SCORE DE PROPENSION EN GRANDE DIMENSION : APPARIEMENT A UNE POPULATION TEMOIN DANS DES DONNEES NATIONALES DE SANTE

Jonathan Cottenet[1] & Catherine Quantin[2] & Camelia Goga[3]

[1] *Service de Biostatistiques et d'Information Médicale, CHU Dijon Bourgogne, France, jonathan.cottenet@chu-dijon.fr*
[2] *Service de Biostatistiques et d'Information Médicale, CHU Dijon Bourgogne, INSERM, Université de Bourgogne, CIC 1432, Module Épidémiologie Clinique, 21000 Dijon, Université Paris-Saclay, UVSQ, Inserm, CESP, 94807 Villejuif, catherine.quantin@chu-dijon.fr*
[3] *Laboratoire de Mathématiques de Besançon, Université de Franche-Comté, 25000 Besançon, France, camelia.goga@univ-fcomte.fr*

Résumé. Compte-tenu de la complexité des règles de recueil de données dans les bases médico-administratives, il est nécessaire de prendre en compte et de modéliser ce type de données de la manière la plus fiable possible, en particulier dans le cadre de sélection de population témoin. Dans un premier temps, il s'agira de tester différentes méthodologies d'échantillonnage pour sélectionner des populations témoins par rapport à des patients atteints de pathologies spécifiques : échantillonnage raisonné, échantillonnage probabiliste, score de propension en grande dimension. Pour ce dernier, nous testerons plusieurs méthodes pour identifier les variables à inclure dans le modèle du score de propension : méthode de régularisation (notamment régression bayésienne, Lasso, Ridge, Elastic Net) ou méthode non paramétrique de type machine learning (notamment Random Forest et Boosting). Cette étude permettra de coupler différentes méthodologies et d'illustrer les différentes stratégies analytiques à partir de données en vie réelle, s'appuyant sur un appariement pour lequel plusieurs témoins pourraient être sélectionnés, rendant les populations les plus comparables possibles à une échelle nationale. Ces méthodes seront appliquées en faisant varier le nombre de témoins à sélectionner et sur plusieurs jeux de données.

Mots-clés. population témoin, échantillonnage, score de propension en grande dimension, base médico-administrative, santé publique

Abstract. Given the complexity of the rules for collecting medico-administrative data, it is necessary to consider and model this type of data in the most reliable way possible, particularly in the context of selecting control populations. The first step will be to determine the correct methodology for selecting control populations related to patients with specific conditions: purposive sampling, probability sampling, high-dimensional propensity score. For the latter, we will test several methods for identifying variables to be included in the propensity score model: shrinkage methods (including Bayesian regression, Lasso, Ridge, Elastic Net) or nonparametric method based on machine learning (including Random Forest and Boosting). This study will make it possible to combine different methodologies and illustrate different analytical strategies using real-life data, based on matching for which several controls could be selected, making the populations as comparable as possible on a national scale. These methods will be applied according to the number of controls to be selected and on several data sets.

Keywords. control population, sampling, high-dimensional propensity score, medico-

administrative database, public health

1 Introduction

L'utilisation des données de santé médico-administratives à des fins d'évaluation des pratiques médicales, mais aussi d'épidémiologie et de recherche clinique prend davantage d'ampleur. Ces données, à l'échelle nationale, intéressent non seulement les chercheurs mais aussi, de plus en plus, les décideurs en matière de santé publique. Compte-tenu de la complexité des règles de recueil de ces données, il est nécessaire de prendre en compte et de modéliser ce type de données de la manière la plus fiable possible, en particulier dans le cadre de sélection de population témoin.

