

# Identifying incident Parkinson's disease using administrative diagnostic codes: a validation study

Brett J. Peterson<sup>a</sup>, Walter A. Rocca<sup>a,b</sup>, James H. Bower<sup>b</sup>, Rodolfo Savica<sup>a,b</sup>, Michelle M. Mielke<sup>a,b,\*</sup>

<sup>a</sup> Division of Epidemiology, Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA

<sup>b</sup> Department of Neurology, Mayo Clinic, Rochester, MN, USA

## ARTICLE INFO

### Article history:

Received 17 January 2020

Received in revised form 20 May 2020

Accepted 2 June 2020

Available online 2 June 2020

### Keywords:

Parkinson's disease  
Health administrative data  
Diagnostic code  
Validation  
Incidence

## ABSTRACT

**Background:** Administrative databases that capture diagnostic codes are increasingly being used worldwide for research because they can save time and reduce costs. However, assessing validity is necessary before defining diseases using only diagnostic codes in research applications.

**Objective:** Our objective was to assess the validity of using diagnostic codes to identify incident Parkinson's disease (PD) cases in Olmsted County, Minnesota using an established standard for comparison (1976–2005).

**Methods:** Cases were identified solely using computer programs applied to administrative diagnostic code indexes from the Rochester Epidemiology Project (REP). Two codes >30 days apart or one code on the death certificate constituted PD. The standard was a clinical diagnosis by movement disorders specialists based on medical record review. Validity was assessed using positive predictive value (PPV) and sensitivity. Numbers of incident cases and incidence rates were compared between the two ascertainment methods by sex.

**Results:** The codes only method over-counted the number of incident PD cases by 73% (804 versus 464), and this over-counting generally increased with calendar year. Sensitivity was 80% (95% CI [76%, 84%]) and PPV was 46% (95% CI [34%, 50%]). Disease status misclassification accounted for two-thirds of falsely identified cases, where individuals were found to not have PD (43%) or even parkinsonism (23%) after medical record review. The codes only method also over-estimated the incidence rate time trend for men and women by approximately two-fold.

**Conclusion:** In our context, using administrative diagnostic codes only to identify incident PD cases is not recommended unless more accurate algorithms are developed.

## 1. Introduction

Parkinson's disease (PD) is a neurological disorder projected to affect 770,000 individuals in the United States by 2040, and increase of 56% from 2005 [1]. Understanding trends in PD incidence within and across populations may lead to the identification of modifiable risk factors [2]. Some studies have reported that the risk of PD may be increasing with time [3–5], whereas other studies suggest the risk may be decreasing with time [6–8]. Differences may be real or due to the way the diagnosis is assessed. Administrative databases that capture diagnoses, prescriptions, and procedures, are increasingly being used worldwide for clinical research. Using these databases can save time and reduce costs because the information is already collected, but they are generally designed for billing purposes and not for research. Therefore, assessing the questionable validity of diagnostic codes for correctly classifying disease status is recommended [9,10]. Manually abstracting data from medical records using accepted criteria is considered the most accurate way to retrospectively identify disease cohorts but takes more time. Our objective was to assess

the validity of only using diagnostic codes to identify incident PD cases and to calculate PD incidence. We hypothesized that using diagnostic codes only would over-estimate incident PD.

## 2. Methods

### 2.1. Study design and setting

The Rochester Epidemiology Project (REP) was used to identify cases of incident PD in Olmsted County, Minnesota from 1976 through 2005. The REP links virtually all medical records across all local medical care providers for individuals residing in Olmsted County, including outpatient, inpatient, emergency room, nursing home, residency, and death certificate information [11]. Physicians providing medical care assign administrative diagnostic codes per routine medical practice. Medical diagnoses, surgical interventions, and other information are routinely abstracted, coded, indexed, and stored in electronic datasets. The REP identifies 102% of the United States census estimates for the Olmsted County population, thereby

\* Corresponding author at: Department of Health Sciences Research, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, USA.  
E-mail address: [Mielke.Michelle@mayo.edu](mailto:Mielke.Michelle@mayo.edu). (M.M. Mielke).

allowing for reliable incidence studies [12]. Approximately 98% of the population provides authorization to use their medical records for research [11]. The diagnostic coding systems used in the REP are the Berkson system (1966–1975), the Hospital Adaptation of the International Classification of Diseases, Eighth Revision (HICDA, 1976–2010), the International Classification of Diseases, Ninth Revision (ICD-9, 1995–2015), and the Tenth Revision (ICD-10, 2015–present).

