Data linkage between the French multiple sclerosis cohort (OFSEP) and the French national health insurance database (SNDS)

E. Leray, Erwan Drézen, Romain Casey, Sandra Vukusic

Available online at
ScienceDirect
www.sciencedirect.com

Elsevier Masson France
EM|consulte
www.em-consulte.com

Original article

Data linkage between the French multiple sclerosis cohort (OFSEP) and the French national health insurance database (SNDS)

E. Leray [a,*], *E. Drezen* [b], *R. Casey* [c,d,e,f], *S. Vukusic* [c,d,e,f]
on behalf of Ofsep

[a] *École des hautes études en santé publique (EHESP), CNRS, Inserm, ARENES UMR 6051, RSMS U 1309, University of Rennes, avenue du Pr Léon Bernard, 35033 Rennes cedex, France*
[b] *CUBR, 5, allée William-Loth, 35000 Rennes, France*
[c] *Service de neurologie, sclérose en plaques, pathologies de la myéline et neuro-inflammation, hôpital neurologique Pierre-Wertheimer, hospices civils de Lyon, 59, boulevard Pinel, 69677 Bron, France*
[d] *Centre de recherche en neurosciences de Lyon, Observatoire français de la sclérose en plaques, Inserm 1028 et CNRS UMR 5292, 59, boulevard Pinel, 69003 Lyon, France*
[e] *Université de Lyon, Université Claude-Bernard Lyon 1, 43, boulevard du 11 novembre 1918, 69100 Villeurbanne, France*
[f] *Eugène Devic EDMUS Foundation Against Multiple Sclerosis, State-Approved Foundation, 59, boulevard Pinel, 69677 Bron, France*

INFO ARTICLE

Article history:
Received 2 September 2024
Received in revised form
18 April 2025
Accepted 4 May 2025
Available online 6 June 2025

Keywords:
Multiple sclerosis
Record linkage
Clinical data
Administrative data
Observational studies

ABSTRACT

Background. – Linking disease registries to nationwide healthcare administrative databases increases the research opportunities. Recent guidelines emphasize the need of transparency in this process.
Objective. – Our aims were to describe the process of record linkage between the French multiple sclerosis (MS) cohort (OFSEP) and the national health insurance database (SNDS) and to evaluate the linkage quality.
Methods. – As no unique identifier was available in the two databases, the OFSEP-SNDS data linkage was performed by indirect matching using the following sixteen patient variables to create a unique key: sex, dates of birth and death, of visits to a neurologist, of MS-related hospitalizations, of MRI, and use of disease-modifying therapies. Three indicators were computed to assess the linkage quality.
Results. – Among the 52,034 eligible patients in the OFSEP registry, 42,603 (81.9%) were matched with patients in the SNDS database, with good overall quality (robustness = 3.19; this is the number of linkage variables that can be removed without losing the uniqueness of the linked pair; 87.8% of common information). Comparison of the linked and unlinked populations revealed no major selection bias regarding age and sex distributions.

* *Corresponding author*: École des hautes études en santé publique (EHESP), avenue du Pr Léon Bernard, CS 74312, 35043 Rennes cedex, France.
E-mail address: emmanuelle.leray@ehesp.fr (E. Leray).
https://doi.org/10.1016/j.neurol.2025.05.002

Conclusion. – The successful linkage of more than 40,000 patients with MS broadens the research perspectives by allowing access to a wide range of clinical and administrative data (e.g., comorbidities, care pathways) over a long mean disease duration (> 15 years).

1. Abbreviations

CIS	clinically isolated syndrome
CSF	cerebrospinal fluid
DMT	disease-modifying therapies
EHESP	*École des hautes études en santé publique*
ICD	international classification of diseases
LTD	long-term disease status
MOGAD	anti-myelin oligodendrocyte glycoprotein antibody-associated disease
MRI	magnetic resonance imaging
MS	multiple sclerosis
NMOSD	neuromyelitis optica spectrum disorders
OFSEP	Observatoire français de la sclérose en plaques
RIS	radiologically isolated syndrome
SNDS	*système national des données de santé*

2. Introduction

Record linkage can be defined as ''a process of pairing records from two files and trying to select the pairs that belong to the same entity''. It is increasingly used in medical and public health research [1,2] because the linkage of records from large populations across disparate sources and over time broaden the research possibilities [3].

