scientific reports

Check for updates

OPEN

Performance of administrative databases for identifying individuals with multiple sclerosis

Pauline Ducatel[1,2]✉, Marc Debouverie[1,3], Marc Soudant[2], Francis Guillemin[2,3], Guillaume Mathey[1,3] & Jonathan Epstein[2,3]

Administrative databases are an alternative to disease registries as a research tool to study multiple sclerosis. However, they are not initially designed to fulfill research purposes. Therefore, an evaluation of their performance is necessary. Our objective was to assess the performance of the French administrative database comprising hospital discharge records and national health insurance databases in identifying individuals with multiple sclerosis, in comparison with a registry that exhaustively compiles resident multiple sclerosis cases in Lorraine, northeastern France, as reference. We recorded all individuals residing in the Lorraine region who were identified by the administrative database or the registry as having multiple sclerosis from 2011 to 2016. We calculated the Matthews correlation coefficient and other concordance indicators. For identifying individuals with multiple sclerosis, the Matthews correlation coefficient by the administrative database was 0.79 (95% CI 0.78–0.80), reflecting moderate performance. The mean time to identification was 5.5 years earlier with the registry than the administrative database. Administrative databases, although useful to study multiple sclerosis, should be used with caution because results of studies based on them may be biased. Our study highlights the value of regional registries that allow for a more exhaustive and rapid identification of cases.

Multiple sclerosis (MS) is a neuroinflammatory disorder affecting the central nervous system. This disease has high socio-economic impact[1,2] and is considered a public health priority in many countries.

Disease registries provide researchers and policymakers with accurate information to conduct observational study, monitor conditions and plan healthcare[3,4]. However, they require substantial resources and careful planning[5]. Therefore, administrative databases (ADs) represent an alternative tool that may be preferred.

ADs are designed for financial and administrative purposes and are increasingly used in medico-economic studies or pharmacoepidemiology or epidemiological research[6]. Recent MS descriptive studies have used ADs for both prevalence and incidence estimates[7–9]. In France, the AD combines hospital discharge records and national health insurance databases. It recognizes MS cases by the presence of at least one of 1) reimbursement for a disease-modifying treatment considered specific to this condition, 2) declared disease-specific payment exemption for MS (in France, this registration allows patients with a particular illness to be 100% reimbursed for their outpatient and inpatient health care costs) and/or 3) hospital discharge with MS diagnosis.

However, the registered prevalence of disease-specific payment exemptions has been found systemically lower than the actual prevalence of illnesses[10]. MS exemptions can also be misreported for a related condition[11]. Moreover, benign MS cases do not necessarily require the initiation of disease-modifying therapy or hospitalization. Indeed, the number of cases without treatment is estimated at 31%[12]. Furthermore, the accuracy of coding diagnoses in the hospital discharge records for MS is uncertain[13,14]. Therefore, the use of ADs outside the designed financial and administrative objectives[6,15] could lead to a misclassification of individuals with MS and affect the conclusions of studies relying on these databases.

To determine the reliability of a data source, one of the most commonly used methods is comparison with a reference of recognized validity[16]. The *Registre Lorrain de la Sclérose en plaques* (ReLSEP) collects MS cases in the Lorraine region, in northeastern France. It was created to meet medical research and MS incidence monitoring objectives. To identify cases, ReLSEP relies on the reports of all neurologists in the region who are an essential part of patients' healthcare pathways. Moreover, multiple sources are also solicited, including the AD.

[1]Department of Neurology, Nancy University Hospital, 29 Avenue du Maréchal de Lattre de Tassigny, Nancy, France. [2]CIC-EC 1433, CHRU, Inserm, Université de Lorraine, 9 Av. de La Forêt de Haye, Vandoeuvre-Lès-Nancy, France. [3]Université de Lorraine, EA 4360 Apemac, 9 Av. de La Forêt de Haye, Vandoeuvre-Lès-Nancy, France. ✉email: p.ducatel@chru-nancy.fr

The main objective of this study was to assess the performance of the French AD in identifying individuals with MS in comparison with the ReLSEP reference. Our secondary objectives were to determine factors associated with misidentification by the AD; assess and compare the proportion of cases recognized for the first time by ReLSEP or the AD, among cases identified in common; and ascertain the performance of different combinations of criteria to identify MS cases in ADs.