La sélection d'une population témoin peut dépendre de l'unité statistique choisie. En particulier, nous pouvons considérer la sélection d'une population témoin pour une pathologie définie. En effet, l'une des méthodes d'étude les plus fréquentes en recherche sur la santé est l'étude cas-témoins, dans laquelle des échantillons distincts sont tirés parmi les « cas » (disons, les personnes présentant une maladie d'intérêt) et parmi les « témoins » (personnes n'ayant pas la maladie) (Breslow, 1996). Nous nous intéressons ici aux études sur la population dans lesquelles les témoins sont sélectionnés en utilisant des méthodes standard d'échantillonnage. L'échantillonnage résulte de méthodes suivies pour tirer des échantillons représentatifs à partir d'une population cible, il est donc devenu naturel de penser aux méthodes de sondage pour obtenir les témoins (Wacholder, 1991). De plus en plus fréquemment, au cours des quelque 30 dernières années, les témoins (et parfois les cas également) ont été sélectionnés en utilisant des plans de sondage complexes stratifiés à plusieurs degrés (Di Gaetano, 2002 ; Scott, 2006 ; Janes, 2008 ; Knol 2008 ; Laara, 2011 ; Rundle, 2012 ; Jorgenson, 2019 ; Mezei, 2020 ; Liew, 2022 ; Byanu, 2023).

Il est toujours compliqué de s'assurer que les témoins soient réellement sélectionnés à partir de la même population, selon les mêmes protocoles, que les cas. Pour ce type de sélection, il est d'usage d'utiliser un échantillonnage raisonné puis si besoin d'apparier les cas aux témoins sur certains critères (comme l'âge et le sexe), et éventuellement à l'aide d'un score de propension (Rosenbaum & Rubin, 1983). Ces scores permettent de minimiser le biais de sélection des études observationnelles afin de se rapprocher des conditions de comparabilité d'un essai randomisé. Afin de prendre en compte un très grand nombre de variables qui, collectivement, peuvent être des proxies pour des facteurs de confusion non disponibles dans les données, un score de propension en grande dimension (hdPS-*high dimensional propensity score*) a été développé (Schneeweiss, 2009). En effet, à la différence du score de propension, la méthode de l'hdPS peut sélectionner un nombre important de facteurs confondants observés dans des données médico-administratives grâce à un algorithme de sélection standardisé.

D'autres approches, par exemple la régression bayésienne (Gelman, 1995), le Lasso (Tibshirani, 1996), le Ridge (Hoerl & Kennard, 1970), l'Elastic Net (Zou & Hastie, 2005) peuvent modéliser l'association entre l'évènement à étudier et de nombreuses covariables simultanément en réduisant les coefficients de régression extrêmes et estimés de manière imprécise. Il s'agit ainsi de contraindre et régulariser les estimations des coefficients vers zéro, permettant ainsi de réduire la variance (Keller, 2018). D'autres méthodes non paramétriques peuvent également être utilisées pour étudier cette association et estimer l'importance des variables en utilisant des techniques de machine learning à partir d'arbres de décision comme le bagging (Breiman, 1996) (exemple du Random Forest (Breiman, 2001)) ou le boosting (Freund & Schapire, 1996). L'utilisation de ces méthodes (en particulier Lasso, Ridge, Random Forest) a été effectuée dans des études de cohorte utilisant le hdPS

(Schneeweiss, 2017) mais sans prendre en compte des populations de patients les plus comparables possibles via un appariement. Cela a été proposé dans plusieurs études, comme dans le papier de Demailly et al. (2020). Néanmoins les appariements étudiés sont en général de 1 cas pour 1 témoin, et sont basés sur des échantillons de faibles effectifs. Or augmenter le nombre de témoins par cas (lorsque cela est possible) augmente la puissance de l'étude et par conséquent augmente les chances de mettre en évidence une association s'il y en a une, même si le gain en puissance s'avère faible au-delà de 3 témoins pour un cas (Ury, 1975 ; Taylor, 1986).

Il apparait donc important de pouvoir coupler ces différentes méthodologies et d'illustrer les différentes stratégies analytiques à partir de données en vie réelle, s'appuyant sur un appariement pour lequel plusieurs témoins pourraient être sélectionnés, rendant les populations les plus comparables possibles à une échelle nationale.

