Démarche d'identification de facteurs de confusion, dans le cadre d'une étude clinique observationnelle visant à identifier les facteurs de risque (ou protecteur) d'une maladie.

Remarque préliminaire

Il est indispensable d'avoir lu et compris tout le contenu du polycopié d'épidémiologie clinique de 3ème année de l'EnvA, disponible ici, avant de vous embarquer dans cette démarche (sans cette compréhension de tout ce polycopié, vous ne pourrez pas parfaitement saisir les tenants et les aboutissants de ce qui est écrit ci-dessous).

Contexte de la démarche

Dans l'étude de l'association causale entre une exposition d'intérêt (E) et une maladie (M), une des conditions pour pouvoir faire de l'inférence causale est celle de la prise en compte de tous les facteurs de confusion [1]. La méthode de référence de prise en compte ces facteurs de confusion est celle de l'ajustement à l'aide d'un modèle multivarié. Cependant, la construction d'un modèle multivarié (c'est-à-dire, choisir les variables pertinentes à inclure dans un modèle) est l'une des choses les plus compliquées en épidémiologie clinique [2], et ne répond pas à une seule et unique règle, mais à une juxtaposition de règles, mêlant les aspects statistiques, épidémiologiques, et médicaux [3]. Il n'est par conséquent pas envisageable de construire un modèle multivarié sans avoir reçu une solide formation en épidémiologie clinique. L'objectif de la démarche décrite ci-dessous permet (ou tente) d'identifier les facteurs de confusion dans l'estimation de l'association causale entre E et M. Elle repose sur la méthode CIE (« Change in estimate ») [4], qui est une démarche purement statistique, alors qu'elle devrait s'accompagner d'une réflexion *a priori* sur l'existence de facteurs de confusion [5, 6]. Pour appliquer cette démarche, il sera nécessaire de réaliser des modèles bivariés (modèles incluant deux variables), et par conséquent d'utiliser un logiciel de statistique capable de réaliser de tels modèles (par exemple, le logiciel gratuit Epi Info[1]). Cette démarche s'applique davantage aux études observationnelles qu'interventionnelles. En effet, dans les études interventionnelles, l'identification des facteurs de confusion s'effectuera en dressant un tableau comparant les individus exposés à l'exposition d'intérêt à ceux non exposés, sur toutes les expositions connues pour être associées à M, ou suspectées de l'être.

Démarche d'identification des facteurs de confusion

La démarche ci-dessous repose sur la méthode CIE qui est la suivante : une exposition X joue un rôle de confusion dans l'association entre E et M si l'ajustement sur X modifie de façon importante (au moins de 20%) l'association initiale, brute, entre E et M.

Si vous avez plusieurs expositions d'intérêt, vous devez répéter cette démarche pour chaque exposition d'intérêt. Dans cet démarche, je vais nommer « facteur de confusion candidat » (FCC) une exposition dont on sait ou dont on pense (pour des raisons physiopathologiques, entre autres) qu'elle est associée à M *et* qui a été recueillie dans l'étude, et je vais nommer « facteur de confusion potentiel » (FCP) un FCC associé à M dans l'échantillon avec un degré de signification inférieur ou égal à 0,20. Notez qu'une exposition dont on a de bonnes raisons de penser qu'elle n'est pas associée à M ne devrait pas faire partie des FCC, donc *a fortiori* ne devrait pas faire partie des FCP. Nous allons supposer que l'on a *P* FCC. Par notation, « OR* » indique « OR ou RR ». Les OR doivent être estimés à l'aide d'une régression logistique, et les RR à l'aide d'un modèle de Cox.

1) Calculez tout d'abord l'$OR^*_{E\rightarrow M}$ brut à l'aide d'un modèle de régression univarié. Notez cet OR, son $IC_{95\%}$, et son degré de signification.

2) Pour chaque FCC_j ($j \in \{1, \ldots, P\}$), calculez l'$OR^*_{FCCj\rightarrow M}$ brut à l'aide d'un modèle de régression univarié, et notez les degrés de signification des *P* $OR^*_{FCCj\rightarrow M}$ bruts dans un tableau.

3) Soient FCP_k les *K* FCC ($k \in \{1, \ldots, K\}$) dont le degré de signification de l'OR* brut estimé à l'étape précédente est inférieur ou égal à 0,20.

[1] Qui ne fonctionne cependant que sur PC.

4) Faites tourner *K* modèles bivariés contenant l'exposition FCP_k et E. Notez chacune des *K* différences relatives des OR entre l'$OR^*_{E\rightarrow M}$ brut et l'$OR^*_{E\rightarrow M}$ ajusté sur FCP_k (Δ_k) :

$$\Delta_k(exprimée\ en\ \%) = \frac{|OR^*_{E\rightarrow M} brut - OR^*_{E\rightarrow M} ajusté\ sur\ FCP_k|}{OR^*_{E\rightarrow M} brut} \times 100$$

5) Si $\Delta \geq 20\%$, alors on peut considérer que l'exposition FCP correspondante a joué un fort rôle de confusion dans l'association entre E et M.

