MÉTHODE D'IDENTIFICATION DES PERSONNES ATTEINTES D'UNE PATHOLOGIE À PARTIR DES BASES MÉDICO-ADMINISTRATIVES : EXEMPLE DE LA MALADIE DE PARKINSON

Frédéric Moisan[1] & Alexis Elbaz[2]

[1]*Institut de veille sanitaire, département santé travail ; 12 rue du Val d'Osne - 94415 Saint-Maurice Cedex France ; f.moisan@invs.sante.fr*
[2] *INSERM U1018, Equipe 11 ; Hôpital Paul Brousse, Bâtiment 15/16 - 16 avenue Paul Vaillant-Couturier - 94807 Villejuif Cedex France ; alexis.elbaz@inserm.fr*

Résumé. La disponibilité croissante des bases médico-administratives fait qu'elles représentent une opportunité pour identifier des patients atteints de certaines maladies. Avant de les utiliser pour cet objectif, il est nécessaire d'évaluer leur validité. Nous avons étudié la validité des bases de remboursements de médicaments pour identifier des patients traités pour maladie de Parkinson dans cinq caisses de la Mutualité sociale agricole en 2007. Nous avons comparé les profils de remboursement d'antiparkinsoniens chez des patients dont le diagnostic de maladie de Parkinson était confirmé par un neurologue et chez des patients traités pour une autre raison. En suivant une approche structurée, nous avons développé un modèle permettant d'estimer la probabilité qu'ont les personnes avec un remboursement d'antiparkinsonien d'être traités pour la maladie de Parkinson. Parmi 1 114 participants avec au moins un remboursement d'antiparkinsonien en 2007, un diagnostic de maladie de Parkinson a été confirmé pour 320 (29 %) d'entre eux tandis que 794 (71 %) étaient traités pour une autre raison. Le modèle logistique incluant les doses cumulées des différents types d'antiparkinsoniens et la régularité du traitement est caractérisé par de bonnes performances (statistique c=0,953 ; sensibilité=92,5 % ; spécificité=86,4 %). Nos résultats montrent qu'il est possible d'identifier à partir des bases médico-administratives les personnes atteintes de certaines pathologies sans recueil d'informations supplémentaires. En plus d'une utilisation à des fins de surveillance épidémiologique, ces données étant recueillies de manière systématique, elles sont disponibles pour les répondants et les non-répondants d'une étude, et peuvent être utilisées pour le traitement de la non-réponse.

Mots-clés. Maladie de Parkinson, épidémiologie, bases médico-administratives, modèle prédictif, courbe ROC.

Abstract. The increasing availability of medico-administrative databases represents an opportunity to identify patients with certain diseases. Before using them for this purpose, it is necessary to assess their validity. We investigated their validity to identify patients treated for Parkinson's disease among members of the Mutalité sociale agricole in five French districts in 2007. We compared patterns of antiparkinsonian drugs use in patients for whom a neurologist confirmed the diagnosis and in patients treated for another reason, and developed prediction models in order to estimate the probability that antiparkinsonian drug users were treated for Parkinson's disease. Among 1,114 persons who used antiparkinsonian drugs in 2007, 320 (29%) were considered to have Parkinson's disease, while 794 (71%) were treated for another reason. Different models were compared in terms of quality of predictions. A logistic model including cumulative doses of different antiparkinsonian drugs and regularity of treatment displayed good performance (c statistic=0.953, sensitivity=92.5%, specificity=86.4%). Our results show that it is possible to identify people affected by some disease from the medico-administrative databases and without gathering additional information. In addition to use for epidemiological surveillance purpose, data from medico-administrative databases are available for both respondents and non-respondents, and can also be used for the treatment of non-response.

Keywords. Parkinson's disease, epidemiology, medico-administrative databases, predictive model, ROC curve.

1 Introduction

En France, comme dans de nombreux pays, les bases médico-administratives sont de plus en plus accessibles au niveau national et elles représentent une source de données supplémentaires lors de la mise en œuvre d'une étude. Parmi ces données, les informations de l'Assurance maladie présentent un intérêt tout particulier pour les enquêtes de santé car elles contiennent notamment des informations détaillées sur les remboursements de médicaments (type, dosage, fréquence, etc.).