Pour ce faire, il s'agira dans un premier temps de tester différentes méthodologies d'échantillonnage pour sélectionner des populations témoins par rapport à des patients atteints de pathologie spécifique : échantillonnage raisonné, échantillonnage probabiliste, hdPS. Pour ce dernier, nous testerons plusieurs méthodes pour identifier les variables à inclure dans le modèle du hdPS.
Ces méthodes seront appliquées en faisant varier le nombre de témoins sélectionnées (entre 1 et 3 lorsque cela est possible) et sur plusieurs jeux de données.

2 Méthodologie

2.1 Bases nationales du SNDS

Le Système National des Données de Santé (SNDS) est le système d'information national français qui contient des données individuelles, exhaustives et chainées mais anonymes sur le recours aux soins pour environ 97% de la population française (Bezin, 2017). Le SNDS agrège les données de deux bases de données nationales (Goldberg, 2012) liées par un identifiant unique du patient : le Programme de Médicalisation des Systèmes d'Information (PMSI) et la base de données de l'assurance maladie (DCIR).

La base de données PMSI fournit des informations médicales détaillées sur toutes les admissions dans les hôpitaux publics et privés en France, y compris les diagnostics de sortie selon la dixième édition de la Classification internationale des maladies (codes CIM-10) et les actes médicaux codés selon la Classification commune des actes médicaux (CCAM).

La base de données DCIR contient toutes les demandes de soins individualisées et anonymes remboursées hors hôpital par l'Assurance maladie française. Ces demandes de remboursement comprennent les consultations de médecins généralistes ou spécialistes, les médicaments délivrés, qui sont codés selon la classification anatomique thérapeutique chimique (ATC), et les actes médicaux ambulatoires, codés selon la CCAM. La base DCIR recueille également des données sur les patients, telles que l'âge, le sexe, l'éligibilité à la couverture maladie universelle complémentaire (CMU-C), qui permet un accès gratuit aux soins pour les personnes à faibles revenus, un score de désavantage (Rey, 2009) et le type de régime d'assurance (général ou autre). L'éligibilité à la prise en charge à 100% par l'assurance maladie des affections de longue durée (ALD) graves et coûteuses, codées selon la CIM-10, est également enregistrée dans la base DCIR.

Depuis 20 ans, les données hospitalières sont utilisées à des fins de recherche médicale et la

qualité de la base PMSI a été confirmée par des études récentes (Piroth, 2020 ; Simon, 2022 ; Chauvet-Gelinier, 2023 ; Viennet, 2023), notamment dans le cadre des domaines liés à la pneumologie (Maitre, 2021) ou au cancer (Cottenet, 2022). Elle fournit un grand nombre d'informations épidémiologiques sur les patients hospitalisés en France et peut être utilisée pour créer des cohortes suffisamment grandes pour détecter des événements rares (Cottenet, 2019 ; Loiseau, 2023).

2.2 Score de propension et score de propension en grande dimension

Le score de propension (PS-*propensity score*) permet de minimiser le biais de sélection des études observationnelles afin de nous rapprocher des conditions de comparabilité d'un essai randomisé. Les individus ayant le même PS ont la même probabilité d'être exposés. Par conséquent, une paire d'individus (un exposé et un non exposé) avec le même PS aura des caractéristiques individuelles différentes mais, en moyenne, toutes les variables prises en compte dans le calcul du score seront équilibrées entre les groupes exposés et non exposés, comme dans une étude randomisée.
Rosenbaum et Rubin (1983) ont défini le PS d'un patient comme la probabilité conditionnelle d'être traité (exposé) compte tenu des covariables observées. La variable dépendante ne peut donc prendre que deux valeurs, selon que l'événement (traitement) s'est produit ou non. Le résultat de la régression logistique est la probabilité (comprise entre 0 et 1) qu'un événement se produise en fonction des valeurs des variables indépendantes.
Les résultats de la régression logistique peuvent ensuite être utilisés pour calculer le PS selon la formule suivante :

$$PS = P(Z=1|X) = \frac{\exp(\beta 0 + \beta 1 X1 + \beta 2 X2 + \cdots + \beta n Xn)}{1+\exp(\beta 0 + \beta 1 X1 + \beta 2 X2 + \cdots + \beta n Xn)}$$

