American Journal of Epidemiology
Vol. 174, No. 3
DOI: 10.1093/aje/kwr081
Advance Access publication:
May 23, 2011

Practice of Epidemiology

Prediction Model of Parkinson's Disease Based on Antiparkinsonian Drug Claims

Frédéric Moisan*, Véronique Gourlet, Jean-Louis Mazurie, Jean-Luc Dupupet, Jean Houssinot, Marcel Goldberg, Ellen Imbernon, Christophe Tzourio, and Alexis Elbaz

* Correspondence to Frédéric Moisan, INSERM Unité 708–Neuroepidemiology, Hôpital de la Salpêtrière, 47 Blvd. de Hôpital, 75651 Paris Cedex 13, France (e-mail: frederic.moisan@upmc.fr).

Initially submitted October 4, 2010; accepted for publication February 25, 2011.

Drug claims databases are increasingly available and provide opportunities to investigate epidemiologic questions. The authors used computerized drug claims databases from a social security system in 5 French districts to predict the probability that a person had Parkinson's disease (PD) based on patterns of antiparkinsonian drug (APD) use. Clinical information for a population-based sample of persons using APDs in 2007 was collected. The authors built a prediction model using demographic variables and APDs as predictors and investigated the additional predictive benefit of including information on dose and regularity of use. Among 1,114 APD users, 320 (29%) had PD and 794 (71%) had another diagnosis as determined by study neurologists. A logistic model including information on cumulative APD dose and regularity of use showed good performance (*c* statistic = 0.953, sensitivity = 92.5%, specificity = 86.4%). Predicted PD prevalence (among persons aged ≥18 years) was 6.66/1,000; correcting this estimate using sensitivity/specificity led to a similar figure (6.04/1,000). These data demonstrate that drug claims databases can be used to estimate the probability that a person is being treated for PD and that information on APD dose and regularity of use improves models' performances. Similar approaches could be developed for other conditions.

antiparkinsonian agents; Parkinson disease; prediction; predictive value of tests; prescriptions; prevalence

Abbreviations: APD, antiparkinsonian drug; ICD, *International Classification of Diseases*; LED, levodopa equivalent dose; MSA, Mutualité Sociale Agricole; PD, Parkinson's disease; SD, standard deviation.

Drug claims databases are increasingly available and offer the potential to identify patients with specific conditions for epidemiologic studies. If proven to be a valid source, they would represent an inexpensive approach to evaluating disease frequency. We explored the feasibility of using drug claims databases to estimate the probability that a person had Parkinson's disease (PD) based on patterns of antiparkinsonian drug (APD) use.

PD is the most frequent cause of parkinsonism (1). Its main clinical features are resting tremor, bradykinesia, rigidity, and postural instability. Diagnosis is mainly based on medical history and neurologic examination. APDs improve PD symptoms, and several APDs are available; the most commonly used agent is levodopa (L-3,4-dihydroxyphenylalanine), but other drugs are increasingly being prescribed. While some APDs are mainly used for PD (e.g., levodopa, selegiline), others are frequently used for other conditions (e.g., dopamine agonists for restless leg syndrome; anticholinergic agents for drug-induced parkinsonism; piribedil for tinnitus), often less regularly and at lower doses (2, 3). Previous studies have estimated PD prevalence based on APD use (4–10), but the reliability of this approach is unknown.

To estimate the probability that APD users have PD using drug claims databases, we identified a population-based sample of persons who had used any type of APD in 2007 from drug claims databases and obtained clinical information for them. We then used demographic variables and APDs as potential predictors to develop a PD prediction model based on APD claims, and investigated the benefit of including information on dose and regularity of use.

Table 1. Antiparkinsonian Drugs Available in France in 2007

Class of Antiparkinsonian Drug	Antiparkinsonian Drug	Levodopa Equivalent Dose, mg/100 mg levodopa[a]
Amantadine	Amantadine	100
Anticholinergic agents	Trihexyphenidyl	NA
	Biperiden	NA
	Tropatepine	NA
Catechol-*O*-methyl transferase inhibitors	Entacapone	NA
	Tolcapone	NA
Levodopa	Levodopa + carbidopa/benserazide	100; 75[b]; 133[c]
Piribedil	Piribedil	100
Selegiline	Selegiline	10
Type 1 dopamine agonists[d]	Pramipexole (salt)	1
	Ropinirole	5
	Pergolide	1
Type 2 dopamine agonists[e]	Apomorphine	10
	Bromocriptine	10
	Lisuride	1

Abbreviation: NA, not available.

[a] The definition of levodopa equivalents was based on a comprehensive review of 56 studies (12).

[b] If catechol-*O*-methyl transferase inhibitors (entacapone; no users of tolcapone were identified) were prescribed on the same date as levodopa.

[c] Slow-release form of levodopa.

[d] Pramipexole, ropinirole, and pergolide, which are often used for treatment of Parkinson's disease, were grouped as type 1 dopamine agonists.

[e] Lisuride, bromocriptine, and apomorphine are infrequently used for treatment of Parkinson's disease and were grouped as type 2 dopamine agonists.

MATERIALS AND METHODS

This study was conducted among adult members (aged $\geq$18 years) of the Mutualité Sociale Agricole (MSA) in 5 French districts (Charente-Maritime, Côte-d'Or, Gironde, Haute-Vienne, and Mayenne). MSA is responsible for the reimbursement of health-related expenses to workers in agriculture and related occupations (farmers; workers in farms, silos, seed shops, and agricultural cooperatives; professional gardeners; and employees of the MSA, an insurance company, and a bank). Workers benefit from health insurance both while employed and when retired. MSA covers their spouses' (if unemployed) and children's health expenses. In 2007, MSA covered approximately 4 million persons.

The Ethical Committee of the Pitié-Salpêtrière University Hospital approved the study protocol.

