A predictive model to identify Parkinson disease from administrative claims data

Susan Searles Nielsen,
PhD
Mark N. Warden, MS
Alejandra Camacho-Soto,
MD
Allison W. Willis, MD,
MS
Brenton A. Wright, MD
Brad A. Racette, MD

Correspondence to
Dr. Racette:
racetteb@wustl.edu

ABSTRACT

Objective: To use administrative medical claims data to identify patients with incident Parkinson disease (PD) prior to diagnosis.

Methods: Using a population-based case-control study of incident PD in 2009 among Medicare beneficiaries aged 66–90 years (89,790 cases, 118,095 controls) and the elastic net algorithm, we developed a cross-validated model for predicting PD using only demographic data and 2004–2009 Medicare claims data. We then compared this model to more basic models containing only demographic data and diagnosis codes for constipation, taste/smell disturbance, and REM sleep behavior disorder, using each model's receiver operator characteristic area under the curve (AUC).

Results: We observed all established associations between PD and age, sex, race/ethnicity, tobacco smoking, and the above medical conditions. A model with those predictors had an AUC of only 0.670 (95% confidence interval [CI] 0.668–0.673). In contrast, the AUC for a predictive model with 536 diagnosis and procedure codes was 0.857 (95% CI 0.855–0.859). At the optimal cut point, sensitivity was 73.5% and specificity was 83.2%.

Conclusions: Using only demographic data and selected diagnosis and procedure codes readily available in administrative claims data, it is possible to identify individuals with a high probability of eventually being diagnosed with PD. *Neurology*® 2017;89:1448–1456

GLOSSARY

AUC = area under the receiver operator characteristic curve; **BASF** = beneficiary annual summary file; **CPT** = Current Procedural Terminology; **ICD-9** = *International Classification of Diseases-9*; **PD** = Parkinson disease; **RBD** = REM sleep behavior disorder.

Clinical trials in patients with Parkinson disease (PD) have failed to demonstrate efficacy of numerous putative neuroprotective therapies.^{1–5} One possible reason is that by the time of diagnosis, patients with PD have lost nearly a third of substantia nigra neurons.⁶ Consequently, there is a need to identify patients with PD earlier.⁷ Equally important, because motor abnormalities can be present several years before diagnosis,^{8–10} earlier treatment of patients with prodromal PD symptoms may reduce the risk of falls and related traumas. Thus, identification of patients with PD during the prodromal disease period is an urgent priority for physicians and researchers.

Most attempts to identify patients with PD in the prodromal period, such as in the Parkinson Progression Markers Initiative predictive model, require biomarker and clinical data.⁷ PD predictive methodologies generally take into account age, sex, race/ethnicity, smoking,^{11,12} and several prodromal PD symptoms, such as constipation,¹³ REM sleep behavior disorder (RBD),¹⁴ and anosmia.¹⁰ RBD is rare, and the other symptoms lack specificity. More cost-effective approaches are needed to screen large populations. A predictive model that only uses administrative health care data could be used by researchers or insurers to identify those with the greatest probability of having prodromal PD. In this study, we sought to develop a predictive model using comprehensive medical claims data, under the hypothesis that patients with

Supplemental data
at Neurology.org

From the Department of Neurology (S.S.N., M.N.W., A.C.-S., B.A.W., B.A.R.), Washington University School of Medicine, St. Louis, MO; Departments of Neurology and Biostatistics and Epidemiology (A.W.W.), University of Pennsylvania School of Medicine, Philadelphia; and School of Public Health, Faculty of Health Sciences (B.A.R.), University of the Witwatersrand, Parktown, South Africa.

Go to Neurology.org for full disclosures. Funding information and disclosures deemed relevant by the authors, if any, are provided at the end of the article.

prodromal PD could be distinguished from the general population using diagnosis and procedure codes in the 5 years prior to PD diagnosis.

METHODS **Institutional review board approval.** The study was approved by the Washington University institutional review board and Centers for Medicare & Medicaid Services.

Study overview. We used detailed claims data from Medicare, the only nationwide, population-based health care program in the United States. We constructed a case-control study with all eligible incident PD cases and a random sample of all eligible controls in 2009 using all carrier (physician/supplier part B), outpatient, inpatient, skilled nursing facility, durable medical equipment, and home health care claims from 2004 to 2009. We used these data, as detailed below, to determine PD case status, PD diagnosis date, and medical predictor variables. We used the 2009 beneficiary annual summary file (BASF) to obtain demographic data, enumerate all potential participants, and determine study eligibility.