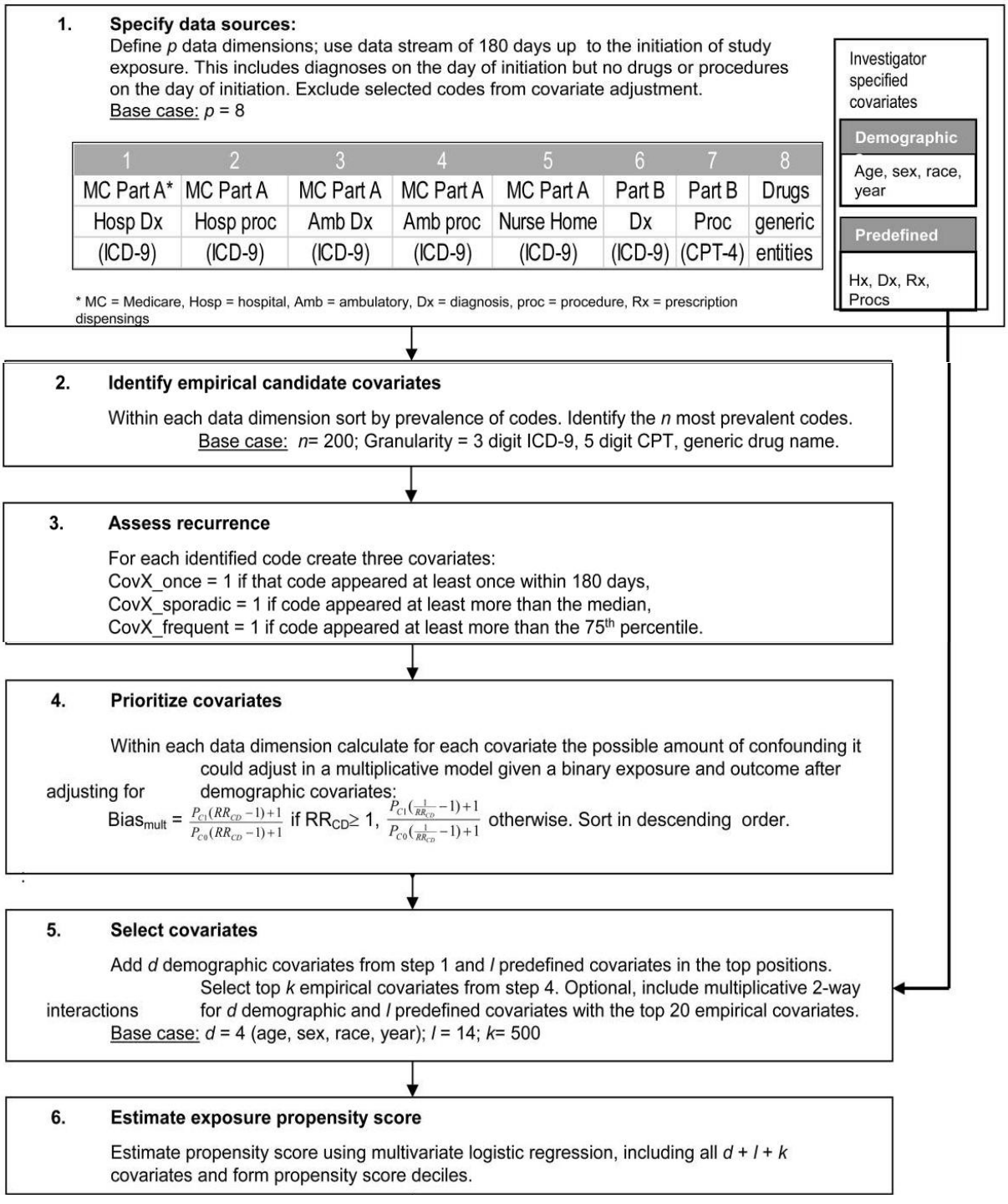
où

- Z est la variable dépendante valant 1 si le patient est traité et 0 sinon.
- x1, x2 à xn sont des covariables indépendantes.
- β0, β1 à βn sont les coefficients de régression correspondants, le coefficient β0 représente l'influence d'une composante absolue (la valeur de la probabilité lorsque toutes les covariables sont égales à zéro).
- Les paramètres β0, β1 à βn sont estimés à partir des données à l'aide de la méthode du maximum de vraisemblance.