Identification of PD patients from drug claims databases

In France, APDs cannot be obtained without medical prescription, and their delivery is electronically registered in drug claims databases. We used computerized MSA drug claims databases to identify persons from the 5 districts who had bought any APD (defined as any drug that can be used to treat PD; Table 1) in 2007 and met the following criteria: age $\leq$80 years on January 1, 2007; disease duration $\leq$15 years (if receiving free health care for PD); and no free health care for dementia or psychiatric disease. All subjects who had filled at least 1 prescription for levodopa, entacapone, tolcapone, ropinirole, pramipexole, apomorphine, bromocriptine, or selegiline were invited to be examined by a neurologist in order to confirm PD using standardized criteria (1), unless they reported taking small doses of dopamine agonists for restless leg syndrome, treatment was discontinued after $\leq$1 month, or there was a documented history of drug-induced parkinsonism. Patients who used only piribedil, amantadine, or anticholinergic agents (which are rarely used for PD) were first contacted by mail; they were asked why these drugs had been prescribed, and those who answered PD/parkinsonism or did not know were invited to be examined. We did not contact women aged $\leq$50 years using small doses of bromocriptine for short periods (lactation suppression) or persons using anticholinergic agents with neuroleptics (drug-induced parkinsonism). Persons institutionalized in nursing homes with indoor pharmacies were not identified, because drugs delivered to them are not included in French drug claims databases.

Predictors

We selected the following variables from the databases: sex, age, number of visits to a neurologist or general

practitioner per year, and APD use. In France, the reason for prescribing a drug is not coded for outpatient visits. Thus, *International Classification of Diseases* (ICD) codes were not available; however, such codes are considered to be inaccurate for PD (11). We combined APDs belonging to the same classes to reduce the number of predictors and to define variables with sufficient numbers of subjects exposed (Table 1). For APDs, we defined "ever/never" variables and computed quantitative indices: number of prescriptions filled per year; number of boxes of medication bought per year; cumulative dose per year (number of boxes bought per year × number of tablets per box × tablet strength); mean daily dose (cumulative dose per year divided by duration; see Web Figure 1, part A, which appears on the *Journal*'s Web site (http://aje.oxfordjournals.org/)). We used levodopa equivalent doses (LEDs) to express cumulative and mean daily doses for APDs with known LEDs (Table 1) (12). For classes of APDs, cumulative and mean daily doses were computed in LEDs when data on all individual APDs were expressed in LEDs (levodopa, piribedil, type 1 dopamine agonists).

In order to assess whether participants received APDs for short periods or were treated regularly, we estimated the proportion of time during which each subject had been treated (total duration of treatment divided by time between date of first prescription and end of follow-up; see Web Figure 1, part B).

Model development and internal validation

To develop the prediction model, we used logistic regression with PD as the outcome. We followed a structured step-by-step approach that is described in more detail in the Web Appendix (13).

Briefly, continuous variables (age, proportion of time treated, APD cumulative doses) were fitted with smooth and flexible transformations using multivariable fractional polynomials selected by means of an iterative algorithm including all predictors (fracpoly/mfp commands in Stata 10; StataCorp LP, College Station, Texas) (14, 15). This approach was used for APD classes including more than 55 persons; for other classes, we used binary variables. Numbers of neurologist and general practitioner visits per year were defined as 3- and 4-level variables.

We first included all candidate predictors ($n = 18$) in the model. We then used a backward selection procedure, with a 2-sided P value ≤ 0.20 for retention of variables in the model. We computed Z ratios (regression coefficient/standard error) to compare the strength of the association across predictors. Since the number of observations per predictor was large ($n = 62$), overfitting was not an issue, and shrinkage or penalized estimation was not necessary (16).

We estimated several measures to assess different aspects of the model's performance (Web Appendix). Nagelkerke's R^2 and scaled Brier score were used to assess overall model performance. The area under the receiver operating characteristic curve (c statistic), sensitivity, specificity, positive predictive value, negative predictive value, and discrimination slope were used to assess discriminative ability. Calibration plots and the le Cessie and van Houwelingen test (17) were used to assess calibration and goodness of fit.

For internal validation of the final model, we used bootstrapping in order to estimate optimism, which was used to correct the model's performance (Nagelkerke's R^2, c statistic) (18). Two measures of overfitting (calibration-in-the-large, calibration slope) were estimated (19).

Performance was measured among persons who had received any APD. In some cases, it may be useful to estimate specificity and negative predictive value at the total population level; for this purpose, we considered as true negatives persons who verified inclusion criteria and did not receive APDs in 2007.

Our main analyses were based on cumulative doses of APDs delivered in 2007. To assess whether including quantitative information on APD doses, proportion of time treated, and number of neurologist and/or general practitioner visits improved model performance in comparison with binary variables, we followed identical steps to develop another prediction model with all variables coded as binary variables, except age (fractional polynomial).

In sensitivity analyses, we used alternative quantitative variables (number of prescriptions per year, number of boxes per year, mean daily dose) and evaluated the impact of using shorter time periods to define drug use (3, 6, and 9 months).

Analyses were performed using the Design and DiagnosisMed packages and the val.prob.ci function in R, version 2.11.0 (R Foundation for Statistical Computing, Vienna, Austria). P values were 2-sided, and the significance level was set at 0.05.

PD prevalence

We used the prediction model based on cumulative dose of APDs to predict the probability that a person was treated for PD and to estimate the prevalence of PD among MSA affiliates from the 5 districts on December 31, 2007.

The model was applied to affiliates who were alive and aged $\geq$18 years on December 31, 2007, who had used any APD in 2007. The logit of the probability that a person has PD is computed by summing the intercept and the estimates from the logistic model multiplied by the value of each variable for this person; the probability of PD is the inverse of $(1 + e^{-\text{logit}})$. Persons with a predicted probability equal to or above the probability cutoff that maximized the Youden index were considered to have been treated for PD; this is the cutoff that maximizes the number of correctly classified persons (20). Prevalence was computed by dividing the number of predicted PD cases by the number of MSA affiliates who were alive on December 31, 2007.

