and LVEF. LVEF was collected as a continuous variable, and then classified into 3 categories (≤20%, 21-29%, and ≥30%).

Therapeutic characteristics

Therapeutic characteristics consisted of the main HF medications including ACE inhibitors, ARBs, β-blockers, mineralocorticoid receptor antagonists (MRAs), and diuretics.

Prescription adherence to recommendations at discharge

From the 2008 ESC guidelines [10], the recommendations available at the time of the EPICAL2 study, patients with prescription of both an ACE inhibitor (or an ARB) and a β-blocker were considered under guideline-consistent discharge prescription (GCDP) for HFREF. Patients with prescription of an ACE inhibitor (or an ARB) alone or a β-blocker alone or none of these medications were considered under guideline-inconsistent discharge prescription (GIDP).

Outcome

The outcome studied was all-cause mortality during the year after hospital discharge. The 1-year vital status for each patient and date of death, if appropriate, were collected from registries. Survival time was calculated from the date of hospital discharge. Surviving patients were censored at 1-year follow-up.

Statistical analysis

First, age, sex and LVEF at admission and main HF prescriptions at discharge were described overall by numbers and percentages. Socio-demographic, medical history, and clinical characteristics at admission were described in the GCDP and GIDP groups by proportions. Characteristics at admission for the 2 groups were compared by calculating absolute standardised differences (|StDiffs|), which indicate the degree of systematic differences in covariates between groups. Empirically, a |StDiff| <10% indicates a negligible difference in proportion of covariates between groups [17]. Finally, the two groups were compared for prescription adherence to recommendations at hospital admission and 6 months after discharge.

Second, the number of deaths during follow-up was described overall and for the 2 groups. Survival during the first year after hospital discharge was described by the Kaplan-Meier method and compared by a Logrank test.

Third, to assess the effect of guideline adherence on all-cause mortality, a propensity score (PS) analysis was used to account for imbalances between the 2 groups in characteristics at admission [18]. PS, representing the likelihood of GCDP conditional on socio-demographic, medical history and clinical characteristics of patients at inclusion, were calculated by using a logistic regression model based on a dependant variable (GCDP vs. GIDP) and independent variables. Independent variables considered in the model were 1) characteristics empirically imbalanced between groups (|StDiff| ≥10%) and 2) additional characteristics previously described in the literature as prognostic factors (i.e., sex, history of acute coronary syndrome with or without ST elevation, history of cardiac resynchronisation therapy or cardiac stimulation, and wide QRS complex). Resulting PS distributions were then graphically plotted for the GCDP and GIDP groups to assess the magnitude of their overlap [19]. A 1:1 PS

matching without replacement was used, with a calliper of 0.02 [20]. To assess the quality of PS matching, the overlap of PS distribution was graphically plotted for the 2 matched groups. After matching, balance of patient characteristics at admission was assessed by estimating |StDiff| between the 2 groups. Finally, shared frailty Cox regression models [21] were used in paired samples to assess the effectiveness of guideline adherence at hospital discharge on all-cause mortality. We adjusted for post-matching residual unbalanced characteristics at admission (post-match |StDiff|≥10%). Proportional-hazard assumption was assessed graphically by plotting the Log(-Log(survival) vs Log(time)). Data are reported as hazard ratios (HRs) and 95% confidence intervals (95% CIs). A p-value of $P < 0.05$ was considered statistically significant.

Variables with 20% to 30% missing data (i.e., BMI, previous hospitalisation for HF, low systolic BP and wide QRS complex) were handled by multiple imputation methods [22], which resulted in 5 imputed datasets. All steps of the analyses described above were used for analysis of the 5 imputed datasets, and then results were pooled.

An additional logistic regression model, including guideline adherence (GCDP vs. GIDP) as the dependent variable and all characteristics with |StDiff| ≥10% as the independent variables, was used with a backward selection procedure to identify the characteristics associated with guideline adherence.

All analyses involved use of SAS© 9.3 (SAS Inst., Inc, Cary, NC, USA).

RESULTS

Socio-demographic, medical history, clinical characteristics and main HF medications at discharge

Among the 632 patients included in the present analysis, the mean (SD) age was 73.8 (12.9) years and 411 (65.0%) were men. The mean LVEF was 28.2±6.2%. At discharge, a total of 427 (67.6%) patients were prescribed an ACE inhibitor, 79 (12.5%) an ARB, and 501 (79.3%) a β-blocker, for 414 (65.5%) patients in the GCDP group and 218 (34.5%) in the GIDP group. In the GIDP group, 72 (33.0%) patients were prescribed an ACE inhibitor alone, 18 (8.3%) an ARB alone, 87 (39.9%) a β-blocker alone, and 41 (18.8%) none of these medications. In addition, 167 (26.4%) patients were prescribed MRAs: 136 (32.9%) in the GCDP group and 31 (14.2%) in the GIDP group (p<0.001). A total of 555 (87.8%) patients were prescribed diuretics: 356 (86.0%) in the GCDP group and 199 (91.3%) in the GIDP group (p=0.053).

Characteristics at admission in the 2 groups along with their |StDiff| values are in Table 1. As compared with GIDP patients, GCDP patients were younger (|StDiff|=48.91%), less frequently lived in a retirement or nursing home (|StDiff|=17.18%), and were more often obese (|StDiff|=12.54%). Moreover, they more often had an ischemic factor precipitating HF decompensation (|StDiff|=14.22%) and dyslipidaemia (|StDiff|=20.18%), were smokers (|StDiff|=18.44%), abused alcohol (|StDiff|=14.63%) and had a family history of cardiovascular disease (|StDiff|=34.07%), acute pulmonary oedema (|StDiff|=12.86%), orthopnoea (|StDiff|=13.00%), hepatojugular reflux (|StDiff|=16.86%), HBP (|StDiff|=14.85%), and left bundle branch block (|StDiff|=11.11%).

As compared with GIDP patients, GCDP patients less often had a history of HF hospitalisation (|StDiff|=14.20%), stroke or transient ischemic attack (|StDiff|=15.62%), other cardiovascular disease (|StDiff|=34.06%), and severe chronic respiratory insufficiency

(|StDiff|=29.36%). Finally, they had a lower Charlson Index (|StDiff|=55.78%) and presented lower LVEF (|StDiff|=21.96%).

Guideline adherence at different times (Figure 2) and history of HF

Among the 632 patients included in the present investigation, 63 (10.0%) had missing therapeutic data at admission, and 305 (44.1%) had missing therapeutic data at 6 months. A total of 239 (42.0%) had guideline-consistent prescription at admission and 216 (61.2%) had guideline-consistent prescription at 6 months.

Among the 414 patients (65.5%) with GCDP, 253 (61.1%) presented previous HF hospitalisation(s) at admission, 200 (55.1%) already had guideline-consistent prescription at admission, and 193 (81.8%) still had guideline-consistent prescription at 6-month follow-up.

Among the 218 patients (34.5%) with GIDP, 148 (67.9%) presented previous HF hospitalisation(s) at admission, 39 (18.9%) had guideline-consistent prescription at admission, and 23 (19.7%) switched to guideline consistent prescription at 6-month follow-up.

Mortality

A total of 165 (26.1%) patients died during follow-up, 78 (18.8%) in the GCDP group and 87 (39.9%) in the GIDP group. No patient was lost to follow-up. Figure 3 shows the Kaplan Meier survival curves by guideline adherence. Survival was impaired in the GIDP group as compared to the GCDP group ($p<0.001$).

Effect of prescription adherence on mortality

PS matching resulted in 171 pairs of patients in the first dataset, 170 in the second, 173 in the third, 169 in the fourth, and 175 in the fifth. PS overlap before and after matching is presented in the appendix (Figures S1-S2). After PS matching, the pooled |StDiffs| was between 0.03%

and 11.68%. Slight residual imbalances were observed for BMI (|StDiff|=11.68%) and history of hypertension (|StDiff|=11.07%).

After PS matching, Kaplan Meier curves still showed better survival for GCDP than GIDP patients (Appendix, Figure S3), and guideline adherence (GCDP vs GIDP) was associated with reduced mortality rate (pooled HR= 0.53, 95% CI [0.35-0.81], p=0.003). Additional adjustment for BMI and history of hypertension did not modify this result (pooled HR= 0.54, 95% CI [0.35-0.83], p=0.005) (Table 2).

As compared with GIDP, GCDP (i.e., better guideline adherence) was associated with younger age, dyslipidaemia, family history of cardiovascular disease, absence of severe chronic insufficiency, lower Charlson index, acute pulmonary oedema, and hepatojugular reflux after adjustment for potential confounders (Appendix, Table S4).

DISCUSSION

Prescription of ACE inhibitors (or ARB) and β-blockers, as recommended by the ESC in HFREF, was suboptimal (65.5%) among HFREF patients in the EPICAL2 cohort despite the evidence for morbidity and mortality improvement with these medications [23, 24]. However, our investigation suggests that guideline-consistent prescription in HFREF remained effective in reducing 1-year mortality in an unselected community-based sample of HFREF patients.

Suboptimal rates of adherence to ESC guidelines concerning HF medications

The moderate adherence to recommendations for ACE inhibitors and β-blockers observed in EPICAL2 (65.5% of patients were in the GCDP group) was similar to that observed in Germany, with 60% optimal prescription [24]; in Sweden, with 42% optimal prescription [25]; and in France, with 65% of patients with ACE inhibitors and 30% with β-blockers (i.e., close to the European prescription rates [62% and 37%, respectively] at the time of the Euro Heart Survey) [26]. Interestingly, we observed an even lower rate of optimal prescription at admission (42% of HFREF patients with both an ACE inhibitor and a β-blocker), which was partly explained by the recruitment of incident cases but also suggested improved optimal prescription rates during a hospital stay in prevalent cases. Maggioni et al. also noted improved HF medication prescription considered separately between hospital admission and hospital discharge, especially for ACE-inhibitors and β-blockers, with prescription rates improving from 60% to 80% for each medication, without details on the evolution of rates of prescription for their combination [27]. We observed similar prescription rates for ACE-inhibitors (67.6%) and β-blockers (79.3%) at discharge. Accordingly, the low rates for optimal prescription we observed could be explained by need to consider prescription of both medications (i.e. ACE-inhibitors [or ARBs] and β-blockers) as adherent to the guidelines. In addition, part of the low optimal prescription rates at admission might be explained by

incident cases of HFREF enrolment. Indeed, only 61% (68%) of the patients in the GCDP (GIDP) group presented a history of HF hospitalisation. However, because the rates of guideline-consistent prescriptions at admission in both groups (55% and 19%, respectively) were lower than the rates of history of HF hospitalisation (i.e., prevalent cases), de novo HF cannot be the only explanation for low guideline-consistent prescription rates at admission.

The guideline-consistent prescription rate decreased only slightly from 65.5% at discharge to 61.2% at 6-month follow-up, whereas HF medication adherence to ACE inhibitors decreased from 70% to 53% over the first year after discharge in US Medicare patients [28]. Our findings could be explained by a shorter follow-up for HF medication at the time of our investigation. However, these results suggest that the decrease in guideline-consistent prescription rates might occur late after discharge. Accordingly, to improve the observed mortality decrease over the last decades for HFREF patients [2], guideline-consistent prescription of ACE inhibitors and β-blockers might need to be improved during hospital stay and also late after discharge. Focus should be on older HFREF patients and those presenting dyslipidaemia and high Charlson index, who actually are at high risk of guideline-inconsistent prescription when hospitalised for HF, as we observed. In fact, comorbidities confer significantly increased mortality risk [29]. Although medication adherence to β-blockers has improved during the past decade in older HF patients in the United States [30], efforts are still required to better understand barriers to medication adherence and should focus on interventions improving evidence-based medication prescriptions in HF patients.

Barriers to guideline implementation are common across Europe [31]. A previous work found that co-morbidities affect how recommended drugs are prescribed such that physicians must take into account competing therapeutic requirements and drug interactions [32].

Another explanation for barriers to guideline adherence could be contraindications to ACE inhibitors (or ARBs) (i.e., hypersensitivity, hyperkalaemia, hypotension) or to β-blockers (i.e., anaphylactic reactions, bradycardia, cardiogenic shock [Appendix, Table S5]) but unlikely explain the proportion of guideline-inconsistent prescriptions we observed. Actually, these contraindications are relatively rare, as reported in from the IMPROVE-HF study [33] in which less than 10% of patients had contraindications for (or intolerance to) these drugs. Moreover, the presence of contraindications such as chronic conditions justifying non-adherence to recommended therapy is not consistent with the observed switch of patients from the GIDP group to guideline-consistent prescription at 6-month follow-up (19.7%). Accordingly, we urgently need to identify barriers that prevent physicians from prescribing recommended medications for HFREF patients.

As previously mentioned, among the 218 patients in the GIDP group, 19.7% switched to guideline-consistent prescription at 6 months. An improvement of renal function likely explained such a switch because acute renal failure was frequently observed in patients with GIDP (data not shown). Unfortunately, follow-up data on glomerular filtration rate or creatinine levels were not collected at discharge, so we are unable to confirm this hypothesis.

Among the 414 patients with GCDP, 81.8% still had guideline-consistent prescription at 6 months. The reason for the switch to guideline-inconsistent prescription (18.2%) (i.e., whether side effects could be responsible) remains for investigation. Another plausible explanation for the switch could be poor transmission of prescriptions between hospital staff and general practitioners. Maintaining guideline adherence far beyond hospital discharge might be achieved by a shared information system like that used in some disease management programs involving health professionals specialised in HF management that were found efficient for HF readmissions [34].

Effect of recommended HF medications on prognostic in a real-world setting

By using methods that mimic experimental conditions [18] (i.e., PS matching), in the sense that they lead to balanced baseline characteristics between the groups compared, we found that prescription of an ACE inhibitor (or an ARB) and a β-blocker was associated with improved 1-year survival in an unselected sample of real-life HFREF patients. Our results are consistent with improved prognostic in unselected HFREF patients exposed to recommended medications reported in Austria [23] and Germany [24]. The main limitation reported by the authors of these studies was residual confounding due to lack of available clinical information in the administrative databases used for investigation. To overcome this limitation, we thoroughly compared socio-demographic and clinical baseline characteristics to identify potential indication bias. Then, we controlled the identified bias by using a PS based on the unbalanced characteristics and additional characteristics of prognostic value. Even if PS matching eliminated a significant number of patients, its use moved experimental conditions closer and minimised confounding bias [18].

In addition, the consecutive recruitment in EPICAL2 minimised selection bias and increased external validity. In fact, in a community-based sample of HF patients, the mean (SD) age was 73 (12) years for men and 79 (11) years for women [35]. In the present investigation, we reported a mean (SD) age of 74 (13) years between these limits that supports the representativeness of our sample at a community level. HFREF patients in EPICAL2 were older than those included in trials, for whom mean age varied from 62 to 71 years [7-8-9], which suggests a benefit of ACE inhibitors and β-blockers on survival in older patients.

Limitations

First, we did not include MRAs in our GCDP scheme, despite evidence of their efficacy in HFREF patients [36]. ESC guidelines recommend considering MRAs for patients with

residual uncontrolled symptoms (i.e., New York Heart Association [NYHA] class III to IV) after prescription of ACE inhibitors and β-blockers [10-11]. Unfortunately, NYHA data were not collected after admission during the hospital stay, so we were not able to identify patients with uncontrolled symptoms who needed MRAs in addition to ACE inhibitors and β-blockers. However, GCDP patients were more often prescribed MRAs than GIDP patients, which suggests that once the guidelines are applied for 2 medications, optimal prescriptions might extend to other recommended medications.

Second, rates of missing data for HF medication prescription at 6-month follow-up should temper the interpretation of a low decrease of guideline-consistent prescription rate over time, even if comparison of baseline characteristics of patients with and without 6-month missing data did not highlight differences supporting a major selection bias (data not shown).

Third, we considered medication prescriptions as a measure of exposure to ACE inhibitors (or ARBs) and β-blockers, which raises concerns about the actual patient medication compliance. However, a study of HF patients' compliance to medications reported that 72% of patients adhere to their prescribed medications [32]. In addition, despite the likely non-differential misclassification bias induced by the use of prescriptions as a marker of exposure to medication, which should have biased the HR towards the null hypothesis, GCDP remained associated with reduced mortality.

CONCLUSIONS

In a real-life setting of HFREF patients, prescription of ACE inhibitors and β blockers at hospital discharge was associated with halved mortality rate during the first year. Attention should be paid to HFREF patients with guideline-inconsistent prescription of ACE inhibitors (or ARBs) and/or β-blockers at discharge, especially older patients and those presenting

dyslipidaemia and comorbidities. Further research is needed to investigate the reason for under prescription among HF patients during a hospital stay and after discharge.

ACKNOWLEGEMENTS

The authors thank all patients and physicians from the 21 centres participating in EPICAL2 (Hôpitaux de Brabois et Hôpital Central, CHU Nancy; CH Lunéville; Espace Chirurgical Ambroise Paré Nancy; CH Alpha Santé Mont-Saint-Martin; CH Pont-à-Mousson; CH Saint-Nicolas Verdun; Hôpital Bon-Secours CHR Metz; CH Freyming Merlebach: Hôpital Sainte-Blandine Metz; Hôpital Bel Air CHR Thionville; CH Marie-Madeleine Forbach; Hôpital Alpha Santé Hayange; CH Saint-Nicolas Sarrebourg; Hôpital Lemire Saint-Avold; Hôpital des Armées Legouest Metz; Clinique Claude Bernard Metz; CH Saint-Charles Saint-Dié; CH Jean Monnet Epinal; CH Neufchateau; CH Vittel).