Two methods were used to identify incident PD cases. The first used electronic diagnostic codes for screening followed by medical record review [13,14], hereafter referred to as standard [3]. The second method used electronic diagnostic codes only without medical record review. The number of incident cases and incidence rates were determined using both methods and then compared.

All medical record information for both methods was collected passively so informed consent was not required. This study was approved by the Mayo Clinic and Olmsted Medical Center institutional review boards.

### 2.2. Identification of incident PD cases by the standard

The standard method used a two phase approach to identify incident PD cases; a screening phase and a clinical confirmation phase [3,13–16]. First, a computerized screening of the diagnostic code index identified individuals with at least one code indicative of parkinsonism from 1976 through 2010. A broad set of codes, including non-specific codes such as tremor, was used to maximize sensitivity. Second, a movement disorders specialist reviewed the medical records of individuals who screened positive and confirmed or refuted parkinsonism as a syndrome and PD as a specific type of parkinsonism using pre-specified criteria (J.H.B in 1998–1999 for codes from 1976 through 1990 and R.S in 2012–2013 for codes from 1991 through 2010). Parkinsonism required two of four cardinal signs: resting tremor, bradykinesia, rigidity, and impaired postural reflexes. PD required a parkinsonism diagnosis with no secondary cause, no documentation of levodopa unresponsiveness at doses of at least 1 g/day in combination with carbidopa (only for patients who were prescribed the medications; use of the medications was not necessary for a diagnosis), and no prominent or early signs of more extensive nervous system involvement not explained otherwise [3,13]. The specialist classified individuals as having parkinsonism or not and as having PD or not, and determined the onset year and residency status at onset. The standard method was shown to be reliable and valid [13,16,17].

### 2.3. Identification of incident PD cases using diagnostic codes only

Incident PD cases were identified solely using computer programs applied to the administrative diagnostic code index. All parkinsonism diagnostic codes were retrieved (online Supplementary Table 1). PD codes were a subset of parkinsonism codes: HICDA codes 03420110 (Disease, Parkinson's), 03420111 (Paralysis, Agitans), and 03420112 (Palsy, Shaking); ICD-9 codes 332 (Parkinson's Disease) and 332.0 (Paralysis Agitans); and ICD-10 code G20 (Parkinson's Disease). Individuals with two PD codes that occurred >30 days apart or who were given one PD code on their death certificate were classified as having PD. A 30-day timeframe was chosen to reduce false positives by assuming that these individuals were evaluated for PD and found not to have it. Longer timeframes may introduce bias due to death or moving out of the county prior to classification. The onset date of PD was defined as the earliest parkinsonism diagnostic code date. Individuals residing in Olmsted County for at least one year before their first parkinsonism diagnostic code were considered incident PD cases. A one year timeframe was chosen to avoid including individuals with prevalent PD who recently moved into Olmsted County but had their onset while residing elsewhere.

### 2.4. Statistical analysis

The difference in the number of cases identified per calendar year by the codes only and standard methods were tested using one-sample Wilcoxon Signed Rank tests. The number of PD codes were counted for each case

and compared across methods using the Kruskal-Wallis test. If an individual had more than one PD code in a calendar day, it was counted as one code. Sensitivity and PPV were calculated requiring 2, 3, 4, 5, and 10 total PD codes with exact 95% binomial confidence intervals (CI). Analysis of incidence rates and time trends followed the same methods as the standard [3]. Sex-specific incidence rates were calculated for each calendar year standardizing by age using the 1990 United States population. Negative binomial regression was used to estimate the relative PD incidence rate change over time for men and women. All statistical tests were two-tailed with an alpha of 0.05. SAS software version 9.4 (SAS institute Inc., Cary, NC, USA) and R version 3.4.2 were used for analyses.