We computed prevalence in subjects aged $\geq$18 years, overall and by sex and 10-year age group. Assuming that there were no PD cases in persons under age 18 years, we estimated the sex- and age-standardized prevalence (direct standardization) on the basis of the age/sex distribution of the 2007 French population (21). We computed a corrected number of predicted PD cases by using the model's sensitivity and specificity to assess the impact of diagnostic misclassification (22).

RESULTS

Among 202,087 subjects meeting the inclusion criteria, 1,540 received 1 or more APDs in 2007 (Web Figure 2). Depending upon which drug they had used, 723 of these persons were directly contacted to be interviewed and 817 were first contacted by mail to obtain additional information; 52 of these 817 persons answered that they used APD for parkinsonism or did not know why they used it, and 188 (23%) did not respond. Nonresponders were younger (64 years; standard deviation (SD), 15) than responders (68 years (SD, 11); $P < 0.001$) but used similar types of APDs ($P = 0.579$). In total, 775 (723 + 52) persons were contacted to be interviewed; 4 persons died before the study began, and 69 persons could not be contacted. Of the 702 remaining persons, 74 did not meet the inclusion criteria and were excluded, 119 were treated for restless leg syndrome or drug-induced parkinsonism or discontinued treatment after $\leq$1 month, and 509 were invited to be examined by one of the study neurologists. Among these 509 persons, 91 (18%) refused; persons who refused were older (75 years (SD, 4)) than those who accepted (72 years (SD, 6); $P < 0.001$) and were less likely to use type 1 dopamine agonists (15% vs. 29%; $P = 0.006$). Of the remaining 418 persons, 320 had PD and 98 did not. Therefore, our analyses were based on 1,114 participants: 320 with PD and 794 (98 + 119 + 577) without PD (Web Figure 2). Thus, a large proportion of participants (71%) who had received at least 1 APD did not have PD.

Table 2 shows participants' characteristics. In univariate analyses, PD patients were older and more likely to be male than other subjects. They also more often saw a neurologist and had a greater number of neurologist visits. While similar proportions of patients with and without PD saw a general practitioner at least once a year, PD patients had more visits per year. The proportion of time treated was greater in subjects with PD than in those without PD. PD patients were more often treated with levodopa, amantadine, selegiline, dopamine agonists, and catechol-*O*-methyl transferase inhibitors than non-PD patients; PD patients received higher cumulative doses of all of these drugs. Anticholinergic agents and piribedil were more often delivered to non-PD patients, but the cumulative dose of piribedil was higher in PD patients. Web Table 1 shows other quantitative measures of APD use. Web Table 2 shows the performance of each APD among persons who received APDs in 2007; levodopa had the best combination of sensitivity (86.6%) and specificity (82.9%).

Table 3 shows the parameter estimates from the multivariable prediction model including information on APD cumulative doses. Age and amantadine were not retained in the final model. Variables with the strongest association with PD were (by decreasing *Z* ratio): levodopa, type 1 dopamine agonists, piribedil, proportion of time treated, number of neurologist visits, and selegiline.

Web Table 3 shows the model's performance. The proportion of the variance explained (R^2) was equal to 71.4%. Figure 1 shows the receiver operating characteristic curve (part A) and box plots of predicted probabilities (part B). The model displayed excellent performance in terms of discrimination (*c* statistic = 0.953, discrimination slope = 0.625) and represents a clear improvement over models that included individual APDs (Web Table 2). For the optimal cutoff, sensitivity was 92.5%, while specificity was lower (86.4%). Web Figure 3 shows discrimination performance for other probability cutoffs. The calibration plot showed good agreement between observed and predicted probabilities (Figure 2). According to the le Cessie–van Houweligen test, the model displayed adequate fit.

In order to compute specificity and negative predictive value at the level of all MSA affiliates, we considered persons who verified inclusion criteria and did not use APDs in 2007 as true negatives ($n = 200{,}547$); using the same cutoff as in the previous analysis, specificity was 99.95% and the negative predictive value was 99.99%.

We investigated whether specific diagnoses were more frequently falsely identified as PD (Web Table 4). Among true-negative cases, memory complaints, vertigo/tinnitus/hypoacusis, essential tremor, restless leg syndrome, and drug-induced parkinsonism represented more than 80% of diagnoses. Among false-positive cases, essential tremor and parkinsonism (except drug-induced) represented more than 70% of diagnoses. Most patients with parkinsonism (except drug-induced), approximately one-quarter of patients with essential tremor, and fewer than 10% of patients with restless leg syndrome or drug-induced parkinsonism were classified as false positives. The majority of patients with memory complaints, vertigo/tinnitus/hypoacusis, vascular disease, or other causes were correctly classified.

The optimisms of R^2 and the *c* statistic obtained through bootstrapping were very low (<3%; Web Table 3). Model performance remained excellent after correction for optimism. Predicted and calibrated values were very close (calibration-in-the large = −3.8%), and overfitting was limited (calibration slope = 92.2%).

The model based on binary covariates is shown in Web Table 5; its performance (Web Table 6) was slightly lower than that of the main model ($R^2 = 67.2\%$, *c* statistic = 0.939, le Cessie–van Houweligen test: $P = 0.591$). Optimism and overfitting were low.

When alternative quantitative variables were used (Web Table 6), models based on the number of boxes of APDs bought per year and mean daily doses performed similarly to the model based on cumulative doses, whereas the number of prescriptions per year yielded slightly lower performance. Worse performances were observed for models based on shorter time periods (Web Table 7).