L'objectif de ce travail est de présenter une démarche permettant, d'une part, de développer un modèle pour identifier les personnes atteintes d'une pathologie à travers leurs profils de remboursements de médicaments, d'autre part, d'évaluer la performance de ce modèle.

Compte tenu des difficultés méthodologiques pour identifier les patients parkinsoniens (méthode de référence – étude en porte-à-porte – couteuse, certificat de décès peu fiable, etc.) et l'intérêt d'étudier la maladie de Parkinson dans le monde agricole, la démarche proposée a été utilisée pour d'identifier les patients parkinsoniens parmi les affiliés de la Mutualité Sociale Agricole (MSA).

2 Matériels et méthodes

Cette étude a été menée auprès des affiliés de la MSA, âgés de 18 ans ou plus, dans cinq départements français (Charente-Maritime, Côte-d'Or, Gironde, Haute-Vienne, Mayenne). La MSA assure la protection sociale du monde agricole et rural. En 2007, la MSA constituait le deuxième régime de protection sociale en France, avec environ 3 800 000 personnes protégées en maladie (assurés et ayants-droits).

Le protocole de l'étude a reçu un avis favorable du Comité de protection des personnes de l'Hôpital de la Pitié-Salpêtrière et a fait l'objet d'une déclaration à la Commission nationale de l'informatique et des libertés (Cnil).

2.1 Constitution des groupes de comparaison

A partir des bases médico-administratives disponibles à la MSA de chaque département, tous les individus avec au moins un remboursement de médicaments antiparkinsoniens en 2007 ont été identifiés i) s'ils étaient âgés de 80 ans ou moins ; ii) s'ils n'étaient pas en affection longue durée (ALD) pour démence ou maladie psychiatrique. De plus, parmi les individus identifiés, ont été exclu ceux ayant une ALD pour maladie de Parkinson (MP) dont la durée était supérieure à 15 ans.

Nous avons obtenu des renseignements cliniques détaillés pour ces personnes afin de pouvoir les classer en deux groupes : 1) les participants traités pour une MP confirmée par un examen clinique réalisé par un neurologue ; 2) les participants traités pour une autre raison.

Nous avons comparé les profils de remboursement de médicaments antiparkinsoniens entre ces deux groupes afin de développer et évaluer les performances d'un modèle prédictif du statut parkinsonien.

2.2 Définition des prédicteurs

L'ensemble des prédicteurs ont été définis à partir des informations disponibles dans les bases médico-administratives. Ils incluaient des informations démographiques (âge, sexe), le nombre de consultations avec un neurologue ou avec un médecin généraliste, la proportion de temps pendant laquelle une personne avait été traitée avec un médicament antiparkinsonien durant l'année et la dose cumulée annuelle de huit classes de médicaments antiparkinsoniens. Les quinze molécules antiparkinsoniennes disponibles sur le marché en 2007 ont été regroupées en huit classes

(tableau 1), d'une part, car les médicaments appartenant à la même classe sont habituellement utilisés de façon identique, d'autre part, car le regroupement permet de limiter le nombre prédicteurs et permet d'avoir des variables avec un nombre de sujets exposés suffisant.

Tableau 1: Médicaments antiparkinsoniens disponibles en France en 2007.

Famille	Molécule	Principaux médicaments
Agonistes dopaminergiques de type 1[a]	Pramipéxole	Sifrol
	Ropinirole	Requip
	Pergolide	Celance
Agonistes dopaminergiques de type 2[b]	Apomorphine	Apokinon
	Bromocriptine	Parlodel
	Lisuride	Dopergine
Amantadine	Amantadine	Mantadix
Anticholinergiques	Trihexyphenidyl	Artane
	Bipéridène	Akineton
	Tropatépine	Lepticur
Inhibiteurs de la COMT	Entacapone	Comtan
	Tolcapone	Tasmar
Inhibiteurs de la MAO	Sélégiline	Sélégiline
Lévodopa	Lévodopa +/- Carbidopa / Bensérazide	Modopar, Sinemet
Piribedil	Piribédil	Trivastal

Abréviations : COMT, catéchol-O-méthyl transférase ; MAO, monoamine oxydase.