La méthode du score de propension en grande dimension (hdPS-*high dimensional propensity score*) cherche à améliorer la performance des scores de propension vis-à-vis du biais de confusion en identifiant un nombre important de facteurs confondants normalement ignorés par les investigateurs. Tout comme le score de propension, l'hdPS implique la création d'un modèle de régression permettant d'identifier la probabilité d'un patient à être initié sur un traitement plutôt que sur un autre conditionnellement à une liste de facteurs confondants potentiels observés au moment d'initier le traitement. Cependant, à la différence du score de propension, la méthode de l'hdPS n'est pas limitée aux covariables sélectionnées par l'investigateur, mais comprend aussi une liste de covariables (par défaut ce nombre est défini à 500) identifiées grâce à un algorithme standardisé. Il s'agit d'un ensemble de covariables provenant d'une même et unique base de données identifiée selon un système de classification

spécifique, dont les plus fréquemment identifiées dans les bases de données médico-administratives sont : 1) liste de codes diagnostics posés par un médecin, 2) liste de procédures médicales posées par un médecin et 3) liste des médicaments administrés à un patient. Même si son utilisation reste complexe, cette méthode commence à être de plus en plus utilisée afin d'étudier la comparaison de 2 interventions ou de 2 traitements à partir des bases médico-administratives (Blin, 2020 ; Payet, 2021).

Voici un schéma issu du papier princeps de Schneeweiss et al. (2009) permettant de résumer les différentes étapes de la construction du hdPS :



Par rapport à la priorisation des covariables, le calcul est basé sur la formule de Bross (Bross, 1966) :

- P_{C1} représente la prévalence du facteur de confusion dans le groupe exposé.
- P_{C0} représente la prévalence du facteur de confusion dans le groupe non exposé.
- RR_{CD} représente le risque relatif entre le facteur de confusion et l'évènement étudié.

Nous proposons de coupler ce score à des approches de régularisation, comme par exemple la régression bayésienne, le Lasso, le Ridge, ou bien l'Elastic Net, afin de pouvoir identifier les variables à inclure dans le modèle du hdPS. La régularisation de type "Elastic Net" permet de combiner les approches Ridge et Lasso afin d'éviter la sélectivité trop forte que peut proposer Lasso tout en conservant des variables fortement corrélées comme avec le Ridge (Zou & Hastie, 2005).
Les méthodes de machine learning ont été proposées comme des alternatives prometteuses à la régression logistique pour l'estimation des scores de propension, principalement à partir d'étude de simulation (Lee, 2010 ; Cannas, 2019), même si quelques études de cohorte font référence à ces méthodes (Schneeweiss, 2017 ; Demailly, 2020). Nous proposons d'utiliser ces approches non paramétriques (adaptées aux données de grande dimension) et notamment les Random Forest et le boosting pour estimer ce score, et déterminer l'importance des variables permettant cette estimation.

2.3 Appariement

En règle générale, l'échantillon des témoins est réparti entre des strates définies par des variables démographiques de base de façon telle que la distribution de ces variables dans l'échantillon de témoins soit la même que celle prévue dans l'échantillon de cas.

Concernant le score hdPS obtenu à partir du modèle de régression logistique, celui-ci est utilisé pour apparier les patients les plus similaires des deux groupes (cas et témoins) afin d'obtenir un ensemble de données équilibré et le moins biaisé possible à partir des étapes suivantes :

1. Le seuil de la différence maximale acceptable entre les paires traitées et non traitées est fixé. Sa valeur la plus courante est de 0,01.

2. Pour chaque paire de patients traités et non traités, la différence du hdPS est calculée.

3. La paire présentant la plus petite différence du hdPS est sélectionnée et représente la première paire de l'ensemble de données équilibré.