Among 239,123 affiliates (aged $\geq$18 years) who were alive on December 31, 2007, we identified 3,337 affiliates who bought 1 or more APDs in 2007; the model predicted that 1,593 persons had PD (crude prevalence = 6.66/1,000). Correcting for the model's sensitivity and specificity, we obtained a slightly lower prevalence (6.04/1,000). Prevalence increased with age and was higher in men than in women (Web Figure 4). Assuming no PD cases under age 18 years, age- and sex-standardized prevalence was 2.93/1,000; the marked decrease is due to the older age structure of the MSA population as compared with the French population.

Table 2. Characteristics of Study Participants With and Without Parkinson's Disease, France, 2007

Characteristic	Persons With PD (n = 320)			Persons Without PD (n = 794)			Odds Ratio[a]	95% Confidence Interval[a]
	No.	%	Mean (SD)	No.	%	Mean (SD)		
Age, years			71.7 (6.7)			68.7 (11.3)	1.4***	1.2, 1.7
Male sex	185	57.8		353	44.5		1.7***	1.3, 2.2
Neurologist visits								
≥1 visit	131	40.9		36	4.5		14.6***	9.8, 21.8
No. of visits			2.4 (1.4)			1.8 (1.0)	1.8*	1.1, 2.9
General practitioner visits								
≥1 visit	275	85.9		714	89.9		0.7	0.5, 1.0
No. of visits			5.9 (3.1)			4.8 (3.2)	1.4***	1.2, 1.6
Proportion of time treated								
100%	294	91.9		483	60.8		7.3***	4.8, 11.1
Mean			98.8 (7.4)			80.7 (31.7)	5.2***	3.1, 8.8
Antiparkinsonian drug or class								
Amantadine[b] (≥1 claim)	19	5.9		8	1.0		6.2***	2.7, 14.3
Anticholinergic agents[c] (≥1 claim)	14	4.4		83	10.5		0.4**	0.2, 0.7
Catechol-*O*-methyl transferase inhibitors (≥1 claim)	84	26.3		11	1.4		25.3***	13.3, 48.3
Levodopa								
≥1 claim	277	86.6		136	17.1		31.2***	21.5, 45.1
Cumulative LEDs[d]			162,461.8 (115,900.7)			74,796.3 (68,197.3)	3.8***	2.7, 5.5
Piribedil								
≥1 claim	80	25.0		531	66.9		0.2***	0.1, 0.2
Cumulative LEDs[d]			40,683.8 (21,793.0)			14,666.7 (11,844.5)	4.0***	3.1, 5.2
Selegiline[b] (≥1 claim)	40	12.5		11	1.4		10.2***	5.1, 20.1
Type 1 dopamine agonists								
≥1 claim	110	34.4		69	8.7		5.5***	3.9, 7.7
Cumulative LEDs[d]			81,002.1 (51,696.4)			19,467.3 (36,934.4)	6.9***	3.6, 13.0
Type 2 dopamine agonists[b] (≥1 claim)	15	4.7		9	1.1		4.3***	1.9, 9.9

Abbreviations: LED, levodopa equivalent dose; PD, Parkinson's disease; SD, standard deviation.

* $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$.

[a] Odds ratios, 95% confidence intervals, and P values were computed using logistic regression. For continuous variables, the odds ratio for a 1-SD increase is shown.

[b] The cumulative dose per year (in LEDs) was not computed because there were fewer than 55 subjects treated.

[c] No LED data were available for these classes of antiparkinsonian drugs.

[d] The cumulative dose over 1 year was computed among treated subjects and is expressed in milligrams of LED.

DISCUSSION

In this population-based study, we illustrated how patterns of APD use derived from drug claims databases can be used to estimate the probability that a person has PD. By obtaining clinical information for a large number of persons who used APDs during a 1-year period, we built a prediction model and assessed its performance. Addition of quantitative information on dose and regularity to the model improved performance. Assessment of APD use over longer periods was associated with better performance. The prediction model was used to estimate PD prevalence (6.66/1,000); correcting this figure using the model's sensitivity and specificity yielded an estimate (6.04/1,000) that was slightly lower but close to the crude estimate.

Few studies have used drug claims databases to estimate PD prevalence. Some of them relied on levodopa (5), while others used several APDs (4, 6, 8, 10, 23) or a combination of APDs and ICD codes (7, 9). None of these studies used information on dose or regularity of use. Diagnosis was not verified in the majority of studies, while a few studies used different gold standards—for example, neurologic examination (5, 10), expert chart review (8, 9), or self-reports (7). Only 1 PD prediction model has been reported; it was part of the Rotterdam Study and was based on a small number of APD users (n = 63) (24). Discrimination performance was computed among all participants (c statistic = 0.93). The

Table 3. Prediction Model for Parkinson's Disease Based on the Cumulative Dose of Antiparkinsonian Drugs Used Over a 1-Year Period, France, 2007[a]

Characteristic and Model Coding	Estimate (SE)	*Z* Ratio[b]	Odds Ratio[c]	95% Confidence Interval[c]
Intercept	−5.332 (1.03)	−5.2		
Age, years (FP: *x*)	NR[d]			
Gender (male vs. female)	0.293 (0.22)	1.3	1.3	0.9, 2.1
No. of neurologist visits				
1 or 2 vs. 0	1.396 (0.35)	4.0	4.0***	2.0, 8.0
>2 vs. 0	0.171 (0.50)	0.3	1.2	0.4, 3.2
No. of general practitioner visits				
1 or 2 vs. 0	−0.502 (0.48)	−1.0	0.6	0.2, 1.6
3–6 vs. 0	−0.759 (0.44)	−1.8	0.5	0.2, 1.1
>6 vs. 0	−1.245 (0.46)	−2.7	0.3**	0.1, 0.7
Proportion of time treated (FP: *x*)	1.182 (0.29)	4.1	3.3***	1.9, 5.7
Antiparkinsonian drug or class[e]				
Amantadine (ever vs. never)	NR			
Anticholinergic drugs (ever vs. never)	−0.897 (0.49)	−1.9	0.4	0.2, 1.0
Catechol *O*-methyl transferase inhibitors (ever vs. never)	0.752 (0.41)	1.9	2.1	1.0, 4.7
Levodopa (FP[f]: log(*x*))	0.448 (0.04)	10.2	1.6***	1.4, 1.7
Piribedil (FP[f]: $x + \log(x)$)	1.220 (0.26)	4.7	3.4***	2.0, 5.6
	−0.643 (0.20)	−3.2	0.5*	0.4, 0.8
Selegiline (ever vs. never)	2.104 (0.53)	4.0	8.2***	2.9, 23.1
Type 1 dopamine agonists (FP[f]: $x + x^2$)	1.879 (0.39)	4.8	6.5***	3.0, 14.1
	−0.303 (0.08)	−3.9	0.7***	0.6, 0.9
Type 2 dopamine agonists (ever vs. never)	1.048 (0.76)	1.4	2.9	0.6, 13.0