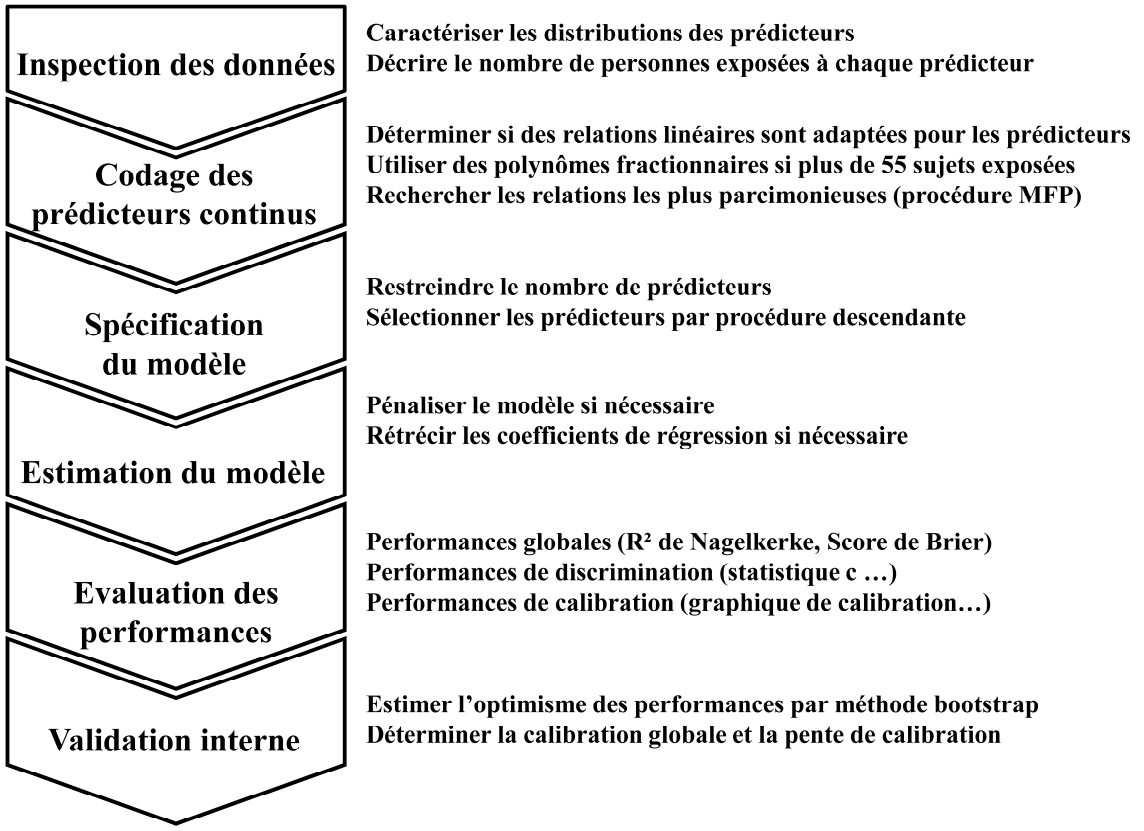
[a]Souvent utilisé pour traiter la maladie de Parkinson.

[b]Rarement utilisé pour traiter la maladie de Parkinson.

2.3 Démarche pour développer le modèle prédictif

Nous avons utilisé la régression logistique afin de développer un modèle prédictif du statut parkinsonien (oui/non) en suivant une approche pas à pas structurée [1] (figure 1).

Figure 1 : Démarche suivie pour développer le modèle prédictif



Nous avons choisi de présenter comme principal modèle celui qui a été développé à partir de la dose cumulée des médicaments antiparkinsoniens prescrits pendant un an (2007).

Nous avons utilisé les polynômes fractionnaires pour transformer les variables continues (âge, proportion de temps traité ; doses cumulées des différentes classes d'antiparkinsoniens) à l'aide des commandes fracpoly et mfp du logiciel Stata® (version 10, StataCorp LP, collège Station, Texas). Cette approche permet d'obtenir une transformation flexible des variables dans un modèle multivarié à l'aide d'un algorithme itératif [2 ; 3]. Pour les variables caractérisant l'utilisation des classes d'antiparkinsoniens, cette approche a été utilisée uniquement pour ceux qui étaient utilisés par au moins 55 personnes ; pour les autres classes, nous avons utilisé des variables binaires (oui/non). Les variables correspondant au nombre de consultations avec un neurologue ou un généraliste ont été définies comme des variables à trois ou quatre classes respectivement.