Eligibility criteria. All participants were born by February 1, 1942, to exclude disabled beneficiaries who were not yet age-eligible for Medicare by January 1, 2007, and were required to meet the following criteria for 2009: Medicare parts A or B coverage; no Medicare Advantage Plan coverage; US residence; and age ≤ 90 years at year end (or death earlier in 2009). The eligibility criteria were designed to ensure (1) comparability of cases and controls; (2) population-based sample; (3) complete capture of diagnoses and procedures; (4) restriction to beneficiaries most likely to seek medical care if they had PD symptoms (age ≤ 90 years); and (5) at least 2 full years of claims data before diagnosis/reference (age ≥ 66 years, 11 months). A total of 22,546,965 Medicare beneficiaries met all the above criteria in 2009, representing 46.1% of all 2009 beneficiaries of any age (including disabled beneficiaries), 71.6% of beneficiaries in the targeted (age-eligible) birth cohort, and 73.3% of beneficiaries who were both in this birth cohort and living in the United States at diagnosis/reference. The remainder were ineligible due to Medicare Advantage Plan coverage. Among Medicare beneficiaries in our study population, which was defined by age/birth cohort and region of residence (United States), those who were and were not eligible were similar with regard to age (mean of 75.7 vs 75.4 years, respectively) and sex (56.6% vs 57.3% female, respectively). They were also fairly similar with regard to race/ethnicity (88.3% vs 85.3% non-Hispanic white, respectively) and vital status at year end (4.5% vs 3.7% died before year end, respectively).

PD case identification. Medicare claims records include ICD-9 diagnosis codes, ICD-9 procedure codes, and Healthcare Common Procedure Coding System codes, including Current Procedural Terminology (CPT) codes. As previously,¹⁵ we included all Medicare beneficiaries who met the above eligibility criteria as a potential (incident) case if they had at least one ICD-9 code for PD (332 or 332.0) in 2009, but no prior year. We only required one occurrence to ensure that we did not restrict our case group to those who used Medicare most. We excluded 4,365 (4.6%) potential cases whose PD case status was uncertain, including prior diagnosis of Lewy body dementia (ICD-9 331.82), other neurodegenerative disease of the basal ganglia (ICD-9 333 or 333.0), or probable typographic error (non-pyogenic meningitis [ICD-9 322 or 322.0] without diagnostic lumbar puncture [CPT 62270]). A total of 89,790 cases

remained, representing all 2009 beneficiaries with incident PD who met eligibility criteria.

Control selection. Controls met the same eligibility criteria but had no ICD-9 code for PD. We also required that they had not died in 2009 prior to their randomly assigned reference date in the same year, to ensure that they were comparable to cases. We obtained claims data for a random sample of all beneficiaries enumerated from the 2009 BASF who met these criteria. We avoided matching so that model coefficients for known PD risk factors would be unbiased when applying the model to future cohorts. Of those sampled, we excluded 200 potential controls with the above exclusion diagnosis codes. In total, there were 118,095 controls, representing a 0.5% sample of all controls otherwise available for analysis.

Calculation of predictor variables. We used all diagnosis or procedure code variables from all claims files to create a dichotomous variable for each code observed in 2004–2009 prior to PD diagnosis (first date with PD diagnosis code) or reference (randomly assigned date for controls) in 2009. In calculating these variables, 96.9% of cases and 92.1% of controls linked to all possible years and 99.7% of cases and 94.7% of controls had at least one valid code. Cases and controls each had up to 441 unique diagnosis codes (median 94 for cases, 58 for controls). Cases had up to 544 (median 108) unique procedure codes. Controls had up to 479 (median 73).

We calculated age and obtained sex and race/ethnicity from the 2009 BASF. Age and sex were available for all participants. Race/ethnicity was unknown for 0.09% but retained in a race/ethnicity unknown category. Using demographic and claims data, we estimated the probability that each participant had ever regularly smoked tobacco. We assigned participants with a code specific to tobacco (ICD-9 V15.82, ICD-9 305.1, CPT 99406, CPT 99407) a probability of 100%. We calculated other participants' probability using a logistic regression predictive model for ever/never smoking built from nationwide data¹⁶ for Americans aged 66–90 years. The model had 17 predictor variables made from more than 600 diagnosis/procedure codes, including cancers with high attributable risk due to smoking.¹⁷ We then applied this predictive model to Medicare data and validated smoking probability against (1) the above tobacco-specific codes before factoring them in; (2) county level smoking¹⁸; (3) diagnoses associated with smoking but in Medicare data only.