Materials and methods

Design

We conducted a study targeting residents in the Lorraine region (comprising four departments: Meurthe-et-Moselle, Meuse, Moselle and Vosges). Cases were identified by the AD or ReLSEP as MS from January 1, 2011 to December 31, 2016. This time frame was chosen to allow for homogeneity in AD and ReLSEP identification criteria, as the MS 2010 McDonald criteria[17] were modified in 2017[18]. Also, we determined in previous analyses that a newly diagnosed case of multiple sclerosis was identified by the ReLSEP within 5 years[19]. Therefore, the end of the study period was set 5 years previous to this study analysis to provide sufficient hindsight to enable a reliable identification of MS cases and comparison of the identification time between the ReLSEP and AD.

ReLSEP

The ReLSEP was created in 2003. It exhaustively compiles and verifies all MS cases of patients residing in Lorraine. It relies on the reports of all neurologists in the region (hospital and private practice). It also benefits from an annual extraction from multiple sources: the AD, the two regional biology laboratories that analyze cerebral spinal fluid (CSF) samples and two regional networks of MS professionals (LORSEP and ALSASEP).

To ensure the accuracy of the diagnosis, each case is confirmed by a neurologist specialized in MS and is validated only if the 2010 McDonald criteria are met[17].

AD

Data are collected annually from the hospital discharge records and the national health insurance databases and are nominative. National health insurance data contain two criteria: reimbursement for a disease-modifying treatment considered specific to MS and MS exemption. For each health institution in France (including outside Lorraine), hospital discharge data for patients residing in Lorraine related to a MS diagnosis are transmitted annually.

Data from annual extractions are cross-referenced with data for individuals already in the ReLSEP and matched by last name, first name, date of birth, sex and department of residence. The search for duplicates is performed using a semi-manual method. When files match on all criteria, they are considered as duplicates. In case of doubt, a source check can be carried out to conclude.

Each reported case is subjected to a validation procedure before being included in the ReLSEP. The patient's file is retrieved from the general practitioner or specialist by clinical research nurses. A neurologist specialized in MS validate the diagnosis if the 2010 McDonald criteria are met.

Data collected

For the individuals in the registry, the variables of interest were age at onset, sex, department of residence, first healthcare facility, associated comorbidities, first Expanded Disability Status Scale (EDSS) assessing the level of disability, number of MS relapses over the first 2 years after diagnosis, results of the first brain MRI, cerebral spinal fluid analysis performed and initial treatment. Data regarding individuals known only to the AD were limited to available data: age at identification by the AD, sex, department of residence, year of identification by the AD and first healthcare facility (Supplementary information SI1).

Statistical analysis

The individuals recognized by the two sources were true positives (TPs). We defined as false positives (FPs) the individuals detected by the AD but not the ReLSEP, and as false negatives (FNs), the individuals identified by the ReLSEP but not the AD. The true negatives (TNs) referred to the Lorraine's population and were estimated from data reported annually by the French national institute for statistical and economic studies[20].

We chose the Matthews correlation coefficient (MCC) as the primary endpoint[21]. The MCC corresponds to the Pearson correlation coefficient applied to a binary classification and is interpreted in the same way[22,23]. It is popular in the machine learning field[24] and has previously been used to evaluate the performance of ADs[25]. It is a useful metric for unbalanced datasets and allows for considering all the parameters of the contingency table[26]. The MCC ranges from − 1 to + 1, indicating perfect agreement (+ 1) or disagreement (-1), and 0 means no relationship. A correlation ≥ 0.8 is considered strong and below moderate[27].

Categorical variables were described with number and percentage and quantitative variables with mean and standard deviation.

The primary analysis included the calculation of the MCC and other indicators: sensitivity (Se), specificity (Sp), positive predictive value (PPV), negative predictive value (NPV), accuracy, Cohen's kappa and F1-score. We used bootstrapping to compute the confidence intervals of the F1-score and MCC.

Using the TPs as a reference, we compared the TPs and FPs, and the TPs and FNs. For the bivariate analyses, Kruskal–Wallis, chi-square, Wilcoxon and Fisher's exact test were used as appropriate, with a 0.05 two-sided significance level.

For the multivariable analyses, we used three hierarchical logistic regression models with the department of residence as a random variable and, when appropriate, the year of identification by the AD (Supplementary Information SI2). The first model studied factors associated with misidentification as MS cases by the AD.