Abbreviations: FP, fractional polynomial; NR, not retained; SE, standard error.

* $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$.

[a] The logit of the probability that a person has Parkinson's disease is computed by summing the intercept and the estimates from the model multiplied by the value of each variable for that person. The probability that the person has Parkinson's disease is then calculated as the inverse of $(1 + e^{-\text{logit}})$.

[b] *Z* ratios were calculated by dividing regression coefficients by their standard error.

[c] Odds ratios, 95% confidence intervals, and *P* values were computed using logistic regression. For continuous variables, the odds ratio for a 1-standard-deviation increase is shown (standard deviations: proportion of time treated, 28.2; levodopa, 110,563.2; piribedil, 16,141.6; type 1 dopamine agonists, 55,314.4). A backward selection procedure was used, with a 2-sided *P* value of 0.20 for retention in the model.

[d] Not retained in the model.

[e] For antiparkinsonian drugs, the fractional polynomial term (*x*) represents the cumulative dose.

[f] Antiparkinsonian drugs coded using fractional polynomials were expressed in levodopa equivalents.

model did not include quantitative information, and calibration and internal validity were not assessed.

As expected, we observed large differences between persons with and without PD regarding APD use. The prediction model including several APDs performed considerably better than models based on individual APDs, particularly levodopa alone. Most PD patients were treated with higher cumulative doses of APDs, and adding this information to the model improved its performance. In addition to APDs, a key variable was the proportion of time treated: A higher proportion is a surrogate for more regular use, and regular APD users were more likely to have PD than persons who used APDs for short periods.

False-positive findings were more frequent for some diagnoses. Patients with neurodegenerative parkinsonism (progressive supranuclear palsy, multiple system atrophy, corticobasal degeneration, parkinsonism following dementia, dementia with Lewy bodies) were incorrectly classified as having PD in over 75% of cases because, in the absence of any other treatment, they often receive levodopa at doses similar to those of PD patients. In addition, some of these patients were actually considered to have PD by their physicians and were treated as such. A number of patients with essential tremor were incorrectly diagnosed and treated for PD and were therefore identified by the model as having PD. Therefore, our prediction model did not perform well for

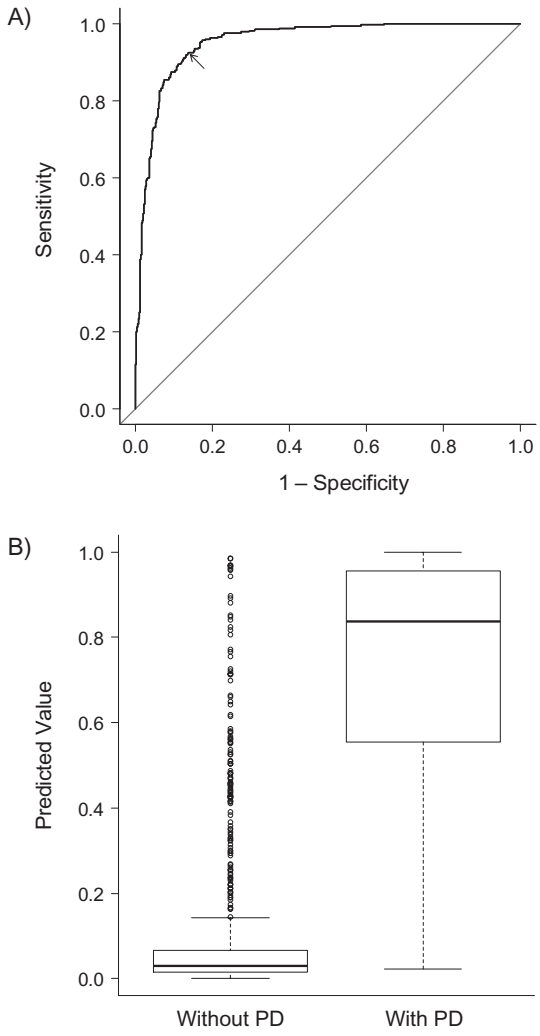

Figure 1. A) Receiver operating characteristic curve and B) box plot of predicted probabilities for a Parkinson's disease (PD) prediction model based on cumulative dose of antiparkinsonian drugs, France, 2007. In part A, the *c* statistic (area under the receiver operating characteristic curve) is 0.953. Sensitivity (92.5%), specificity (86.4%), positive predictive value (73.3%), and negative predictive value (96.6%) can be computed for the optimal cutoff ($P = 0.255$) that maximizes the Youden index among persons who used any antiparkinsonian drug in 2007 (shown as an arrow). In part B, the discrimination slope (absolute difference in average predictions for persons with and without PD) is 0.625.