Statistical analysis. We used SAS version 9.4 (Cary, NC) for data handling and R version 3.2.4 (including the package pROC¹⁹ and boot²⁰) and Stata 13.1 (College Station, TX) for statistical analysis. We employed a 2-step approach to build the PD predictive model. In the first step, we identified codes significantly associated with PD. For each code, we generated a logistic regression model to estimate association with PD, while adjusting for age, sex, race/ethnicity, ever smoking, and number of unique diagnosis codes, as a general measure of usage of medical care. The case-control dataset included 26,468 valid codes (11,063 diagnoses and 15,405 procedures), so we used a Bonferroni-corrected α of 1.889E-6. In the second step, we built a predictive model using codes that were identified as significant in the first step and also occurring in both sexes, had >10 cases and >10 controls, and did not simply reflect secondary parkinsonism (e.g., ICD-9 332.1 or codes for schizophrenia or bipolar disorder that might require use of neuroleptics). We built the model by applying the elastic net algorithm^{21,22} to the selected diagnosis and procedure codes. The elastic net algorithm is a method that optimally blends ridge and lasso penalized regression. These 2 types of penalized regressions improve upon

generalized linear models by regularizing coefficients based on penalties for possible overfitting. The elastic net algorithm is particularly suitable for high dimensional data, allows for a grouping effect of variables (e.g., similar adjacent codes), includes a cross-validation step, and shrinks estimates toward null to avoid overfitting and increase external validity. For the k-fold cross-validation step, we used 5-fold cross-validation of the model using area under the receiver operator characteristic curve (AUC) as the measure of model quality. For comparison to our primary predictive model, we also generated a simple model with age, sex, and race/ethnicity only, and then a basic model enhanced by the addition of total number of unique diagnosis codes, smoking, constipation, RBD, and anosmia/hyposmia. We compared the models' performance using recommended²³ metrics: (1) difference between the mean predicted probability of PD for cases and controls; (2) AUC (*c* statistic); (3) sensitivity and specificity at the cut point that maximized percentage of participants classified correctly.

RESULTS Basic predictive model. PD was positively associated with age, was markedly more common in men than in women, and was more common in non-Hispanic white than non-Hispanic nonwhite participants (tables 1 and 2). PD also was inversely associated with tobacco smoking, after accounting for the number of unique diagnosis codes. Constipation, RBD, and anosmia/hyposmia were strongly positively associated with PD, although the latter, based on a code that also includes taste disturbance, did not meet the Bonferroni-corrected threshold for significance (table 2). When we simultaneously stratified

PD cases according to whether a neurologist diagnosed them and according to the total number of ICD-9 codes for PD in 2009 (1, 2, 3–5, >5 [approximate quartiles]), we observed all of the above associations for each PD case group vs all controls (data not shown in tables).

The age-sex-race only model provided relatively little separation between cases and controls, and was clearly improved by the addition of known predictors of PD and the total number of diagnosis codes (table 2). Almost all of this improvement, however, was due to the latter (AUC 0.670 [95% confidence interval 0.668–0.673], without total number of diagnosis codes).

Full predictive model. In the first step, models that included adjustment for all of the basic covariates, a total of 2,291 diagnosis or procedure codes, had a *p* value that met the Bonferroni-corrected α , including 2,142 retained to build the predictive model. The elastic net algorithm's procedure indicated that 100% lasso regression (no ridge regression) provided the best predictive model. At the recommended error level (fewest predictors such that the model's error was within 1 SD of the best model's error),²¹ the model had 536 codes (table e-1 at Neurology.org). All established demographic predictors and smoking associations remained, with the exception of some race/ethnicity categories with

Table 1 Characteristics of incident Parkinson disease (PD) cases and controls, Medicare 2009

	PD cases (n = 89,790), %	Controls (n = 118,095), %	Unadjusted OR (95% CI)	Mutually adjusted OR (95% CI) ^a
Age, y				
66–69	8.1	18.3	1.0 (reference)	1.0 (reference)
70–74	19.5	28.9	1.53 (1.48–1.58)	1.48 (1.43–1.53)
75–79	24.8	22.5	2.51 (2.44–2.60)	2.33 (2.26–2.41)
80–84	26.9	18.2	3.36 (3.25–3.47)	2.99 (2.90–3.09)
85–90	20.7	12.1	3.89 (3.75–4.02)	3.37 (3.26–3.49)
Female	50.2	57.0	0.76 (0.75–0.77)	0.57 (0.56–0.58)
Race/ethnicity				
White	88.8	86.4	1.0 (reference)	1.0 (reference)
Black	6.0	7.5	0.78 (0.75–0.80)	0.83 (0.80–0.86)
Pacific Islander/other	1.0	1.5	0.63 (0.58–0.68)	0.68 (0.63–0.74)
Asian	1.7	2.2	0.78 (0.74–0.83)	0.73 (0.69–0.78)
Hispanic	2.1	1.9	1.11 (1.04–1.18)	0.98 (0.92–1.05)
Native American	0.3	0.4	0.75 (0.65–0.87)	0.85 (0.73–0.99)
Unknown	0.09	0.09	0.95 (0.71–1.27)	0.88 (0.66–1.19)
Smoking index ^b ≥ median	42.9	55.4	0.60 (0.59–0.61)	0.57 (0.56–0.59)

Abbreviations: CI = confidence interval; OR = odds ratio.