The second model investigated factors associated with failure to identify MS cases by the AD. For these two multivariate models, we retained factors with $p < 0.15$ on bivariate analysis. Age and sex were also included because of their role in the use of healthcare described in the literature[28,29]. Odds ratios and 95% confidence intervals (CIs) were estimated. We calculated the time to identify individuals with MS by the ReLSEP and AD. With a third model, we examined factors associated with early identification by the AD versus the ReLSEP among TPs. Because the AD is extracted annually, cases were considered identified first by the ReLSEP if they were recognized at least 1 year before the AD. This was the most unfavorable situation for the registry.

We used the VIM package to provide a visual representation of the proportion of missing data for each variable, as well as their distribution (Supplementary information SI3, SI4 and SI5). Multiple imputations were performed under the missing at random assumption, using the predictive mean matching method of the MICE package[30] (15 imputations). The variables imputed were the outcome and covariates with missing data. After generating imputed datasets, we ran our models using the with() function on each dataset. Finally, we pooled the results over the imputed datasets using the pool() function to compute multilevel (residence + /− year of identification by the AD as random variables) imputed (average effect relative to each imputed dataset) odds ratio. We also conducted complete-case sensitivity analyses to confirm the results of our secondary analyses.

To assess the performance of different combinations of the three criteria (reimbursement for a disease-modifying treatment, MS exemption and hospital discharge) for identifying MS cases using the AD, we calculated, for each combination, the same metrics used beforehand.

All analyses were performed with R 4.1.2.

Ethical approval and Informed consent

The National Commission for Data Protection and Liberties and the Consultative Committee for the data processing in health research gave both a favorable opinion for the collection of data by the RELSeP (no. 913001 the 01/06/2014 and no. 10–258 the 05/06/2010). All methods were performed in accordance with the relevant guidelines and regulations. Informed consent was obtained from all subjects and/or their legal guardian(s).

Results

Of the 2084 individuals identified by ReLSEP, 1152 (55%) were declared by neurologists/LORSEP, 759 by biology laboratories (36.2%) and 37 by ALSACEP (1.8%).

During the considered period, 1826 cases were TP, 258 were FN and 725 were FP. In total, 28.4% (725/2551) of individuals reported by the AD were wrongly identified as having MS. The estimated number of TNs was 2,245,196 (Supplementary information SI6).

The characteristics of TPs and FNs are in Table 1 (and Supplementary information SI7). The variable with the highest rate of missing data was the first EDSS (8.5%). For FPs, the mean age at identification by the AD was 50.8 (18.6) years, 68.5% were women. Most patients resided in Moselle (49.8%) and were followed in a hospital (80.2%) (Supplementary information SI8).

The concordance indicators assessing the performance of the AD are summarized in the last row of Table 2. The MCC was 0.79 (95% CI 0.78–0.80), reflecting moderate performance by the AD in identifying MS cases. Other global metrics such as kappa (0.79 [95% CI 0.78–0.80]) and F1-score (0.79 [95% CI 0.77–0.80]) confirmed

Status		TP	FN	OR (multilevel imputed) (95% CI, *p* value)
		N = 1826	N = 258	
Age at onset, years	Mean (SD)	33.4 (11.4)	33.5 (12.0)	1.01 (0.99–1.02, $p = 0.460$)
Sex	Male	538 (29.8)	66 (25.6)	–
	Female	1270 (70.2)	192 (74.4)	1.05 (0.77–1.45, $p = 0.749$)
First healthcare facility	Hospital	1409 (77.4)	215 (83.3)	–
	Private practice	411 (22.6)	43 (16.7)	0.72 (0.50–1.04, $p = 0.078$)
Associated comorbidities	No	611 (33.5)	104 (40.3)	–
	Yes	1215 (66.5)	154 (59.7)	0.85 (0.63–1.13, $p = 0.266$)
First EDSS	Mean (SD)	2.7 (2.0)	1.8 (1.6)	0.68 (0.61–0.75, $p < 0.001$)
Number of relapses over the first 2 years	Mean (SD)	1.2 (0.9)	1.1 (0.6)	0.73 (0.59–0.91, $p = 0.005$)
First brain MRI	Normal	83 (4.7)	28 (10.9)	–
	Abnormal	1678 (95.3)	229 (89.1)	0.47 (0.28–0.76, $p = 0.002$)
CSF analysis performed	Yes	1312 (71.9)	208 (80.6)	–
	No	514 (28.1)	50 (19.4)	0.73 (0.51–1.05, $p = 0.089$)
Initial treatment	None	315 (17.3)	88 (34.1)	–
	Moderate	1211 (66.3)	139 (53.9)	0.33 (0.23–0.46, $p < 0.001$)
	Highly	300 (16.4)	31 (12.0)	0.49 (0.30–0.79, $p = 0.003$)

Table 1. Factors associated with failure to identify multiple sclerosis by the administrative database. Number (multilevel imputed): 2084, EDSS = Expended Disability Status Scale, CSF = cerebrospinal fluid, TP = false positive, FN = true negative, OR = odds ratio, 95% CI = 95% confidence interval.