Nous avons tout d'abord inclus dans le modèle l'ensemble des prédicteurs (n=18). Nous avons ensuite utilisé une procédure de sélection descendante pas-à-pas ($p \leq 0,20$ pour retenir les variables dans le modèle). Pour chaque variable retenue dans le modèle multivarié, nous avons calculé le rapport entre le coefficient de régression et son erreur standard (Z-ratio) ; la comparaison des Z-ratio associés à différentes variables permet de comparer la force de l'association entre l'évènement d'intérêt et ces variables. Puisque le nombre d'observations par prédicteur était élevé (n=62), le surajustement aux données (overfitting) n'était pas un problème et des méthodes de pénalisation ou de rétrécissement (shrinkage) des coefficients n'ont pas été nécessaires [4].

Plusieurs mesures ont été utilisées pour évaluer différents aspects des performances du modèle : le R^2 de Nagelkerke et le score de Brier évaluent les performances globales du modèle ; l'aire sous la courbe (statistique c), la sensibilité, la spécificité, la valeur prédictive positive et négative et la pente de discrimination permettent d'évaluer la capacité de discrimination du modèle ; le test de le Cessie et Van Houwelingen évalue l'ajustement du modèle aux données.

Pour la validation interne du modèle, nous avons utilisé la méthode du bootstrap pour estimer

l'optimisme du modèle, qui a ensuite été utilisé pour corriger les performances (R^2 de Nagelkerke, statistique c) [5]. Deux mesures de surajustement aux données ont été estimées : la calibration globale (calibration in the large) et la pente de calibration [6].

Les performances du modèle ont été estimées parmi l'ensemble des personnes ayant reçu un antiparkinsonien en 2007. Dans certains cas, il est intéressant de calculer la spécificité et la valeur prédictive négative pour l'ensemble de la population ; pour cela, nous avons considéré que les affiliés de la MSA des cinq départements qui n'ont reçu aucun antiparkinsonien en 2007 et vérifiaient les critères d'inclusion étaient de vrais négatifs.

Nous avons réalisé des tests bilatéraux qui ont été considérés significatifs au seuil de signification de 0,05.

Ces analyses ont été effectuées à l'aide du package DiagnosisMed et de la fonction val.prob.ci implémentées dans le logiciel R (version 2.11.0, R Foundation for Statistical Computing, Vienna, Austria, Vienne, Autriche).

3 Résultats

3.1 Inclusion et caractéristiques des participants

Parmi les 202 087 affiliés vérifiant les critères d'inclusion, 1 540 personnes ont au moins un remboursement de médicament antiparkinsonien en 2007. Parmi ces derniers, le motif de prise de médicament antiparkinsonien est connu pour 1 114 participants : 320 sont traités pour MP et 794 sont traités par antiparkinsonien pour une autre raison. Une proportion élevée (71 %) des participants qui ont reçu une prescription d'antiparkinsonien en 2007 ne sont donc pas traités pour la MP.

Les 320 patients parkinsoniens sont plus âgés et plus souvent des hommes que les 794 patients non parkinsoniens. Ils ont vu plus souvent un neurologue avec, de plus, un plus grand nombre de consultations par an. En revanche, une proportion similaire de patients avec et sans MP ont vu un médecin généraliste au moins une fois par an, même si, les patients parkinsoniens ont consulté plus fréquemment le généraliste que les autres participants. La proportion de temps traité est plus élevée chez les patients parkinsoniens que chez les autres patients. Les médicaments consommés sont très différents entre les deux groupes. Les patients atteints de MP sont plus souvent traités par lévodopa, amantadine, sélégiline, agonistes dopaminergiques de type 1 et inhibiteurs de la COMT que les patients non parkinsoniens. Ils ont également reçu des doses cumulées plus élevées pour l'ensemble de ces médicaments. A l'inverse, les anticholinergiques et le piribédil sont plus souvent prescrits aux patients non parkinsoniens, mais les doses cumulées de piribédil sont plus élevées chez les patients parkinsoniens.