^aMutually adjusted indicates that all ORs are adjusted for all other variables in the column.

^bPredicted probability of ever smoking divided by the person's total number of unique diagnosis codes (or 1 for 292 cases and 6,227 controls without any diagnosis codes).

Table 2 Predictive models for identifying Parkinson disease and their performance

	Age-sex-race model, OR (95% CI) ^a	Basic model, OR (95% CI) ^a	Full model, OR (CIs NA) ^a
Age, per year ^b	1.087 (1.085–1.089)	1.056 (1.054–1.058)	1.036
Female	0.71 (0.70–0.72)	0.56 (0.55–0.58)	0.49
Race/ethnicity			
White	1.0 (reference)	1.0 (reference)	1.0 (reference)
Black	0.84 (0.81–0.87)	0.86 (0.83–0.90)	0.77
Pacific Islander/other	0.73 (0.67–0.80)	0.83 (0.76–0.90)	0.90
Asian	0.78 (0.73–0.83)	0.85 (0.79–0.91)	1.07
Hispanic	1.07 (1.005–1.14)	0.97 (0.90–1.04)	0.95
Native American	0.82 (0.71–0.96)	0.87 (0.75–1.03)	1.01
Unknown	0.88 (0.65–1.18)	1.05 (0.77–1.44)	0.95
Smoking ^c	—	0.70 (0.67–0.73)	0.78
Constipation ^d	—	1.39 (1.36–1.42)	3 codes, OR 1.02–1.20
RBD ^e	—	7.25 (4.87–10.8)	3.35
Taste/smell disturbance ^f	—	1.24 (1.10–1.40)	NA ^g
Total diagnosis codes, per unique code	—	1.0127 (1.0125–1.0129)	1.00
Other diagnoses or procedures	—	—	536 codes ^h
Model performance (95% CI) ⁱ			
Difference between mean predicted probabilities	0.057 (0.056–0.058)	0.145 (0.143–0.146)	0.362 (0.360–0.364)
AUC ^j	0.638 (0.636–0.640)	0.724 (0.722–0.726)	0.857 (0.855–0.859)
Sensitivity, % ^k	46.4 (46.1–46.8)	60.2 (59.9–60.5)	73.5 (73.2–73.8)
Specificity, % ^k	71.7 (71.5–72.0)	72.1 (71.8–72.3)	83.2 (83.0–83.4)
% Classified correctly ^k	60.8 (60.6–61.0)	66.9 (66.7–67.2)	79.0 (78.9–79.2)

Abbreviations: AUC = area under the receiver operator characteristic curve; CI = confidence interval; NA = not applicable; OR = odds ratio; RBD = REM sleep behavior disorder.

^a ORs and 95% CIs adjusted for all variables in the respective columns. CIs cannot be calculated when applying penalized logistic regression, including the elastic net algorithm, to develop a predictive model. All models based on 89,790 Parkinson disease cases and 118,095 controls.

^b Per year estimate for beneficiaries aged 66–85 years; the estimate for beneficiaries aged 86–90 years was lower for all models.

^c Predicted probability of ever smoking tobacco (potential range of 0.00–1.00); coefficient thus interpreted as ever/never smoking.

^d ICD-9 564, 564.0, 564.00, 564.01, 564.02, or 564.09.

^e ICD-9 327.42.

^f ICD-9 781.1; this single code combines both smell and taste disturbances in one code.

^g Not included because failed to reach statistical significance after Bonferroni correction.

^h A total of 536 diagnosis and procedure codes in the full model include 3 ICD-9 codes for constipation and ICD-9 327.42 (RBD). These 536 codes are listed in table e-1 and summarized in table 3.

ⁱ 95% CIs, calculated with 2,000 stratified bootstrap replicates.²⁰

^j Paired tests comparing the full model to the other 2 models indicated significant differences between the models (both $p < 2.2\text{E-}16$).

^k At the cut point that maximizes the percent accurately classified.

relatively few participants (table 2). This full model performed markedly better than the age-sex-race only and basic models (table 2 and figures 1–2; table e-2).

