Validation de facteurs environnementaux grâce à l'estimation du risque génétique chez les patients de maladie complexe

Félix Balazard [1,2], Sophie Le Fur[2], Alain-Jacques Valleron[2], Pierre Bougnères[2] & Collaboration Isis-Diab[2]

[1] *Sorbonne Université, Laboratoire de Statistique Théorique et Appliquée*
[2] *Université Paris Saclay, Inserm u1169, Hopital Bicêtre, felix.balazard@inserm.fr*

Résumé. *Contexte* : L'identification de facteurs environnementaux des maladies complexes à l'aide d'études cas-témoins est sujette à des biais. En particulier, la sélection des contrôles est une source potentielle de confusion. Lorsque la maladie a une forte composante génétique, une estimation du risque génétique peut quantifier la prédisposition d'un individu. Le biais de collision apparaît entre deux causes quand on conditionne par une conséquence partagée.

Méthodes : Nous proposons Collision avec la Maladie(Disease As Collider DAC), une nouvelle méthodologie pour valider les facteurs environnementaux en utilisant le risque génétique chez les patients. La maladie est un collisionneur entre les facteurs génétiques et environnementaux et, sous des hypothèses raisonnables, une association chez les cas entre risque génétique et environnement est la signature d'un véritable facteur de risque environnemental. Une telle association permettrait de valider les résultats des études cas-témoins. Nous appliquons DAC chez 831 patients ayant le diabète de type 1 (DT1) de la cohorte Isis-Diab en évaluant l'association entre le risque génétique et 7 facteurs environnementaux venant des résultats de l'étude cas-témoins. Nous effectuons des simulations pour estimer la puissance de notre méthodologie dans notre cas ainsi que dans des scénarios alternatifs.

Résultats : L'hygiène bucco-dentaire était associée au risque génétique. Cependant, les simulations montrent que le pouvoir était faible dans notre cadre. DAC a une puissance raisonnable dans les scénarios d'incidence plus élevée, avec une plus grande taille d'échantillon et une meilleure estimation du risque génétique.

Conclusions : Bien que DAC ait une faible puissance dans notre cadre, cette nouvelle méthodologie peut apporter de l'information pour identifier les facteurs environnementaux des maladies complexes. Nous exposons les circonstances nécessaires pour que DAC puisse participer à la triangulation des causes environnementales de maladie.

Mots-clés. Médecine épidémiologie, Modèles graphiques.

Abstract. *Background* : The identification of environmental factors of complex human diseases using case-control studies is prone to biases. In particular, selection of controls is a potential source of confusion. When the disease has a strong genetic component,

a genetic risk estimation can quantify the predisposition of an individual. Collider bias appears between two causes when conditioning on a shared consequence.

Methods : We introduce Disease As Collider (DAC), a new methodology to validate environmental factors using genetic risk in cases. Disease is a collider between genetic and environmental factors and, under reasonable assumptions, an association in cases between genetic risk and environment is a signature of a genuine environmental risk factor. Such an association would validate results from case-control studies. We used DAC in 831 patients of the Isis-Diab cohort on type 1 diabetes (DT1) by assessing association between genetic risk and 7 environmental factors from the results of the case-control study. We perform simulations to estimate the power of our methodology in our case as well as in alternative scenarios.

Results : Oral hygiene was nominally associated with genetic risk. However, simulations show that power was low in our setting. DAC has reasonable power in higher incidence scenarios, with larger sample size and better genetic risk estimation. Conclusions: While DAC had low power in our setting, this new methodology could provide a new line of evidence for environmental factors of complex diseases. We expose the circumstances needed for DAC to participate in the triangulation of environmental causes of disease.

Keywords. Medecine epidemiology, Graphical models

1 Difficultés pour les facteurs environnementaux

L'identification des déterminants environnementaux des maladies complexes est difficile. Pour les maladies dont la prévalence est inférieure à 1%, comme le DT1, les études de cohortes prospectives sont coûteuses car elles impliquent de suivre une population importante. En comparaison, le design cas-témoin permet d'obtenir une grande population de cas à coût réduit. Il est toutefois sensible à des problèmes spécifiques: le biais de rappel et le choix des contrôles.

De nombreuses études cas-témoins ont été menées sur le DT1, mais des associations robustes avec des facteurs environnementaux restent évasives[1]. Par conséquent, les traitements de prévention évalués par des essais randomisés contrôlés (RCT) n'ont pas démontré d'effet protecteur[2-3]. Ces résultats négatifs soulignent l'importance d'avoir des candidats solides avant de passer par le processus coûteux et long d'un RCT de prévention.