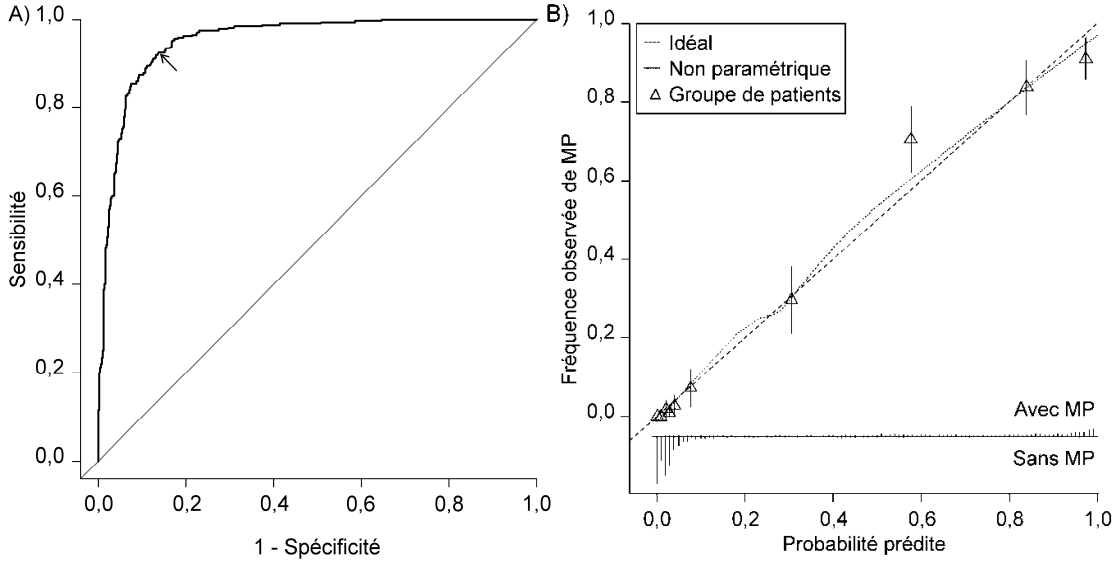
3.2 Performances du modèle

Dans le modèle développé à partir des doses cumulées d'antiparkinsoniens, l'âge et l'amantadine n'ont pas été retenus et les variables les plus fortement associées au statut parkinsonien sont : lévodopa, agonistes dopaminergiques de type 1, piribédil, la proportion de temps traité, nombre de consultations avec un neurologue et les inhibiteurs de la MAO.

Pour ce modèle, la proportion de variance expliquée (R^2 de Nagelkerke) est égale à 71,4 %. Le modèle est caractérisé par de bonnes performances en terme de discrimination (statistique c = 0,953 ; pente de discrimination = 0,625 ; partie A de la figure 2). Pour le seuil optimal (p = 0,255), la sensibilité est de 92,5 %, tandis que la spécificité est un peu plus faible (86,4 %). Malgré les bonnes performances du modèle, 108 participants sans MP sont identifiés par le modèle comme ayant la MP (faux positifs) et 24 participants avec un diagnostic de MP ne sont pas identifiés par le modèle

(faux négatifs).

Figure 2 : Courbe ROC (A) et courbe de calibration (B) du modèle prédictif développé à partir des doses cumulées de médicaments antiparkinsoniens.



Abréviation : MP, maladie de Parkinson.

La flèche dans la partie A indique le seuil optimal (p = 0,255) qui maximise le nombre d'individus correctement classés parmi les personnes ayant reçu au moins un médicament antiparkinsonien en 2007 et pour lequel les performances de discrimination sont calculées.

La courbe de calibration montre qu'il existe une bonne corrélation entre les valeurs observées et prédites (partie B de la figure 2). Le test de le Cessie de Van Houweligen montre que le modèle est caractérisé par un bon ajustement aux données (p = 0,794).

L'optimisme du R^2 de Nagelkerke et de la statistique c estimés par la méthode boostrap sont très faibles (< 3 %). Les performances du modèle sont peu modifiées après prise en compte de l'optimisme. En moyenne, les valeurs prédites et observées après la méthode boostrap sont très proches (calibration globale = -3,8 %). De plus, le sur-ajustement est limité (pente de calibration = 92,2 %).