Diagnostic and procedure codes that were positive predictors of PD included those suggestive of PD motor symptoms (such as tremor and abnormalities of posture), psychiatric or cognitive dysfunction, autonomic dysfunction (gastrointestinal, urinary, orthostatic hypotension), sleep disturbance (including RBD), fatigue, and trauma (including falls) at a median of 1.4–3.1 years prior to diagnosis (table 3). In parallel with these symptoms, there was a surge in diagnostic testing approximately 1.7 years prior to diagnosis, including brain imaging and various blood

tests including those for syphilis, vitamin deficiencies, and thyroid conditions. PD-associated conditions such as idiopathic peripheral neuropathy and restless legs syndrome were also positive predictors. A variety of codes indicating weight loss were positive predictors of PD, appearing at a median of 1.1 years prior to PD diagnosis. Diabetes type II without any complications was positively associated with PD, in contrast to several other diabetes codes that were negative predictors of PD (table 3).

Other diagnosis and procedure codes that were negative predictors of PD included numerous codes indicating that controls were more active than cases, such as codes for hip replacement, osteoarthritis,

Figure 1 Distribution of predicted probability of Parkinson disease (PD) by true PD case status

(A) Basic predictive model: predicted by age (2 linear splines), sex, race/ethnicity (7 categories), ever/never smoking, total number of unique ICD-9 diagnosis codes, constipation (ICD-9 564, 564.0, 564.00, 564.01, 564.02, 564.09), REM sleep behavior disorder (RBD) (ICD-9 327.42), and anosmia/hyposmia (included in ICD-9 781.1 as smell and taste disturbances). (B) Full predictive model: predicted by age (2 linear splines), sex, race/ethnicity (7 categories), ever/never smoking, total number of unique ICD-9 diagnosis codes, and 536 diagnosis or procedure codes including codes for RBD (ICD-9 327.42) and selected codes for constipation (table e-1).

carpal tunnel syndrome, and a variety of tendonopathies (table 3). Cancer, cardiovascular disease, and obesity-related codes were also negatively associated with PD. Gout and a variety of codes indicating systemic inflammation (e.g., inflammatory bowel disease and scleroderma) or allergy were negatively associated with PD.

In sensitivity analyses with the 50,395 (56.1%) PD cases diagnosed by a neurologist or with ≥ 3 PD diagnosis codes in 2009, model performance

was essentially identical to the model with all PD cases for all performance measures that would not be affected by our altered case:control ratio (i.e., difference in mean predicted probability between PD cases and controls, AUC, and sensitivity and specificity at the cut point where sensitivity and specificity are approximately equal). Moreover, only 14 (2.6%) diagnosis or procedure codes both dropped out of the model and suggested any difference in results ($>10\%$ change in the odds ratio, as opposed to potentially

Figure 2 Parkinson disease predictive model receiver operator characteristic (ROC) curve

Nonparametric ROC curve for the full predictive model: predicted by age (2 linear splines), sex, race/ethnicity (7 categories), ever/never smoking, total number of unique ICD-9 diagnosis codes, and 536 diagnosis or procedure codes including codes for RBD (ICD-9 327.42) and selected codes for constipation (table e-1). RBD = REM sleep behavior disorder.

dropping out of the model due to the smaller sample size for this subanalysis).

DISCUSSION This study demonstrates that Medicare beneficiaries with prodromal PD can be identified using a predictive model that includes diagnosis and procedure codes in the 5 years prior to PD diagnosis. By a number of metrics,²³ this model was far superior to the model that included only previously identified PD nonmotor comorbidities.^{13,14,24,25} Moreover, because the model generates a probability of PD, the model can be applied in different settings, and the probability cut points selected as needed to maximize sensitivity or specificity. Strengths of this study are the large population-based sample that included all eligible incident PD cases in patients over age 65 in the United States, a randomly selected comparable control group, and rigorous analytic approach.

The large sample size enabled us to discover hundreds of novel associations with PD that contributed to the predictive model. Many of the diagnoses identified as positive predictors of PD are important disease comorbidities, such as falls and trauma. Subtle motor dysfunction occurring several years prior to PD diagnosis may contribute to falls,^{8–10} suggesting that these comorbidities may be preventable if patients with prodromal PD were identified earlier, using this predictive model, and treated earlier. Dopaminergic therapies are currently used only in patients with PD with bothersome motor symptoms²⁶;

however, it is possible that starting dopaminergic therapies during the prodromal period in selected patients could prevent falls and associated trauma. In addition, given the failure of numerous neuroprotective therapies in PD, there is growing consensus that it may be necessary to use these therapies earlier, before the majority of dopaminergic neurons degenerate.²⁷ If the results of our study are confirmed, it may be possible to implement neuroprotective studies in patients with prodromal PD prior to irreversible loss of dopaminergic neurons.