6) A partir de ces différences relatives Δ, plusieurs possibilités.

6.1) Si aucune des Δ_k n'est supérieure ou égale à 20%, alors l'$OR^*_{E\rightarrow M}$ brut n'est pas fortement biaisé par du biais de confusion dû aux FCC envisagés. Cela ne veut pas dire que cet $OR^*_{E\rightarrow M}$ brut est proche de l'$OR^*_{E\rightarrow M}$ causal, car chacun des FCP peut jouer un « petit » rôle de confusion, et la somme de ces « petits » rôles de confusion peut conduire à un « grand » rôle de confusion. De plus, il peut exister d'autres facteurs de confusion non recueillis, ou non envisagés, donc de fait non pris en compte dans cette démarche d'identification. Ainsi, il est quand même difficilement envisageable de faire de l'inférence causale à partir de l'$OR^*_{E\rightarrow M}$ brut. Si cet $OR^*_{E\rightarrow M}$ brut est significativement différent de « 1 », vous pourrez tout d'abord conclure à la présence d'une association *statistique* dans la population cible (en prenant cependant toutes les précautions de langage nécessaires [7]). Ensuite, si vous pouvez convaincre votre auditoire que les FCC que vous avez envisagés sont les seules expositions qui auraient pu jouer un rôle de confusion (sous-entendu, il n'y a pas d'autres expositions qui seraient susceptibles d'être des facteurs de confusion), alors vous pourriez envisager de faire de l'inférence causale. Néanmoins, je ne vous conseille pas de le faire (c'est-à-dire de faire de l'inférence causale à partir de cette association brute).

6.2) Si au moins une Δ_k est supérieure ou égale à 20%, alors l'$OR^*_{E\rightarrow M}$ brut est biaisé par du biais de confusion dû à chaque FCP_k correspondant. Soit *L* le nombre de différences relatives supérieures ou égales à 20%. Vous avez donc identifié *L* facteurs de confusion parmi les *P* FCC.

6.2.1) Si tous les *L* $OR^*_{E\rightarrow M}$ ajustés (chacun ajusté sur un seul facteur de confusion) sont plus éloignés de « 1 » que ne l'est l'$OR^*_{E\rightarrow M}$ brut tout en restant dans la même direction[2], alors on peut se dire que si l'$OR^*_{E\rightarrow M}$ avait pu être ajusté sur tous ces *L* facteurs de confusion dans un seul modèle multivarié, il y a des chances pour cet $OR^*_{E\rightarrow M}$ ajusté soit plus éloigné de « 1 » que l'$OR^*_{FRPk\rightarrow M}$ brut.

6.2.1.1) Si cet $OR^*_{E\rightarrow M}$ brut est significativement différent « 1 », il y a des chances pour qu'il le fût resté après ajustement sur ces *L* expositions. S'il n'y a pas d'autres facteurs de confusion (hypothèse que vous devez faire), autres que les *L* identifiés, alors il est envisageable de suggérer l'éventuelle possibilité de l'existence d'une relation causale entre E et M.

6.2.1.2) Si cet $OR^*_{E\rightarrow M}$ brut n'est pas significativement différent « 1 », mais si chacun des *L* $OR^*_{E\rightarrow M}$ ajustés est significativement différent de « 1 », alors on peut se dire que si l'$OR^*_{E\rightarrow M}$ avait pu être ajusté sur tous ces *L* facteurs de confusion dans un seul modèle multivarié, il y a des chances pour cet $OR^*_{E\rightarrow M}$ ajusté soit significativement différent de « 1 ». Alors il est envisageable de suggérer l'éventuelle possibilité de l'existence d'une relation causale entre E et M.

6.2.1.3) Si cet $OR^*_{E\rightarrow M}$ brut n'est pas significativement différent « 1 », et au moins 1 parmi les *L* $OR^*_{E\rightarrow M}$ ajustés n'est pas significativement différent de « 1 », alors il est difficile de dire qu'il le serait *devenu* après ajustement sur ces *L* expositions (c'est possible, mais on ne pourra jamais en être certain). On pourra alors suggérer un éventuel manque de puissance statistique, et une étude comprenant plus d'individus serait nécessaire.