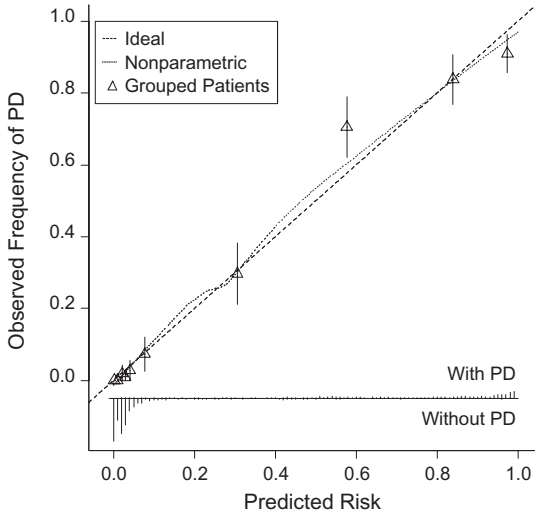

Figure 2. Calibration plot for a Parkinson's disease (PD) prediction model based on cumulative dose of antiparkinsonian drugs, France, 2007.

conditions that are treated similarly to PD (neurodegenerative parkinsonism, which is considerably less frequent than PD) or for persons who had an incorrect PD diagnosis.

The performance of prediction models based on drug claims databases is influenced by multiple factors. First, by definition, only treated patients are present in drug claims databases; if the aim is to identify patients with a disease, the main assumption is that the majority of patients are being treated. In France, this assumption is reasonable for PD, except in the oldest age groups. In a previous population-based study (1988–1989), 11% of prevalent PD patients identified using a 2-phase approach were not diagnosed; newly diagnosed patients were considerably more common above 80 years of age (25). The difficulty in identifying PD cases among the oldest subjects results from a variety of factors (e.g., diagnostic uncertainty, concern about more serious comorbidity, delayed diagnosis) (26). Altogether, prediction models may be less reliable in the oldest populations. Second, drug claims need to be recorded exhaustively. Conditions that are treated using over-the-counter drugs cannot be studied using this approach. Furthermore, prescriptions for some segments of the population may not be included in the databases. We did not identify persons institutionalized in nursing homes with indoor pharmacies; we estimate that a minority (<2%) of APD users were in this situation (27). Third, drugs used to develop the model need to be relatively specific to the disease. Even if similar drugs are used for the disease of interest and for other disorders, the pattern of use (i.e., dose, frequency) may help to improve the model's performance. We found that including cumulative dose and a variable assessing regularity of prescription improved model performance in terms of discrimination and calibration.

Drug claims databases are increasingly available in many countries. While they were not primarily designed for this purpose, they provide opportunities to study epidemiologic questions. Prediction models based on these databases have a number of advantages when attempting to estimate disease frequency. For diseases that are underdeclared on death certificates, such as PD (28), they represent a clear improvement over mortality studies (29). They are also useful for diseases that cannot be traced through laboratory test results or other investigations. Regarding PD, population-based door-to-door surveys are considered the reference method (25, 30, 31). However, this approach is costly and difficult to implement; in addition, because PD is not frequent, only a small number of cases are usually identified. Drug claims databases are usually available on a large scale (country,

state, etc.) and can identify larger numbers of patients, thus providing more stable frequency estimates and higher power to study risk factors. Another advantage of these models is that, once their validity has been assessed, and assuming that there are no major changes in medical care, they can be used repeatedly over time; if new drugs became available or therapeutic strategies were modified, their performance would need to be reevaluated.

Therefore, this approach may be particularly helpful for disease surveillance because it allows for the study of temporal trends. Further, if the place of residency is recorded in the databases, it allows investigators to study the spatial distribution of diseases; the relation between disease frequency and environmental factors can be studied using semi-ecologic designs. If data on characteristics such as sex or occupation are available, differences in disease frequency according to these characteristics can be investigated.

In order to conduct analytical studies, prediction models can be used in different ways. First, patients with the disease of interest can be identified using the prediction model based on a given probability cutoff, and the relation between exposure and disease can be evaluated using different study designs (e.g., case-control study, prevalence ratios). In our PD example, the cutoff that maximized sensitivity (92.5%) and specificity (86.4%) among APD users was 0.255, while specificity was greater than 99% at the total population level; when the number of false-positive cases for disease status is negligible, as in this example, association estimates between exposure and disease are unbiased, provided that misclassification is independent from exposure (32). For less specific models, it is possible to use cutoffs associated with higher specificity (Web Figure 3) in order to reduce the number of false positives and to correct association measures using standard formulas (32). Second, the prediction model can be used as a screening tool (2-stage designs), in order to identify patients likely to have the disease and for whom additional information (e.g., clinical data, laboratory tests) will be collected at the second stage to confirm diagnosis. This approach is appropriate for studies that require ''pure phenotypes,'' such as genetic association studies; it could also be used to recruit participants into clinical trials. Investigators can select the first-stage probability cutoff depending on how representative they want patients to be and how many false positives are acceptable at the second stage, particularly in terms of costs and time. Our PD prediction model has higher specificity than other PD screening tools at the general population level (33–36); therefore, the number of false positives that will need to be examined at the second stage is considerably reduced, resulting in sizeable savings.

Our study had limitations. First, our prediction model was developed among persons meeting specific inclusion criteria. All subjects had $\leq$15 years of disease duration; however, model sensitivity was similar among persons with short (<6 years) and long (6–15 years) disease durations (data not shown). It is therefore unlikely that the model's performance changed significantly with increasing disease duration. We did not include subjects with dementia or psychosis; the frequency of dementia and psychosis increases with PD duration and, as discussed above, we found no evidence that the model's performance was modified by duration. Finally, we included only subjects aged $\leq$80 years. Among subjects over 80 years of age, levodopa (56%) and piribedil (46%) represented the main APDs. For both drugs, there was no evidence of any interaction with age; therefore, it is unlikely that the model's performance was significantly different in older subjects. Second, we did not perform external validation of the model, which would have required collecting similar data in another population. In France, all persons have access to health care in a similar manner, and affiliates of different health-care systems have access to the same physicians, including neurologists. In addition, the number of predictors that we included in the models was small compared with the number of subjects, and overfitting was not an issue. It is therefore unlikely that a study in another French population would have yielded significantly different findings. However, an obvious limitation of this approach is that prediction models based on drug claims databases are not universal. A prediction model developed in one country probably cannot be applied in another country without modification, unless there are no differences in access to medical care, medical practices, or drug availability. However, this is unlikely for many conditions, and regarding PD, heterogeneity in APD use has been described in Europe (37, 38). Nevertheless, a model developed in one country may be a useful aid in the development of similar models in other countries. Lastly, we did not consider interactions between APDs; assessment of whether simultaneous prescription of 2 or more drugs predicts the outcome differently compared with each of the drugs considered separately could be performed, but more complex models and larger sample sizes would be necessary.