The diagnosis and procedure codes associated with a lower probability of PD were highly informative. Many cancer-related codes were inversely associated with PD, consistent with most prior studies of PD, perhaps due to an underlying difference in propensity for programmed cell death in those with PD. This may provide some insight into cellular mechanisms that could serve as neuroprotective targets. Interestingly, obesity was also a negative predictor, consistent with weight loss being a positive predictor. Weight loss may represent an early symptom of PD or could reflect a difference in systemic metabolism in patients with prodromal PD. As in prior studies, both gout and smoking were negative predictors.^{28–31} Cardiovascular disease, complications of diabetes, and obesity all were negative predictors, highlighting how different patients with PD are from the general population. Understanding whether these marked differences are driven by life-long choices, made due to difference in brain chemistry in at-risk people, or are due to specific metabolic or pharmaceutical protective factors associated with these diseases deserves further study.

Many of the diagnosis and related procedure codes that were associated with a higher probability of PD represent early nonmotor and motor symptoms of PD. Several nonmotor symptoms associated with prodromal PD, such as RBD, taste/smell disturbance, and constipation, were either not common (RBD) or were very common in both cases and controls (constipation), making these insufficiently predictive when used alone. Not surprisingly, disorders of mood and cognition were strongly predictive of PD, representing early symptoms in patients with prodromal PD. There were also a large number of diagnosis codes suggestive of motor dysfunction in patients with prodromal PD, including signs of bradykinesia, gait impairment, and tremor. Associations between prodromal PD and therapies such as gait training, psychotherapy, and prescriptions for assistive devices are likely in response to PD motor and nonmotor symptoms. Finally, restless legs syndrome appeared to be associated strongly with PD, which is intriguing given that dopaminergic medications are highly effective treatments for this condition.

Table 3 Categories of positive and negative diagnosis/procedure code predictors of Parkinson disease (PD)

Category	No. of codes	Odds ratio			Median no. of years between first occurrence and PD diagnosis ^{a,b}
		25th percentile	Median	75th percentile	
Positive predictors					
Diagnostic testing	37	1.02	1.05	1.15	1.7
Psychiatric conditions	29	1.03	1.07	1.16	2.1
PD symptoms	20	1.06	1.15	1.36	1.4
Autonomic symptoms	19	1.04	1.08	1.11	2.1
Trauma/falls	16	1.02	1.04	1.07	1.9
Gait/balance	15	1.04	1.07	1.18	1.6
Renal	13	0.98	1.02	1.03	2.2
Cognitive	13	1.05	1.09	1.16	1.6
Immobility	11	1.02	1.03	1.15	1.4
Endocrine	7	1.01	1.01	1.05	2.7
PD-associated conditions	7	1.02	1.03	1.23	2.4
Weight loss/nutrition	6	1.02	1.05	1.15	1.1
Integument	6	1.01	1.04	1.10	2.9
Vision	6	1.01	1.04	1.10	2.9
Speech	4	1.02	1.08	1.16	1.7
Fatigue	3	1.00	1.05	1.19	3.1
Sleep	2	1.06	2.21	3.35	1.9
Diabetes	1	1.10	1.10	1.10	4.5
Negative predictors					
Activity	77	0.94	0.97	0.98	NA
Cardiovascular	57	0.96	0.99	1.00	NA
Cancer	48	0.95	0.97	0.99	NA
Infectious	26	0.95	0.98	1.00	NA
Tobacco	21	0.94	0.96	0.99	NA
Physician visit	15	0.97	0.99	0.99	NA
Other medical	10	0.96	0.99	0.99	NA
Diabetes	10	0.93	0.95	0.97	NA
Allergy	10	0.93	0.97	0.99	NA
Inflammation	8	0.95	0.99	1.00	NA
Health maintenance	7	0.96	0.97	0.99	NA
Gastrointestinal	6	0.94	0.98	0.98	NA
Dermatologic	6	0.97	0.98	0.99	NA
Nursing home physician visit	5	0.96	0.98	1.00	NA
Therapy	3	0.89	0.97	0.99	NA
Tinnitus/hearing	3	0.89	0.95	0.98	NA
Other surgery	3	0.93	0.94	0.99	NA
Other neurologic	2	0.89	0.93	0.98	NA
Hepatic	2	0.97	0.97	0.98	NA
Obesity	1	0.97	0.97	0.97	NA
Gout	1	0.98	0.98	0.98	NA

Abbreviation: NA = not applicable.

^aFor positive predictors only.^bAmong PD cases only; median of the median time between first occurrence of each of the represented diagnosis or procedure codes in the respective category and PD diagnosis.

Numerous procedure codes and resulting diagnosis codes predicted a higher probability of PD, largely reflecting health-seeking behaviors in response to early symptoms of PD. Clearly, patients with prodromal PD and their doctors recognize that something is wrong, resulting in physicians ordering diagnostic tests to determine the cause of the symptoms in patients with prodromal PD. This includes diagnostic tests commonly ordered by physicians for nonspecific neurologic symptoms, such as tests for vitamin B₁₂, syphilis, and thyroid function. The phenomenon of ongoing testing may be exacerbated by lack of a gold standard diagnostic test for PD, since PD is ultimately a clinical diagnosis. In aggregate, these tests for conditions other than PD represent a substantial cost to the health care system that could be reduced by profiling claims data, as done by pharmacy benefit management companies to reduce drug costs and drug interactions.