4. L'algorithme sélectionne la paire suivante de patients traités et non traités la plus proche et la place dans les ensembles de données équilibrés.

5. Ce calcul est répété jusqu'à ce que l'une des règles d'arrêt suivantes s'applique
   - tous les patients traités ont déjà leur paire
   - il n'y a pas de paire qui satisfasse le seuil initial

6. On obtient un ensemble de données équilibré avec des patients appariés sur la base du hdPS. Les patients non appariés sont retirés de l'ensemble de données final équilibré.

7. L'ensemble de données équilibré est utilisé dans les analyses ultérieures, par exemple pour l'estimation non biaisée des effets du traitement.

Alors que l'appariement est généralement utilisé pour 1 témoin pour 1 cas, notamment du fait des effectifs des cohortes étudiées, nous proposons ici d'augmenter le nombre de témoins afin de pouvoir augmenter la puissance statistique, ce qui nous semble possible grâce à la taille des données du SNDS.

3 Application

Afin de tester les différentes méthodes et stratégies analytiques, nous allons utiliser deux jeux de données réelles qui ont déjà été étudiées, pour pouvoir se comparer aux effets trouvés. Nous chercherons ainsi à confronter nos approches aux analyses statistiques précédemment utilisées sur ces cohortes (Quantin 2021, Viennet 2023).
Dans le 1er exemple (Quantin 2021), il s'agit d'évaluer le risque de naissance prématurée associé aux anti-inflammatoires non stéroïdiens (AINS) en se concentrant sur l'exposition précoce entre la conception et la 22e semaine de gestation. Dans ce papier, parmi les 1 598 330 grossesses singletons identifiées entre 2012 et 20214, 130 815 étaient exposées aux AINS (8,18%).
Dans le 2ème exemple (Viennet 2023), nous nous intéresserons à évaluer la fréquence des hospitalisations pour cancer du côlon après une appendicectomie. Dans ce papier, durant la période 2010 à 2015, 230 349 patients ont effectué une appendicectomie,

Bibliographie

Bezin J, Duong M, Lassalle R, Droz C, Pariente A, Blin P, Moore N. (2017). The national healthcare system claims databases in France, SNIIRAM and EGB: Powerful tools for pharmacoepidemiology. *Pharmacoepidemiol Drug Saf*.;26(8):954-962. doi: 10.1002/pds.4233.

Blin P, Dureau-Pournin C, Jové J, Lassalle R, Droz C, Moore N. (2020). Secondary prevention of acute coronary syndrome with antiplatelet agents in real life: A high-dimensional propensity score matched cohort study in the French National claims database. *MethodsX*.;7:100796.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24, 123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32.

Breslow, N.E. (1996). Statistics in epidemiology: The casecontrol study. *Journal of the American Statistical Association*, 91, 1428.

Byanju R, Kandel RP, Poudyal B, Bhandari S, Ligal A, Pradhan S, Gautam M, Shrestha P, Sah RK, Gonzales JA, Porco TC, Whitcher JP, Srinivasan M, Upadhyay MP, Lietman TM, Keenan JD, O'Brien KS; Village-Integrated Eye Worker Trial Group. (2023). Risk factors for corneal ulcers: a population-based matched case-control study in Nepal. *Br J Ophthalmol*.;107(12):1771-1775. doi: 10.1136/bjo-2022-322141.

Cannas M, Arpino B. (2019). A comparison of machine learning algorithms and covariate balance measures for propensity score matching and weighting. *Biom J.* Jul;61(4):1049-1072. doi: 10.1002/bimj.201800132.

Chauvet-Gelinier JC, Cottenet J, Guillaume M, Endomba FT, Jollant F, Quantin C. (2023). Risk of hospitalization for self-harm among adults hospitalized with SARS-CoV-2 in France: A nationwide

retrospective cohort study. *Psychiatry Res*.;324:115214. doi: 10.1016/j.psychres.2023.115214.