6.2.2) Si au moins un $OR^*_{E\rightarrow M}$ ajusté, parmi les *L*, est moins éloigné de « 1 » que ne l'est l'$OR^*_{E\rightarrow M}$ brut tout en restant dans la même direction, ou bien si au moins un des $OR^*_{E\rightarrow M}$ ajusté est de l'autre côté de « 1 » par rapport à l'$OR^*_{E\rightarrow M}$ brut, alors on peut se dire que si l'$OR^*_{E\rightarrow M}$ avait pu être ajusté sur tous ces *L* facteurs de confusion dans un seul modèle multivarié, il y a des risques pour cet $OR^*_{E\rightarrow M}$ ajusté soit moins éloigné de « 1 » que ne l'est l'$OR^*_{FRPk\rightarrow M}$ brut. Et auquel cas, si cet $OR^*_{E\rightarrow M}$ brut est significativement différent « 1 », il y a des risques pour qu'il ne le fût pas resté après ajustement sur ces *L* facteurs de confusion. Ainsi, il n'est pas possible de suggérer l'éventuelle possibilité de l'existence d'une relation causale entre E et M, à partir de cet $OR^*_{E\rightarrow M}$ brut.

[2] C'est-à-dire qu'ils sont tous du même côté de « 1 » que l'$OR^*_{E\rightarrow M}$ brut

Problématique des données manquantes

La démarche ci-dessus repose sur le fait qu'aucune donnée ne manque sur E ni sur tous les FCP. En effet, un modèle comprenant par exemple deux expositions X_1 et X_2 ne va tourner que sur les individus pour lesquels la valeur de X_1 *et* la valeur de X_2 est renseignée. Admettons un échantillon de 100 individus, admettons qu'il n'y ait aucune donnée manquante pour la variable X_1, mais qu'il y en 3 pour X_2, alors le modèle brut comprenant X_1 seulement va tourner sur les 100 individus mais le modèle bivarié comprenant X_1 et X_2 ne va tourner que sur les 97 individus pour lesquels aucune donnée ne manque pour X_1 *et* X_2. Ainsi, dans la mesure où l'échantillon d'individus n'est pas le même entre le modèle brut et le modèle bivarié, les résultats ne doivent pas être comparés (notamment pour calculer Δ) ! Alors que faire ? S'assurer que les deux modèles tournent sur le *même* échantillon d'individus. Ainsi, le modèle brut ne comprenant que X_1 ne doit tourner que sur les individus pour lesquels il n'y a pas de donnée manquante pour X_2 (c'est-à-dire sur les 97 individus).

Exemple

Soit une étude clinique qui vise à montrer qu'une élévation en ALAT (> *versus* ≤ 100 U/L), en PAL (> *versus* ≤ 1500 U/L), et que la présence (*versus* absence) de PU-PD soient causalement associées à la survenue d'un décès de cause rénale (l'étude a donc trois expositions d'intérêt). Supposons que l'on ait des raisons de penser que l'âge, le sexe, le poids, les concentrations en ALAT, en PAL, et la présence de PU-PD soient associées à la survenue d'un décès de cause rénale [3]. Dans cette situation, il y a donc 7 FCC. Supposons qu'il n'y a aucune donnée manquante sur toutes les expositions étudiées. Voici les résultats ci-dessous issus de modèles bruts et bivariés (présentation de la valeur des OR assortie du degré de signification entre parenthèses).

	Exposition d'intérêt (E)					
	ALAT		PAL		PU-PD	
	OR (p)	Δ	OR (p)	Δ	OR (p)	Δ
$OR_{E\rightarrow M}$ brut	2,0 (p=0,02)	--	3,0 (p=0,07)	--	4,0 (p=0,01)	--
$OR_{E\rightarrow M}$ ajusté sur l'âge	2,2 (p=0,02)	10%	2,8 (p=0,09)	7%	3,0 (p=0,12)	25%
$OR_{E\rightarrow M}$ ajusté sur sexe	2,3 (p=0,01)	15%	3,1 (p=0,07)	3%	3,3 (p=0,08)	18%
$OR_{E\rightarrow M}$ ajusté sur poids	2,1 (p=0,04)	5%	3,0 (p=0,07)	0%	4,9 (p=0,01)	23%
$OR_{E\rightarrow M}$ ajusté sur ALAT	--		3,2 (p=0,06)	7%	4,4 (p=0,01)	10%
$OR_{E\rightarrow M}$ ajusté sur PAL	2,6 (p=0,01)	30%	--		5,1 (p=0,01)	28%
$OR_{E\rightarrow M}$ ajusté sur PU-PD	2,4 (p=0,02)	20%	2,7 (p=0,10)	10%	--	

D'après les résultats ci-dessus, voici ce que l'on peut dire.