FUNDING

The EPICAL2 study was funded as part of the 2009 national *Programme Hospitalier de la Recherche Clinique* (PHRC) from the French *Direction Générale de l'Offre de Soins* (DGOS).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

1. Saudubraya T, Saudubray C, Viboud, et al. Prevalence and management of heart failure in France: national study among general practitioners of the Sentinelles network. *La revue de médecine interne* 2005;(26):845–850.
2. Gabet A, Lamarche-Vadel A, Chin F, et al. Mortalité due à l'insuffisance cardiaque en France, évolutions 2000-2010. *Bull EpidémiolHebd* 2014;(21-22):386-94.
3. Rey G, Lamarche-Vadel A, Jougla E. Comment mesure-t-on les causes de décès en France?. *Questions de santé publique* 2013:(21).
4. DREES. L'état de santé de la population en France : Suivi des objectifs annexés à la loi de santé publique. Rapport 2011 :296-2999. http://www.sante-jeunesse-sports.gouv.fr/maladies-cardiovasculaires.html
5. Merlierre J, Couvreux C, Smadja L. Caractéristiques et trajets de soins des insuffisants cardiaques du Régime Général. *Point de repère* 2007:(38).
6. The Public Health Policy Act of 9 August 2004, WA [statute on the Internet]. C2016 [cited 2016 may 15]. Available from: https://www.legifrance.gouv.fr/
7. The CONSENSUS Trial study group. Effects of enalapril on mortality in severe congestive heart failure: results of the Cooperative North Scandinavian Enalapril Survival Study (consensus). *N Engl J Med* 1987; 316:1429-35.
8. Packer M. Fowler MB, Roecker EB, and al. Effects of carvedilol on the morbidity of patients with severe chronic heart failure: resuls of the carvedilol prospective randomized cumulative survival (COPERNICUS) study. *Circulation* 2002;106:2194-2199.
9. Maggioni AP, Anand I, Gottlieb SO, et al. Effects of valsartan on morbidity and mortality in patients with heart failure not receiving angiotensin-converting enzyme inhibitors. *J Am Coll Cardiol* 2002 Oct 16;40(8):1414-21.
10. Dickstein K, Cohen-Solal A, Filippatos G, et al. ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure 2008: the Task Force for the Diagnosis and Treatment of Acute and Chronic Heart Failure 2008 of the European Society of Cardiology. Developed in collaboration with the Heart Failure Association of the ESC (HFA) and endorsed by the European Society of Intensive Care Medicine (ESICM). *Eur Heart J* 2008;29(19):2388-442.
11. McMurray JJ, Adamopoulos S, Anker SD, et al. ESC guidelines for the diagnosis and treatment of acute and chronic heart failure 2012: The Task Force for the Diagnosis and Treatment of Acute and Chronic Heart Failure 2012 of the European Society of Cardiology. Developed in collaboration with the Heart Failure Association of the ESC (HFA) and endorsed by the European Society of Intensive Care Medicine (ESICM). *EurHeart J* 2012 Aug;14(8):803-69.
12. Komajda M, Forette F, JPetit JF, et al. Recommendations for the diagnosis and management of cardiac failure in the elderly subject. *Arch Mal Cœur Vaiss* 2004:97:803-822.
13. Braunstein JB, Anderson GF, Gerstenblith G, et al. Noncardiac comorbidity increases preventable hospitalizations and mortality among medicare beneficiaries with chronic heart failure. *J Am CollCardiol* 2003;42:1226–33.
14. Lang CC, Mancini DM. Non-cardiac comorbidities in chronic heart failure. Heart 2007;93:665-671.
15. Murray MD, Young J, Hoke S, et al. Pharmacist Intervention to Improve Medication Adherence in Heart Failure: A Randomized Trial. *Ann Intern Med* 2007;146(10):714-725.
16. Charlson ME, Pompei P, Ales KL, et al. A New Method of Classifying Prognostic Comorbidity in Longitudinal Studies: Development and Validation. *Journal of Chronic Diseases* 1987:40(5):373-383.

17. 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. *Statist. Med* 2007 Feb 20;26(4):734-53.

18. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research* 2011; 46(3):399-424.

19. Baser O. Propensity Score Matching with Limited Overlap. Economics Bulletin 2007; 9(8):1-8.

20. 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 Feb;51(1):171-84.

21. Gutierrez RG. Parametric frailty and shared frailty survival models. *The Stata J* 2002;2(1);22–44.

22. Rubin, D. B. Multiple Imputation for Nonresponse in Surveys, Wiley, New York, 1987.

23. Marzluf BA, Reichardt B, Neuhofer LM, et al. Influence of drug adherence and medical care on heart failure outcome in the primary care setting in Austria. *Pharmacoepidemiol Drug Saf* 2015;24: 722–730.

24. Neubauer S, Schilling T, Zeidler J, and al. Impact of guideline adherence on mortality in treatment of left heart failure. *Herz* 2016 Feb 16.

25. Dahlstrom U, Hakansson J, Swedberg K, et al. Adequacy of diagnosis and treatment of chronic heart failure in primary health care in Sweden. *Eur J Heart Fail* 2009 Jan;11(1):92-8.

26. Komajda M, Follath F, Swedberg K, and al. The EuroHeart Failure Survey programme--a survey on the quality of care among patients with heart failure in Europe. Part 2: treatment. *Eur Heart J* 2003 Mar;24(5):464-74.

27. Maggioni AP, Dahlström U, Filippatos G, et al. EURObservational Research Programme: the Heart Failure Pilot Survey (ESC-HF Pilot). *Eur J Heart Fail* 2010 Oct;12(10):1076-84.

28. Sueta CA, Rodgers JE, Chang PP, et al. Medication Adherence Based on Part D Claims for Patients With Heart Failure After Hospitalization (from the Atherosclerosis Risk in Communities Study). *Am J Cardiol* 2015 Aug 1;116(3):413-9.

29. Ahluwalia SC, Gross CP, Chaudhry SI. Impact of comorbidity on mortality among older persons with advanced heart failure. *J Gen Intern. Med* 2012;27:513-519.

30. Setoguchi S, Choudhry NK, Levin R, et al. Temporal trends in adherence to cardiovascular medications in elderly patients after hospitalization for heart failure. *Clin Pharmacol Ther* 2010 Oct;88(4):548-54.

31. Shoukat S, Gowani SA, Taqui AM, et al. Adherence to the European Society of Cardiology (ESC) guidelines for chronic heart failure – A national survey of the cardiologists in Pakistan. *BMC Cardiovascular Disorders* 2011,11:68.

32. Laperche T. Impact des comorbidités dans la prise en charge thérapeutique de l'insuffisance cardiaque. *La revue du praticien* 2010;60: 955-9.

33. Fonarow GC, Albert NM, Curtis AB, and al. Improving evidence-based care for heart failure in outpatient cardiology practices. Primary results of the registry to Improve the Use of Evidence-Based Heart Failure Therapies in the Outpatient Setting (IMPROVE HF). *Circulation* 2010; 122: 585-596.

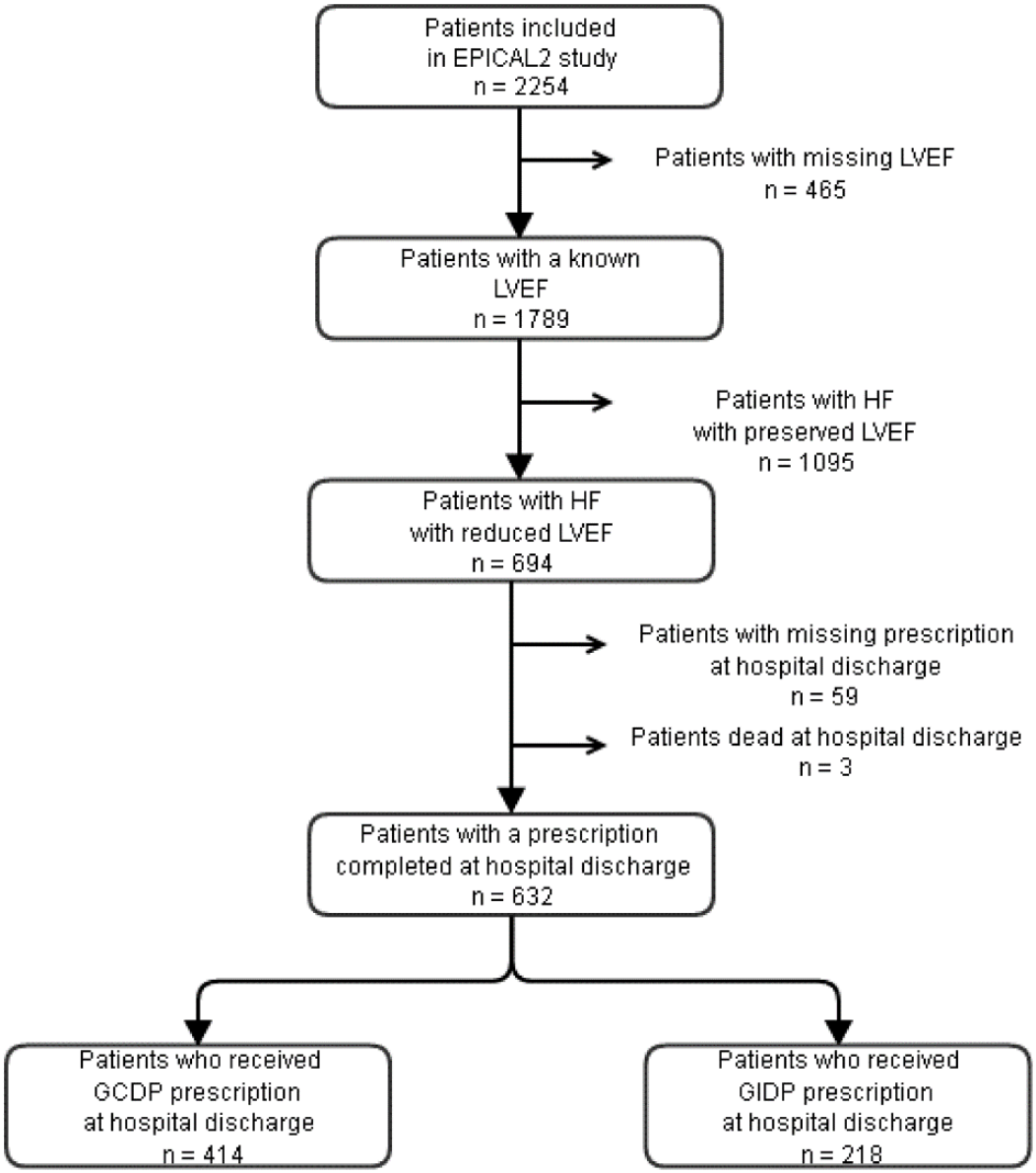
34. Agrinier N, Altieri C, Alla F, et al. Effectiveness of a multidimensional home nurse led heart failure disease management program--a French nationwide time-series comparison. *Int J Cardiol* 2013 Oct 9;168(4):3652-8.

35. Roger VL, Weston SA, Redfield MM, et al. Trends in Heart Failure incidence and survival in a community-based population. *JAMA* 2004;292(3):344-350.

36. Zannad F, McMurray J, Krum H, et al. Eplerenone in Patients with Systolic Heart Failure and Mild Symptoms. *N Engl J Med* 2011; 364:11-21.

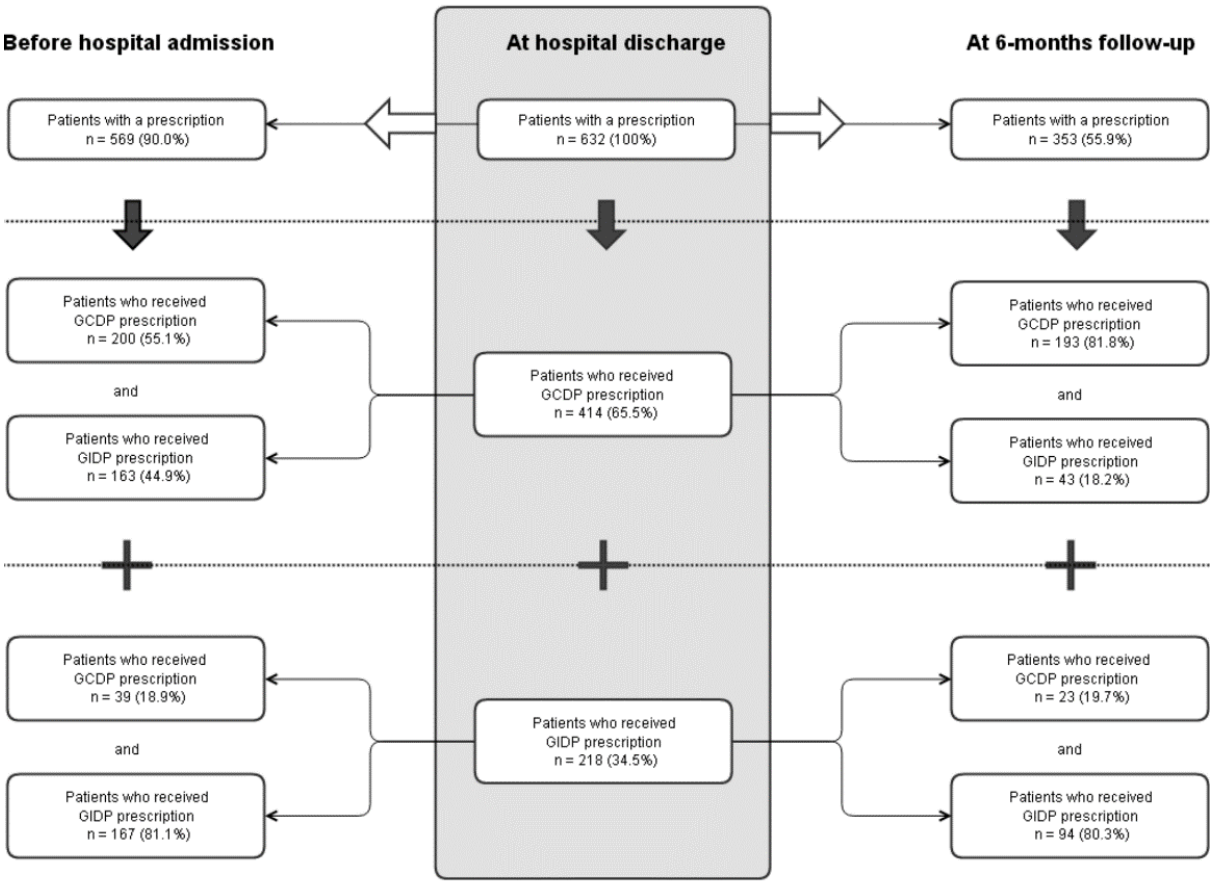
FIGURES AND TABLES

Figure 1: Flow of patients with heart failure and reduced ejection fraction from the EPICAL2 cohort for evaluating guideline adherence effectiveness



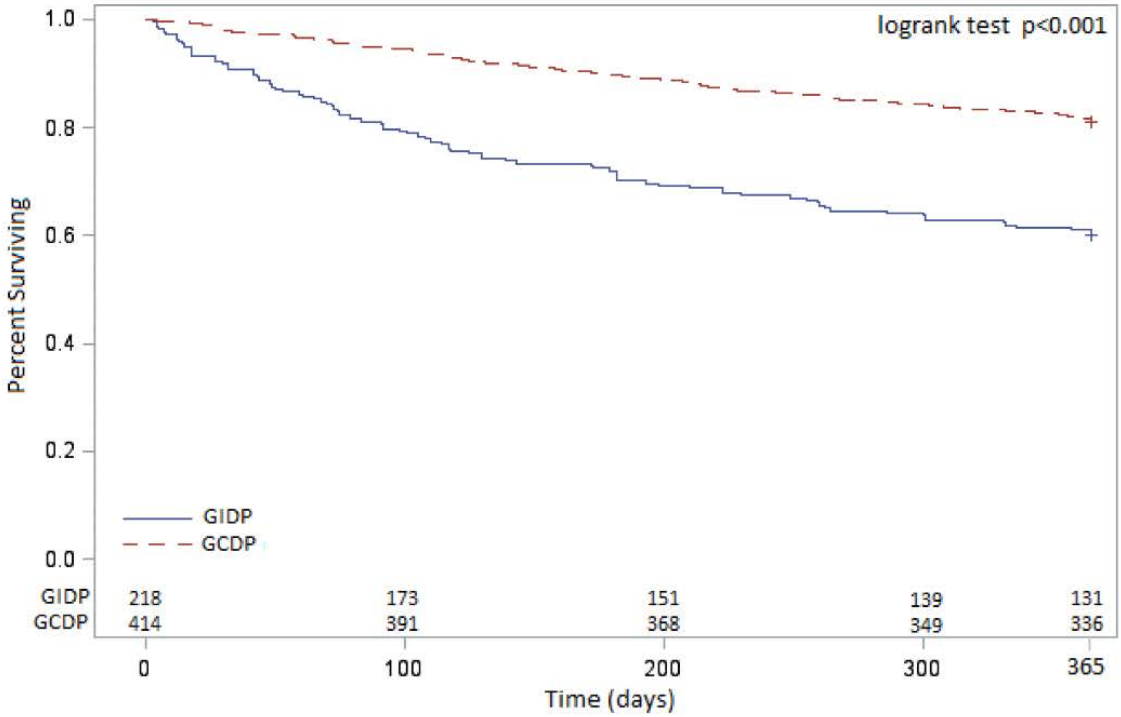
Notes: HF, heart failure; LVEF, left ventricular ejection fraction; GCDP, guideline-consistent discharge prescription; GIDP, guideline-inconsistent discharge prescription

Figure 2: Guideline adherence evolution from before hospital admission to 6-month follow-up



Notes: GCDP, guideline-consistent discharge prescription; GIDP, guideline-inconsistent discharge prescription,

Figure 3: Kaplan-Meier curves for 1-year mortality by guideline adherence



Notes: GCDP, guideline-consistent discharge prescription; GIDP, guideline-inconsistent discharge prescription.