In conclusion, prediction models based on drug claims databases can be used to estimate disease frequency. We illustrated this approach in the context of a prevalence study, but a similar strategy could be used for incidence. Use of drug claims databases for analytical studies can also be considered. Several features of the prediction model (e.g., dose, regularity) are likely to be relevant for other conditions. Development of prediction models should take into account specific aspects of the disease of interest. We used clinical examination as a gold standard, but, for other diseases, questionnaires, biologic samples, or administrative databases could be used, and ICD codes, if available, could be included to improve performance (39).

ACKNOWLEDGMENTS

Author affiliations: Unité 708–Neuroepidemiology, INSERM, Paris, France (Frédéric Moisan, Véronique Gourlet, Christophe Tzourio, Alexis Elbaz); Unité Mixte de Recherche en Santé 708, Université Pierre et Marie Curie, University of Paris VI, Paris, France (Frédéric Moisan, Véronique Gourlet, Christophe Tzourio, Alexis Elbaz); Département Santé Travail, Institut de Veille Sanitaire, Saint-Maurice, France (Marcel Goldberg, Ellen Imbernon, Alexis Elbaz); Caisse Centrale de la Mutualité Sociale Agricole, Bagnolet, France (Jean-Luc Dupupet, Jean Houssinot);

Caisse Départementale de la Gironde, Mutualité Sociale Agricole, Bordeaux, France (Jean-Louis Mazurie); Centre for Research in Epidemiology and Population Health, Unité 1018–Epidemiology of Occupational and Social Determinants of Health, INSERM, Villejuif, France (Marcel Goldberg); Unité Mixte de Recherche en Santé 1018, University of Versailles Saint-Quentin-en-Yvelines, Versailles, France (Marcel Goldberg); and Unité Mixte de Recherche en Santé 1018, University of Paris XI, Villejuif, France (Marcel Goldberg).

This work was supported by l'Institut National de la Santé et de la Recherche Médicale, l'Agence Nationale de la Recherche, l'Agence Française de Sécurité Sanitaire de l'Environnement et du Travail, and France Parkinson. Frédéric Moisan was supported by a scholarship from the Ministère de l'Enseignement Supérieur et de la Recherche and the Fondation pour la Recherche Médicale.

The authors thank the Mutualité Sociale Agricole physicians and personnel at each site (Drs. Jacques Aimedieu, Daniel Albert, Catherine Bolut, Christophe Fuzeau, Virginie Gaussères, Maryline Grandjean, Jean Houssinot, Marine Jeantet, Bernard Ladépèche, Didier Menu, Omar Tarsissi, Joël Gourgues, Sandrine Nogues, Emilie Richard, and Pierre Vannier); the study interviewers (Véronique Dumay, Viviane Palleau, Frédérique Pellerin, Estelle Seguin, and Sophie Sinibaldi); the study neurologists (Drs. Irina Balaboi, Isabelle Benatru, Julien Dumurgier, Elsa Krim, and Danièle Ranoux); Yann Hamon, who was involved in data management; and Aïcha Soumaré, who helped coordinate the study.

This work was presented in part as a poster at the 62nd Annual Meeting of the American Academy of Neurology, Toronto, Ontario, Canada, April 10–17, 2010.

Conflict of interest: none declared.