Criteria	Sensitivity (95% CI)	PPV (95% CI)	F1-score (95% CI)	Kappa (95% CI)	MCC (95% CI)
DMT	0.38 [0.36–0.40]	0.86 [0.84–0.88]	0.53 [0.51–0.55]	0.53 [0.51–0.55]	0.57 [0.55–0.59]
DSPE	0.58 [0.56–0.60]	0.82 [0.80–0.84]	0.68 [0.66–0.70]	0.68 [0.66–0.70]	0.69 [0.67–0.71]
DMT + DSPE	0.61 [0.59–0.63]	0.80 [0.78–0.82]	0.69 [0.67–0.71]	0.69 [0.67–0.71]	0.70 [0.68–0.71]
HD	0.68 [0.66–0.70]	0.74 [0.72–0.76]	0.71 [0.69–0.73]	0.71 [0.70–0.72]	0.71 [0.69–0.73]
DMT + HD	0.79 [0.78–0.81]	0.74 [0.72–0.76]	0.77 [0.75–0.78]	0.77 [0.75–0.78]	0.77 [0.75–0.78]
DSPE + HD	0.87 [0.85–0.88]	0.73 [0.71–0.74]	0.79 [0.78–0.80]	0.79 [0.78–0.80]	0.79 [0.78–0.81]
DMT + DSPE + HD	0.88 [0.86–0.89]	0.72 [0.70–0.73]	0.79 [0.77–0.80]	0.79 [0.78–0.80]	0.79 [0.78–0.80]

Table 2. Performance of different combinations of administrative database criteria to identify individuals with multiple sclerosis. DMT = disease-modifying treatment, DSPE = disease-specific payment exemption, HD = hospital discharge, PPV = positive predictive value, MCC = Matthews correlation coefficient, 95% CI = 95% confidence interval.

this finding. The Sp, NPV and accuracy were 0.99 (95% CI 0.99–0.99). The Se was 0.88 (95% CI 0.86–0.99) and PPV 0.72 (95% CI 0.70–0.73). Because most of the FPs and FNs resided in Moselle, we performed a sensitivity analysis by removing individuals from Moselle (Supplementary Information SI9). The results of the main analysis were not modified, suggesting that because Moselle is Lorraine's most populous department (almost half of Lorraine's population), the largest number of VPs, FPs and FNs reside there.

Table 2 also summarizes the performance of different combinations of the three identification criteria in the AD. When using unique criteria, hospital discharges had the best Se (0.68 [range 0.38–0.68]) and reimbursement for a disease-modifying treatment the best PPV (0.86 [range 0.74–0.86]). The addition of reimbursement for a disease-modifying treatment to the combination of the two other criteria did not bring any benefit for all metrics. Unlike FPs, most TPs were identified by 2 or 3 criteria in the AD. The HD criterion detected the greatest number of TPs and FPs, while the DMT criterion identified the fewest (Supplementary information SI10).

The only factor associated with probability of FP as compared with TP was age (OR = 1.03 [95% CI 1.02–1.04] for 1 more year) (Table 3).

MS cases with factors reflecting benign forms [31] were less likely to be identified by the AD. Indeed, the following characteristics were associated with the probability of FN: low EDSS (OR = 1.49 [95% CI 1.33–1.64] for 1 unit decrease), low number of MS relapses over the first 2 years after diagnosis (OR = 1.35 [95% CI 1.10–1.79] for 1 unit decrease), first brain MRI normal (abnormal OR = 0.46 [95% CI 0.28–0.76]) and no treatment initially (moderate activity OR = 0.33 [95% CI 0.23–0.45] and high activity OR = 0.49 [95% CI 0.30–0.79]) (Table 1).

Among MS cases identified by the two sources, 60.5% were identified first by the ReLSEP. The mean time to identification was 5.5 years earlier with the ReLSEP than the AD (median 1 year) (Supplementary information SI11). Factors associated with early identification by the AD are in Table 4.