Table 1: Socio-demographic, medical history, and clinical characteristics of patients with heart failure and reduced ejection fraction by guideline adherence before and after propensity score (PS) matching

	Before PS matching			After PS matching		
	GIDP %	**GCDP %**	**|StDiffs| %**	**GIDP %**	**GCDP %**	**|StDiffs| %**
Socio-demographic characteristics						
Males	61.9	66.7	9.91	59.8	60.9	2.37
Age, years			**48.91**			3.86
≤ 65	13.8	32.9		16.4	17.7	
66–80	39.4	36.5		40.0	38.6	
> 80	46.8	30.7		43.6	43.7	
Area of residence			**23.56**			6.59
Meurthe-et-Moselle	26.1	34.0		29.0	32.0	
Moselle	47.7	36.5		42.0	40.6	
Meuse or Vosges	26.1	29.5		29.0	27.4	
Living in a retirement or nursing home	10.2	5.6	**17.18**	9.4	10.4	3.15
BMI			**12.54**			**11.68**
Underweight or normal	39.9	35.7		40,3	35.0	
Overweight	34.9	33.6		33,3	37.9	

Obese	25.2	30.7		26,4	27.1	
Medical history						
Ischemic factor precipitating HF decompensation	11.9	16.9	**14.22**	13.8	15.0	3.62
Cardiovascular risk factors						
Dyslipidaemia	35.3	45.2	**20.18**	37.3	35.5	3.66
Hypertension	69.7	68.4	2.96	40.7	41.9	**11.07**
Smoking	39.9	49.0	**18.44**	9.9	9.8	2.56
Alcohol abuse	10.6	15.5	**14.63**	3.4	2.8	0.44
Family history of cardiovascular disease	2.8	11.4	**34.07**	69.0	74.0	3.35
Previous HF hospitalisation(s)	67.9	61.1	**14.20**	57.4	62.2	9.88
Acute coronary syndrome ST+ or ST-	49.1	49.5	0.87	48.1	47.0	2.34
Stroke / TIA	13.3	8.5	**15.62**	10.5	10.5	0.03
Peripheral arterial disease	15.6	12.8	8.01	12.5	12.4	0.33
Other cardiovascular disease[1]	62.4	45.7	**34.06**	57.9	56.0	3.92
Cardiac resynchronization therapy or cardiac stimulation	20.2	20.3	0.26	20.4	18.3	5.34
Severe chronic respiratory insufficiency	9.6	2.7	**29.36**	5.6	5.7	0.52
Charlson Index[2]			**55.78**			9.43
≤ 5	19.7	39.1		24.2	28.3	

6–8	29.4	34.3		33.6	31.2	
≥ 9	50.9	26.6		42.2	40.5	
Clinical characteristics at admission						
Acute pulmonary oedema	21.1	26.6	**12.86**	22.0	21.0	2.58
Peripheral oedema	61.9	58.0	7.51	59.5	61.8	4.75
Orthopnoea	19.3	24.4	**13.00**	20.5	18.5	5.00
Hepatojugular reflux	8.3	13.3	**16.86**	9.3	8.7	1.99
HBP[3]	17.9	23.9	**14.85**	19.1	19.5	0.86
Wide QRS complex[4]	19.7	19.1	1.63	18.6	19.7	2.72
Left bundle branch block	25.2	30.2	**11.11**	26.1	24.7	3.22
LVEF (%)			**21.96**			6.87
≤ 20	17.9	20.8		19.8	20.4	
21–29	23.4	31.1		24.2	26.8	
≥ 30	58.7	48.1		56.0	52.8	

Notes: GCDP, guideline-consistent discharge prescription; GIDP, guideline-inconsistent discharge prescription; |Stdiffs|, absolute standardised differences; BMI, body mass index; HBP, high blood pressure; HF, heart failure; TIA, transient ischemic attack; LVEF, left ventricular ejection fraction. 1: Other cardiovascular disease considered were valvular heart disease, pulmonary embolism and arrhythmia; 2: The Charlson index calculation was based on age, diabetes mellitus, chronic kidney failure, depressive disorder, haematological malignancy, asthma or chronic obstructive pulmonary disease, cancer, cirrhosis, history of hospitalisation for acute heart failure, acute coronary syndrome ST+, peripheral arterial disease, stroke or TIA, peptic ulcer, and AIDS (Charlson ME, 1987); 3: systolic BP≥140 mmHg or diastolic BP≥90mmHg; 4: QRS > 12 s.

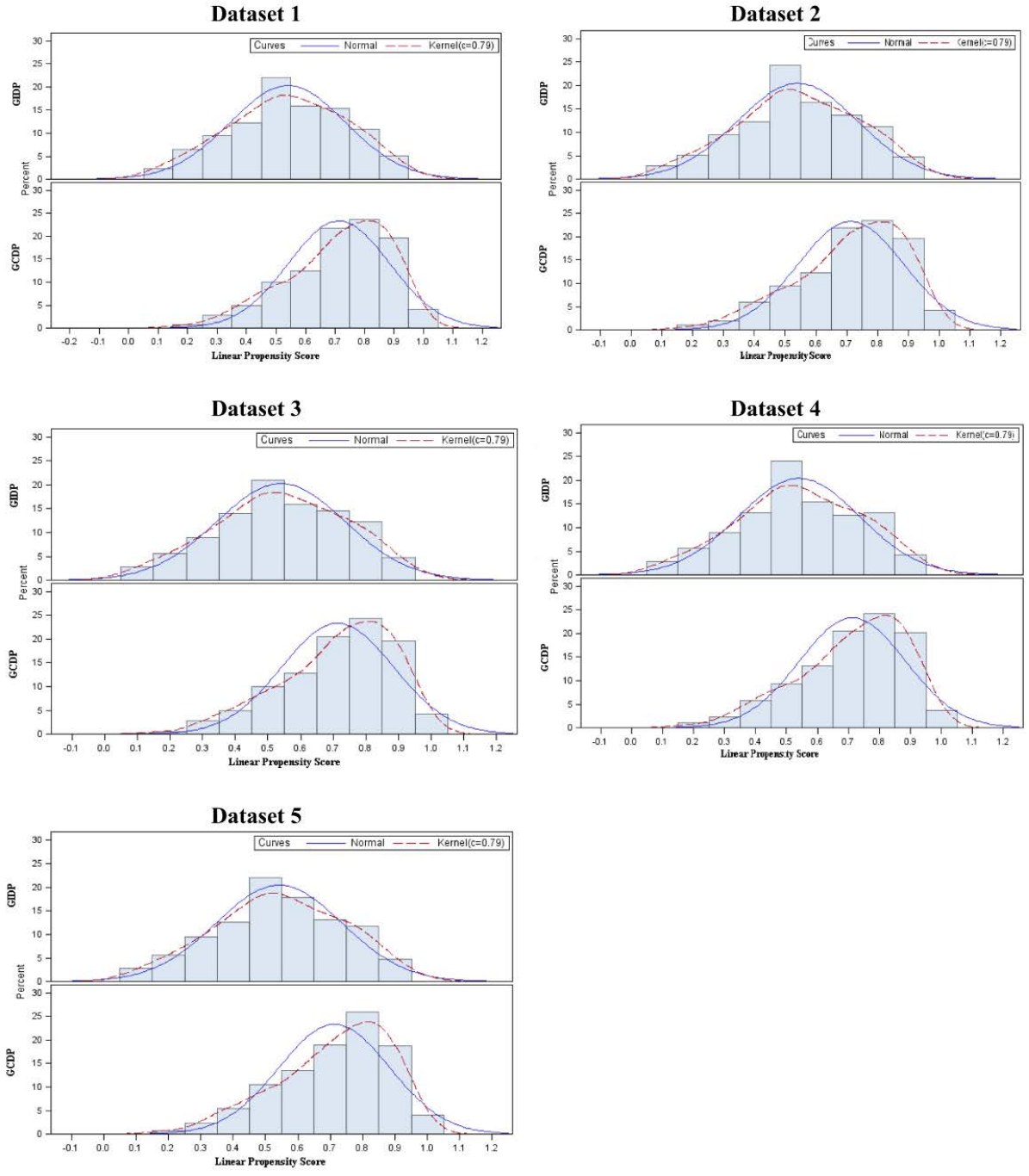
Table 2: Effect of guideline adherence on mortality in HFREF patients after PS matching and adjustment for residual unbalanced characteristics

	Dataset 1 (171 pairs)		Dataset 2 (170 pairs)		Dataset 3 (173 pairs)		Dataset 4 (169 pairs)		Dataset 5 (175 pairs)		Pool Estimates	
	HR [95% CI]	p-value	HR [95% CI]	p-value	HR [95% CI]	p-value	HR [95% CI]	p-value	HR [95%CI]	p-value	HR [95% CI]	p-value
Model 1												
GCDP vs. GIDP	0.57 [0.39;0.85]	p=0.005	0.56 [0.38;0.83]	p=0.004	0.52 [0.34;0.75]	p<0.001	0.54 [0.36;0.80]	p=0.002	0.49 [0.33;0.72]	p<0.001	0.53 [0.35;0.81]	p=0.003
Model 2												
GCDP vs. GIDP	0.59 [0.40;0.87]	p=0.008	0.57 [0.39;0.85]	p=0.006	0.52 [0.35;0.77]	p=0.001	0.54 [0.37;0.80]	p=0.002	0.49 [0.33;0.73]	p<0.001	0.54 [0.35;0.83]	p=0.005
BMI overweight vs. underweight or normal	0.65 [0.41;1.02]	p=0.059	0.57 [0.36;0.92]	p=0.021	0.54 [0.34;0.86]	p=0.010	0.61 [0.39;0.97]	p=0.038	0.490 [0.30;0.790]	p=0.003	0.57 [0.34;0.96]	p=0.038
BMI obese vs. underweight or normal	0.53 [0.32;0.89]	p=0.016	0.53 [0.32;0.88]	p=0.014	0.46 [0.27;0.78]	p=0.004	0.65 [0.39;1.08]	p=0.094	0.624 [0.38;1.03]	p=0.066	0.56 [0.31;1.00]	p=0.057
Hypertension	0.97 [0.64;1.49]	p=0.909	1.03 [0.66;1.61]	p=0.905	1.01 [0.65;1.58]	p=0.946	0.93 [0.60;1.45]	p=0.760	1.077 [0.68;1.69]	p=0.747	1.00 [0.64;1.59]	p=0.984

Notes: GCDP, guideline-consistent discharge prescription; GIDP, guideline-inconsistent discharge prescription; BMI, body mass index

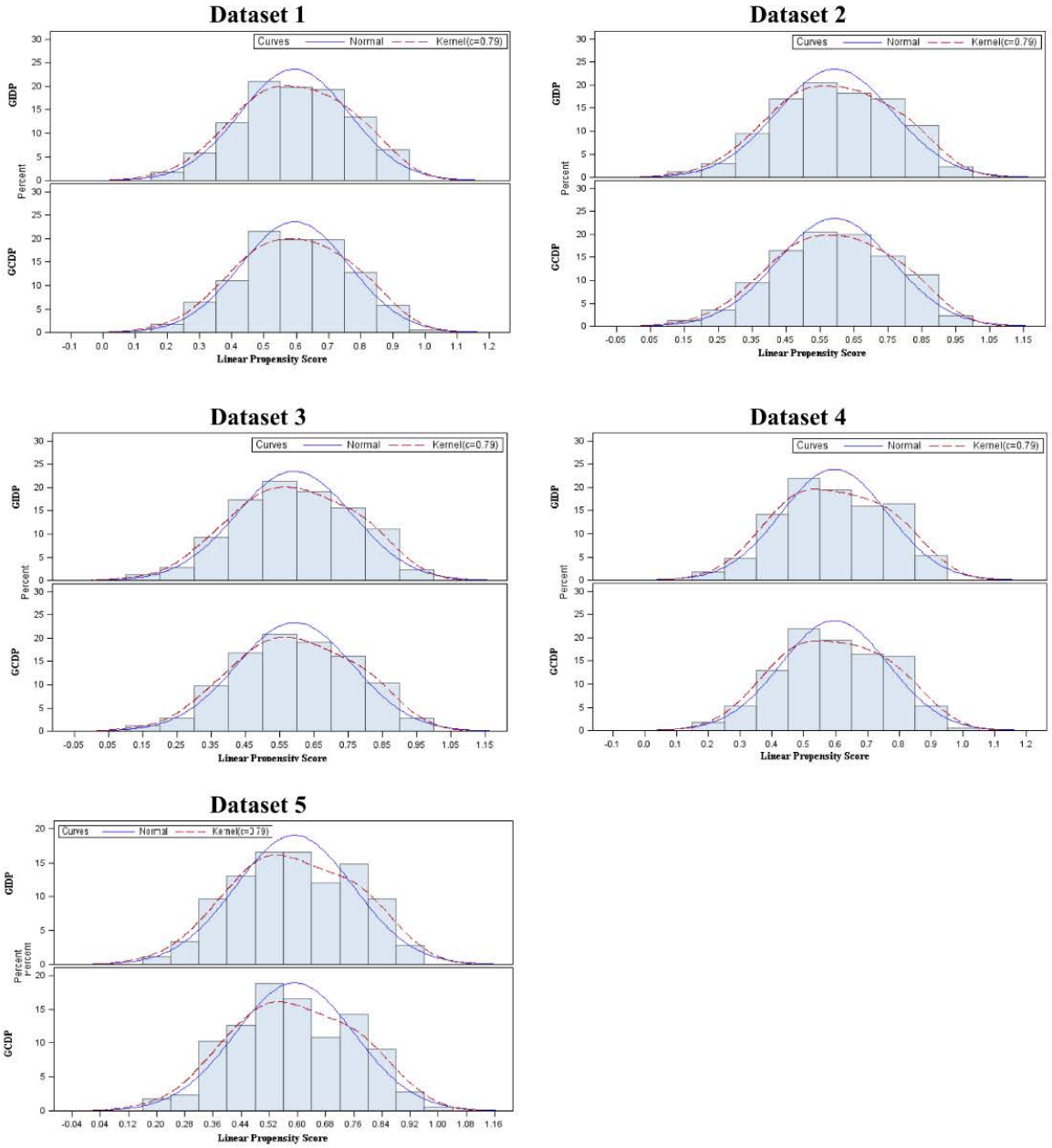
APPENDICES

Figure S1: Distribution of propensity score before matching by adherence to prescription recommendations in the 5 imputed datasets for HFREF patients



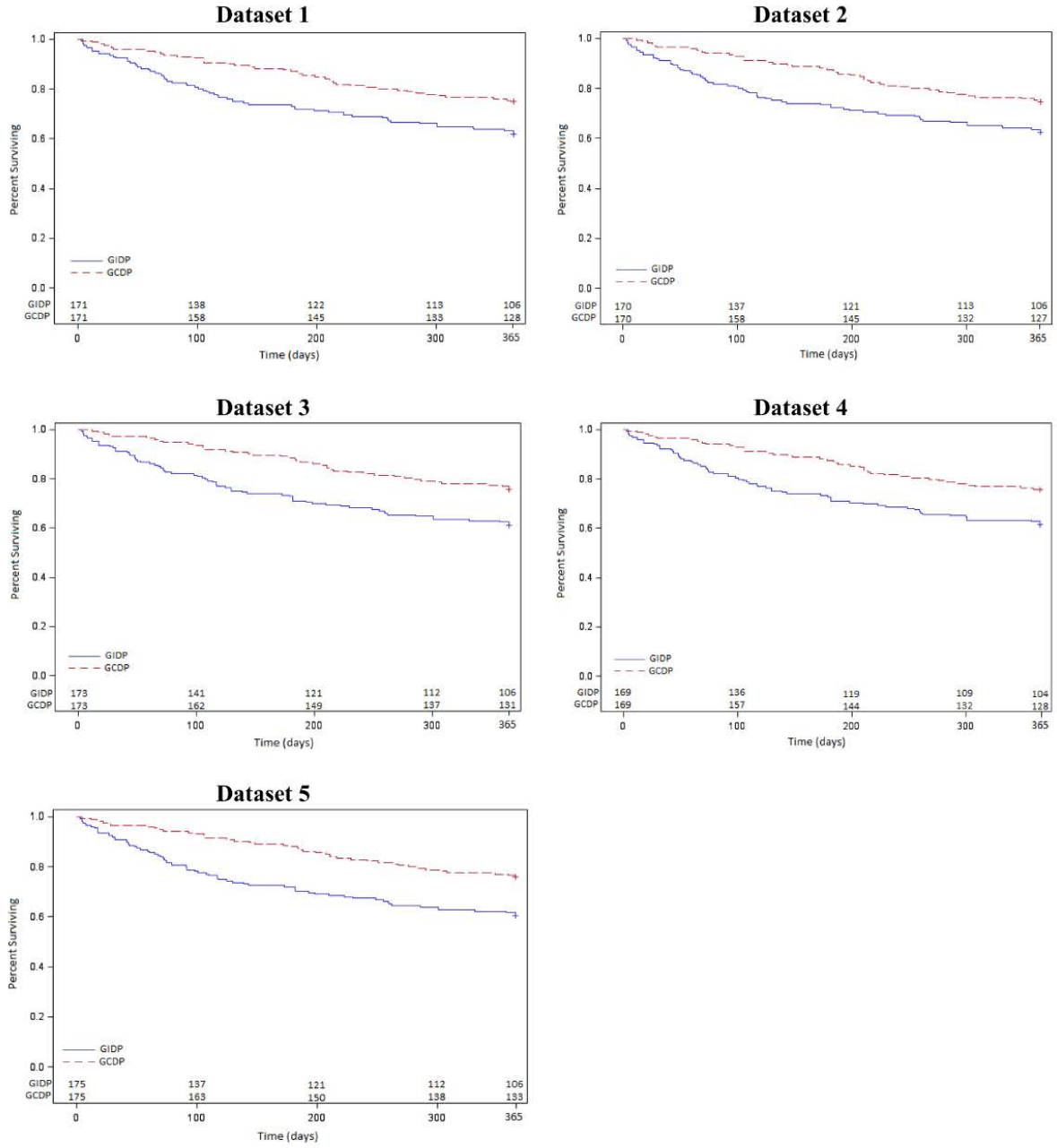
Notes: GCDP, guideline-consistent discharge prescription; GIDP, guideline-inconsistent discharge prescription.

Figure S2: Distribution of propensity score after matching by adherence to prescription recommendations in the 5 imputed datasets for HFREF patients



Notes: GCDP, guideline-consistent discharge prescription; GIDP, guideline-inconsistent discharge prescription.

Figure S3: Kaplan-Meier curves for 1-year mortality by guideline adherence after propensity-score matching



Notes: GCDP, guideline-consistent discharge prescription; GIDP, guideline-inconsistent discharge prescription.