REFERENCES

1. Bower JH, Maraganore DM, McDonnell SK, et al. Incidence and distribution of parkinsonism in Olmsted County, Minnesota, 1976–1990. *Neurology.* 1999;52(6):1214–1220.
2. Ferini-Strambi L, Manconi M. Treatment of restless legs syndrome. *Parkinsonism Relat Disord.* 2009;15(suppl 4):S65–S70.
3. de Azevedo AA, Langguth B, de Oliveira PM, et al. Tinnitus treatment with piribedil guided by electrocochleography and acoustic otoemissions. *Otol Neurotol*. 2009;30(5):676–680.
4. Menniti-Ippolito F, Spila-Alegiani S, Vanacore N, et al. Estimate of parkinsonism prevalence through drug prescription histories in the Province of Rome, Italy. *Acta Neurol Scand*. 1995;92(1):49–54.
5. Chiò A, Magnani C, Schiffer D. Prevalence of Parkinson's disease in Northwestern Italy: comparison of tracer methodology and clinical ascertainment of cases. *Mov Disord*. 1998;13(3):400–405.
6. van de Vijver DA, Roos RA, Jansen PA, et al. Estimation of incidence and prevalence of Parkinson's disease in the elderly using pharmacy records. *Pharmacoepidemiol Drug Saf*. 2001;10(6):549–554.
7. Noyes K, Liu H, Holloway R, et al. Accuracy of Medicare claims data in identifying Parkinsonism cases: comparison with the Medicare current beneficiary survey. *Mov Disord*. 2007;22(4):509–514.
8. Newman EJ, Grosset KA, Grosset DG. Geographical difference in Parkinson's disease prevalence within West Scotland. *Mov Disord*. 2009;24(3):401–406.
9. Szumski NR, Cheng EM. Optimizing algorithms to identify Parkinson's disease cases within an administrative database. *Mov Disord*. 2009;24(1):51–56.
10. Masalha R, Kordysh E, Alpert G, et al. The prevalence of Parkinson's disease in an Arab population, Wadi Ara, Israel. *Isr Med Assoc J*. 2010;12(1):32–35.
11. Swarztrauber K, Anau J, Peters D. Identifying and distinguishing cases of parkinsonism and Parkinson's disease using ICD-9 CM codes and pharmacy data. *Mov Disord*. 2005;20(8):964–970.
12. Tomlinson CL, Stowe R, Patel S, et al. Systematic review of levodopa dose equivalency reporting in Parkinson's disease. *Mov Disord*. 2010;25(15):2649–2653.
13. Steyerberg EW. Developing valid prediction models. In: Steyerberg EW, ed. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. New York, NY: Springer Publishing Company; 2009:113–331.
14. Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *J R Stat Soc Ser C Appl Stat*. 1994;43(3):429–467.
15. Royston P, Sauerbrei W. Building multivariable regression models with continuous covariates in clinical epidemiology—with an emphasis on fractional polynomials. *Methods Inf Med.* 2005;44(4):561–571.
16. Peduzzi P, Concato J, Kemper E, et al. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol*. 1996;49(12):1373–1379.
17. le Cessie S, van Houwelingen JC. A goodness-of-fit test for binary regression models, based on smoothing methods. *Biometrics*. 1991;47(4):1267–1282.
18. Harrell FE. Resampling, validating, describing and simplifying the model. In: Harrell FE, ed. *Regression Modeling Strategies With Applications to Linear Models, Logistic Regression and Survival Analysis*. New York, NY: Springer Publishing Company; 2001:87–103.
19. Harrell FE. Binary logistic regression. In: Harrell FE, ed. *Regression Modeling Strategies With Applications to Linear Models, Logistic Regression and Survival Analysis*. New York, NY: Springer Publishing Company; 2001:215–267.
20. Gerds TA, Cai T, Schumacher M. The performance of risk prediction models. *Biom J*. 2008;50(4):457–479.
21. National Institute for Statistics and Economic Studies. *Pyramides des âges 2007*. Paris, France: National Institute for Statistics and Economic Studies, 2007. (http://www.insee.fr). (Accessed May 5, 2010).
22. Couris CM, Colin C, Rabilloud M, et al. Method of correction to assess the number of hospitalized incident breast cancer cases based on claims databases. *J Clin Epidemiol*. 2002;55(4):386–391.
23. Lai BC, Schulzer M, Marion S, et al. The prevalence of Parkinson's disease in British Columbia, Canada, estimated by using drug tracer methodology. *Parkinsonism Relat Disord*. 2003;9(4):233–238.
24. van de Vijver DA, Stricker BH, Breteler MM, et al. Evaluation of antiparkinsonian drugs in pharmacy records as a marker for Parkinson's disease. *Pharm World Sci.* 2001;23(4):148–152.
25. de Rijk MC, Tzourio C, Breteler MM, et al. Prevalence of parkinsonism and Parkinson's disease in Europe: The EUROPARKINSON collaborative study. European Community Concerted Action on the Epidemiology of Parkinson's disease. *J Neurol Neurosurg Psychiatry.* 1997;62(1):10–15.

26. Van Den Eeden SK, Tanner CM, Bernstein AL, et al. Incidence of Parkinson's disease: variation by age, gender, and race/ethnicity. *Am J Epidemiol*. 2003;157(11):1015–1022.
27. Elbaz A, Leleu H, Houssinot J. Increased use of potentially inappropriate medications in patients with Parkinson's disease [abstract]. Presented at the 62nd Annual Meeting of the American Academy of Neurology, Toronto, Canada, April 10–17, 2010.
28. Phillips NJ, Reay J, Martyn CN. Validity of mortality data for Parkinson's disease. *J Epidemiol Community Health*. 1999; 53(9):587–588.
29. Mylne AQ, Griffiths C, Rooney C, et al. Trends in Parkinson's disease related mortality in England and Wales, 1993–2006. *Eur J Neurol*. 2009;16(9):1010–1016.
30. de Rijk MC, Breteler MM, Graveland GA, et al. Prevalence of Parkinson's disease in the elderly: the Rotterdam Study. *Neurology*. 1995;45(12):2143–2146.
31. Zhang ZX, Roman GC, Hong Z, et al. Parkinson's disease in China: prevalence in Beijing, Xian, and Shanghai. *Lancet*. 2005;365(9459):595–597.
32. Rothman KJ, Greenland S, Lash TL. Bias analysis. In: *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2008:345–380.
33. Tanner CM, Gilley DW, Goetz CG. A brief screening questionnaire for parkinsonism [abstract]. Presented at the 115th Annual Meeting of the American Neurological Association, Atlanta, Georgia, October 14–17, 1990.
34. Rocca WA, Maraganore DM, McDonnell SK, et al. Validation of a telephone questionnaire for Parkinson's disease. *J Clin Epidemiol*. 1998;51(6):517–523.
35. Pramstaller PP, Falk M, Schoenhuber R, et al. Validation of a mail questionnaire for parkinsonism in two languages (German and Italian). *J Neurol*. 1999;246(2):79–86.
36. Kim JH, Cheong HK, Lee CS, et al. The validity and reliability of a screening questionnaire for Parkinson's disease in a community. *J Prev Med Public Health*. 2010;43(1): 9–17.
37. de Pedro-Cuesta J, Petersen IJ, Vassilopoulos D, et al. Epidemiological assessment of levodopa use by populations. *Acta Neurol Scand*. 1991;83(5):328–335.
38. Rosa MM, Ferreira JJ, Coelho M, et al. Prescribing patterns of antiparkinsonian agents in Europe. *Mov Disord*. 2010;25(8): 1053–1060.
39. Henderson T, Shepheard J, Sundararajan V. Quality of diagnosis and procedure coding in ICD-10 administrative data. *Med Care*. 2006;44(11):1011–1019.