2 Progrès en génétique

D'autre part, la génétique de la maladie est mieux comprise au cours de la dernière décennie. Les études d'association à l'échelle du génome (GWAS) ont abouti à plus d'un millier d'associations validées entre des phénotypes et des loci dont plus de 60 loci associé au diabète de type 1[5]. Cela a permis d'explorer l'interaction entre le gène et

l'environnement. Les efforts ont porté principalement sur l'identification des interactions GxE[6] et sur la randomisation mendélienne[7] qui utilise des gènes comme variables instrumentales pour étudier les relations entre des phénotypes intermédiaires et des phénotypes d'intérêt.

Une autre utilisation des ensembles de données GWAS a été d'estimer à un niveau individuel le risque génétique de développer la maladie en utilisant des techniques d'apprentissage statistiques[8]. Ceci fournit un résumé unidimensionnel des données génétiques importantes pour l'épidémiologie. Cet article examine l'interaction entre une telle estimation du risque génétique et les facteurs de risque environnementaux et il le fait en considérant le biais de collision.

3 Biais de collision

Le biais de collision est la corrélation négative qui apparaît entre deux causes lorsqu'on conditionne par leur conséquence partagée[9]. Elle peut induire en erreur les enquêtes épidémiologiques [10,11]. Un exemple classique est le biais de Berkson dans lequel deux maladies sont associées négativement dans une population hospitalisée même si elles sont indépendantes dans la population générale[12,13]. Dans cet exemple, le collisionneur est l'hospitalisation, la conséquence partagée des deux maladies. En regardant seulement les patients à l'hôpital, c'est-à-dire en conditionnant par l'hospitalisation, une corrélation négative apparaît entre les deux maladies.

Dans ce travail, nous tirons partie du biais de collision entre le risque génétique et les facteurs environnementaux quand on conditionne par la maladie pour valider les facteurs de risque environnementaux putatifs. D'une part, l'estimation du risque génétique est bien validée[8]. En revanche, les preuves d'une cause environnementale spécifique du DT1 restent limitées. La maladie est une conséquence commune de la génétique et de l'environnement. L'association entre risque génétique et environnement est une preuve de l'authenticité du lien causal entre le facteur environnemental et la maladie. Cette idée est résumée dans la figure 1. Nous nous référons à cette méthode par Collision avec la Maladie (Disease As Collider DAC).

Formellement, si on considère deux variables aléatoires G, et E avec $G \perp\!\!\!\perp E$ et $D = f(G, E, U)$ où U est un aléa et f une application, alors $G \not\perp\!\!\!\perp E|D$.

4 Collision avec la Maladie (DAC)

Nous illustrons l'intérêt de cette nouvelle méthodologie avec l'étude Isis-Diab, une étude cas-témoins du DT1 basé sur un long questionnaire[14]. Cette étude exploratoire a abouti à une dizaine d'associations environnementales jugées significatives même si un certain nombre de ces associations ne semblent pas plausibles. Une partie des patients de l'étude avait été génotypée et nous avons utilisé une estimation de risque génétique basée sur

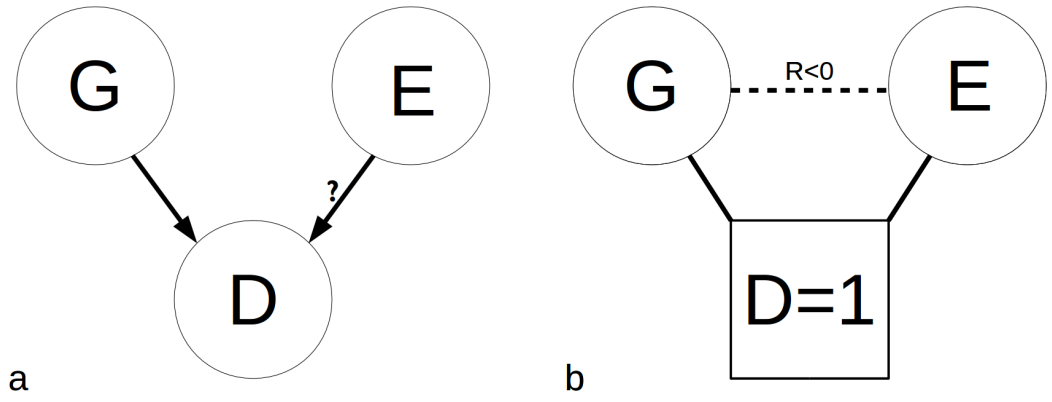

Figure 1: Méthodologie DAC. a: La maladie est une conséquence de causes génétiques et environnementales. Ceux-ci sont souvent indépendants dans la population générale. L'association du facteur environnement avec la maladie nécessite une confirmation. b: Lorsqu'on conditionne par la maladie et s'il existe une association réelle entre le facteur environnemental et la maladie, une association négative apparaît entre le risque génétique et le facteur environnemental. Cela confirme que le facteur environnemental est associé à la maladie.

un support vecteur machine entraînée sur les données du WTCCC1[8,15] pour estimer le risque génétique chez ces patients. Nous avons testé l'association de 7 questions significatives dans le grand questionnaire avec le risque génétique chez les patients afin de confirmer leur association à la maladie. L'association est testé par des tests classiques (régression linéaire).