Table S4: Characteristics associated with guideline-consistent discharge prescription

	Dataset 1 (n=632)		Dataset 2 (n=632)		Dataset 3 (n=632)		Dataset 4 (n=632)		Dataset 5 (n=632)	
	OR	**[95%CI]**	**OR**	**[95%CI]**	**OR**	**[95%CI]**	**OR**	**[95%CI]**	**OR**	**[95%CI]**
Age										
66–80 vs. ≤65	0.536	[0.312;0.921]	0.536	[0.312;0.921]	0.536	[0.312;0.921]	0.536	[0.312;0.921]	0.536	[0.312;0.921]
>80 vs. ≤65	0.497	[0.279;0.886]	0.497	[0.279;0.886]	0.497	[0.279;0.886]	0.497	[0.279;0.886]	0.497	[0.279;0.886]
Dyslipidaemia	1.806	[1.236;2.637]	1.806	[1.236;2.637]	1.806	[1.236;2.637]	1.806	[1.236;2.637]	1.806	[1.236;2.637]
Family history of cardiovascular disease	3.013	[1.212;7.491]	3.013	[1.212;7.491]	3.013	[1.212;7.491]	3.013	[1.212;7.491]	3.013	[1.212;7.491]
Severe chronic respiratory insufficiency	0.298	[0;131;0.679]	0.298	[0;131;0.679]	0.298	[0;131;0.679]	0.298	[0;131;0.679]	0.298	[0;131;0.679]
Charlson Index										
6-8 vs. ≤5	0.799	[0.479;1.382]	0.799	[0.479;1.382]	0.799	[0.479;1.382]	0.799	[0.479;1.382]	0.799	[0.479;1.382]
≥9 vs. ≤5	0.346	[0.205;0.584]	0.346	[0.205;0.584]	0.346	[0.205;0.584]	0.346	[0.205;0.584]	0.346	[0.205;0.584]
Acute pulmonary oedema	1.635	[1.062;2.518]	1.635	[1.062;2.518]	1.635	[1.062;2.518]	1.635	[1.062;2.518]	1.635	[1.062;2.518]
Hepatojugular reflux	2.236	[1.216;4.110]	2.236	[1.216;4.110]	2.236	[1.216;4.110]	2.236	[1.216;4.110]	2.236	[1.216;4.110]

Notes: odds ratios (ORs) are from multivariate logistic regression model with a backward selection using guideline adherence as the dependent variable and unbalanced characteristics as independent variables: sex, age, area of residence, living in a retirement or nursing home, BMI, ischemic factor precipitating HF, dyslipidaemia, smoking, alcohol abuse, family history of cardiovascular disease, severe chronic respiratory insufficiency, and the Charlson index.

Table S5: Contraindications to angiotensin-converting enzyme (ACE) inhibitors, angiotensin II receptor blockers (ARBs), and β-blockers in HF

ACE inhibitors [*1]	ARBs [*2]	β-blockers [*3]
Hypersensitivity to the active substance or to any of the excipients		
Hypersensitivity to any other ACE inhibitor	Hypersensitivity to sulfonamide derived substances or dihydropyridine Concomitant use of ACE inhibitor (ARB should be administered 36 h after discontinuation of the ACE inhibitor)	A history of anaphylactic reactions
Association with drug-containing aliskiren in patients with diabetes or renal impairment (GFR <60 ml/min/1.73m²)		
Association with potassium-sparing diuretics, potassium salts, estramustine, lithium Extracorporeal treatment involving contact of blood with negatively charged surfaces		Combination with cimetidine, the class I antiarrhythmic (except lidocaine), fluoxetine, paroxetine, to floctafenine, and sultopride
A history of angioedema (angioedema) associated with previous treatment with ACE inhibitor or ARB Hereditary or idiopathic angioedema		
	Cardiogenic shock	
	Obstruction of the outflow tract of the left ventricle (e.g., high-grade aortic stenosis) Hemodynamically unstable HF after myocardial infarction in acute phase	Acute HF, severe HF, or episode of decompensation with fluid overload of signs and/or requiring treatment with positive inotropic venous, vasodilator or intravenous (edema, ascites, pulmonary rattle stasis) Sinus heart disease including sino-atrial block Atrioventricular block of II and III degrees Variant angina Symptomatic bradycardia or severe (heart rate <60 beats/min before start of treatment) Peripheral arterial circulatory disorders or occlusive devices or in severe form Raynaud's phenomenon of severe forms Chronic to severe asthma Chronic obstructive pulmonary disease with severe or a history of bronchospasm Pheochromocytoma untreated
	Achievement / alteration / failure of acute liver function, chronic , or moderate to severe	
	Biliary cirrhosis Biliary obstruction	Clinical signs of liver dysfunction
Significant stenosis of bilateral renal	Severe renal impairment	

artery or a single functioning kidney	(creatinine clearance <30 ml/min)	Acidosis
Hyperkalaemia	Cholestasis Refractory hypokalaemia Hypercalcemia Hyponatremia Symptomatic hyperuricaemia	
Hypotension	Severe hypotension	Symptomatic or severe hypotension (systolic BP <85-100 mmHg)
Hemodynamic instability 1st-trimester pregnancy		
2nd- or 3rd-trimester pregnancy		
Breastfeeding		Breastfeeding
	Age <1 year	

Notes. *Source : https://www.vidal.fr/; 1. List from contraindications for Renitec, Lopril, Fositec, Acuitec, Triatec, Odrik, Justor, Zestril, Coversyl; 2. List from contraindications for Cozaar, Atacand, Kenzen, Aprovel, Coaprovel, Teveten, Alteis, Alteisduo, Axeler, Olmetec, Coolmetec, Sevikar, Micardis, Micardisplus, Twynsta, Entresto, Nisis, Tareg. 3. List from contraindications for Nebilox, Temerit, Cardensiel, Kredex, Selozok, Trandate.

B.2.2. Discussion

Comme le précédent, ce travail illustre le risque de biais d'indication dans les études observationnelles. En effet, les individus des deux groupes comparés avaient des caractéristiques initiales différentes, caractéristiques qui sont des facteurs pronostiques de l'événement étudié (la mortalité). Par exemple, les individus du groupe « prescription conforme » étaient plus jeunes (32,9% étaient âgés de 65 ans ou moins *vs* 13,8%) et avaient moins de comorbidités (score de Charlson < 9 chez 73,4% *vs* 49,1%) que ceux de l'autre groupe.

Pour limiter le biais d'indication, nous avons utilisé la méthode d'appariement sur un SP. Notre choix a été d'évaluer l'effet moyen chez les traités (*ATT*), dans une configuration mimant l'expérimentation. L'appariement a été réalisé sur le mode 1:1, sans remise, avec un '*caliper*' de 0,02, permettant d'apparier environ 78% des individus du groupe « prescription non-conforme » mais seulement 41% de ceux du groupe « prescription conforme ». Si la perte de puissance produite par la restriction de l'échantillon étudié lors de l'appariement n'a pas semblé problématique (puisque le résultat obtenu était statistiquement significatif), on peut toutefois s'interroger sur la représentativité de l'échantillon apparié par rapport à l'échantillon initial (biais de sélection potentiel).

Un biais de confusion résiduel est possible dans les résultats de l'analyse avec SP. Il n'a pas été possible dans cette étude d'utiliser une méthode basée sur une VI, la taille limitée de l'échantillon empêchant de l'envisager. Toutefois l'importance de l'effet mesuré (HR 0,54) nécessiterait qu'un potentiel facteur de confusion non mesuré soit particulièrement fort pour qu'il explique à lui-seul l'effet observé.

B.3. Évaluation de l'effet des stratégies antithrombotiques chez les patients hémodialysés sur le risque hémorragique (article 4)

B.3.1. Contexte

Les patients hémodialysés chroniques présentent un risque élevé d'hémorragies par rapport à la population générale (Fischer 2007). Ce risque pourrait être lié à la maladie rénale elle-même, mais également aux traitements antiagrégants plaquettaires (AAP) et/ou anticoagulants per os (AC) souvent administrés à ces patients, dans le cadre des dialyses ou en raison de comorbidités cardio-vasculaires. À ce jour, les études expérimentales et méta-analyses publiées n'ont pas permis de conclure formellement à l'existence ou non d'un risque hémorragique induit par les AAP et AC chez les patients hémodialysés chroniques.

Pour tenter de préciser ce risque hémorragique, nous avons exploité les données de l'étude T2HD (*Thrombosis and Hemorrhage in HemoDialysis patients*), étude de cohorte rétrospective qui a inclus 502 adultes ayant initié une suppléance de la fonction rénale par hémodialyse pour insuffisance rénale chronique en Lorraine en 2009 et 2010. Les patients ont été suivis jusqu'au 30 juin 2013, avec un recueil de données démographiques, cliniques et thérapeutiques, et l'enregistrement de leur statut vital et des événements thrombotiques et hémorragiques graves.

Dans le cadre de l'étude T2HD, notre objectif était d'évaluer le risque d'hémorragies graves associé à la prescription d'AAP et/ou d'AC chez les patients hémodialysés chroniques.

Ce travail a fait l'objet d'un article publié dans *Pharmocoepidemiology and Drug Safety* reproduit dans les pages suivantes (Collette, Clerc-Urmes et al. 2016).

PHARMACOEPIDEMIOLOGY AND DRUG SAFETY 2016; **25**: 935–943
Published online 1 May 2016 in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/pds.4002

ORIGINAL REPORT

Antiplatelet and oral anticoagulant therapies in chronic hemodialysis patients: prescribing practices and bleeding risk

Camille Collette[1†], Isabelle Clerc-Urmès[2†], Hervé Laborde-Castérot[1,3], Luc Frimat[1,4], Carole Ayav[2], Nicolas Peters[4], Alexandre Martin[4], Nelly Agrinier[1,2] and Nathalie Thilly[1,2*]

[1] *Lorraine University, Paris-Descartes University, Apemac EA 4360 Apemac Nancy, France*
[2] *Clinical Epidemiology and Evaluation, INSERM CIC 1433—Clinical Epidemiology, University Hospital of Nancy, Nancy, France*
[3] *Department of Occupational Medicine and Occupational Pathology, University Hospital of Bordeaux, Bordeaux, France*
[4] *Department of Nephrology, University Hospital of Nancy, Nancy, France*

ABSTRACT

Purpose Results of previous studies assessing the risk of bleeding associated with prescription of antiplatelet (AP) and/or oral anticoagulant (AC) therapy to hemodialysis patients are conflicting. Our purpose was to describe practices for prescription of AP and AC in hemodialysis patients in the Lorraine region, and to assess their effect on the risk of major bleeding events.

Methods All adults with chronic kidney disease who began a first renal replacement therapy by hemodialysis in 2009 or 2010 in one of the 12 dialysis centers in Lorraine were included in the Thrombosis and Hemorrhage in HemoDialysis patients (T2HD) study and followed up until 30 June 2013. The association of each treatment (AP, AC, AP + AC) with the risk of major bleeding was estimated by three Cox proportional hazard models with an inverse probability of treatment weighting on a propensity score, considering the untreated patients as the reference.

Results Among 502 patients included, 227 (45.2%) received an AP, 68 (13.5%) an AC, 81 (16.1%) a combination AP + AC, and 126 (25.1%) were untreated. As compared with untreated patients, those given AP (HR 5.52, 95% CI [3.11–9.80]), AC (HR: 4.15, 95% CI: [3.46–4.99]), and AP + AC (HR: 5.59, 95% CI [2.62–11.91]) were at greater risk of major bleeding events.

Conclusions The risk of major bleeding is higher in patients receiving an oral AC compared with untreated patients and those receiving an AP agent. A combination of the two drugs does not seem to increase the risk. Copyright © 2016 John Wiley & Sons, Ltd.

KEY WORDS—antiplatelet agent; bleeding events; hemodialysis; oral anticoagulant; prescribing practices; pharmacoepidemiology

Received 20 August 2015; Revised 19 January 2016; Accepted 07 March 2016

INTRODUCTION

Chronic hemodialysis patients are at a high risk of bleeding compared with the general population. This may be explained by the accumulation of uremic toxins, anemia, secondary hyperparathyroidism, platelet activation associated with extracorporeal circulation, and anticoagulation in dialysis.[1,2] Despite the risk of bleeding, antiplatelet (AP) agents and/or oral anticoagulants (AC) are often prescribed to hemodialysis patients, because of the high prevalence of cardiovascular comorbidities they exhibit.[3,4] Indeed, cardiovascular mortality is ten times greater in dialysis patients and 40 times greater in diabetic patients with renal failure.[5,6] Furthermore, these patients regularly receive heparin to prevent clotting in the extracorporeal system during the dialysis sessions. All these medications increase the baseline risk of bleeding in this population.

Results of previous studies assessing the risk of major bleeding events associated with AP and/or AC prescription are conflicting. Two randomized controlled trials[7,8] found no relationship between bleeding and the use of clopidogrel and warfarin in hemodialysis patients.

A meta-analysis by Coleman *et al.*[9] concluded that AP were not associated with an increased risk of

*Correspondence to: N. Thilly, PharmD, PhD, Clinical Epidemiology and Evaluation, INSERM CIC 1433—Clinical Epidemiology, University Hospital of Nancy, Nancy F-54000, France. E-mail: n.thilly@chu-nancy.fr

†Joint first authors

#The results of this manuscript have never been presented previously, neither in postings nor in public presentations.

bleeding; in contrast Hiremath *et al.*[10] reported that it depended on the number and type of AP agents used. Regarding the risk of bleeding associated with AC, Elliott *et al.*[11] showed an increased risk in hemodialysis patients.

Studies comparing the respective bleeding risk of AP and AC are also conflicting: Holden *et al.*[12] showed a higher risk with AP, whereas Sood *et al.*[13] and Chan *et al.*[14] showed a higher risk with AC.

We sought to describe prescription practices for AP and AC in chronic hemodialysis patients in the Lorraine region, and to assess the effect of these drugs on the risk of major bleeding in this population.

METHODS

Setting, design, and patients

The Thrombosis and Hemorrhage in HemoDialysis patients (T2HD) study was an observational, retrospective cohort study involving the 12 private and public dialysis centers operating in the administrative region of Lorraine, North-East France (population 2 350 000, according to the 2011 census). All the adults with a chronic kidney disease who began a first renal replacement therapy by hemodialysis in one of these 12 dialysis centers between 1 January 2009 and 31 December 2010 were identified from the regional end stage renal disease registry (REIN registry) and considered eligible for the cohort. Patients with progressive cancer at start of dialysis and those who died during the first 45 days of replacement therapy, not considered as chronic hemodialysis patients by the REIN registry, were not included. Patients were then followed until death, change of replacement therapy modality, relocation outside Lorraine or until 30 June 2013 if none of the aforementioned events occurred. The T2HD cohort study conforms to the principles outlined in the Declaration of Helsinki.

Data collection

A standardized form was used to collect demographic, clinical, and therapeutic data from the REIN registry and outpatient medical records. Demographic and clinical data included age, sex, body mass index (BMI), smoking status, date, modality and circumstances of dialysis initiation, primary renal disease, and comorbid conditions. Comorbid conditions collected were: (i) history of acute coronary syndrome, stroke or transient ischemic attack (TIA), pulmonary embolism, deep venous or arterial thrombosis, (ii) comorbidities at dialysis initiation or during the follow-up such as hypertension, diabetes, coronary insufficiency, atrial fibrillation, peripheral arterial disease, mitral valve insufficiency or prosthetic valve, and congestive heart failure. Therapeutic data included chronic treatment by AP and AC at dialysis initiation and during the follow-up, and heparin use during dialysis sessions. Indications of AP and/or AC prescription were recorded if available. The following events occurring during the follow-up were also recorded: relocation out of Lorraine, change of replacement therapy modality, bleeding events, and death. All data were collected and checked for completeness, according to French Good Practices in Epidemiology[15], by two specialized clinical research assistants. Patient enrolment and quality of data collection were regularly controlled by a steering committee of two epidemiologists and two nephrologists. Ten per cent of completed standardized forms were audited by an independent clinical research assistant by comparing, for each form, data collected and the patient medical record.

Outcome of interest

The outcome of interest was the first major bleeding event during the follow-up. A major bleeding event was defined as intracranial hemorrhage, gastrointestinal bleeding, or any other bleeding that required hospitalization.

Statistical analyses

Patients were split into four treatment groups according to their AP/AC prescription during the follow-up period: patients receiving AP agent(s) and no AC (AP group), patients receiving an AC and no AP (AC group), patients receiving a combination of AP and AC (AP + AC group), and patients receiving neither an AP nor an AC at any time during follow-up (untreated group). A patient was considered as being treated with AP or AC if he (she) had received the drug for at least one day during follow-up. Two situations might be found concerning prescribing of these treatments: (i) the AP or/and AC is prescribed from dialysis initiation; maybe some patients were already treated before dialysis, but the dialysis initiation led the physician to reassess the risk/benefit ratio of the prescription, which could then be considered a new prescription; (ii) the AP or/and AC is introduced during the follow-up period.

Demographic and clinical characteristics, heparin treatment, and events occurring during follow-up were first described overall and then compared between each of the three treated groups (AP, AC, AP + AC) and the untreated group using bivariate analysis—Pearson chi-square test or Fisher exact test for

DOI: 10.1002/pds

qualitative variables, ANOVA F-statistic for quantitative variables. Three Cox proportional hazard models were used to evaluate the effect on bleedings of: (i) AP treatment, (ii) AC treatment, and (iii) AP + AC treatment, as compared with no treatment, the follow-up time beginning at dialysis initiation in all cases. To minimize potential bias and confounding effects, a propensity score (PS) was calculated by using logistic regressions for each of the three models.[16] The PS represents the likelihood of AP, AC, or AP + AC prescription conditional on demographic and clinical characteristics. The following covariates were used to calculate PS: sex, age, BMI, smoking status, hypertension, diabetes, coronary insufficiency, atrial fibrillation, peripheral arterial disease, mitral valve insufficiency or prosthetic valve, congestive heart failure, history of acute coronary syndrome, stroke or TIA, pulmonary embolism, and deep venous and arterial thrombosis. We carefully checked that each comorbidity considered in the PS occurred before the AP or/and AC initiation. The inverse probability of treatment weighting (IPTW) method was then applied by using the PS to assign individual weights to all observations, which allows some characteristics of randomized controlled trials to be mimicked in an observational study. With this approach, the contributions of each patient are weighted by 1/PS for patients treated by AP, AC, or AP + AC and by $1/(1-PS)$ for untreated patients.[17,18] To address confounding by frailty, both treated and untreated patients below the 1st percentile of the PS distribution in the treated group and the 99th percentile in the untreated group were trimmed as recommended by Stürmer *et al.*[19] To assess whether the PS models were adequately specified, standardized differences were calculated.[17] They indicate the degree of systematic differences in covariates between groups compared using IPTW and PS. Empirically, an absolute standardized difference of $<10\%$ indicates a negligible difference in mean or percentage of the covariates between treated and untreated patients.[18] There were no missing values for variables considered in this study, with the exception of treatment indications. The effects of treatments on risk of bleeding are reported as adjusted hazard ratio (HR) with 95% confidence interval (95%CI) and p value. A p value of <0.05 for two-sided tests was considered significant. All analyses were performed using SAS® software version 9.3 (SAS Institute, Cary, NC, USA).