In France, the Observatoire français de la sclérose en plaques (OFSEP; French Multiple Sclerosis Registry) promotes the prospective, standardized, high-quality, and multimodal collection of data from individuals with multiple sclerosis (MS) and related diseases [4]. Data collection focuses on the disease description (clinical, treatment, MRI and laboratory data) during routine visits and is made by neurology teams who are responsible for gathering information from all the care teams including radiologists and biologists. The French national health insurance database, called *Système national des données de santé* (SNDS), covers 98% of the French population and routinely collects information on their healthcare utilization and expenditures [5]. The SNDS database contains a comprehensive range of data and outcomes (e.g., comorbidities, care consumption by patients with MS), but not clinical data (e.g., MS phenotype, relapses, disability, MRI details and cerebrospinal fluid [CSF] results) [6–9]. Therefore, data linkage between the OFSEP and SNDS databases will allow combining their respective strengths and increasing the research scope. This linkage has two main objectives: enriching the OFSEP data with SNDS data, which are not MS-specific, and assessing the recruitment bias in OFSEP. This linkage will provide among the largest cross-sectional and longitudinal MS databases worldwide, open to the scientific community. As such, it could become a valuable resource to support research and explore unmet needs, with the aim of improving the lives of people living with MS. Linked clinical and administrative data are a powerful resource, and transparency throughout the entire linkage pathway is important to ensure that the validity of this resource is fit-for-purpose, as recently highlighted in the GUILD recommendations [1]. Therefore, the aims of the present manuscript were to describe the process of record linkage and evaluate the linkage quality.

3. Materials and methods

3.1. Data sources

In France, OFSEP is a national prospective cohort that collects clinical data from patients with MS followed at expert or tertiary MS centers [4]. In 2011, after several years of informal existence and functioning, OFSEP was formally labeled and funded for 10 years by the ''Investments for the Future'' program from the French Agency of Research. Patients are included if they receive one of the following diagnoses: MS according to the latest criteria, radiologically or clinically isolated syndrome (RIS or CIS) suggestive of MS, or MS-related condition, i.e. neuromyelitis optica spectrum disorders (NMOSD) and anti-myelin oligodendrocyte glycoprotein antibody-associated disease (MOGAD). For each patient, clinical and imaging data are retrospectively collected at the time of the first visit and then prospectively during routine follow-up visits, usually once per year. Data are recorded using the European Database on Multiple Sclerosis (EDMUS) software with a standardized language [10]. The following characteristics are available: sex, birth date, date of MS clinical onset, relapse dates and symptoms, dates of neurological visits and disability scores, disease-modifying therapy (DMT) types and dates, occurrence of severe adverse events, MRI dates and results, CSF results, date of birth of children. Each patient is identified with a record number specific to the OFSEP registry. In the case of people who have several records in the OFSEP database because they have been to several centres, a single record has been kept for linkage purposes.

The French national health insurance database [5] (SNDS) covers 98% of the French population without any age or wealth criteria, whatever the insurance scheme, with the exception of a few special insurance schemes such as those of the Senate and the National Assembly. It prospectively and exhaustively collects anonymous individual data on the reimbursement of ambulatory healthcare (e.g., consultations, drug prescriptions, medical devices, exams) and all public and private hospital activity (dates and ICD-10 codes). Each individual is identified

by a unique lifelong identifier, based on the pseudonymized *numéro d'inscription au répertoire* (NIR; number of registration in the directory) that corresponds to the individual's social security number to which a specific, irreversible, two-level pseudonymisation algorithm is applied. The following individual characteristics are available: sex, year of birth, date of death, insurance scheme (general scheme, agricultural workers, self-employed, and other schemes), and long-term disease status (LTD, which allows 100% reimbursement of disease-related care), coded using the ICD-10 codes and the starting year, if applicable.