The findings obtained with complete-case analyses did not differ from multiple imputation, reinforcing and strengthening the results of our secondary analyses (Supplementary information SI12, SI13 and SI14).

Discussion

The performance of the AD for identifying individuals with MS in comparison with the ReLSEP was moderate, with an MCC of 0.79. This finding may have consequences for the results of the studies using ADs, depending on their objectives. A previous study conducted in 2012 by Foulon et al.[9] investigated the prevalence of MS in each region of France using the AD. As an illustration, one can correct the prevalence observed in this study according to the sensitivity and specificity obtained in our study[32]. For the highest (lowest) regional prevalence observed at 200.2 (125.6) per 100 000 inhabitants, the corrected prevalence would be 191.8 (105.9). The total prevalence in France would be overestimated by 11.3% (151.2 vs 135.8 per 100 000 inhabitants). Thus, based solely on the AD, the MS prevalence would be overestimated, varying from 4.4% to 18%. Also, individuals with benign MS were more likely to be missed by the AD, because they use less healthcare. This finding could affect the results of medico-economic studies, for example, by overestimating the average costs per patient or misestimating the costs of the disease[33–35]. Similarly, the results of any study in which differential identification of the most severe

Status		TP	FP	OR (multilevel imputed) (95% CI, p value)
		N = 1826	N = 725	
Age at onset, years	Mean (SD)	43.4 (15.0)	50.8 (18.6)	1.03 (1.02–1.04, $p < 0.001$)
Sex	Male	538 (29.8)	219 (31.5)	–
	Female	1270 (70.2)	476 (68.5)	0.98 (0.80–1.19, $p = 0.827$)

Table 3. Factors associated with misidentification as multiple sclerosis by the administrative database. Number (multilevel imputed): 2551, TP = true positive, FP = false positive, OR = odds ratio, 95% CI = 95% confidence interval.

First identification		ReLSEP	AD	OR (multilevel imputed) (95% CI, *p* value)
		N = 1094	N = 714	
Age at onset, years	Mean (SD)	33.5 (11.5)	33.1 (11.3)	1.00 (0.99–1.01, $p = 0.781$)
Sex	Male	301 (27.5)	237 (33.2)	–
	Female	793 (72.5)	477 (66.8)	0.69 (0.54–0.88, $p = 0.002$)
First healthcare facility	Hospital	799 (73.2)	596 (83.6)	–
	Private practice	293 (26.8)	117 (16.4)	0.58 (0.44–0.76, $p < 0.001$)
Associated comorbidities	No	365 (33.4)	228 (31.9)	–
	Yes	729 (66.6)	486 (68.1)	1.02 (0.81–1.29, $p = 0.851$)
First EDSS	Mean (SD)	2.9 (2.1)	2.4 (1.8)	1.04 (0.96–1.13, $p = 0.348$)
Number of relapses over the first 2 years	Mean (SD)	1.2 (0.9)	1.3 (0.9)	1.09 (0.94–1.26, $p = 0.260$)
Form of MS	RR	647 (59.2)	576 (80.9)	–
	PP	170 (15.6)	108 (15.2)	0.51 (0.32–0.81, $p = 0.005$)
	SP	276 (25.3)	28 (3.9)	0.08 (0.05–0.13, $p < 0.001$)
First brain MRI	Normal	70 (6.6)	13 (1.9)	–
	Abnormal	990 (93.4)	688 (98.1)	4.16 (2.22–7.82, $p < 0.001$)
CSF analysis performed	Yes	721 (65.9)	591 (82.8)	–
	No	373 (34.1)	123 (17.2)	0.50 (0.38–0.66, $p < 0.001$)
Initial treatment	None	211 (19.3)	86 (12.0)	–
	Moderate	716 (65.4)	495 (69.3)	1.59 (1.14–2.21, $p = 0.007$)
	Highly	167 (15.3)	133 (18.6)	3.06 (2.02–4.64, $p < 0.001$)

Factors associated with early identification of individuals with multiple sclerosis by the administrative database. Number (multilevel imputed): 1826 (the first source of identification was missing for 18 true positives), MS = multiple sclerosis, EDSS = Expended Disability Status Scale, CSF = cerebrospinal fluid, AD = administrative database, ReLSEP = *Registre Lorrain de la Sclérose en plaques*, RR = relapse remitting, PP = primary progressive, SP = secondary progressive, OR = odds ratio, 95% CI = 95% confidence interval.

forms may have an impact could be affected. Moreover, we determined that among the mutual cases known by both sources, 60% were identified by the ReLSEP at least 1 year before the AD. Thus, not only were not all individuals with MS identified by the AD, they were also known later than by ReLSEP. This observation challenges the preconceived notion that ADs allow for a timely monitoring of the disease[36] and raises the question of their relevance for incidence studies[7,37,38].