Afin de calculer la spécificité et la valeur prédictive négative parmi l'ensemble des affiliés, les sujets qui n'ont pas utilisé d'antiparkinsonien en 2007 et qui vérifiaient les critères d'inclusion ont été considérés comme de vrais négatifs (n = 200 547). En utilisant le même seuil que précédemment, la spécificité est égale à 99,95 % et la valeur prédictive négative est égale à 99,99 % parmi l'ensemble des affiliés.

4 Discussion

Cette étude, conduite en 2007 parmi les affiliés à la MSA de cinq départements français, montre que les bases de remboursement de médicaments représentent un outil intéressant pour estimer la probabilité qu'une personne a d'être traitée pour la maladie de Parkinson. Un modèle prédictif a été développé à partir des profils de remboursement de médicaments pouvant être utilisés pour traiter la maladie de Parkinson ; il est caractérisé par des performances satisfaisantes. Cela repose sur le fait,

que comme attendus les profils de remboursements de médicaments (type, dose, fréquence) des patients parkinsoniens étaient très différents de ceux des personnes ayant eu des médicaments antiparkinsoniens pour une autre raison.

Un seul modèle prédictif de maladie de Parkinson a été développé antérieurement dans le cadre de l'étude de cohorte de Rotterdam à partir d'un petit nombre d'utilisateurs de médicaments antiparkinsoniens (n=63) [7]. Les performances en termes de discrimination ont été calculées parmi l'ensemble des participants avec une statistique c égale à 0,93. Aucune information quantitative sur la consommation médicamenteuse n'a été prise en compte dans le modèle et la calibration et la validité interne n'ont pas été évaluées.

Plusieurs facteurs sont susceptibles d'influencer les performances du modèle prédictif. Tout d'abord, par définition, seuls les patients traités sont présents dans les bases de remboursement de médicaments. Cette approche nécessite donc de considérer que la majorité des patients sont traités. En France, cette hypothèse est raisonnable pour la MP, sauf chez les sujets les plus âgés. Dans une étude en population générale conduite en Gironde (Paquid, 1988-89), 11 % des patients parkinsoniens n'avaient pas été préalablement diagnostiqués et ont été identifiés au moment de l'étude alors qu'ils n'étaient pas traités ; ils étaient, pour la plupart, âgés de plus de 80 ans [8]. La difficulté d'identifier les patients parkinsoniens chez les sujets les plus âgés résulte de différents facteurs (incertitude diagnostique, comorbidités, diagnostic plus tardif et difficile, etc.). Les modèles prédictifs établis à partir de profils de remboursements de médicaments pourraient donc être moins fiables chez les personnes les plus âgées, en particulier en raison d'une moindre sensibilité.

Ensuite, le modèle repose uniquement sur les données enregistrées dans les bases de remboursement. Ainsi, les prescriptions pour certains sous-groupes de la population peuvent ne pas être incluses dans les bases de remboursement. Dans cette étude, les personnes institutionnalisées dans des maisons de retraite avec des pharmacies à usage interne n'ont pas été identifiées. Toutefois, d'après une étude précédente, une minorité des consommateurs de médicaments antiparkinsoniens (< 2 %) est institutionnalisée dans des maisons de retraite avec des pharmacies à usage interne avant 80 ans.

Parmi les faux positifs identifiés par le modèle prédictif, certains diagnostics sont plus fréquents (paralysie supranucléaire progressive, atrophie multisystématique, dégénérescence cortico-basale, syndrome parkinsonien secondaire à un syndrome démentiel, démence à corps de Lewy). En l'absence de tout autre traitement, les patients atteints de ces pathologies reçoivent de la lévodopa aux mêmes doses que les patients parkinsoniens. De plus, un certain nombre de patients atteints de tremblement essentiel sont traités comme la MP (erreur de diagnostic), et ils sont identifiés par le modèle comme ayant la MP. En conséquence, le modèle prédictif entraine des erreurs de classement (faux positifs) plus particulièrement pour les personnes qui ont un diagnostic erroné de MP ou pour les personnes présentant des pathologies qui sont traitées comme la MP, essentiellement les autres syndromes parkinsoniens neurodégénératifs ; ces pathologies sont toutefois considérablement plus rares que la MP. Concernant les faux négatifs le modèle identifie moins bien les patients avec des formes moins évoluées et qui nécessitent des doses moins élevées de médicaments antiparkinsoniens.