RESULTS

In 2009 and 2010, 809 adults began hemodialysis treatment for chronic kidney disease in Lorraine. Among them, 307 were excluded for the following reasons: hemodialysis was not the first renal replacement therapy ($n=97$), death occurred during the first 45 days of dialysis ($n=82$), a progressive cancer was diagnosed at the start of dialysis ($n=76$), or the patient medical record was not found ($n=52$). Accordingly, 502 patients were included in the T2HD study. The mean follow-up was 2.4 ± 1.2 years and 1223 person-years of exposure were observed.

Therapeutic characteristics

A total of 227 patients (45.2%) were treated by AP, 68 (13.5%) by AC, and 81 (16.1%) by AP + AC. In the AP group, 76 (33.5%) patients received dual AP therapy, whereas all patients in the AC group received monotherapy. In addition, during the dialysis sessions, 455 (90.6%) received heparin, with a significant difference between those given AC and those untreated (95.2% vs. 79.4%, $p=0.038$).

Among the 227 patients in the AP group, 170 (74.9%) already received an AP at the start of dialysis and 213 (93.8%) were still being treated at the end of follow-up. The mean duration of AP prescription was 19.8 ± 15.8 months (median = 16.9; range = [0.1–53.7]), representing 66% of the time of the mean follow-up. Main medications used were acetylsalicylic acid (70.9%) and clopidogrel (22.9%). The most frequent indications of AP prescriptions were secondary prevention of acute coronary syndrome (47.1%) and peripheral arterial disease (34.4%).

Among the 68 patients in the AC group, 44 (64.7%) already received an AC at the start of dialysis and 62 (91.2%) were still being treated at the end of follow-up. The mean duration of AC prescription was 26.2 ± 13.6 months (median = 23.4; range = [1.5–52.1]), representing almost 100% of the time of the mean follow-up. Main medication used was fluindione (72.1%). The most frequent indications of AC prescriptions were atrial fibrillation (60.3%), and secondary prevention of deep venous thrombosis (17.7%).

Among the 81 patients in the AP + AC group, 31 (38.3%) already received both drugs at the start of dialysis and 60 (74.1%) were still being treated at the end of follow-up. The mean duration of AP + AC prescription was 16.6 ± 14.5 months (median = 13.1; range = [0.1–53.5]), representing 54% of the time of the mean follow-up. The main combination used was acetylsalicylic acid–fluindione (65.4%).

Demographic and clinical characteristics

Demographic and clinical characteristics of included patients are presented in Table 1, overall and

Table 1. Baseline characteristics of included patients according to their treatment group

	Overall	Untreated	Antiplatelet (AP)		Anticoagulant (AC)		Antiplatelet + Anticoagulant (AP + AC)	
	N = 502	*N* = 126	*N* = 227	*p* Value* AP vs. Untreated	*N* = 68	*p* Value* AC vs. Untreated	*N* = 81	*p* Value* AP + AC vs. Untreated
Male sex (%)	63.3	68.3	64.8	0.506	55.9	0.087	58.0	0.134
Age, year (mean ± SD)	68.7 ± 13.7	59.7 ± 16.7	70.9 ± 11.7	<0.001	72.5 ± 10.0	<0.001	73.1 ± 9.7	<0.001
Body mass index, kg/m^2 (mean ± SD)	29.2 ± 6.6	28.6 ± 7.2	28.7 ± 6.1	0.736	30.2 ± 6.2	0.090	30.4 ± 6.8	0.063
Smoking status, current or former (%)	37.8	37.3	37.4	0.979	35.3	0.781	42.0	0.501
Dialysis initiation in 2009 (%)	47.8	45.2	49.8	0.413	45.6	0.963	48.1	0.682
Place of dialysis (%)				0.046		0.036		0.004
Hospital-based hemodialysis	92.2	85.7	92.1		97.0		98.8	
Dialysis unit under medical supervision	6.0	9.5	7.0		1.5		1.2	
Home hemodialysis	1.8	4.8	0.9		1.5		0.0	
Emergency first dialysis (%)	17.3	9.5	16.7	0.062	20.6	0.031	28.4	<0.001
Arteriovenous fistula (%)	76.1	78.6	79.3	0.873	69.1	0.145	69.1	0.127
Primary renal disease (%)				<0.001		0.032		<0.001
Glomerulonephritis	7.6	15.9	6.2		4.4		1.2	
Diabetic nephropathy	22.3	12.7	30.0		14.7		22.2	
Hypertensive nephropathy	15.9	15.1	16.3		10.3		21.0	
Hereditary nephropathy	10.8	17.5	8.4		11.8		6.2	
Others/Unknown	43.4	38.8	39.1		58.8		49.4	
Comorbid conditions (%)								
History of:								
Acute coronary syndrome	27.5	4.0	40.5	<0.001	14.7	0.007	38.3	<0.001
Stroke/TIA	12.8	2.4	16.7	<0.001	16.2	<0.001	14.8	<0.001
Pulmonary embolism	4.6	3.2	3.5	1.000	5.9	0.190	8.6	0.114
Deep venous thrombosis	9.0	2.4	9.3	0.014	14.7	0.002	13.6	0.003
Arterial thrombosis	7.6	2.4	11.9	0.001	0.0	0.553	9.9	0.026
At dialysis initiation or during the follow-up:								
Hypertension	88.6	83.3	91.6	0.018	92.6	0.069	85.2	0.722
Diabetes	51.8	29.4	61.7	<0.001	47.1	0.014	63.0	<0.001
Coronary insufficiency	7.0	1.6	7.5	0.019	10.3	0.006	11.1	0.004
Atrial fibrillation	34.7	7.1	25.1	<0.001	67.6	<0.001	76.5	<0.001
Peripheral arterial disease	32.1	4.0	42.7	<0.001	19.1	<0.001	56.8	<0.001
Mitral valve insufficiency/ prosthetic valve	27.1	20.6	28.2	0.119	25.0	0.485	35.8	0.016
Congestive heart failure	34.5	12.7	37.9	0.003	45.6	<0.001	49.4	<0.001
Heparin treatment	90.6	95.2	90.7	0.146	79.4	0.038	92.6	0.641

TIA, transient ischemic attack; SD, standard deviation.
*Pearson chi-square test or Fischer exact test (as appropriate) for qualitative variables and Anova F-test for quantitative variables.

according to the treatment group. Mean age was 68.7 ± 13.7 years and 318 (63.3%) were male. As compared with patients who received AP or/and AC, untreated patients were younger and more likely to receive dialysis treatment outside of hospital sites. Their primary renal disease was most often glomerulonephritis or hereditary nephropathy, and they had fewer comorbid conditions than did patients given AP and/or AC.

In analyses using IPTW, 88, 33, and 55 patients respectively were trimmed in the AP, AC, and AP + AC groups. The distribution of weights after trimming in all groups compared is presented in Figure 1. The standardized differences in baseline patient characteristics between treatment groups after IPTW presented in Table 2 were all <10%, showing a good balance in potential confounding factors between the groups compared.

Events during the follow-up

The mean duration of follow-up was 29.3 ± 14.8 months, and did not differ between the treatment groups (Table 3). As compared with patients who received AP or/and AC, untreated patients switched replacement therapy modality more frequently. Almost one-third switched to kidney transplantation. Likewise, they moved out of Lorraine during the follow-up more often than did treated patients.

During the follow-up, deaths occurred in 28 (22.2%) of untreated patients, 87 (38.3%) of the AP group, 34 (50.0%) of the AC group, and 37 (45.7%) of the AP + AC group. Major bleeding events were also significantly more frequent in the AP group (29 (12.8%)), the AC group (12 (17.6%)) and the AP + AC group (22 (27.2%)), as compared with untreated patients (7 (5.6%)). These differences principally concerned gastro-intestinal bleeding in the AP + AC group (13.6%) and bleeding other than intracranial or gastro-intestinal for the three treated groups.

Impact of treatment on major bleeding events

In crude analyses, patients in the AP and AP + AC groups presented a significantly higher risk of major bleeding than untreated patients (respectively for AP: HR = 2.77, 95% CI = [1.12;6.87], $p = 0.028$; for AP + AC: HR = 5.13, 95% CI = [1.20;23.58], $p = 0.036$). In the AC group, the bleeding risk was higher than the untreated group but remained under the significance level (HR = 3.62, 95% CI = [0.85;15.43], $p = 0.082$).

In IPTW analyses, a higher risk of major bleeding was found in all three treatment groups as compared with untreated patients. The HR for time to first major bleeding event was 5.52 for the AP group, 4.15 for the AC group, and 5.59 for the AP + AC group (Table 4).

DISCUSSION

The T2HD study shows that chronic hemodialysis patients are widely prescribed AP and/or oral AC therapies, with only one quarter of them not receiving these medications. The very frequent use of AP agents (61.3% of patients) and oral AC (29.7% of patients) may be explained by a high prevalence of cardiovascular comorbidities/events (i.e. atrial fibrillation: 34.7%; peripheral arterial disease: 32.1%; history of acute coronary syndrome: 27.5%) in this population, leading to a high mortality. The mean durations of prescriptions, representing more than 50% of the time of follow-up, indicate a long-term usage consistent with the chronic condition of cardiovascular comorbidities. As compared with untreated patients, the risk of major bleeding was five-fold higher in AP and AP + AC patients, and four-fold higher in AC patients.

In the DOPPS study[13] variations in antithrombotic agents used were studied in 44 144 hemodialysis patients recruited in the late 1990s from 12 countries. There was wide variation between countries in use of oral AC (0.3–18%), AP agents (3–25%), and acetylsalicylic acid (8–36%). In France, DOPPS reported prescribing rates of these medications of 1%, 25%, and 23%, respectively. In our study, the antithrombotic prescribing rate is actually much higher than the rate recorded in DOPPS in Canada, which was the country where antithrombotic use was highest. Indeed, ten years after inclusion in DOPPS, the antithrombotic prescribing rate in hemodialysis patients has greatly increased in France. The main reason is most likely to be an increase in the prevalence of dialysis patients with cardiovascular comorbidity, rising from 49.7% in 2003 to 64.6% in 2013 in Lorraine. This trend was also reported in the rest of the country.[20] In addition, nephrologists probably have fewer reservations now than in the past about prescribing AP, and particularly AC, in dialysis patients. Concerning the characteristics of patients receiving AP and/or AC, our results showing that treated patients were older and more often had comorbid conditions than did untreated patients are consistent with those reported by Chan *et al.*[14] in 41 425 hemodialysis patients. Comorbid conditions were cardiovascular diseases (i.e. atrial fibrillation, peripheral arterial disease, congestive heart failure), well-known to increase risk of thromboembolic events. Patients receiving AP and/or AC present then a risk of bleeding because of

DOI: 10.1002/pds

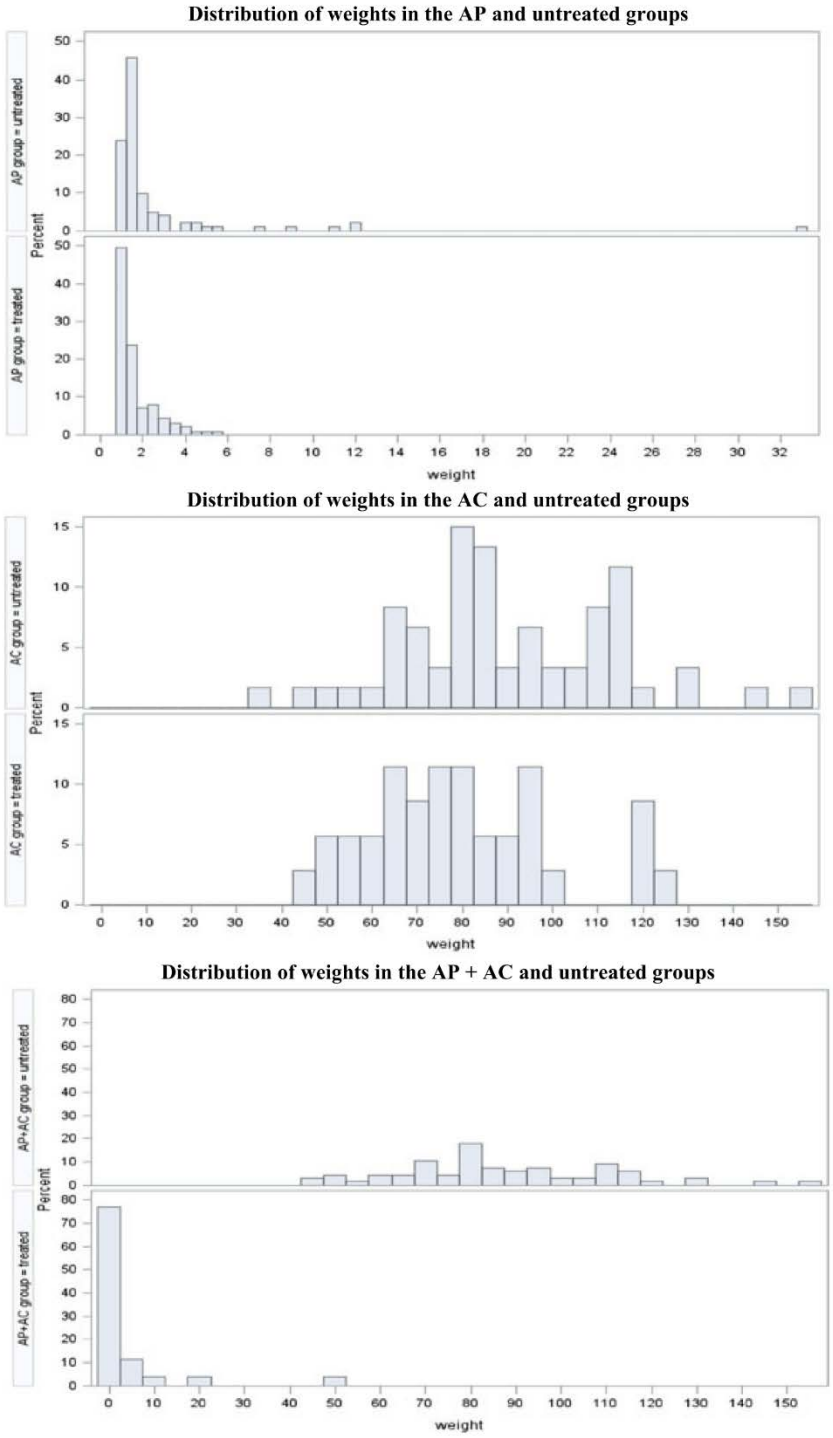


Figure 1. Distribution of weights in the groups (AP, AC, AP + AC), compared to untreated group after trimming

Pharmacoepidemiology and Drug Safety, 2016; **25**: 935–943
DOI: 10.1002/pds

Table 2. Proportion (%) of standardized differences before and after IPTW

Covariates	AP ($n=139$) vs. untreated ($n=101$) Before IPTW	AP ($n=139$) vs. untreated ($n=101$) After IPTW	AC ($n=35$) vs. untreated ($n=61$) Before IPTW	AC ($n=35$) vs. untreated ($n=61$) After IPTW	AP + AC ($n=26$) vs. untreated ($n=68$) Before IPTW	AP + AC ($n=26$) vs. untreated ($n=68$) After IPTW
Male Sex	14.3	1.3	13.0	8.4	3.8	0.9
Age						
≤60 versus 60–75 years	38.2	0.9	42.1	1.5	11.8	1.3
≤60 versus ≥ 75 years	49.0	6.7	72.2	8.1	19.8	4.0
60–75 versus ≥75 years	10.8	6.1	30.2	1.3	11.7	9.7
Body mass index						
Skinny/normal versus overweight	10.8	8.7	32.8	4.6	10.9	0.8
Skinny/normal versus obese	14.4	4.4	2.8	8.4	44.1	2.6
Overweight versus obese	3.6	2.1	36.4	6.2	17.1	9.1
Smoking status, current or former	1.5	9.3	20.7	3.2	15.0	6.5
Acute coronary syndrome	16.7	7.0	5.0	7.0	0.9	2.7
Stroke/TIA	30.3	8.5	96.0	0.9	12.1	8.9
Pulmonary embolism	56.4	9.9	31.2	1.3	7.6	6.2
Deep venous thrombosis	15.4	4.3	23.5	4.0	1.6	1.2
Arterial thrombosis	19.2	7.7	30.3	5.4	9.1	5.2
Hypertension	36.0	1.0	2.3	9.7	4.5	1.3
Diabetes	19.4	4.7	20.6	6.8	3.4	9.9
Coronary insufficiency	56.4	9.2	41.1	6.2	31.9	7.7
Atrial fibrillation	29.1	4.1	13.8	6.1	9.6	4.6
Peripheral arterial disease	0.5	3.8	7.2	5.3	7.6	2.5
Mitral valve insufficiency/prosthetic valve	24.4	3.8	2.8	6.0	9.6	6.2
Congestive heart failure	19.3	3.9	1.5	7.7	0.0	0.0

TIA, transient ischemic attack; AP, antiplatelet; AC, anticoagulant; IPTW, inverse probability of treatment weighting.

their treatments, but also a risk of thromboembolic events because of cardiovascular comorbidities. This double risk explains the differences found between crude and adjusted analyses: when comorbidities were taken into account by the IPTW method, HR for time to first major bleeding are more significant in all three groups treated by AP and/or AC.