Cottenet J, Dabakuyo-Yonli TS, Mariet AS, Roussot A, Arveux P, Quantin C. (2019). Prevalence of patients hospitalised for male breast cancer in France using the French nationwide hospital administrative database. *Eur J Cancer Care (Engl)*.;28(5):e13117. doi: 10.1111/ecc.13117.

Cottenet J, Tapia S, Arveux P, Bernard A, Dabakuyo-Yonli TS, Quantin C. (2022). Effect of Obesity among Hospitalized Cancer Patients with or without COVID-19 on a National Level. *Cancers (Basel)*.;14(22):5660. doi: 10.3390/cancers14225660.

Demailly R, Escolano S, Haramburu F, Tubert-Bitter P, Ahmed I. (2020). Identifying Drugs Inducing Prematurity by Mining Claims Data with High-Dimensional Confounder Score Strategies. *Drug Saf*.;43(6):549-559. doi: 10.1007/s40264-020-00916-5.

D iGaetano, R., et Waksberg, J. (2002). Trade¬offs in the development of a sample design for case¬control studies. *American Journal of Epidemiology*, 155, 771¬775.

Freund Y., Schapire R. (1996). Experiments with the new boosting algorithm. *International Conference on Machine Learning*, p. 148-156.

Gelman A, Carlin J, Stern H, Rubin D. (1995). Bayesian Data Analysis. *New York, NY: Chapman Hall*.

Goldberg M, Jougla E, Fassa M, Padieu R, Quantin C. (2012). The French public health information system. *Journal of the International Association for Official Statistics*; 28: 31–41.

Hoerl, A. E., and Kennard, D.J. (1970), Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics*, 12, 55-67.

Janes H, Pepe MS. (2008). Matching in studies of classification accuracy:implications for analysis, efficiency, and assessment of incremental value. *Biometrics*.;64(1):1-9. doi: 10.1111/j.1541-0420.2007.00823.x.

Jorgenson MR, Descourouez JL, Lyu B, Astor BC, Garg N, Smith JA, Mandelbrot DA. (2019). The risk of cytomegalovirus infection after treatment of acute rejection in renal transplant recipients. *Clin Transplant*.;33(8):e13636. doi: 10.1111/ctr.13636.

Keller JP, Rice KM (2018). Selecting Shrinkage Parameters for Effect Estimation: The Multi-Ethnic Study of Atherosclerosis. *Am J Epidemio.*. Feb 1;187(2):358-365. doi: 10.1093/aje/kwx225.

Knol MJ, Vandenbroucke JP, Scott P, Egger M. (2008). What do case-control studies estimate? Survey of methods and assumptions in published case-control research. *Am J Epidemiol*. Nov 1;168(9):1073-81. doi: 10.1093/aje/kwn217.

Läärä E. (2011). Study designs for biobank-based epidemiologic research on chronic diseases. *Methods Mol Biol*.;675:165-78. doi: 10.1007/978-1-59745-423-0_6.

Lee BK, Lessler J, Stuart EA. (2010). Improving propensity score weighting using machine learning. *Stat Med.* Feb 10;29(3):337-46. doi: 10.1002/sim.3782.

Liew JW, Peloquin C, Tedeschi SK, Felson DT, Zhang Y, Choi HK, Terkeltaub R, Neogi T. (2022). Proton-Pump Inhibitors and Risk of Calcium Pyrophosphate Deposition in a Population-Based Study. *Arthritis Care Res (Hoboken)*.;74(12):2059-2065. doi: 0.1002/acr.24876.

Loiseau M, Cottenet J, François-Purssell I, Bechraoui-Quantin S, Jud A, Gilard-Pioc S, Quantin C. (2023). Hospitalization for physical child abuse: Associated medical factors and medical history since birth. *Child Abuse Negl*.;146:106482. doi: 10.1016/j.chiabu.2023.106482.

Maitre T, Cottenet J, Godet C, Roussot A, Abdoul Carime N, Ok V, Parrot A, Bonniaud P, Quantin C, Cadranel J. (2021). Chronic pulmonary aspergillosis: prevalence, favouring pulmonary diseases and prognosis. *Eur Respir J*.;58(2):2003345. doi: 10.1183/13993003.03345-2020.