## 3. Results

### 3.1. Incident Parkinson's disease cases identified

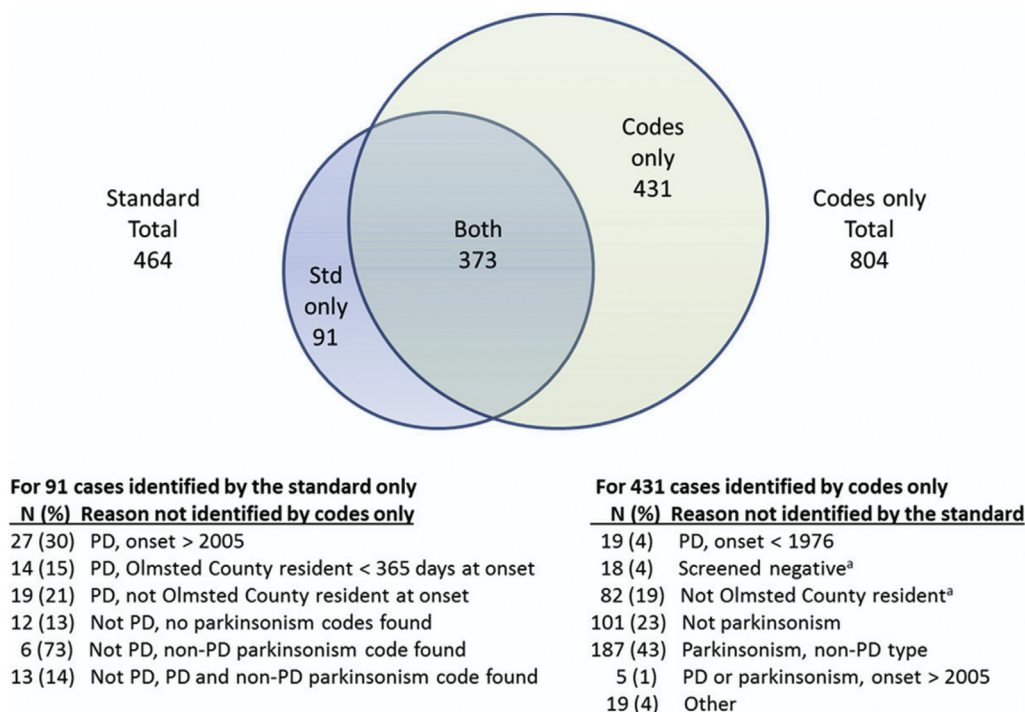
The codes only method identified 804 incident PD cases in Olmsted County from 1976 through 2005 compared to 464 identified by the standard (Fig. 1). There were 373 individuals with incident PD identified by both methods, 91 identified by the standard only and 431 identified by codes only. Therefore, using codes only identified 373 of 464 'true' PD cases (80% sensitivity) while erroneously identifying 431 (PPV 46%, Table 1). The negative predictive value was 99.98% (CI [99.97%, 99.98%]) and the specificity was 99.89% (CI [99.88%, 99.90%]). The most frequent reason individuals were classified as incident PD by the codes and not by the standard was a false positive for disease status. Of the 431 cases identified by codes only, 187 (43%) were classified as parkinsonism but not PD and 101 (23%) were classified as free of parkinsonism by the standard. The most frequent reason for an individual being classified as incident PD by the standard and not the codes were differing onset years. Of the 91 cases identified by the standard only, 27 (30%) were classified as PD with an onset year after 2005 by the codes. The same PD onset year was found in 40% of overlapping cases and 82% were within 2 years. Compared to the standard, 3% of cases identified using codes only had an earlier onset year and 57% had a later onset year. The median difference was 1 year (interquartile range [IQR] 0–2, range – 4–10). Fig. 1 displays other reasons for non-overlapping cases. When residency status was ignored, the PPV for using codes only to identify PD cases from 1976 through 2005 was 54% and the sensitivity was 84%.

The median number of PD codes significantly differed between PD cases identified by the codes only (10, IQR 4–29), the standard only (20, IQR 1–53), and both methods (39, IQR 14–72,  $P < .0001$ ). However, each incremental increase in the number of PD codes required for the codes method resulted in only a 2–3% increase for PPV and a 2–3% decrease for sensitivity (Table 1). Twelve of the 804 cases found by the codes were identified because they only had one code on their death certificate. Nine cases were reviewed as part of the standard, and one was classified as PD.

### 3.2. Incident Parkinson's disease cases identified by onset year

Fig. 2 shows the number of incident PD cases identified by each method per calendar year for men and women. The difference in the number of cases identified using codes only minus the standard per calendar onset year ranged from – 3 to 15 for women (median 5) and – 2 to 24 for men (median 4.5). Both medians differed significantly from 0,  $P < .0001$ . The differences increased over time for men and women starting around 1980 with the codes only method identifying more cases than the standard (Fig. 2). The ratio of the number of cases identified using codes only versus the standard per calendar onset year ranged from 0.8 to 4.3 for women (median 2.0) and 0.7 to 3.4 for men (median 1.5). Both medians differed significantly from 1,  $P < .0001$ .