The effect of AP on the risk of bleeding in hemodialysis patients was evaluated in the review by Hiremath *et al.*[10] which reported relative risks ranging from 1.98 to 5.24. An HR of 5.24, reported by the Holden study,[12] is consistent with our finding. However, studies considered in the Hiremath review have given conflicting results and variations appear to be related to the methodology used, the exposure, and events considered. A formal conclusion on the bleeding risk of AP in dialysis patients cannot then be drawn, and the risk depends on the

Table 3. Follow-up events according to treatment group

	Overall	Untreated	Antiplatelet		Anticoagulant		Antiplatelet + Anticoagulant	
	$N=502$	$N=126$	$N=227$	*p* Value*	$N=68$	*p* Value*	$N=81$	*p* Value*
Follow-up duration, month (mean ± SD)	29.3 ± 14.8	28.5 ± 14.1	30.2 ± 15.3	0.215	26.2 ± 15.2	0.373	30.9 ± 14.4	0.167
Move out of Lorraine (%)	2.4	5.7	1.3	0.022	1.5	0.133	1.3	0.094
Change of replacement therapy (%)	14.5	30.2	11.5	<0.001	7.4	<0.001	4.9	<0.001
Deaths (%)	37.1	22.2	38.3	0.002	50.0	<0.001	45.7	<0.001
At least one major bleeding event (%)	13.9	5.6	12.8	0.032	17.6	0.007	27.2	<0.001
At least one intracranial hemorrhage (%)	0.8	0.0	1.3	0.555	0.0	0.000	1.2	0.391
At least one gastro-intestinal hemorrhage (%)	7.0	4.7	6.2	0.584	5.9	0.743	13.6	0.024
At least one other bleedings (%)	7.6	0.8	7.5	0.005	13.2	<0.001	13.6	<0.001

*Pearson chi-square test or Fischer exact test (as appropriate) for qualitative variables and Anova F-test for quantitative variables.

Table 4. Effect of treatment on major bleeding events

Treatments	Number of patients	Number of bleeding events	Adjusted HR	95% CI	*P* value*
Antiplatelet agents					
Untreated	101	27	1		
Treated	139		5.52	[3.11–9.80]	<0.0001
Anticoagulant					
Untreated	61	9	1		
Treated	35		4.15	[3.46– 4.99]	<0.0001
Antiplatelet + Anticoagulant					
Untreated	68	9	1		
Treated	26		5.59	[2.62–11.91]	<0.0001

HR, hazard ratio; CI, confidence interval.
*Resulting from three Cox proportional hazard models after IPTW in trimmed samples.

characteristics of the treatment and the type of bleeding considered.

The effect of AC on the risk of bleeding in hemodialysis patients was evaluated in the review of Elliott *et al.*[11] which reported relative risks ranging from 0.8 to 2.36 depending on the intensity of warfarin anticoagulation. With an HR of 4.15, our result is consistent with the HR of 3.59 (95% CI=[0.95–13.60]) reported by Holden *et al.* in their study on 255 hemodialysis patients.[12] In contrast to AP, the bleeding risk associated with AC use seems to be well established and is confirmed by our study. Indeed, Elliot *et al.*[11] conclude in their review that low and full intensity anticoagulation use in hemodialysis patients is associated with a significant bleeding risk.

Holden *et al.*[12] is the only study, to our knowledge, that has investigated the risk of bleeding associated with the combination of AP and AC. The HR for time to first major bleeding event of 6.19 (95% CI=[1.60–23.97]) reported is not far from the HR of 5.59 (95% CI=[2.62–11.91]) found in our study. Our result supports the contention that the combination of AP and AC appears not to increase the risk associated with each as monotherapy. A likely reduction in the dose of each treatment (AP and AC) when used in combination as compared with monotherapy may explain this result, but the lack of information about dosages prohibits conclusions being drawn.

Concerning the respective risks of major bleeding events with AC and AP, our results are, again, consistent with those described by Holden[12] who found that HR for time to first bleeding was 3.59 (95% CI=[0.95–13.60]) for warfarin, but higher for aspirin with an HR of 5.24 (95% CI=[1.64–16.79]). However, these results contrast with those found by the DOPPS study[13] and Chan's study[14]. For example, in the DOPPS study, HR for time to first bleeding was 1.39 (95% CI=[1.20–1.61]) for AC, 1.13 (95% CI=[1.02–1.26]) for aspirin, and 1.25 (95% CI=[1.10–1.42]) for AP.

Our results should be interpreted in the light of some limitations. First, history of bleeding is a well-known predictor of current bleeding that should have been taken into account in our analyses. Unfortunately, this information is not available in the REIN registry and is rarely mentioned in the outpatient medical records. However, when a bleeding event occurs, physicians often become reluctant to prescribe (or to continue the prescription of) AP and/or AC. Then, the probability of bleeding history is presumably higher in the untreated group as compared with the three treated groups, leading to a risk of under-estimating the HR for time to first bleeding. The second limitation is the limited number of events of interest, i.e. major bleeding events ($n=70$), particularly in the untreated group ($n=7$). Although significant results are obtained, their sensitivity to the impact of unmeasured confounders, such as bleeding history, is sizeable. Third, the observational design of the T2HD study allows us to measure associations between antithrombotic prescriptions and bleeding events, but cannot demonstrate strictly causal relationships. However, as AC and AP therapies are already widely prescribed to hemodialysis patients because of the high prevalence of cardiovascular comorbidities in this specific population, a randomized controlled trial would clearly be impractical for ethical reasons. To address potential bias and confounding effects, we applied the IPTW method using the PS, which is recommended by Austin *et al.* [21] for estimating risk differences in observational studies. As illustrated by our results, this method helps reconstitute groups of patients with similar baseline characteristics. Fourth, the T2HD study did not record dosages used for AC and AP agents. Differences in dosages prescribed may explain

differences in HR estimations for time to first major bleeding event reported in previous studies. This information would be of particular interest to assess the dose–response relationship, and to interpret the result found with the combination of AP and AC.

The main strengths are the high reliability and completeness of the data collected. All patients initiating a hemodialysis treatment in Lorraine in 2009–2010 and meeting the inclusion criteria were considered here and no missing data were recorded, eliminating the potential bias because of selective reporting.

In conclusion, the present study adds to our understanding of impact of AP and/or AC use on the risk of major bleeding events in hemodialysis patients. Prescriptions of AP and AC are associated with a significant bleeding risk, but the risk seems to be higher with AP. The combination of AP and AC appears not to increase the risk associated with each treatment prescribed in monotherapy. The management of the double risk of cardiovascular events and bleeding in hemodialysis patients is complex and the bleeding risk has to be balanced against benefits of therapy. Further research should investigate how to optimize AP and/or AC use, by minimizing the risks with which they are associated.

CONFLICT OF INTEREST

All authors report no relevant potential conflicts of interest related to this manuscript.

KEY POINTS

- In the Lorraine region of France, chronic hemodialysis patients are widely prescribed antiplatelet or/and oral anticoagulant therapies (61.3% and 29.7% of patients, respectively) for cardiovascular comorbidities, only one quarter do not receive such medications.
- As compared with untreated patients, the risk of major bleeding is 5.5-fold higher in those receiving antiplatelet agents and those receiving the combination of antiplatelet agents and oral anticoagulants, and four-fold higher in those given oral anticoagulants.
- The management of the double risk of cardiovascular events and bleeding in hemodialysis patients is complex, and the bleeding risk has to be balanced against benefits of therapy.

ACKNOWLEDGEMENTS

The T2HD study was supported by a grant from the Biomedicine Agency and the Lorraine Region.

REFERENCES

1. Aggarwal A, Kabbani SS, Rimmer JM, *et al.* Biphasic effects of hemodialysis on platelet reactivity in patients with end-stage renal disease: a potential contributor to cardiovascular risk. *Am J Kidney Dis* 2002; **40**: 315–322. doi:10.1053/ajkd.2002.34510.
2. Fischer KG. Essentials of anticoagulation in hemodialysis. *Hemodial Int* 2007; **11**: 178–189. doi:10.1111/j.1542-4758.2007.00166.x.
3. Dennis VW. Coronary heart disease in patients with chronic kidney disease. *J Am Soc Nephrol* 2005; **16**: 103–106. doi:10.1681/ASN.2005060665.
4. Wetmore JB, Shireman TI. The ABCs of cardioprotection in dialysis patients: a systematic review. *Am J Kidney Dis* 2009; **53**: 457–466. doi:10.1053/j.ajkd.2008.07.037.
5. Parfrey PS, Foley RN. The clinical epidemiology of cardiac disease in chronic renal failure. *J Am Soc Nephrol* 1999; **10**: 1606–1615.
6. Kessler M, Zannad F, Lehert P, *et al.* Predictors of cardiovascular events in patients with end-stage renal disease: an analysis from the Fosinopril in dialysis study. *Nephrol Dial Transplant* 2007; **22**: 3573–3579. doi:10.1093/ndt/gfm417.
7. Dember LM, Beck GJ, Allon M, *et al.* Effect of clopidogrel on early failure of arteriovenous fistulas for hemodialysis: a randomized controlled trial. *JAMA* 2008; **299**: 2164–2171. doi:10.1001/jama.299.18.2164.
8. Wilkieson TJ, Ingram AJ, Crowther MA, *et al.* Low-intensity adjusted-dose warfarin for the prevention of hemodialysis catheter failure: a randomized, controlled trial. *Clin J Am Soc Nephrol* 2011; **6**: 1018–1024. doi:10.2215/CJN.07240810.
9. Coleman CI, Tuttle LA, Teevan C, Baker WL, White CM, Reinhart KM. Antiplatelet agents for the prevention of arteriovenous fistula and graft thrombosis: a meta-analysis. *Int J Clin Pract* 2010; **64**: 1239–1244. doi:10.1111/j.1742-1241.2009.
10. Hiremath S, Holden RM, Fergusson D, Zimmerman DL. Antiplatelet medications in hemodialysis patients: a systematic review of bleeding rates. *Clin J Am Soc Nephrol* 2009; **4**: 1347–1355. doi:10.2215/CJN.00810209.
11. Elliott MJ, Zimmerman D, Holden RM. Warfarin anticoagulation in hemodialysis patients: a systematic review of bleeding rates. *Am J Kidney Dis* 2007; **50**: 433–440. doi:10.1053/j.ajkd.2007.06.017.
12. Holden RM, Harman GJ, Wang M, Holland D, Day AG. Major bleeding in hemodialysis patients. *Clin J Am Soc Nephrol* 2008; **3**: 105–110. doi:10.2215/CJN.01810407.
13. Sood MM, Larkina M, Thumma JR, *et al.* Major bleeding events and risk stratification of antithrombotic agents in hemodialysis: results from the DOPPS. *Kidney Int* 2013; **84**: 10. doi:10.1038/ki.2013.170.
14. Chan KE, Lazarus JM, Thadhani R, Hakim RM. Anticoagulant and antiplatelet usage associates with mortality among hemodialysis patients. *J Am Soc Nephrol* 2009; **20**: 872–881. doi:10.1681/ASN.2008080824.
15. Recommendations deontology and good practices in epidemiology (v France—2007). *Epidemiol Public Health Rev* 2008; **56**: 121–S148.
16. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**: 41–55. doi:10.1093/biomet/70.1.41.
17. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivar Behav Res* 2011; **46**: 399–424. doi:10.1080/00273171.2011.568786.
18. 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**: 734–775. doi:10.1002/sim.2580.
19. Stürmer T, Wyss R, Glynn RJ, Brookhart MA. Propensity scores for confounder adjustment when assessing the effects of medical interventions using nonexperimental study designs. *J Intern Med* 2014; **275**: 570–580. doi:10.1111/joim.12197.
20. The Renal Epidemiology and Information Network (REIN): Annual Report 2013, Biomedicine Agency. Available: www.agence-biomedecine.fr/IMG/pdf/rapport_rein2013.pdf.
21. Austin PC. The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. *Stat Med* 2010; **29**: 2137–2148. doi:10.1002/sim.3854.

B.3.2. Discussion

Cette troisième application dans le cadre d'une étude observationnelle se confronte également au risque de biais d'indication. En effet, les patients traités par AAP et/ou AC avaient des caractéristiques initiales différentes des patients ne recevant pas ces traitements, en étant par exemple plus âgés et en ayant davantage de comorbidités que les autres patients, caractéristiques qui peuvent influencer directement le risque hémorragique.

Pour limiter le biais d'indication, notre choix s'est porté sur une analyse pondérée sur un SP. Nous souhaitions évaluer le risque hémorragique moyen dans la population (*ATE*). Nous avons mis en évidence une augmentation du risque d'hémorragie grave chez les patients traités, qu'il s'agisse d'un traitement par AAP seul (HR 5,52 [IC95% 3,11 – 9,80]), par AC seul (HR 4,15 [IC95% 3,46 – 4,99]) ou d'une association AAP+AC (HR 5,59 [IC95% 2,62 – 11,91]). On remarque que le risque d'hémorragie grave lié aux traitements a semblé plus important dans l'analyse pondérée sur le SP que dans l'analyse brute (HR respectifs pour AAP, AC et AAP+AC : 2,77 [IC95% 1,12 – 6,87], 3,62 [IC95% 0,85 – 15,43] et 5,13 [IC95% 1,20 – 23,58]). Cela s'explique probablement par le fait que le risque hémorragique est diminué par des comorbidités pro-thromboemboliques plus fréquentes chez les patients traités ; ainsi, lorsque l'analyse prend en compte ces comorbidités, le risque hémorragique lié aux traitements apparaît plus important.

Dans cette étude, le faible nombre d'événements observés (70 hémorragies graves au total) n'autorisait pas la prise en compte par une méthode d'ajustement conventionnel d'un nombre aussi important de facteurs de confusion que ne l'a permis l'analyse avec SP. En effet, en se référant à la recommandation de ne pas avoir moins de 10 événements observés par covariables d'ajustement (Peduzzi, Concato et al. 1995), seules 7 variables auraient pu être utilisées dans un ajustement conventionnel. Or, 15 variables ont été utilisées pour calculer le SP.

Concernant le choix de la méthode d'utilisation du SP, hormis la justification théorique relative type d'effet que l'on a souhaité évaluer (évaluation de l'effet moyen dans la population *ATE* par pondération inverse sur le SP), il faut souligner qu'un appariement sur le SP aurait vraisemblablement diminué encore les effectifs analysés, entrainant une diminution de puissance possiblement rédhibitoire.

DISCUSSION ET PERSPECTIVES

Comme nous l'avons vu dans la première partie de ce mémoire, le recours aux études observationnelles s'avère nécessaire (i) pour conforter, en situation réelle, les résultats issus des ECR dont la validité externe est limitée, et (ii) dans des situations, notamment lorsqu'il s'agit d'interventions complexes, où l'ECR n'est pas toujours pertinent ou réalisable pour des questions éthiques et/ou organisationnelles. Les trois applications réalisées dans la seconde partie ont bien illustré la pertinence des études observationnelles pour l'évaluation des interventions en santé. Toutes trois avaient l'objectif de compléter les données d'évaluation insuffisantes ou non concluantes fournies par les ECR. Dans la première application, il s'agissait d'évaluer l'efficacité d'une intervention complexe sur la survie, un réseau de soins spécialisé dans l'insuffisance cardiaque. La généralisation des résultats obtenus par ECR pour évaluer de telles interventions, fortement dépendantes du contexte, est sujette à caution, et l'évaluation menée dans le cadre de la pratique courante est essentielle. Dans la deuxième application, l'objet de l'évaluation était la stratégie médicamenteuse recommandée dans l'insuffisance cardiaque. Au-delà de la prescription, l'effet évalué intègre l'ensemble de la chaîne causale complexe qui débute par la prescription du médicament et se termine par les conséquences mesurées sur la santé des individus. De nombreux facteurs intermédiaires influent sur le résultat mesuré, comme ceux qui modulent l'observance du traitement (un traitement prescrit n'est pas nécessairement pris par le patient). Il était donc intéressant de chercher à savoir si l'efficacité d'une stratégie médicamenteuse, démontrée dans le cadre particulier des ECR, était également présente lorsqu'elle était utilisée dans la pratique courante. Enfin, la troisième application était consacrée à l'évaluation d'un risque provoqué par la prise de médicaments anticoagulants oraux et antiagrégants plaquettaires par des

patients hémodialysés. Les ECR n'avaient pas permis de clarifier le risque d'hémorragie lié à ces traitements. Cette question se prêtait donc à la réalisation d'une étude observationnelle comparative, d'autant plus que ces traitements sont habituellement prescrits à des patients « complexes » présentant plusieurs pathologies pouvant modifier le risque évalué, qui ne sont pas nécessairement inclus dans des ECR classiques (du fait de l'existence de critères d'exclusion).

Les trois applications ont donc bien montré l'intérêt d'évaluer les interventions de santé dans un cadre observationnel, mais également le risque inhérent que l'analyse soit faussée par un biais d'indication. En effet, les groupes comparés ont été constitués spontanément et n'étaient pas comparables pour de nombreux facteurs confondants potentiels. Pour limiter le biais d'indication, différentes méthodes d'analyse ont été passées en revue dans la partie théorique de ce mémoire. Les deux premières à envisager sont l'ajustement multivarié et l'utilisation d'un SP. La seconde possède quelques avantages théoriques sur la première (*cf.* §A.5.2.1.5, p.53), alors que les deux fournissent souvent des estimations très proches, parce qu'elles prennent en compte les mêmes facteurs de confusion mesurés. Si le SP ne paraît pas améliorer résolument la justesse de l'estimation de l'effet par rapport aux méthodes conventionnelles, la philosophie de la méthode n'en demeure pas moins séduisante : elle met au cœur de la réflexion le biais d'indication, met en avant la discussion sur le type et l'homogénéité de l'effet évalué, et rend compte du redressement obtenu dans la comparabilité des groupes après prise en compte du score. Nous avons utilisé dans les trois applications une méthode basée sur un SP. Cette méthode a fait l'objet de nombreux développements théoriques (*cf.* §A.5.2.1, p.41), les modalités de sa mise en œuvre, ses forces et ses limites étant aujourd'hui précisément établies. Parmi les quatre modalités d'utilisation du SP, les plus performantes sont l'appariement et la pondération inverse. Dans les applications, nous avons donc considéré ces deux méthodes, le choix de l'une ou l'autre ayant été guidé par le type d'effet

que l'on souhaitait évaluer (*cf.* §A.5.2.1.4, p.49). Pour l'évaluation du réseau de soins spécialisé dans l'insuffisance cardiaque et celle des prescriptions médicamenteuses dans cette même pathologie, l'effet d'intérêt était l'effet moyen dans la population (*ATE*), considérant que l'ensemble des individus pouvaient potentiellement bénéficier de l'intervention évaluée. Cet effet a donc été évalué par pondération inverse. L'évaluation du risque hémorragique de certains médicaments prescrits chez des patients hémodialysé est une situation différente : on a voulu comparer le risque des individus traités à celui des non traités. C'est donc l'effet moyen chez les traités (*ATT*) que l'on a cherché à évaluer, en utilisant un appariement sur le SP.