As with any administrative data study, there are several potential limitations. Medicare data are derived from provider charges submitted to Medicare for reimbursement, so our diagnosis and procedure codes rely on competent provider diagnoses and charge entry by billing staff. However, we found all of the expected PD associations,^{10,13,14,24,25} thereby providing strong evidence of the validity of our case ascertainment method and a Medicare-derived PD predictive model. We also did not have pharmaceutical data, such as in Medicare part D, which may improve the performance of the predictive model. Our predictive model was highly influenced by ICD-9 and CPT codes that were consistent with PD symptoms, although there were many other novel predictors. However, it is clear that patients with prodromal PD are different from the general population, suggesting that it may be possible to detect these patients beyond the 5-year window of this study. We anticipate that additional predictors, positive and negative, would be identified with a longer prodromal window. Of course, confirmation of these findings in another large dataset would be ideal, including a dataset with other types of insurance beneficiaries, although we applied an innovative statistical method to cross-validate the model to improve external validity. Nevertheless, our predictive model should serve as a critical first step and benchmark for further model development.

AUTHOR CONTRIBUTIONS

Dr. Searles Nielsen performed data analysis and coauthored the first draft of this manuscript. M.N. Warden performed data analysis and edited this manuscript. Dr. Camacho-Soto assisted with creating variables and edited this manuscript. Dr. Willis obtained funding for this study, assisted with creating variables, and edited this manuscript. Dr. Wright assisted with creating variables and edited this manuscript. Dr. Racette obtained funding for this study, oversaw data analysis, and coauthored the first draft of this manuscript.

STUDY FUNDING

Study funded by the Michael J. Fox Foundation, National Institute for Environmental Health Sciences (K24ES017765), and American Parkinson Disease Association. Dr. Racette takes full responsibility for the data, the analyses and interpretation, and the conduct of the research; has full access to all of the data; and has the right to publish any and all data separate and apart from any sponsor.

DISCLOSURE

Dr. Searles Nielsen reports government research support from the NIH (R21 ES024120 [co-I], R01ES021488 [co-I], R01 ES025991 [co-I]) and research support from the Michael J. Fox Foundation. M.N. Warden reports no disclosures relevant to the manuscript. Dr. Camacho-Soto reports government research support from NCMRR (K12HD00109719). Dr. Willis reports government research support from NINDS (K23NS081087). Dr. Wright reports research support from Teva (sub-I) and Vaccinex (sub-I). Dr. Racette reports research support from Teva (PI), US World Meds (PI), Allergan (PI), and Vaccinex (PI); government research support from the NIH (K24ES017765 [PI], R21ES17504 [PI], R01ES021488 [PI], R01ES021488-02S1 [PI], R01 ES025991 [PI], R21ES024120); research support from the Michael J. Fox Foundation; and consultation income from 86 Pillars for legal consultation. Go to Neurology.org for full disclosures.

Received November 19, 2016. Accepted in final form July 11, 2017.

REFERENCES

1. Parkinson Study Group. Effect of deprenyl on the progression of disability in early Parkinson's disease. *N Engl J Med* 1989;321:1364–1371.
2. Parkinson Study Group PRECEPT Investigators. Mixed lineage kinase inhibitor CEP-1347 fails to delay disability in early Parkinson disease. *Neurology* 2007;69:1480–1490.
3. Olanow CW, Rascol O, Hauser R, et al. A double-blind, delayed-start trial of rasagiline in Parkinson's disease. *N Engl J Med* 2009;361:1268–1278.
4. Writing Group for the NET-PD Investigators, Kieburtz K, Tilley BC, et al. Effect of creatine monohydrate on clinical progression in patients with Parkinson disease: a randomized clinical trial. *JAMA* 2015;313:584–593.
5. Olanow CW, Goetz CG, Kordower JH, et al. A double-blind controlled trial of bilateral fetal nigral transplantation in Parkinson's disease. *Ann Neurol* 2003;54:403–414.
6. Cheng HC, Ulane CM, Burke RE. Clinical progression in Parkinson disease and the neurobiology of axons. *Ann Neurol* 2010;67:715–725.
7. Parkinson Progression Marker Initiative. The Parkinson Progression Marker Initiative (PPMI). *Prog Neurobiol* 2011;95:629–635.
8. Postuma RB, Berg D. Advances in markers of prodromal Parkinson disease. *Nat Rev Neurol* 2016;12:622–634.
9. Nystrom H, Nordstrom A, Nordstrom P. Risk of injurious fall and hip fracture up to 26 y before the diagnosis of Parkinson disease: nested case-control studies in a nationwide cohort. *PLoS Med* 2016;13:e1001954.
10. Schrag A, Horsfall L, Walters K, Noyce A, Petersen I. Prediagnostic presentations of Parkinson's disease in primary care: a case-control study. *Lancet Neurol* 2015;14:57–64.
11. Checkoway H, Powers K, Smith-Weller T, Franklin GM, Longstreth WT Jr, Swanson PD. Parkinson's disease risks associated with cigarette smoking, alcohol consumption, and caffeine intake. *Am J Epidemiol* 2002;155:732–738.
12. Ritz B, Ascherio A, Checkoway H, et al. Pooled analysis of tobacco use and risk of Parkinson disease. *Arch Neurol* 2007;64:990–997.