The estimated relative incidence rate change per 10 calendar years for men was 1.57 (CI [1.38, 1.79]) using codes only compared to 1.24 (CI [1.08, 1.43]) using the standard. The estimated relative incidence rate change per 10 calendar years for women was 1.18 (CI [1.01, 1.38]) using codes only compared to 1.09 (CI [0.87, 1.38]) using the standard. The PD



**Fig. 1.** Incident Parkinson's disease cases identified in Olmsted County, Minnesota from 1976 through 2005 by the codes only and standard methods. Reasons each method did not identify a case when the other method did are enumerated. PD = Parkinson's disease; Std = standard. <sup>a</sup> Medical records not reviewed by specialist to ascertain parkinsonism and PD.

incidence time trends derived separately from the codes only and standard methods diverged over time for both men (Fig. 3A) and women (Fig. 3B).

**4. Discussion**

The codes only method over-counted the number of incident PD cases in Olmsted County, Minnesota from 1976 through 2005 by 73% compared to our standard and this over-counting generally increased with calendar year. Disease status misclassification accounted for the majority of falsely identified PD cases. These individuals were found to not have PD, or even parkinsonism, after medical record review. The codes only method also over-estimated the incidence rate time trend for both men and women.

Administrative health data has been used for parkinsonism studies, including validation assessments [18–24]. A review of 18 articles identifying PD using administrative datasets showed that PPVs ranged from 53 to 90% and sensitivities ranged from 15 to 73% using their respective standard [10]. Our PPV of 46% was below this range and our 80% sensitivity was above this range. Feldman et al. reported 71% PPV and 73% sensitivity when using at least one PD code to identify cases [21]. Our lower PPV may have resulted from differing standards, administrative datasets, or coding practices. Feldman et al. used telephone interviews, clinical examinations,

**Table 1**  
Accuracy of identifying Parkinson's disease incident cases using the codes only method.

No. of PD codes	PPV		Sensitivity	
	%	95% CI	%	95% CI
≥ 2	46.4	42.9–49.9	80.4	76.5–83.9
≥ 3	49.4	45.7–53.1	78.4	74.4–82.1
≥ 4	51.4	47.6–55.2	75.6	71.5–79.5
≥ 5	53.0	49.0–56.9	72.4	68.1–76.4
≥ 10	57.7	53.3–62.0	64.4	59.9–68.8

Number of days with a PD code required to classify someone as an incident PD case using codes only. Additionally in all scenarios, at least two codes must be over 30 days apart, and onset year and residency criteria must be met. CI, confidence interval; PPV, positive predictive value.

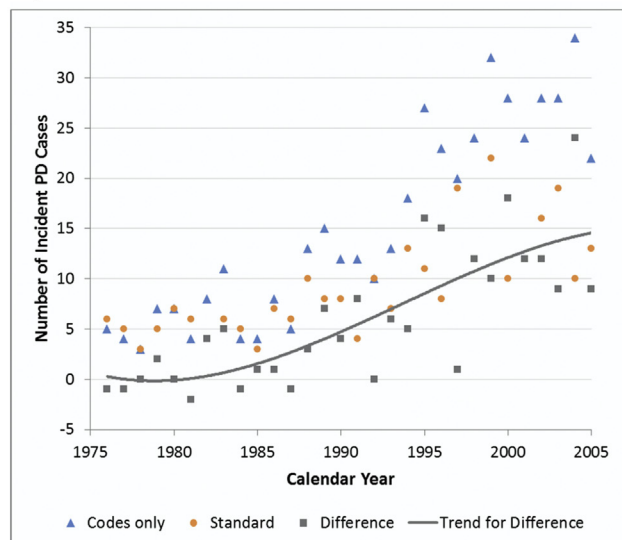
medical records, and in-person interviews to establish the standard compared to medical record review alone. Additionally, PPVs can be over-estimated when the disease prevalence in the validation standard is higher than the database in which the algorithm will be applied [9]. Similar to our findings, another REP study using only diagnostic codes to identify anterior cruciate ligament tears concluded the low accuracy (PPV of 66%) could be improved by adding medical record chart review [25].