Si les méthodes basées sur un SP peuvent être considérées aujourd'hui comme des méthodes éprouvées pour limiter le biais d'indication, leur principale limite est qu'elles ne prennent pas en compte les facteurs de confusion non mesurés. A cet égard, l'importation récente des méthodes utilisant une VI dans le domaine de l'évaluation en santé a suscité beaucoup d'enthousiasme, car ces méthodes permettent théoriquement de prendre en compte l'ensemble des facteurs de confusion, même ceux qui ne sont pas mesurés ni même connus. Cependant, le passage de la théorie à la pratique n'est pas dénué d'obstacles, et il n'est jamais certain de parvenir à identifier un instrument valide qui n'exposerait pas l'analyse à des biais non prévisibles. L'utilisation des VI en épidémiologie est encore balbutiante, des développements sont encore nécessaires pour mieux cerner ses avantages et limites, et donc ses indications. Dans cette perspective, on peut saluer la contribution récente de Jackson et Swanson proposant une méthode pour comparer le biais de confusion de différentes analyses, afin d'interroger la pertinence d'utiliser une VI par rapport à un ajustement conventionnel (Jackson and Swanson 2015). Compte tenu de ces réserves, nous avons testé l'utilisation d'une VI dans la première application en tant qu'analyse de sensibilité. Nous avons identifié un instrument théoriquement valide, du moins sans argument statistique pour l'invalider. La

conclusion de cette analyse allait dans le même sens que celle de l'analyse avec SP, mais l'estimation était accompagnée d'une large incertitude, attendue en raison de l'effectif de notre étude.

L'intérêt des études observationnelles pour l'évaluation des interventions en santé est aujourd'hui largement admis mais la question de la maîtrise du biais d'indication demeure imparfaitement prise en compte. La recherche sur les méthodes d'analyse a récemment opéré des avancées notables (clarification des propriétés des méthodes utilisant un SP, importation de la VI en épidémiologie), mais on ne dispose pas encore de la méthode parfaite qui supprimerait le biais d'indication.

Au-delà des méthodes appliquées pour analyser leurs données, les études observationnelles doivent davantage prendre en compte la complexité des interventions évaluées. Par exemple, nous avons vérifié que, dans la pratique courante, l'inclusion de patients dans un réseau de soin se traduisait bien par un effet bénéfique sur la santé des individus. C'est un résultat intéressant en termes de santé publique, qui apporte une information complémentaire aux résultats d'ECR. Cette évaluation prend implicitement en compte toute la complexité de la chaîne causale entre l'inclusion dans le réseau et la mesure des effets sur la santé. Cependant, cette chaîne causale peut être assimilée à une « boîte noire » dont on ne peut expliquer comment ses effets ont été produits. Pour aller plus loin dans l'interprétation des résultats, il aurait été intéressant de disposer de données supplémentaires concernant les processus de l'intervention, telles que la réalisation des différentes activités du réseau de soins et la participation des individus. Ces données auraient par exemple pu servir à construire un indicateur assimilable à une dose d'intervention et utilisable dans l'analyse pour tester l'amplitude de l'effet en fonction de la dose reçue (Legrand, Bonsergent et al. 2012). Les recommandations du *MRC* sur l'évaluation des processus des interventions complexes, publiées en 2015, suggèrent ainsi d'intégrer des données de l'évaluation des processus à

l'analyse des effets de l'intervention (Moore, Audrey et al. 2015). Une telle évaluation implique donc d'avoir mis en œuvre au préalable un dispositif de recueil de données tout au long de l'étude. De plus, une meilleure connaissance des processus pourrait aussi permettre d'analyser la part respective de chaque composante de l'intervention complexe dans l'effet global observé et d'identifier d'éventuelles interactions entre ces composantes. Dans le cadre expérimental, une proposition encore peu pratiquée est la réalisation de plans factoriels (Montgomery, Astin et al. 2011). En situation observationnelle, nous n'avons pas connaissance qu'une méthodologie cherchant à atteindre un tel objectif n'ait encore été mise en œuvre.

Ainsi, il faudrait envisager l'élaboration de nouveaux schémas d'études quasi-expérimentales dont l'analyse des effets intégrerait des éléments relatifs aux processus des interventions complexes.

CONCLUSION

Ce travail de thèse a permis de souligner l'intérêt des études observationnelles pour l'évaluation des interventions en santé et de dresser un état de l'art sur les méthodes d'analyse applicables pour maîtriser le biais d'indication. Le catalogue des méthodes disponibles, longtemps réduit aux méthodes d'ajustement multivarié, s'est enrichi de méthodes utilisant un SP ou une VI. Si les méthodes utilisant un SP ont permis quelques progrès méthodologiques, elles ne préviennent pas l'existence d'un biais résiduel lié à des facteurs de confusion non mesurés. Les méthodes utilisant une VI pourraient pallier cette limite, mais elles ne sont pas encore suffisamment développées en épidémiologie et leur indication demeure limitée. Les travaux de recherche doivent donc être poursuivis pour améliorer les stratégies d'analyse, notamment en y intégrant des éléments de complexité des interventions.

BIBLIOGRAPHIE

Agrinier, N., C. Altieri, F. Alla, N. Jay, D. Dobre, N. Thilly and F. Zannad (2013). "Effectiveness of a multidimensional home nurse led heart failure disease management program--a French nationwide time-series comparison." Int J Cardiol **168**(4): 3652-3658.

Ali, M. S., R. H. Groenwold, S. V. Belitser, W. R. Pestman, A. W. Hoes, K. C. Roes, A. Boer and O. H. Klungel (2015). "Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: a systematic review." J Clin Epidemiol **68**(2): 112-121.

Ali, M. S., M. J. Uddin, R. H. Groenwold, W. R. Pestman, S. V. Belitser, A. W. Hoes, A. de Boer, K. C. Roes and O. H. Klungel (2014). "Quantitative falsification of instrumental variables assumption using balance measures." Epidemiology **25**(5): 770-772.

Austin, P. C. (2007). "Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: a systematic review and suggestions for improvement." J Thorac Cardiovasc Surg **134**(5): 1128-1135.

Austin, P. C. (2008). "A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003." Stat Med **27**(12): 2037-2049.

Austin, P. C. (2008). "Goodness-of-fit diagnostics for the propensity score model when estimating treatment effects using covariate adjustment with the propensity score." Pharmacoepidemiol Drug Saf **17**(12): 1202-1217.

Austin, P. C. (2008). "Primer on statistical interpretation or methods report card on propensity-score matching in the cardiology literature from 2004 to 2006: a systematic review." Circ Cardiovasc Qual Outcomes **1**(1): 62-67.

Austin, P. C. (2009). "Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples." Stat Med **28**(25): 3083-3107.

Austin, P. C. (2009). "Some methods of propensity-score matching had superior performance to others: results of an empirical investigation and Monte Carlo simulations." Biom J **51**(1): 171-184.

Austin, P. C. (2010). "Statistical criteria for selecting the optimal number of untreated subjects matched to each treated subject when using many-to-one matching on the propensity score." Am J Epidemiol **172**(9): 1092-1097.

Austin, P. C. (2011). "Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies." Pharm Stat **10**(2): 150-161.

Austin, P. C. (2013). "The performance of different propensity score methods for estimating marginal hazard ratios." Stat Med **32**(16): 2837-2849.

Austin, P. C. (2014). "The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments." Stat Med **33**(7): 1242-1258.

Austin, P. C., P. Grootendorst and G. M. Anderson (2007). "A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study." Stat Med **26**(4): 734-753.

Austin, P. C. and M. M. Mamdani (2006). "A comparison of propensity score methods: a case-study estimating the effectiveness of post-AMI statin use." Stat Med **25**(12): 2084-2106.

Austin, P. C., M. M. Mamdani, T. A. Stukel, G. M. Anderson and J. V. Tu (2005). "The use of the propensity score for estimating treatment effects: administrative versus clinical data." Stat Med **24**(10): 1563-1578.

Austin, P. C. and T. Schuster (2014). "The performance of different propensity score methods for estimating absolute effects of treatments on survival outcomes: A simulation study." Stat Methods Med Res.

Austin, P. C. and E. A. Stuart (2015). "Optimal full matching for survival outcomes: a method that merits more widespread use." Stat Med **34**(30): 3949-3967.

Austin, P. C. and E. A. Stuart (2015). "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.

Baiocchi, M., J. Cheng and D. S. Small (2014). "Instrumental variable methods for causal inference." Stat Med **33**(13): 2297-2340.

Barreto, M. L. (2005). "Efficacy, effectiveness, and the evaluation of public health interventions." J Epidemiol Community Health **59**(5): 345-346.

Baser, O. (2009). "Too much ado about instrumental variable approach: is the cure worse than the disease?" Value Health **12**(8): 1201-1209.

Black, N. (1996). "Why we need observational studies to evaluate the effectiveness of health care." BMJ **312**(7040): 1215-1218.

Boef, A. G., O. M. Dekkers, S. le Cessie and J. P. Vandenbroucke (2013). "Reporting instrumental variable analyses." Epidemiology **24**(6): 937-938.

Boef, A. G., O. M. Dekkers, J. P. Vandenbroucke and S. le Cessie (2014). "Sample size importantly limits the usefulness of instrumental variable methods, depending on instrument strength and level of confounding." J Clin Epidemiol **67**(11): 1258-1264.

Boef, A. G., S. le Cessie, O. M. Dekkers, P. Frey, P. M. Kearney, N. Kerse, C. D. Mallen, V. J. McCarthy, S. P. Mooijaart, C. Muth, N. Rodondi, T. Rosemann, A. Russell, H. Schers, V. Virgini, M. W. de Waal, A. Warner, J. Gussekloo and W. P. den Elzen (2016). "Physician's Prescribing Preference

as an Instrumental Variable: Exploring Assumptions Using Survey Data." Epidemiology **27**(2): 276-283.

Bound, J., D. A. Jaeger and R. M. Baker (1995). "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak." Journal of the American Statistical Association **90**(430): 443-450.

Boutron, I., F. Tubach, B. Giraudeau and P. Ravaud (2004). "Blinding was judged more difficult to achieve and maintain in nonpharmacologic than pharmacologic trials." J Clin Epidemiol **57**(6): 543-550.

Breart, G. and J. Bouyer (1991). "[Epidemiological methods in evaluation]." Rev Epidemiol Sante Publique **39 Suppl 1**: S5-14.

Brewin, C. R. and C. Bradley (1989). "Patient preferences and randomised clinical trials." BMJ **299**(6694): 313-315.

Brookhart, M. A., J. A. Rassen and S. Schneeweiss (2010). "Instrumental variable methods in comparative safety and effectiveness research." Pharmacoepidemiol Drug Saf **19**(6): 537-554.

Brookhart, M. A. and S. Schneeweiss (2007). "Preference-based instrumental variable methods for the estimation of treatment effects: assessing validity and interpreting results." Int J Biostat **3**(1): Article 14.

Brookhart, M. A., S. Schneeweiss, K. J. Rothman, R. J. Glynn, J. Avorn and T. Sturmer (2006). "Variable selection for propensity score models." Am J Epidemiol **163**(12): 1149-1156.

Brookhart, M. A., T. Sturmer, R. J. Glynn, J. Rassen and S. Schneeweiss (2010). "Confounding control in healthcare database research: challenges and potential approaches." Med Care **48**(6 Suppl): S114-120.

Cambon, L., L. Minary, V. Ridde and F. Alla (2012). "Transferability of interventions in health education: a review." BMC Public Health **12**: 497.

Cepeda, M. S., R. Boston, J. T. Farrar and B. L. Strom (2003). "Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders." Am J Epidemiol **158**(3): 280-287.

Chen, Y. and B. A. Briesacher (2011). "Use of instrumental variable in prescription drug research with observational data: a systematic review." J Clin Epidemiol **64**(6): 687-700.

Collette, C., I. Clerc-Urmes, H. Laborde-Casterot, L. Frimat, C. Ayav, N. Peters, A. Martin, N. Agrinier and N. Thilly (2016). "Antiplatelet and oral anticoagulant therapies in chronic hemodialysis patients: prescribing practices and bleeding risk." Pharmacoepidemiol Drug Saf **25**(8): 935-943.

Concato, J., E. V. Lawler, R. A. Lew, J. M. Gaziano, M. Aslan and G. D. Huang (2010). "Observational methods in comparative effectiveness research." Am J Med **123**(12 Suppl 1): e16-23.

Connelly, J. B. (2007). "Evaluating complex public health interventions: theory, methods and scope of realist enquiry." J Eval Clin Pract **13**(6): 935-941.

Contandriopoulos, A. P., F. Champagne, J. L. Denis and M. C. Avargues (2000). "[Evaluation in the health sector: concepts and methods]." Rev Epidemiol Sante Publique **48**(6): 517-539.

Conway, P. H. and C. Clancy (2009). "Comparative-effectiveness research--implications of the Federal Coordinating Council's report." N Engl J Med **361**(4): 328-330.

Cook, E. F. and L. Goldman (1989). "Performance of tests of significance based on stratification by a multivariate confounder score or by a propensity score." J Clin Epidemiol **42**(4): 317-324.

Craig, P., C. Cooper, D. Gunnell, S. Haw, K. Lawson, S. Macintyre, D. Ogilvie, M. Petticrew, B. Reeves, M. Sutton and S. Thompson (2012). "Using natural experiments to evaluate population health interventions: new Medical Research Council guidance." J Epidemiol Community Health **66**(12): 1182-1186.

Craig, P., P. Dieppe, S. Macintyre, S. Michie, I. Nazareth, M. Petticrew and G. Medical Research Council (2008). "Developing and evaluating complex interventions: the new Medical Research Council guidance." BMJ **337**: a1655.

Crown, W. H., H. J. Henk and D. J. Vanness (2011). "Some cautions on the use of instrumental variables estimators in outcomes research: how bias in instrumental variables estimators is affected by instrument strength, instrument contamination, and sample size." Value Health **14**(8): 1078-1084.

Cummings, P., B. McKnight and N. S. Weiss (2003). "Matched-pair cohort methods in traffic crash research." Accid Anal Prev **35**(1): 131-141.

D'Agostino, R. B., Jr. (2007). "Propensity scores in cardiovascular research." Circulation **115**(17): 2340-2343.

Dahabreh, I. J., R. C. Sheldrick, J. K. Paulus, M. Chung, V. Varvarigou, H. Jafri, J. A. Rassen, T. A. Trikalinos and G. D. Kitsios (2012). "Do observational studies using propensity score methods agree with randomized trials? A systematic comparison of studies on acute coronary syndromes." Eur Heart J **33**(15): 1893-1901.

Davies, N. M., G. D. Smith, F. Windmeijer and R. M. Martin (2013). "Issues in the reporting and conduct of instrumental variable studies: a systematic review." Epidemiology **24**(3): 363-369.

Deb, S., P. C. Austin, J. V. Tu, D. T. Ko, C. D. Mazer, A. Kiss and S. E. Fremes (2016). "A Review of Propensity-Score Methods and Their Use in Cardiovascular Research." Can J Cardiol **32**(2): 259-265.

Drake, C. (1993). "Effects of Misspecification of the Propensity Score on Estimators of Treatment Effect." Biometrics **49**(4): 1231-1236.

Faillie, J. L. and S. Suissa (2015). "[Immortal time bias in pharmacoepidemiological studies: definition, solutions and examples]." Therapie **70**(3): 259-263.

Fang, G., J. M. Brooks and E. A. Chrischilles (2012). "Apples and oranges? Interpretations of risk adjustment and instrumental variable estimates of intended treatment effects using observational data." Am J Epidemiol **175**(1): 60-65.

Fischer, K. G. (2007). "Essentials of anticoagulation in hemodialysis." Hemodial Int **11**(2): 178-189.

Gail, M. H. (1972). "Does cardiac transplantation prolong life? A reassessment." Ann Intern Med **76**(5): 815-817.

Gale, E. A. (2004). "The Hawthorne studies-a fable for our times?" QJM **97**(7): 439-449.

Garabedian, L. F., P. Chu, S. Toh, A. M. Zaslavsky and S. B. Soumerai (2014). "Potential bias of instrumental variable analyses for observational comparative effectiveness research." Ann Intern Med **161**(2): 131-138.

Glynn, R. J., S. Schneeweiss and T. Sturmer (2006). "Indications for propensity scores and review of their use in pharmacoepidemiology." Basic Clin Pharmacol Toxicol **98**(3): 253-259.

Godwin, M., L. Ruhland, I. Casson, S. MacDonald, D. Delva, R. Birtwhistle, M. Lam and R. Seguin (2003). "Pragmatic controlled clinical trials in primary care: the struggle between external and internal validity." BMC Med Res Methodol **3**: 28.

Greenland, S. (1996). "Basic methods for sensitivity analysis of biases." Int J Epidemiol **25**(6): 1107-1116.

Greenland, S. (2000). "An introduction to instrumental variables for epidemiologists." Int J Epidemiol **29**(4): 722-729.

Grootendorst, P. (2007). "A review of instrumental variables estimation of treatment effects in the applied health sciences." Health Services and Outcomes Research Methodology **7**(3): 159-179.

Guertin, J. R., E. Rahme, C. R. Dormuth and J. LeLorier (2016). "Head to head comparison of the propensity score and the high-dimensional propensity score matching methods." BMC Med Res Methodol **16**: 22.

Hajage, D., F. Tubach, P. G. Steg, D. L. Bhatt and Y. De Rycke (2016). "On the use of propensity scores in case of rare exposure." BMC Med Res Methodol **16**: 38.

Hemming, K., T. P. Haines, P. J. Chilton, A. J. Girling and R. J. Lilford (2015). "The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting." BMJ **350**: h391.