Interestingly, the FPs far outnumbered the FNs. These FPs correspond to individuals identified by the administrative database but whose diagnosis was ultimately refuted after study of their file by a neurologist specialized in MS. Apart from misdiagnosis, several hypotheses can be put forward to explain these FPs. It can happen that a disease-specific payment exemption for MS is wrongly declared when a patient suffers from a condition similar to MS that does not belong to the list of disease-specific payment exemption[14]. In addition, uncertainty about the quality of diagnosis coding for MS in the hospital discharge records has been described previously[11]. Unfortunately, we cannot provide more information regarding the FPs and the reasons behind misidentification. As these individuals do not have MS, they are not part of the ReLSEP and we are unable to collect data about them.

Several countries have conducted validation studies of algorithms used to detect individuals with MS in ADs. The criteria used were close to those of the French AD: mainly inpatient/outpatient encounters[39–41], more rarely drug dispensing records as well or even disease-specific payment exemption[42,43]. However, the chosen references were imperfect in exhaustively identifying MS cases. Indeed, some verified the MS diagnosis of individuals identified by ADs from medical records[39,44]. Therefore, this strategy does not adequately detect the FNs missed by ADs. Additional studies used regional registries populated by one[45] or more[42,43] specialized regional care centers. Nevertheless, in contrast with individuals in the ReLSEP, those not followed up in these institutions could not be identified by the reference. Other works were based on registries in which physician participation was voluntary[40,41], with a diagnosis not always confirmed by a neurologist[41].

The ReLSEP represents the main strength of this work. This registry exhaustively identified validated MS cases in Lorraine. It had the advantage of providing detailed information about clinical elements, complementary examinations, treatments and follow-up, impossible to collect from ADs. Also, the study period selected allowed for reliable identification of individuals with MS with a sufficient hindsight of more than 5 years, consistent with the chronic and progressive nature of the disease. Finally, the combination of the three criteria of the AD considered in our study corresponded to their use in practice in identifying MS cases in France[9,33,46]. To strengthen our external validity, other combinations of criteria, which may more closely approximate the use of ADs in other countries, were assessed.

The AD, whose performance was evaluated in our study, is also part of the ReLSEP sources. This situation could have led to incorporation bias, which can occur when the test under consideration is used to determine the reference. This type of bias is likely to overvalue the performance of the studied procedure[47]. Yet, despite

this potential overestimation, the performance of the AD was moderate, which reinforces the validity of our conclusion. Also, since 2016, new drugs have been reimbursed and were not considered in this study. However, there are still cases of benign MS that do not require any treatment[12]. Therefore, our conclusions are unlikely to change. Finally, missing data could have affected our secondary analyses, so we took them into account using multiple imputation.

This work provides a clearer picture of the limitations of ADs and the potential impact their use may have on MS studies. Indeed, depending on their objective, their use may lead to risk of error in results. Therefore, ADs, although beneficial to investigate MS, should be used with caution. Moreover, MS registries collect rich and high-quality data that is impossible to obtain from ADs. Finally, by highlighting the value of MS registries that allow for a more exhaustive and rapid identification of cases, our findings support their development and the justification of the resources allocated.

Data availability

All data generated or analyzed during this study are included in this article and its supplementary material files. Further enquiries can be directed to the corresponding author.

Received: 30 April 2023; Accepted: 19 October 2023
Published online: 25 October 2023