Two-thirds of false positive cases were disease status misclassifications where 43% had parkinsonism but not PD and 23% did not have parkinsonism at all. Another study similarly found that most false positives had a parkinsonism other than PD according to their standard [21]. Using administrative codes to distinguish between parkinsonism and PD was shown to be ineffective using only codes [18]. Diagnoses made by movement disorder specialists are more accurate compared to non-specialists who can substantially over diagnose [26–28]. Using a PD code assigned by a specialist improves PPV, but can markedly reduce sensitivity [19–21], likely depending on the proportion of individuals with PD who visit a specialist in a given population.

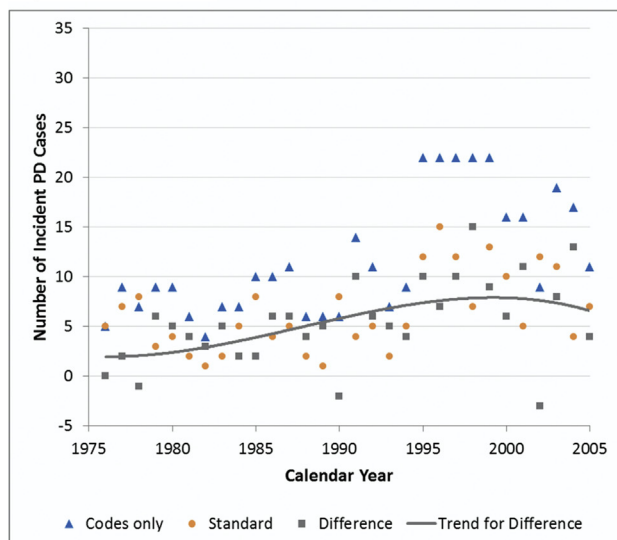
False positives without any type of parkinsonism may result from suspected disease or misdiagnosis. While some reports required one or more codes at any time, we required two PD codes >30 days apart to remove misclassification from rule-out evaluations. Similarly, another study found that a 30 day window between codes was preferable to no window in their algorithms [22]. In addition, our data suggest that requiring more than two PD codes had little effect on overall classification accuracy and only slightly increased PPV and slightly reduced sensitivity. Misdiagnosis may occur when physicians observe symptoms such as tremor and code PD as the diagnosis without applying diagnostic criteria or using the specific codes for those symptoms. This possibility is supported by a study in which 17 of 42 patients who were false positives for parkinsonism identified by ICD-9 codes were affected by other conditions with similar symptoms, mostly tremor [22]. Data entry errors for codes occur infrequently and therefore should have relatively minimal impact on misclassification [29].

An accurate onset date is crucial for incidence studies. An incorrect onset date can affect onset timing and inclusion in the target study population. The codes method used the earliest parkinsonism diagnostic code to

**A) Men**



**B) Women**



**Fig. 2.** Incident Parkinson's disease cases identified per calendar year by the codes only and standard methods. The grey squares represent the number of cases identified by the codes only method minus the number identified by the standard method with a grey cubic regression trend line for men (Panel A) and women (Panel B).

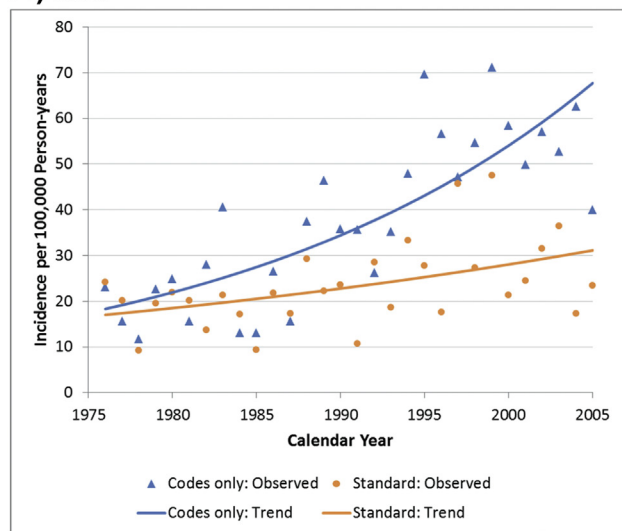
establish the onset date and residency. Reviewing records can determine a more biologically correct onset date that precedes the actual diagnosis. In 27 of 91 false negatives (30%), the codes correctly identified PD but the onset year was after 2005, and therefore outside of the standard study's incidence window. On the other hand, the codes estimated the onset year well with a median difference of 1 year later than the standard, consistent with 0.7 years estimated by Bower et al. [13]. Residency at the time of PD onset accounted for approximately 20% of misclassifications in both directions. When residency status was ignored using the codes method, the PPV increased by 8%, the sensitivity increased by 4%, and the over-counting of cases decreased to 56%.