As the OFSEP and SNDS databases do not have a common identifier, the linkage was performed by indirect matching by combining several patient variables to create a unique key (see details below in cf. Section 3.3) that allowed linking records in the two databases.

3.2. Study population

All patients with a neurological visit at an OFSEP center between January 1st, 2009 and December 31st, 2019 and who did not oppose to the linkage of their personal data with clinical-administrative data were eligible. Data were extracted in June 2022.

Data from January 1st, 2009 to December 31st, 2019 on all people with MS were extracted from the SNDS database using the following criteria: LTD status for MS (G35), OR hospitalization for MS (G35), OR MS-specific drug reimbursements (beta-interferon, dimethyl fumarate, fingolimod, glatiramer acetate, natalizumab, ocrelizumab, and teriflunomide). Off-label drugs were also collected in the database but not used to identify MS cases.

3.3. Linkage strategy

Linkage is achieved using a limited set of factors, called "linkage variables", to identify uniquely and reliably an individual across two datasets [3]. In the present case, the linkage variables were: sex, date of birth (day/month/year), date of death (month/year), dates of visits to a neurologist (maximum four visits), dates of MS-related hospitalizations (maximum two hospitalizations), dates of MRI exams (maximum four exams), and DMT use (maximum two different DMT).

When the number of available events was higher than the maximum number N retained for the study period (e.g., four visits to a neurologist), only the last N (i.e., the most recent ones) were considered. Visits to neurologists, MS-related hospital admissions, and MRI exams were assessed only in a specific geographical area because considering the whole France would have led to too many potential candidates. As OFSEP data are collected at regional MS expert centers, the geographical area was defined as the region where the patient was registered, plus the neighboring regions. Demographic data were combined in two different ways: sex + month/year of birth and sex + year of birth. The second variable was used twice, which is equivalent to put a weight of 2 on this variable.

The set of linkage variables allowed creating a 16-digit "signature" for all eligible individuals that can be seen as a unique key. Each linkage variable was represented by a letter: S for demographic data (3 digits), D for death (1 digit), V for neurological visits (4 digits), I for MRI exams (4 digits), H for hospitalizations (2 digits), M for DMT (2 digits), or "." if a variable was missing. Thus, a patient's signature was the list of known linkage variables for that patient; for instance, SSS.VVVVII.H.MM was the signature of a patient with known sex and date of birth, alive, with 4 visits, 2 MRI exams, 1 hospitalization, and 2 DMT.

3.4. Linkage quality assessment

The linkage rate was defined as the proportion of OFSEP files which were successfully linked to SNDS records (%). Three indicators were used to assess the linkage quality based on the following definitions [11]:

- robustness (R) of a linked pair was the number of linkage variables that could be removed without losing the uniqueness of the linked pair. For example, R = 1 indicates that any linking variable could have been omitted without losing the pair uniqueness, or in other words, that any variable (out of the 16 linkage variables) could have been removed from the signature without losing the uniqueness. R = 4 indicated that four OFSEP variables could be deleted without losing the linkage uniqueness. Therefore, higher R values indicate higher quality of the record linkage. By definition, the robustness was calculated only for linked pairs;
- missing variable: for a linked pair, a missing variable defined information that was available in the OFSEP registry but was not found in the SNDS database;
- percentage of common information: (number of variables known in the OFSEP database)/(number of variables known in the SNDS database) $\times$ 100. For example, if a patient had five known OFSEP variables but was linked to a patient in the SNDS database using four variables, the percentage of common information was $100 \times 4/5 = 80\%$.

These indicators were calculated for all signatures and presented separately for linked and unlinked records. After a first linkage process and calculation of the quality indicators, the linked pairs were reviewed to define the minimum linkage quality. If the number of missed linkage variables was ≥ 6 (out of 16), the risk of error was considered too high. Therefore, the pair was rejected and the records were considered as unlinked. Lastly, to assess the potential bias due to linkage failure, the characteristics of the unlinked and linked records were compared.