References

1. Jennum, P., Wanscher, B., Frederiksen, J. & Kjellberg, J. The socioeconomic consequences of multiple sclerosis: a controlled national study. *Eur. Neuropsychopharmacol.* **22**, 36–43 (2012).
2. Vandhuick, O. *et al.* Economic burden of highly active relapsing-remitting multiple sclerosis patients in the French national health insurance database. *Exp. Rev. Pharmacoecon. Outcomes Res.* **21**, 1135–1144 (2021).
3. *Registries for Evaluating Patient Outcomes: A User's Guide*. (Agency for Healthcare Research and Quality (US), 2020).
4. Pop, B. *et al.* The role of medical registries, potential applications and limitations. *Med. Pharm. Rep.* **92**, 7–14 (2019).
5. Wormald, J. S., Oberai, T., Branford-White, H. & Johnson, L. J. Design and establishment of a cancer registry: a literature review. *ANZ J. Surg.* **90**, 1277–1282 (2020).
6. Gavrielov-Yusim, N. & Friger, M. Use of administrative medical databases in population-based research: Table 1. *J Epidemiol. Commun. Health* **68**, 283–287 (2014).
7. Wnuk, M. *et al.* Multiple sclerosis incidence and prevalence in Poland: Data from administrative health claims. *Multiple Sclerosis Related Dis.* **55**, 103162 (2021).
8. Bakirtzis, C. *et al.* The administrative prevalence of multiple sclerosis in greece on the basis of a nationwide prescription database. *Front Neurol.* **11**, 1012 (2020).
9. Foulon, S. *et al.* Prevalence and mortality of patients with multiple sclerosis in France in 2012: a study based on French health insurance data. *J. Neurol.* **264**, 1185–1192 (2017).
10. Goldberg, M *et al.* Bases de données médico-administratives et épidémiologie : intérêts et limites. *Courrier des Statistiques - INSEE* 59–70 (2008).
11. Grosclaude, P *et al.* Utilité des bases de données médico-administratives pour le suivi épidémiologique des cancers. Comparaison avec les données des registres au niveau individuel. *Bull. Epidémiologique Hebdomadaire* 63–67 (2012).
12. Moisset, X. *et al.* Untreated patients with multiple sclerosis: A study of French expert centers. *Eur. J. Neurol.* **28**, 2026–2036 (2021).
13. Gologorsky, Y., Knightly, J. J., Lu, Y., Chi, J. H. & Groff, M. W. Improving discharge data fidelity for use in large administrative databases. *Neurosurg. Focus* **36**, E2 (2014).
14. Sagnes-Raffy, C. *et al.* La SEP en Haute-Garonne: une sous-estimation importante du nombre de cas. *Revue d'épidémiologie et de santé publique* **58**(1), 23–31. https://doi.org/10.1016/J.RESPE.2009.08.012 (2010).
15. Article L461–1 - Code de la sécurité sociale - Légifrance. https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000036393217/.
16. Behrendt, C.-A. *et al.* Data privacy compliant validation of health insurance claims data: the IDOMENEO Approach. *Gesundheitswesen* **82**, S94–S100 (2020).
17. Polman, C. H. *et al.* Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria. *Ann. Neurol.* **69**, 292–302 (2011).
18. Thompson, A. J. *et al.* Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *The Lancet Neurol.* **17**, 162–173 (2018).
19. Gbaguidi, B. *et al.* Age-period-cohort analysis of the incidence of multiple sclerosis over twenty years in Lorraine. *France. Sci. Rep.* **12**, 1001 (2022).
20. Estimation de la population au 1er janvier 2021 | Insee. https://www.insee.fr/fr/statistiques/1893198.
21. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure* **405**, 442–451 (1975).
22. Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. & Nielsen, H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* **16**, 412–424 (2000).
23. Powers, D. M. W. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv: 2010.16061 *[cs, stat]* (2020).
24. Chicco, D., Warrens, M. J. & Jurman, G. The matthews correlation coefficient (MCC) is more informative than Cohen's Kappa and brier score in binary classification assessment. *IEEE Access* **9**, 78368–78381 (2021).
25. Tawfik, D. S., Gould, J. B. & Profit, J. Perinatal risk factors and outcome coding in clinical and administrative databases. *Pediatrics* **143**, e20181487 (2019).
26. Delgado, R. & Tibau Alberdi, X.-A. Why Cohen's Kappa should be avoided as performance measure in classification. *PLoS ONE* **14**, 1–26 (2019).
27. Zou, K. H., Tuncali, K. & Silverman, S. G. Correlation and simple linear regression. *Radiology* **227**, 617–628 (2003).
28. Deeks, A., Lombard, C., Michelmore, J. & Teede, H. The effects of gender and age on health related behaviors. *BMC Public Health* **9**, 213 (2009).
29. Keene, J. & Li, X. Age and gender differences in health service utilization. *J. Public Health* **27**, 74–79 (2005).
30. van Buuren, S. & Groothuis-Oudshoorn, K. mice: multivariate imputation by chained equations in R. *J. Stat. Softw.* **45**, 1–67 (2011).
31. Reynders, T., D'haeseleer, M., De Keyser, J., Nagels, G. & D'hooghe, M. B. Definition, prevalence and predictive factors of benign multiple sclerosis. *eNeurologicalSci* **7**, 37–43 (2017).
32. Rogan, W. J. & Gladen, B. Estimating prevalence from the results of a screening test. *Am. J. Epidemiol.* **107**, 71–76 (1978).
33. Lefeuvre, D., Rudant, J., Foulon, S., Alla, F. & Weill, A. Healthcare expenditure of multiple sclerosis patients in 2013: A nationwide study based on French health administrative databases. *Mult. Scler. J. Exp. Transl. Clin.* **3**, 2055217317730421 (2017).