The PD incidence rate trends diverged over time, and the estimated relative change from the codes only method was approximately double that that estimated by the standard method for both men and women. Furthermore, the estimated relative change in PD incidence was statistically significant for women using codes only, but not using the standard. One possible explanation for the increasing over-estimation with calendar year is the

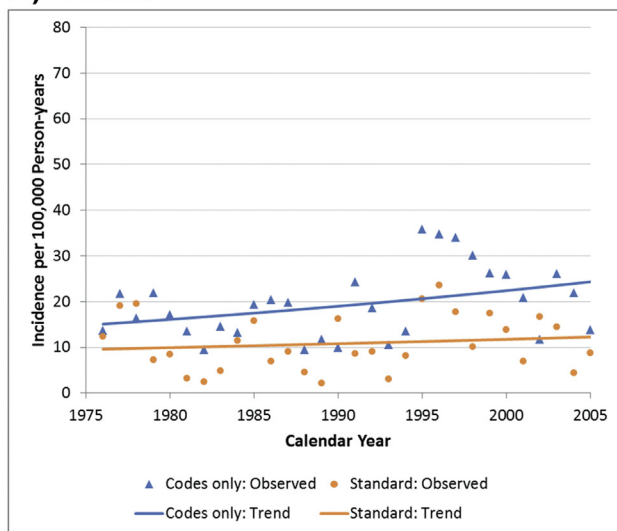
heightened awareness of and screening for PD. Using codes only misclassifies people as having PD who do not. This misclassification weakens analyses of risk factors and distorts clinical characterization of PD populations. Noyes et al. reported that people with claims-based PD diagnoses differed in comorbidities, Medicare expenses, residency location, and income compared to their standard [19].

Strengths of the analyses include the population-based setting, rigor for identifying incident PD cases in the standard study, and the large number of cases. Our standard included all 464 PD incident cases identified in the general population of Olmsted County, Minnesota. The standard included a broad screening and chart review by movement disorder specialists using established diagnostic criteria. A validity study of 321 individuals without a screening code found no one with parkinsonism after chart review [13]. Additionally, all individuals determined to have parkinsonism after a neurologic exam were identified by the standard ascertainment method, demonstrating that these diagnoses were captured in the REP diagnostic coding indices [13]. Fifty-seven of 59 PD cases identified by the standard study

**A) Men**



**B) Women**



**Fig. 3.** Parkinson's disease incidence per 100,000 person-years derived by the codes only and standard methods for each calendar year for men (Panel A) and women (Panel B). Trend lines were estimated using negative binomial regression.

were confirmed to have PD by a movement disorders specialist at a standardized in-person examination [17]. Finally, the clinical PD classifications from the standard showed high agreement with autopsy findings [16].

Limitations include generalizability and the standard time period. The findings from this study are reflective of the Olmsted County population, incident PD rates in Olmsted County, and the REP records-linkage system administrative datasets, coding practices, and coding systems. Therefore, specific results may not translate to other settings, analyses, and diseases. However, our findings demonstrate the importance of validating the use of diagnostic codes for finding disease cases, and that accuracy can be worse when used to identify incident disease cases in a defined geography and timeframe. The standard identified incident PD through 2005 and the specialty of the physician assigning the PD diagnostic codes was rarely documented, therefore we could not assess more recent years or physician specialty. Two studies reported that adding the use of parkinsonism medications with diagnostic codes improves PPV and decreases sensitivity, and can reduce overall accuracy [20,22]. We did not evaluate medications because they are only available from 2003 onward in the REP electronic indices. Future investigations using natural language processing of electronic health records hold promise. A recent study showed that PPV and sensitivity for identifying incident strokes increased when using ICD codes and natural language processing compared to ICD codes alone [30].

Conducting research using administrative health data can be cost-effective and time-saving. Knowing the limitations and how they affect analyses and inferences is critical to creating sound evidence that informs clinical decision making and future planning. In this investigation, using only diagnostic codes over-counted the number of incident PD cases in Olmsted County, Minnesota from 1976 through 2005 by 73% compared to using diagnostic codes for screening with subsequent medical record review and adjudication by a movement disorders specialist. In our context, using administrative diagnostic codes only to identify incident PD cases is not recommended unless more accurate algorithms are developed. Using diagnostic codes to screen for potential cases followed by medical record review remains the recommended approach.