13. Abbott RD, Petrovitch H, White LR, et al. Frequency of bowel movements and the future risk of Parkinson's disease. *Neurology* 2001;57:456–462.
14. Postuma RB, Gagnon JF, Montplaisir J. Clinical prediction of Parkinson's disease: planning for the age of neuroprotection. *J Neurol Neurosurg Psychiatry* 2010;81:1008–1013.
15. Wright Willis WA, Evanoff BA, Lian M, Criswell SR, Racette BA. Geographic and ethnic variation in Parkinson disease: a population-based study of US Medicare beneficiaries. *Neuroepidemiology* 2010;34:143–151.
16. Centers for Disease Control and Prevention. Behavior Risk Factor Surveillance System: BRFSS 2009 Survey Data and Documentation [online]. Available at: cdc.gov/brfss/annual_data/annual_2009.htm. Accessed February 21, 2017.
17. Whitman DC, Wilson LF. The fractions of cancer attributable to modifiable factors: a global review. *Cancer Epidemiol* 2016;44:203–221.
18. Dwyer-Lindgren L, Mokdad AH, Srebotnjak T, Flaxman AD, Hansen GM, Murray CJ. Cigarette smoking prevalence in US counties: 1996–2012. *Popul Health Metr* 2014;12:5.
19. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 2011;12:77.
20. Canty A, Ripley B. Bootstrap R (S-Plus) Functions. R package version 1.3-18. 2016.
21. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Ser B* 2005;67:301–320.
22. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 2010;33:1–22.
23. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010;21:128–138.
24. Gagnon JF, Bedard MA, Fantini ML, et al. REM sleep behavior disorder and REM sleep without atonia in Parkinson's disease. *Neurology* 2002;59:585–589.
25. Schuurman AG, van den Akker M, Ensink KT, et al. Increased risk of Parkinson's disease after depression: a retrospective cohort study. *Neurology* 2002;58:1501–1504.
26. Clarke CE, Patel S, Ives N, Rick C, Wheatley K, Gray R. Should treatment for Parkinson's disease start immediately on diagnosis or delayed until functional disability develops? *Mov Disord* 2011;26:1187–1193.
27. Mahlknecht P, Seppi K, Poewe W. The concept of prodromal Parkinson's disease. *J Parkinsons Dis* 2015;5:681–697.
28. Gao X, Chen H, Choi HK, Curhan G, Schwarzschild MA, Ascherio A. Diet, urate, and Parkinson's disease risk in men. *Am J Epidemiol* 2008;167:831–838.
29. Alonso A, Rodriguez LA, Logroscino G, Hernan MA. Gout and risk of Parkinson disease: a prospective study. *Neurology* 2007;69:1696–1700.
30. De Vera M, Rahman MM, Rankin J, Kopec J, Gao X, Choi H. Gout and the risk of Parkinson's disease: a cohort study. *Arthritis Rheum* 2008;59:1549–1554.
31. Hernan MA, Takkouche B, Caamano-Isorna F, Gestal-Otero JJ. A meta-analysis of coffee drinking, cigarette smoking, and the risk of Parkinson's disease. *Ann Neurol* 2002;52:276–284.

Visit the *Neurology*[®] Website at Neurology.org

- Enhanced navigation format
- Increased search capability
- Highlighted articles
- Detailed podcast descriptions
- RSS Feeds of current issue and podcasts
- Personal folders for articles and searches
- Mobile device download link
- AAN Web page links
- Links to *Neurology Now*[®], *Neurology Today*[®], and *Continuum*[®]
- Resident & Fellow subsite

Find *Neurology*[®] on Facebook: <http://tinyurl.com/neurologyfan>

Follow *Neurology*[®] on Twitter: <https://twitter.com/GreenJournal>