Parmi les limites des modèles prédictifs développés, il y a l'absence de validation externe qui nécessiterait la collecte de données similaires dans une autre population. Toutefois, en France, toutes les personnes qui bénéficient de l'Assurance maladie ont potentiellement le même accès aux soins médicaux et les affiliés des différents systèmes d'Assurance maladie ont accès aux mêmes médecins, y compris les neurologues. Dans notre étude, la distribution des médicaments utilisés pour traiter les patients parkinsoniens et les doses journalières estimées sont classiques et similaires à celles habituellement observées dans les consultations spécialisées. Il est peu vraisemblable qu'une étude réalisée dans une autre population française conduise à des résultats très différents.

5 Conclusion

Finalement, bien qu'elles n'aient pas été conçues dans cet objectif, les bases de remboursement de médicaments, qui sont maintenant de plus en plus de accessibles dans de nombreux pays, peuvent permettre de répondre à certaines questions épidémiologiques. Ainsi la démarche proposée montre qu'il est possible d'utiliser ces bases pour développer – avec une validité connue – un modèle permettant d'identification les personnes atteintes d'une pathologie.

Une fois l'outil développé, plusieurs utilisations peuvent être envisagées. En plus de l'application pour estimer la prévalence d'une maladie ou étudier sa distribution spatiale, les modèles prédictifs peuvent, s'ils sont utilisés de façon répétée, permettre la surveillance épidémiologique de maladies en étudiant les tendances temporelles. Dans le cadre d'études étiologiques, les modèles prédictifs peuvent être utilisés soit pour définir directement l'événement d'intérêt – et la qualité de sa définition est alors connue grâce aux performances du modèle –, soit comme outil de dépistage dans le cadre d'une étude en deux temps. Une approche similaire peut être utilisée pour recruter des participants dans des essais thérapeutiques ou des études d'intervention. Les investigateurs peuvent alors sélectionner le seuil de probabilité lors de la première étape en fonction de la représentativité qu'ils souhaitent et du nombre de faux positifs acceptables en particulier en termes de coûts et de temps. Enfin, dans la mesure où les informations nécessaires pour appliquer des modèles prédictifs proviennent uniquement de données passives, la fréquence de la pathologie peut être connue parmi les non-répondants d'une étude et cette information peut être utilisée pour le traitement de la non-réponse.

Bibliographie

[1] Steyerberg EW (2009). Developing valid prediction models. Dans: *Clinical prediction models: a practical approach to development, validation, and updating*. New York: Springer : 113-331.

[2] Royston P, Altman DG (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *J R Stat Soc Ser C Appl Stat* ; 43(3):429-67.

[3] Royston P, Sauerbrei W (2005). Building multivariable regression models with continuous covariates in clinical epidemiology--with an emphasis on fractional polynomials. *Methods Inf Med* ; 44(4):561-71.

[4] Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR (1996). A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol* ; 49(12):1373-9.

[5] Harrell FE (2001). Resampling, validating, describing and simplifying the model. Dans : *Regression modeling strategies with applications to linear models, logistic regression and survival analysis*. New York: Springer : 87-103.

[6] Harrell FE (2001). Binary logistic regression. Dans : *Regression modeling strategies with applications to linear models, logistic regression and survival analysis*. New York: Springer : 215-67.

[7] Van de Vijver DA, Stricker BH, Breteler MM, Roos RA, Porsius AJ, de Boer A (2001). Evaluation of antiparkinsonian drugs in pharmacy records as a marker for Parkinson's disease. *Pharm World Sci* ; 23(4):148-52.

[8] De Rijk MC, Tzourio C, Breteler MMB, Dartigues JF, Amaducci L, Lopez-Pousa S et al. (1997) Prevalence of parkinsonism and Parkinson's disease in Europe: The EUROPARKINSON collaborative study. J Neurol Neurosurg Psychiatry ; 62(1):10-5.