Hernan, M. A. and J. M. Robins (2006). "Instruments for causal inference: an epidemiologist's dream?" Epidemiology **17**(4): 360-372.

Hershman, D. L. and J. D. Wright (2012). "Comparative effectiveness research in oncology methodology: observational data." J Clin Oncol **30**(34): 4215-4222.

Hirano, K. and G. W. Imbens (2001). "Estimation of Causal Effects using Propensity Score Weighting: An Application to Data on Right Heart Catheterization." Health Services and Outcomes Research Methodology **2**(3): 259-278.

Ho, D. E., K. Imai, G. King and E. A. Stuart (2007). "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." Political Analysis **15**(3): 199-236.

Imbens, G. (2000). "The role of the propensity score in estimating dose-response functions." Biometrika **87**(3): 706-710.

Imbens, G. W. and J. D. Angrist (1994). "Identification and Estimation of Local Average Treatment Effects." Econometrica **62**(2): 467-475.

IOM (2009). Initial national priorities for comparative effectiveness research: report brief.

Jackson, J. W. and S. A. Swanson (2015). "Toward a clearer portrayal of confounding bias in instrumental variable applications." Epidemiology **26**(4): 498-504.

Jamal Uddin, M. and O. Klungel (2015). "Instrumental Variable Analysis in Epidemiologic Studies: An Overview of the Estimation Methods." Pharm Anal Acta **06**(04): 1-9.

Klungel, O. H., M. Jamal Uddin, A. de Boer, S. V. Belitser, R. H. Groenwold and K. C. Roes (2015). "Instrumental Variable Analysis in Epidemiologic Studies: An Overview of the Estimation Methods." Pharm Anal Acta **6**(4): 1-9.

Kurth, T., A. M. Walker, R. J. Glynn, K. A. Chan, J. M. Gaziano, K. Berger and J. M. Robins (2006). "Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect." Am J Epidemiol **163**(3): 262-270.

Laborde-Casterot, H., N. Agrinier and N. Thilly (2015). "Performing both propensity score and instrumental variable analyses in observational studies often leads to discrepant results: a systematic review." J Clin Epidemiol **68**(10): 1232-1240.

Laborde-Casterot, H., N. Agrinier, F. Zannad, A. Mebazaa, P. Rossignol, N. Girerd, F. Alla and N. Thilly (2016). "Effectiveness of a multidisciplinary heart failure disease management programme on 1-year mortality: Prospective cohort study." Medicine (Baltimore) **95**(37): e4399.

Legrand, K., E. Bonsergent, C. Latarche, F. Empereur, J. F. Collin, E. Lecomte, E. Aptel, N. Thilly and S. Briancon (2012). "Intervention dose estimation in health promotion programmes: a framework and a tool. Application to the diet and physical activity promotion PRALIMAP trial." BMC Med Res Methodol **12**: 146.

Lin, D. Y. and L. J. Wei (1989). "The Robust Inference for the Cox Proportional Hazards Model." Journal of the American Statistical Association **84**(408): 1074-1078.

Lunt, M. (2014). "Selecting an appropriate caliper can be essential for achieving good balance with propensity score matching." Am J Epidemiol **179**(2): 226-235.

Marchal, B., G. Westhorp, G. Wong, S. Van Belle, T. Greenhalgh, G. Kegels and R. Pawson (2013). "Realist RCTs of complex interventions - an oxymoron." Soc Sci Med **94**: 124-128.

Martens, E. P., W. R. Pestman, A. de Boer, S. V. Belitser and O. H. Klungel (2006). "Instrumental variables: application and limitations." Epidemiology **17**(3): 260-267.

Martens, E. P., W. R. Pestman, A. de Boer, S. V. Belitser and O. H. Klungel (2008). "Systematic differences in treatment effect estimates between propensity score methods and logistic regression." Int J Epidemiol **37**(5): 1142-1147.

Montgomery, A. A., M. P. Astin and T. J. Peters (2011). "Reporting of factorial trials of complex interventions in community settings: a systematic review." Trials **12**: 179.

Moore, G. F., S. Audrey, M. Barker, L. Bond, C. Bonell, W. Hardeman, L. Moore, A. O'Cathain, T. Tinati, D. Wight and J. Baird (2015). "Process evaluation of complex interventions: Medical Research Council guidance." BMJ **350**: h1258.

Olschewski, M., M. Schumacher and K. B. Davis (1992). "Analysis of randomized and nonrandomized patients in clinical trials using the comprehensive cohort follow-up study design." Control Clin Trials **13**(3): 226-239.

Peduzzi, P., J. Concato, A. R. Feinstein and T. R. Holford (1995). "Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates." J Clin Epidemiol **48**(12): 1503-1510.

Peduzzi, P., J. Concato, E. Kemper, T. R. Holford and A. R. Feinstein (1996). "A simulation study of the number of events per variable in logistic regression analysis." J Clin Epidemiol **49**(12): 1373-1379.

Permutt, T. and J. R. Hebel (1989). "Simultaneous-equation estimation in a clinical trial of the effect of smoking on birth weight." Biometrics **45**(2): 619-622.

Petticrew, M., S. Cummins, C. Ferrell, A. Findlay, C. Higgins, C. Hoy, A. Kearns and L. Sparks (2005). "Natural experiments: an underused tool for public health?" Public Health **119**(9): 751-757.

Pirracchio, R., M. Carone, M. R. Rigon, E. Caruana, A. Mebazaa and S. Chevret (2013). "Propensity score estimators for the average treatment effect and the average treatment effect on the treated may yield very different estimates." Stat Methods Med Res.

Ponikowski, P., A. A. Voors, S. D. Anker, H. Bueno, J. G. Cleland, A. J. Coats, V. Falk, J. R. Gonzalez-Juanatey, V. P. Harjola, E. A. Jankowska, M. Jessup, C. Linde, P. Nihoyannopoulos, J. T. Parissis, B. Pieske, J. P. Riley, G. M. Rosano, L. M. Ruilope, F. Ruschitzka, F. H. Rutten, P. van der Meer and M. Authors/Task Force (2016). "2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: The Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC)Developed with the special contribution of the Heart Failure Association (HFA) of the ESC." Eur Heart J **37**(27): 2129-2200.

Rassen, J. A., S. Schneeweiss, R. J. Glynn, M. A. Mittleman and M. A. Brookhart (2009). "Instrumental variable analysis for estimation of treatment effects with dichotomous outcomes." Am J Epidemiol **169**(3): 273-284.

Relton, C., D. Torgerson, A. O'Cathain and J. Nicholl (2010). "Rethinking pragmatic randomised controlled trials: introducing the "cohort multiple randomised controlled trial" design." BMJ **340**: c1066.

Ridde, V. and S. Haddad (2013). "[Pragmatism and realism for public health intervention evaluation]." Rev Epidemiol Sante Publique **61 Suppl 2**: S95-106.

Robins, J. M. and S. Greenland (1996). "Identification of Causal Effects Using Instrumental Variables: Comment." Journal of the American Statistical Association **91**(434): 456-458.

Rosenbaum, P. and D. Rubin (1983). "The central role of the propensity score in observational studies for causal effects." Biometrika **70**(1): 41-55.

Rosenbaum, P. R. and D. B. Rubin (1984). "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." Journal of the American Statistical Association **79**(387): 516-524.

Rubin, D. B. (1997). "Estimating causal effects from large data sets using propensity scores." Ann Intern Med **127**(8 Pt 2): 757-763.

Rubin, D. B. (2004). "On principles for modeling propensity scores in medical research." Pharmacoepidemiol Drug Saf **13**(12): 855-857.

Salas, M., A. Hofman and B. H. Stricker (1999). "Confounding by indication: an example of variation in the use of epidemiologic terminology." Am J Epidemiol **149**(11): 981-983.

Salkind, N. J. (2010). Encyclopedia of research design. Thousand Oaks, Calif., SAGE Publications.

Sato, T. and Y. Matsuyama (2003). "Marginal structural models as a tool for standardization." Epidemiology **14**(6): 680-686.

Schneeweiss, S., J. A. Rassen, R. J. Glynn, J. Avorn, H. Mogun and M. A. Brookhart (2009). "High-dimensional propensity score adjustment in studies of treatment effects using health care claims data." Epidemiology **20**(4): 512-522.

Shah, B. R., A. Laupacis, J. E. Hux and P. C. Austin (2005). "Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review." J Clin Epidemiol **58**(6): 550-559.

Small, D. S. and P. R. Rosenbaum (2008). "War and wages: the strength of instrumental variables and their sensitivity to unobserved biases." Journal of the American Statistical Association **103**(483): 924-933.

Smith, G. C. and J. P. Pell (2003). "Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials." BMJ **327**(7429): 1459-1461.

Sox, H. C. (2010). "Defining comparative effectiveness research: the importance of getting it right." Med Care **48**(6 Suppl): S7-8.

Sturmer, T., M. Joshi, R. J. Glynn, J. Avorn, K. J. Rothman and S. Schneeweiss (2006). "A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods." J Clin Epidemiol **59**(5): 437-447.

Sturmer, T., K. J. Rothman, J. Avorn and R. J. Glynn (2010). "Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution--a simulation study." Am J Epidemiol **172**(7): 843-854.

Sturmer, T., S. Schneeweiss, J. Avorn and R. J. Glynn (2005). "Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration." Am J Epidemiol **162**(3): 279-289.

Sturmer, T., S. Schneeweiss, M. A. Brookhart, K. J. Rothman, J. Avorn and R. J. Glynn (2005). "Analytic strategies to adjust confounding using exposure propensity scores and disease risk scores: nonsteroidal antiinflammatory drugs and short-term mortality in the elderly." Am J Epidemiol **161**(9): 891-898.

Sturmer, T., R. Wyss, R. J. Glynn and M. A. Brookhart (2014). "Propensity scores for confounder adjustment when assessing the effects of medical interventions using nonexperimental study designs." J Intern Med **275**(6): 570-580.

Suissa, S. (2007). "Immortal time bias in observational studies of drug effects." Pharmacoepidemiol Drug Saf **16**(3): 241-249.

Suissa, S. (2008). "Immortal time bias in pharmaco-epidemiology." Am J Epidemiol **167**(4): 492-499.

Swanson, S. A. and M. A. Hernan (2013). "Commentary: how to report instrumental variable analyses (suggestions welcome)." Epidemiology **24**(3): 370-374.

Swanson, S. A. and M. A. Hernan (2014). "Think globally, act globally: An epidemiologist's perspective on instrumental variable estimation." Stat Sci **29**(3): 371-374.

Swanson, S. A., M. Miller, J. M. Robins and M. A. Hernan (2015). "Definition and evaluation of the monotonicity condition for preference-based instruments." Epidemiology **26**(3): 414-420.

Swanson, S. A., J. M. Robins, M. Miller and M. A. Hernan (2015). "Selecting on treatment: a pervasive form of bias in instrumental variable analyses." Am J Epidemiol **181**(3): 191-197.

Takeda, A., S. J. Taylor, R. S. Taylor, F. Khan, H. Krum and M. Underwood (2012). "Clinical service organisation for heart failure." Cochrane Database Syst Rev(9): CD002752.

Terza, J. V., A. Basu and P. J. Rathouz (2008). "Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling." J Health Econ **27**(3): 531-543.

Torgerson, D. J. and B. Sibbald (1998). "Understanding controlled trials. What is a patient preference trial?" BMJ **316**(7128): 360.

Treweek, S. and M. Zwarenstein (2009). "Making trials matter: pragmatic and explanatory trials and the problem of applicability." Trials **10**: 37.

Trietsch, J., P. Leffers, B. van Steenkiste, R. Grol and T. van der Weijden (2014). "The balanced incomplete block design is not suitable for the evaluation of complex interventions." J Clin Epidemiol **67**(12): 1295-1298.

Uddin, M. J., R. H. Groenwold, M. S. Ali, A. de Boer, K. C. Roes, M. A. Chowdhury and O. H. Klungel (2016). "Methods to control for unmeasured confounding in pharmacoepidemiology: an overview." Int J Clin Pharm **38**(3): 714-723.

Walker, A. M. (1996). "Confounding by indication." Epidemiology **7**(4): 335-336.

Wan, F. and N. Mitra (2016). "An evaluation of bias in propensity score-adjusted non-linear regression models." Stat Methods Med Res.

Weitzen, S., K. L. Lapane, A. Y. Toledano, A. L. Hume and V. Mor (2004). "Principles for modeling propensity scores in medical research: a systematic literature review." Pharmacoepidemiol Drug Saf **13**(12): 841-853.

Westreich, D., S. R. Cole, M. J. Funk, M. A. Brookhart and T. Sturmer (2011). "The role of the c-statistic in variable selection for propensity score models." Pharmacoepidemiol Drug Saf **20**(3): 317-320.

Zelen, M. (1990). "Randomized consent designs for clinical trials: an update." Stat Med **9**(6): 645-656.

LISTE DES TRAVAUX EFFECTUÉS AU COURS DE LA THÈSE

Publications acceptées

Laborde-Casterot, H., N. Agrinier and N. Thilly (2015). "Performing both propensity score and instrumental variable analyses in observational studies often leads to discrepant results: a systematic review." J Clin Epidemiol 68(10): 1232-1240.

Laborde-Casterot, H., N. Agrinier, F. Zannad, A. Mebazaa, P. Rossignol, N. Girerd, F. Alla and N. Thilly (2016). "Effectiveness of a multidisciplinary heart failure disease management programme on 1-year mortality: Prospective cohort study." Medicine (Baltimore) 95(37): e4399.

Collette, C., I. Clerc-Urmes, H. Laborde-Casterot, L. Frimat, C. Ayav, N. Peters, A. Martin, N. Agrinier and N. Thilly (2016). "Antiplatelet and oral anticoagulant therapies in chronic hemodialysis patients: prescribing practices and bleeding risk." Pharmacoepidemiol Drug Saf 25(8): 935-943.

Publication soumise

Busson, A., H. Laborde-Casterot, F. Alla, Z. Messikh, I. Clerc-Urmes, A. Mebazaa, N. Thilly and N. Agrinier. Effectiveness of guideline-consistent heart failure drug prescriptions at hospital discharge on 1-year mortality: results from the EPICAL2 cohort study. BMJ Quality & Safety.

Actes de congrès

Laborde-Castérot, H., N. Agrinier, I. Clerc-Urmès and N. Thilly (2015). Analyse des études observationnelles : score de propension et méthode des variables instrumentales fournissent-ils des résultats concordants ? Revue de la littérature. Rev Epidemiol Sante Publique 63(S2); S71.

Agrinier N., M. Schockmel, N. Thilly, H. Laborde-Castérot, P. Jourdain, C. Leclercq, F. Dany, J. Druelle, G. Mulak and Y. Juillière. Efficacité d'un programme d'éducation thérapeutique sur la survie des patients insuffisants cardiaques à fraction d'éjection préservée. Rev Epidemiol Sante Publique 63(S2); S89.

Résumé

La médecine fondée sur les preuves a conféré à l'essai contrôlé randomisé (ECR) le plus haut niveau de preuve dans l'évaluation de l'effet des médicaments, et par extension de toute intervention en santé. Cependant, le recours aux études observationnelles s'avère également nécessaire (i) pour conforter, en situation réelle, les résultats issus des ECR dont la validité externe est limitée, (ii) dans des situations, notamment lorsqu'il s'agit d'interventions complexes, où l'ECR n'est pas toujours réalisable pour des questions éthiques et/ou organisationnelles. Toutefois, les études observationnelles sont sujettes à différents types de biais, et notamment au biais d'indication. Ce travail de thèse explore les différentes techniques d'analyse statistique des résultats permettant de maîtriser ce biais. Dans une première partie, les aspects théoriques ont été abordés. Les différentes techniques disponibles ont été identifiées, analysées et comparées : les techniques d'ajustement multivarié, celles utilisant un score de propension (SP) et celles utilisant une variable instrumentale (VI). Pour approfondir les connaissances sur la question, une revue systématique de la littérature a été effectuée. Elle a mis en évidence la faible concordance entre les résultats obtenus en utilisant un SP et ceux obtenus en utilisant une VI, lorsque ces deux techniques étaient utilisées dans une même étude pour évaluer la même intervention. Dans une seconde partie, l'utilisation de SP et/ou VI a été testée dans trois exemples d'évaluation d'interventions complexes à partir de données de pratiques courantes recueillies dans le cadre de deux études observationnelles de cohorte : (i) l'évaluation de l'effet d'un réseau de soins spécialisé dans l'insuffisance cardiaque (IC) sur la mortalité ; (ii) l'évaluation de l'effet des stratégies médicamenteuses appropriées dans l'IC sur la mortalité ; (iii) l'évaluation de l'effet des stratégies antithrombotiques chez les patients hémodialysés sur le risque hémorragique.

Mots-clés : études observationnelles – évaluation des résultats – biais d'indication – score de propension – variable instrumentale

Abstract

Evidence-based medicine placed randomized controlled trials (RCT) at the highest level of evidence to evaluate the effects of medications and, by extension, of all health interventions. Nevertheless, observational studies are necessary (i) to support, in real-world settings, the results of RCTs, the external validity of which is limited, and (ii) in situations where RCTs are not feasible for ethical or practical reasons, particularly when evaluating complex interventions. However, observational studies are particularly prone to confounding by indication. This thesis focuses on analytical methods to reduce this bias. In its first part, the theoretical aspects were addressed. Available methods were identified, reviewed and compared: multivariate adjustment methods, methods using a propensity score (PS) and methods using an instrumental variable (IV). To further knowledge on this issue, a systematic literature review was performed. This review revealed that more and more observational studies simultaneously use PS and IV approaches to evaluate the same intervention, often leading to nonconcordant results that may be difficult to interpret. In a second part, the use of PS and/or VI methods was tested in three evaluations of complex interventions in real-world settings, using data from two cohort studies: (i) to evaluate the effectiveness on mortality of a community-based multidisciplinary disease management programme for heart failure (HF) patients; (ii) to evaluate the effectiveness of recommended drug prescriptions on mortality in patients with HF; (iii) to evaluate the effect of antiplatelet and anticoagulant therapies on the risk of major bleeding events in chronic hemodialysis patients.

Keywords: observational studies – outcome evaluation – confounding by indication – propensity score – instrumental variable