Mezei G, Chang ET, Mowat FS, Moolgavkar SH. (2021). Comments on a recent case-control study of

malignant mesothelioma of the pericardium and the tunica vaginalis testis. *Scand J Work Environ Health.* ;47(1):85-86. doi: 10.5271/sjweh.3909.

Payet C, Polazzi S, Obadia JF, Armoiry X, Labarère J, Rabilloud M, Duclos A. (2021). High-dimensional propensity scores improved the control of indication bias in surgical comparative effectiveness studies. *J Clin Epidemiol*.;130:78-86.

Piroth L, Cottenet J, Mariet AS, Bonniaud P, Blot M, Tubert-Bitter P, Quantin C. (2021). Comparison of the characteristics, morbidity, and mortality of COVID-19 and seasonal influenza: a nationwide, population-based retrospective cohort study. *Lancet Respir Med*.;9(3):251-259. doi: 10.1016/S2213-2600(20)30527-0.

Quantin C, Yamdjieu Ngadeu C, Cottenet J, Escolano S, Bechraoui-Quantin S, Rozenberg P, Tubert-Bitter P, Gouyon JB. (2021). Early exposure of pregnant women to non-steroidal anti-inflammatory drugs delivered outside hospitals and preterm birth risk: nationwide cohort study. *BJOG*.;128(10):1575-1584. doi: 10.1111/1471-0528.16670.

Rey G, Jougla E, Fouillet A, et al. (2009) Ecological association between a deprivation index and mortality in France over the period 1997–2001: variations with spatial scale, degree of urbanicity, age, gender and cause of death. *BMC Public Health*;9:33.

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

Rundle A, Ahsan H, Vineis P. (2012). Better cancer biomarker discovery through better study design. *Eur J Clin Invest*.;42(12):1350-9. doi: 10.1111/j.1365-2362.2012.02727.x.

Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. (2009). High-dimensional Propensity Score Adjustment in Studies of treatment effects using health care claims data. *Epidemiology* ;20(4):512-22

Schneeweiss S, Eddings W, Glynn RJ, Patorno E, Rassen J, Franklin JM. (2017). Variable election for Confounding Adjustment in High-dimensional Covariate Spaces When Analyzing ealthcare Databases. *Epidemiology*.;28(2):237-248. doi: 0.1097/EDE.0000000000000581.

Scott, A. (2006). Études cas-témoins basées sur la population. *Techniques d'enquête*, 32, 137-147.

Simon E, Gouyon JB, Cottenet J, Bechraoui-Quantin S, Rozenberg P, Mariet AS, Quantin C. (2022). Impact of SARS-CoV-2 infection on risk of prematurity, birthweight and obstetric complications: A multivariate analysis from a nationwide, population-based retrospective cohort study. *BJOG*.;129(7):1084-1094. doi: 10.1111/1471-0528.17135.

Taylor JMG. (1986). Choosing the number of controls in a matched case-control study, some sample size, power and efficiency considerations. *Stat Med*; 5(1):29–36. doi: 10.1002/sim.4780050106.

Tibshirani R. (1996). Regression shrinkage and selection via the Lasso. *J R Stat Soc.*;58:267–288

Ury HK. (1975). Efficiency of case-control studies with multiple controls per case: continuous or dichotomous data. *Biometrics*; 31:643–649. doi: 10.2307/2529548.

Viennet M, Tapia S, Cottenet J, Bernard A, Ortega-Deballon P, Quantin C. (2023). Increased risk of colon cancer after acute appendicitis: a nationwide, population-based study. *EClinicalMedicine*.;63:102196. doi: 10.1016/j.eclinm.2023.102196.

Wacholder, S., McLaughlin, J.K., Silverman, D.T. et Mandel, J.S. (1991). Selection of controls in case¬control studies. I. Principles. *American Journal of Epidemiology,* 135, 1019¬1028

Zou H. and Hastie T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 67 : 301–320, 2005.