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.prdoa.2020.100061>.

#### Declaration of competing interest

None.

#### Acknowledgements

The authors acknowledge Barbara A. Abbott for assisting with diagnostic code retrieval, Brandon R. Grossardt, MS, for providing datasets with context from the study used as the standard of comparison, and Mary G. Roberts for assisting with manuscript preparation.

#### Funding sources for study

This study was possible using the resources of the Rochester Epidemiology Project, which is supported by the National Institute on Aging of the National Institutes of Health Grants R01 AG034676 and R01 AG052425 and by the Mayo Foundation for Medical Education and Research. However, the content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes Health.

#### Author contributions

BJP: Research project design, organization, and execution; statistical analysis plan and execution; writing the first and subsequent drafts of the manuscript. WAR: Research project design; critical review and critique of the manuscript. JHB: Data collection; critical review and critique of the manuscript. RS: Data collection; critical review and critique of the manuscript. MMM: Critical review and critique of the manuscript. All authors have approved the final article.

#### References

- [1] A. Rossi, K. Berger, H. Chen, D. Leslie, R.B. Mailman, X. Huang, Projection of the prevalence of Parkinson's disease in the coming decades: revisited, *Mov. Disord.* 33 (2018) 156–159.
- [2] W.A. Rocca, The burden of Parkinson's disease: a worldwide perspective, *Lancet Neurol.* 17 (2018) 928–929.
- [3] R. Savica, B.R. Grossardt, J.H. Bower, J.E. Ahlskog, W.A. Rocca, Time trends in the incidence of Parkinson disease, *JAMA Neurol.* 73 (2016) 981–989.
- [4] J. Isotalo, T. Vahlberg, V. Kaasinen, Unchanged long-term rural-to-urban incidence ratio of Parkinson's disease, *Mov. Disord.* 32 (2017) 474–475.
- [5] C.C. Liu, C.Y. Li, P.C. Lee, Y. Sun, Variations in incidence and prevalence of Parkinson's disease in Taiwan: a population-based nationwide study, *Parkinsons Dis.* 2016 (2016) 8756359.
- [6] L. Horsfall, I. Petersen, K. Walters, A. Schrag, Time trends in incidence of Parkinson's disease diagnosis in UK primary care, *J. Neurol.* 260 (2013) 1351–1357.
- [7] W.M. Liu, R.M. Wu, J.W. Lin, Y.C. Liu, C.H. Chang, C.H. Lin, Time trends in the prevalence and incidence of Parkinson's disease in Taiwan: a nationwide, population-based study, *J. Formos. Med. Assoc.* 115 (2016) 531–538.
- [8] S.K. Darweesh, P.J. Koudstaal, B.H. Stricker, A. Hofman, M.A. Ikram, Trends in the incidence of Parkinson disease in the general population: the Rotterdam study, *Am. J. Epidemiol.* 183 (2016) 1018–1026.
- [9] E.I. Benchimol, D.G. Manuel, T. To, A.M. Griffiths, L. Rabeneck, A. Guttmann, Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data, *J. Clin. Epidemiol.* 64 (2011) 821–829.
- [10] Z. Harding, T. Wilkinson, A. Stevenson, S. Horrocks, A. Ly, C. Schnier, D.P. Breen, K. Rannikmae, C.L.M. Sudlow, Identifying Parkinson's disease and parkinsonism cases using routinely collected healthcare data: a systematic review, *PLoS One* 14 (2019), e0198736.
- [11] J.L. St. Sauver, B.R. Grossardt, B.P. Yawn, L.J. Melton 3rd, J.J. Pankratz, S.M. Brue, W.A. Rocca, Data resource profile: the Rochester epidemiology project (REP) medical records-linkage system, *Int. J. Epidemiol.* 41 (2012) 1614–1624.
- [12] J.L. St. Sauver, B.R. Grossardt, B.P. Yawn, L.J. Melton 3rd, W.A. Rocca, Use of a medical records linkage system to enumerate a dynamic population over time: the Rochester epidemiology project, *Am. J. Epidemiol.* 173 (2011) 1059–1068.