Concernant les PAL

(Possibilité 6.1) On peut voir qu'aucun des 6 FCC (les 7 FCC moins bien entendu l'exposition « PAL » qui est l'exposition d'intérêt) n'a joué de *fort* rôle de confusion dans l'étude de l'association causale entre PAL et décès, puisqu'aucune des différences relatives entre l'$OR_{PAL\text{-}décès}$ brut et chacun des 6 $OR_{PAL\rightarrow décès}$ ajusté n'est supérieure à 20% (la différence relative la plus importante est celle après ajustement sur la PU-PD, et elle vaut 10%). Par conséquent, si les 6 FCC sont effectivement les seules expositions qui auraient pu jouer un rôle de confusion, l'$OR_{PAL\rightarrow décès}$ brut (de valeur égale à 3,0) n'est *a priori* pas biaisé par du biais de confusion. Mais dans une étude observationnelle, il est difficile de dire qu'il n'y a aucun biais de confusion du tout… !

[3] Vous vous rendez compte déjà qu'une exposition d'intérêt peut tout à fait jouer le rôle de facteur de confusion candidat dans l'association entre une autre exposition d'intérêt et M.

Concernant les ALAT

(Possibilité 6.2.1.1) On peut voir que deux expositions sont de forts facteurs de confusion (différences relatives supérieures ou égales à 20% ; *L* = 2) : PAL et PU-PD. Après ajustement sur chacune de ces deux expositions, l'$OR_{ALAT \rightarrow décès}$ est plus éloigné de « 1 » que ne l'est l'$OR_{ALAT \rightarrow décès}$ brut (tout en restant dans la même direction). Ainsi, si nous avions pu ajuster sur ces deux expositions à la fois (sur PAL et PU-PD) dans un seul modèle multivarié, il y a des chances pour que l'$OR_{ALAT \rightarrow décès}$ ajusté fût plus éloigné de « 1 » que l'$OR_{ALAT \rightarrow décès}$ brut. Cet $OR_{ALAT \rightarrow décès}$ brut était significativement différent de 1, et les 2 $OR_{ALAT \rightarrow décès}$ ajustés l'étaient aussi. Il y a donc des chances pour que l'$OR_{ALAT \rightarrow décès}$ ajusté sur PAL et PU-PD fût resté significativement différent de « 1 » après ajustement sur ces deux expositions. S'il n'y a pas d'autres facteurs de confusion, alors il est envisageable de suggérer l'éventuelle possibilité de l'existence d'une relation causale entre ALAT et la survenue d'un décès, à partir de cet $OR^*_{ALAT \rightarrow décès}$ brut.

Concernant la PU-PD

(Possibilité 6.2.2) On peut voir que trois expositions sont de forts facteurs de confusion (différences relatives supérieures ou égales à 20% ; *L* = 3) : âge, poids, et PAL. Après ajustement sur l'âge, l'$OR_{PU\text{-}PD \rightarrow décès}$ est moins éloigné de « 1 » que ne l'est l'$OR_{PU\text{-}PD \rightarrow décès}$ brut. Ainsi, si nous avions pu ajuster sur ces trois expositions à la fois (sur âge, poids, et PAL) dans un seul modèle multivarié, il y a des risques pour que l'$OR_{PU\text{-}PD \rightarrow décès}$ ajusté fût moins éloigné de « 1 » que ne l'est l'$OR_{PU\text{-}PD \rightarrow décès}$ brut. Ainsi, si nous avions pu ajuster ces trois expositions à la fois, il y a des risques pour que l'$OR_{PU\text{-}PD \rightarrow décès}$ ajusté fût devenu non significativement différent de « 1 ». Ainsi encore, il n'est pas possible de suggérer l'éventuelle possibilité de l'existence d'une relation causale entre PU-PD et décès, à partir de cet $OR^*_{PU\text{-}PD \rightarrow décès}$ brut.

Littérature

1. Hernan MA, Hernandez-Diaz S, Werler MM, Mitchell AA. **Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology**. *Am J Epidemiol* 2002; 155(2):176-184.

2. Walter S, Tiemeier H. **Variable selection: current practice in epidemiological studies**. *Eur J Epidemiol* 2009; 24(12):733-736.

3. Heinze G, Wallisch C, Dunkler D. **Variable selection - A review and recommendations for the practicing statistician**. *Biom J* 2018; 60(3):431-449.

4. Mickey RM, Greenland S. **The impact of confounder selection criteria on effect estimation**. *Am J Epidemiol* 1989; 129(1):125-137.

5. Weng HY, Hsueh YH, Messam LL, Hertz-Picciotto I. **Methods of covariate selection: directed acyclic graphs and the change-in-estimate procedure**. *Am J Epidemiol* 2009; 169(10):1182-1190.

6. Evans D, Chaix B, Lobbedez T, Verger C, Flahault A. **Combining directed acyclic graphs and the change-in-estimate procedure as a novel approach to adjustment-variable selection in epidemiology**. *BMC Med Res Methodol* 2012; 12:156.

7. Desquilbet L. **Enhancing Clinical Decision-Making: Challenges of making decisions on the basis of significant statistical associations**. *Journal of the American Veterinary Medical Association* 2020; 256(2):187-193.