34. Buijs, S., Krol, M. & de Voer, G. Healthcare utilization and costs of multiple sclerosis patients in the Netherlands: a healthcare claims database study. *J. Comp. Effect. Res.* **7**, 453–462 (2018).
35. Maia Diniz, I. *et al.* The long-term costs for treating multiple sclerosis in a 16-year retrospective cohort study in Brazil. *PLoS ONE* **13**, e0199446 (2018).
36. Brocco, S. *et al.* Monitoring the occurrence of diabetes mellitus and its major complications: the combined use of different administrative databases. *Cardiovas. Diabetol.* **6**, 5 (2007).
37. Fromont, A. *et al.* Geographic variations of multiple sclerosis in France. *Brain* **133**, 1889–1899 (2010).
38. Iljicsov, A. *et al.* Incidence and prevalence of multiple sclerosis in Hungary based on record linkage of nationwide multiple healthcare administrative data. *PLoS ONE* **15**, e0236432 (2020).
39. Teljas, C. *et al.* Validating the diagnosis of multiple sclerosis using Swedish administrative data in Värmland County. *Acta Neuro Scandinavica* **144**, 680–686 (2021).
40. Murley, C., Friberg, E., Hillert, J., Alexanderson, K. & Yang, F. Validation of multiple sclerosis diagnoses in the Swedish National Patient Register. *Eur. J. Epidemiol.* **34**, 1161–1169 (2019).
41. Widdifield, J. *et al.* Development and validation of an administrative data algorithm to estimate the disease burden and epidemiology of multiple sclerosis in Ontario, Canada. *Mult. Scler.* **21**, 1045–1054 (2015).
42. Bezzini, D. *et al.* Prevalence of multiple sclerosis in tuscany (Central Italy): a study based on validated administrative data. *Neuroepidemiology* **46**, 37–42 (2016).
43. Gnavi, R. *et al.* Validation of an algorithm to detect multiple sclerosis cases in administrative health databases in Piedmont (Italy): an application to the estimate of prevalence by age and urbanization level. *Neuroepidemiology* **55**, 119–125 (2021).
44. Culpepper, W. J. *et al.* Validation of an algorithm for identifying MS cases in administrative health claims datasets. *Neurology* **92**, e1016–e1028 (2019).
45. Moccia, M. *et al.* Multiple sclerosis in the campania region (South Italy): algorithm validation and 2015–2017 prevalence. *Int. J. Environ. Res. Public Health* **17**, 3388 (2020).
46. Roux, J., Guilleux, A., Lefort, M. & Leray, E. Use of healthcare services by patients with multiple sclerosis in France over 2010–2015: a nationwide population-based study using health administrative data. *Mult. Scler. J. Exp. Trans. Clin.* **5**, 205521731989609 (2019).
47. Worster, A. & Carpenter, C. Incorporation bias in studies of diagnostic tests: how to avoid being biased about bias. *CJEM* **10**, 174–175 (2008).

Author contributions

P.D, J.E, M.D, G.M contributed to the study conception and design. M.S. led the preparation of the study data. P.D performed the statistical analysis. All authors contributed to interpretation of the results. The first draft of the manuscript was written by P.D and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interests

G.M had travel/accomodations/meeting expenses funded by BIOGEN, NOVARTIS, SANOFI-GENZYME, MERCK, TEVA, and ROCHE. He had contracts for lectures or boards with BIOGEN, SANOFI-GENZYME, ALEXION, ROCHE, MERCK and NOVARTIS without compensation/honoraria. He participated in clinical trials by BIOGEN, ACTELION, ROCHE, MERCK and NOVARTIS without compensation/honoraria. His institution received research grants from BIOGEN, NOVARTIS, ACTELION, SANOFI-GENZYME, MERCK, and ROCHE. The other authors have no conflict of interest.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-45384-w.

Correspondence and requests for materials should be addressed to P.D.