- [13] J.H. Bower, D.M. Maraganore, S.K. McDonnell, W.A. Rocca, Incidence and distribution of parkinsonism in Olmsted County, Minnesota, 1976–1990, *Neurology* 52 (1999) 1214–1220.
- [14] R. Savica, B.R. Grossardt, J.H. Bower, B.F. Boeve, J.E. Ahlskog, W.A. Rocca, Incidence of dementia with Lewy bodies and Parkinson disease dementia, *JAMA Neurol.* 70 (2013) 1396–1402.
- [15] W.A. Rocca, J.H. Bower, S.K. McDonnell, B.J. Peterson, D.M. Maraganore, Time trends in the incidence of parkinsonism in Olmsted County, Minnesota, *Neurology* 57 (2001) 462–467.
- [16] R. Savica, B.R. Grossardt, J.H. Bower, J.E. Ahlskog, W.A. Rocca, Incidence and pathology of synucleinopathies and tauopathies related to parkinsonism, *JAMA Neurol.* 70 (2013) 859–866.
- [17] A. Elbaz, B.J. Peterson, P. Yang, J.A. Van Gerpen, J.H. Bower, D.M. Maraganore, S.K. McDonnell, J.E. Ahlskog, W.A. Rocca, Nonfatal cancer preceding Parkinson's disease: a case-control study, *Epidemiology* 13 (2002) 157–164.
- [18] K. Swartztrauber, J. Anau, D. Peters, Identifying and distinguishing cases of parkinsonism and Parkinson's disease using ICD-9 CM codes and pharmacy data, *Mov. Disord.* 20 (2005) 964–970.
- [19] K. Noyes, H. Liu, R. Holloway, A.W. Dick, Accuracy of Medicare claims data in identifying parkinsonism cases: comparison with the Medicare current beneficiary survey, *Mov. Disord.* 22 (2007) 509–514.
- [20] N.R. Szumski, E.M. Cheng, Optimizing algorithms to identify Parkinson's disease cases within an administrative database, *Mov. Disord.* 24 (2009) 51–56.
- [21] A.L. Feldman, A.L. Johansson, M. Gatz, M. Flensburg, G.M. Petzinger, H. Widner, M.F. Lew, N.L. Pedersen, K. Wirdefeldt, Accuracy and sensitivity of Parkinsonian disorder diagnoses in two Swedish national health registers, *Neuroepidemiology* 38 (2012) 186–193.
- [22] D.A. Butt, K. Tu, J. Young, D. Green, M. Wang, N. Ivers, L. Jaakkimainen, R. Lam, M. Guttman, A validation study of administrative data algorithms to identify patients with parkinsonism with prevalence and incidence trends, *Neuroepidemiology* 43 (2014) 28–37.
- [23] M. Nerijs, A. Fink, G. Doblhammer, Parkinson's disease in Germany: prevalence and incidence based on health claims data, *Acta Neurol. Scand.* 136 (2017) 386–392.
- [24] P. Blin, C. Dureau-Pourmin, A. Foubert-Samier, A. Grolleau, E. Corbillon, J. Jove, R. Lassalle, P. Robinson, N. Poutignat, C. Droz-Perroteau, N. Moore, Parkinson's disease incidence and prevalence assessment in France using the national healthcare insurance database, *Eur. J. Neurol.* 22 (2015) 464–471.
- [25] T.L. Sanders, A. Pareek, V.S. Desai, T.E. Hewett, B.A. Levy, M.J. Stuart, D.L. Dahm, A.J. Krych, Low accuracy of diagnostic codes to identify anterior cruciate ligament tear in orthopaedic database research, *Am. J. Sports Med.* 46 (2018) 2894–2898.
- [26] A.J. Hughes, S.E. Daniel, Y. Ben-Shlomo, A.J. Lees, The accuracy of diagnosis of parkinsonian syndromes in a specialist movement disorder service, *Brain* 125 (2002) 861–870.
- [27] J. Horvath, P.R. Burkhard, C. Bouras, E. Kovari, Etiologies of parkinsonism in a century-long autopsy-based cohort, *Brain Pathol.* 23 (2013) 28–33.
- [28] J. Joutsma, M. Gardberg, M. Roytta, V. Kaasinen, Diagnostic accuracy of parkinsonism syndromes by general neurologists, *Parkinsonism Relat. Disord.* 20 (2014) 840–844.
- [29] S.S. Lloyd, J.P. Rissing, Physician and coding errors in patient records, *JAMA* 254 (1985) 1330–1336.
- [30] Y. Zhao, S. Fu, N.B. Larson, P.A. Decker, A.M. Chamberlain, V.L. Roger, H. Liu, S.J. Bielski, Validation of Phenotyping Algorithms for Stroke from Electronic Health Records Using Natural Language Processing, *Circulation* 139, 2019.