Si DAC utilise les mêmes données environnementales sur les patients que l'étude cas-témoins, les témoins ne sont pas utilisés et les résultats de DAC ne sont donc pas redondants avec les résultats de l'étude cas-témoins. En particulier, DAC ne dépend pas du choix des contrôles, une source importante de confusion dans le design cas-témoins. DAC a donc le potentiel de renforcer les preuves d'une association environnementale.

Pour estimer la puissance de la méthode, nous avons simulé l'apparition de la maladie dans une population. Nous attribuons à chaque individu un risque génétique et environnemental indépendamment avant d'attribuer la maladie en fonction de ces risques en suivant un modèle logistique. Nous regardons alors la corrélation entre le risque génétique et le risque environnemental chez les patients pour savoir si la méthode détecte le biais de collision.

Le biais de collision est un effet subtil et n'est pas facilement domestiqué. DAC n'est pas assez puissant dans le cadre de l'étude Isis-Diab et donc nos résultats ne sont pas informatifs. Cependant, DAC atteint des puissances acceptables dans des situations de prévalence plus élevée, avec de grandes tailles d'échantillon et une meilleure estimation du risque génétique.

Ce travail fait l'objet d'un preprint plus exhaustif[16].

Bibliographie

[1] Rewers M, Ludvigsson J. (2016) Environmental risk factors for type 1 diabetes. The Lancet. ;387(10035):2340–2348.
[2] Gale E. M, Bingley PJ, Emmett CL, Collier T, (2004) European Nicotinamide Diabetes Intervention Trial (ENDIT) Group. European Nicotinamide Diabetes Intervention Trial (ENDIT): a randomised controlled trial of intervention before the onset of type 1 diabetes. Lancet Lond Engl. ;363(9413):925–931.
[3] Knip M, Åkerblom HK, Becker D, et al. (2014) Hydrolyzed infant formula and early β-cell autoimmunity: A randomized clinical trial. JAMA. ;311(22):2279–2287.
[5] Welter D, MacArthur J, Morales J, et al. (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 42(D1):D1001–D1006.
[6] Thomas D. (2010) Gene–environment-wide association studies: emerging approaches. Nat Rev Genet;11(4):259–272.
[7] Davey Smith G, Hemani G. (2014) Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Hum Mol Genet. 23(R1):R89-98.
[8] Wei Z, Wang K, Qu H-Q, et al. (2009) From Disease Association to Risk Assessment:

An Optimistic View from Genome-Wide Association Studies on Type 1 Diabetes. PLOS Genet. ;5(10):e1000678.
[9] Cole SR, Platt RW, Schisterman EF, et al. (2010) Illustrating bias due to conditioning on a collider. Int J Epidemiol. ;39(2):417–420.
[10] Greenland S. (2003) Quantifying Biases in Causal Models: Classical Confounding vs Collider-Stratification Bias: Epidemiology. 14(3):300–306.
[11] Gage SH, Smith GD, Ware JJ, Flint J, Munafò MR. (2016) G = E: What GWAS Can Tell Us about the Environment. PLOS Genet. 12(2):e1005765.
[12] Berkson J. (1946) Limitations of the application of fourfold table analysis to hospital data. Biometrics. Jun;2(3):47–53.
[13] Snoep JD, Morabia A, Hernández-Díaz S, Hernán MA, Vandenbroucke JP. (2014) Commentary: A structural approach to Berkson's fallacy and a guide to a history of opinions about it. Int J Epidemiol. 43(2):515–521.
[14] Balazard F, Le Fur S, Valtat S, et al. (2016) Association of environmental markers with childhood type 1 diabetes mellitus revealed by a long questionnaire on early life exposures and lifestyle in a case-control study. BMC Public Health. 16(1):1021.
[15] Burton PR, Clayton DG, Cardon LR, et al. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 447(7145):661–678.
[16] Félix Balazard, Sophie Le Fur, Pierre Bougnères, Alain-Jacques Valleron, the Isis-Diab collaborative group (2017) Disease as collider: a new case-only method to discover environmental factors in complex diseases with genetic risk estimation. bioRxiv 124560; doi: https://doi.org/10.1101/124560