The linkage rate was provided for the whole OFSEP database, for each MS center, for the RIS patients, the NMOSD/MOGAD patients and for the OFSEP-HD cohort ($n = 2637$). The OFSEP-HD cohort is a well-defined subgroup of patients from the initial OFSEP cohort, called OFSEP High-Definition (https://www.ofsep.org/en/hd-cohort), who were included over the 2018–2020 period and for whom additional data are available (lifestyle factors, socioeconomic factors, quality of life, MS Functional Composite score, additional laboratory and MRI parameters). The objective of the OFSEP-HD cohort study is to identify prognostic factors of disability progression in MS in real life (registered in clinicaltrials.gov under NCT03603457).

3.5. *Standard protocol approvals, registrations, and patient consents*

Ethical and data access approvals for the study were obtained according to the current French legislation. Patients were asked to approve or not for data collection in the OFSEP database and their use for research in France and abroad (www.ofsep.org/en/cohort/ofsep-consent), including a specific item dedicated to data linkage. The OFSEP-SNDS data linkage was approved by the French data protection authority (Commission nationale de l'informatique et des libertés [CNIL]; approval DR-2021-034) to be performed on medical records of patients who did not oppose to it. Data were stored in a secure server at *Eskemm Numérique* (Rennes, France) that guarantees all the security requirements, and only linkage variables were provided. The record linkage was performed using the CUBR-LINK tool [11], under the EHESP (French school of public health) supervision, in collaboration with the OFSEP coordination team. As linkage is not straightforward, a multidisciplinary team with medical, methodological and technical knowledge and skills was required to handle the whole process [12].

3.6. *Data availability*

According to the data protection and French regulation, the authors cannot publicly release SNDS data. However, data reuse can be granted upon request to the OFSEP Scientific Board and after validation by the OFSEP Steering Committee (https://www.ofsep.org/en/data-access). Moreover, data cannot be moved from the secure server located at *Eskemm Numérique* (Rennes, France).

4. Results

Overall, 52,034 patients from the OFSEP database met the inclusion criteria and 167,300 records were extracted from the SNDS database. The age and sex distributions of the two populations (Fig. 1) were almost similar, but older people (year of birth < 1930) were more numerous in the SNDS dataset.

Table 1 – Results of the OFSEP-SNDS data linkage.

Number of eligible OFSEP patients	52,034
Number of eligible SNDS patients	167,300
Ratio of eligible SNDS/OFSEP patients	3.21
Number of patients with successful linkage (%)	42,603 (81.9%)
Overall robustness of the linkage, mean [min–max]	3.19 [0–10]
Percentage of common information	87.8%
Number of patients according to the number of missed variables (out of 16 linkage variables and 42,603 linked records)	
0 (perfect match)	17,551 (41.2%)
1	11,813 (27.7%)
2	6,487 (15.2%)
3	5,652 (13.3%)
4 or 5	1,100 (2.6%)

OFSEP: Observatoire français de la sclérose en plaques (French Multiple Sclerosis registry); SNDS: *Système national des données de santé* (French health insurance database).

The OFSEP-SNDS data linkage was successful for 42,603 patients (i.e., linkage rate 81.9% of the eligible OFSEP population) (Table 1). The R robustness value was 3.19, meaning on average that it would have been possible to remove three linkage variables without losing the linked pair. The percentage of common information in the two datasets was 87.8%. After a first data linkage, the quality assessment led to the exclusion of 816 pairs because the number of missed linkage variables was ≥ 6, strongly reducing the linkage quality confidence. Mean MS duration (from clinical onset to last visit, except for people with RIS where it was computed from index MRI) was 15.7 ± 11.0 years.

This high overall linkage rate of 81.9% concealed disparities, because this rate ranged between 31.1 and 96.2% across the OFSEP centers (median 83.5%; 1st quartile 77.7%; 3rd quartile 87.4%). It increased to 96.9% (2554/2637) for the OFSEP-HD cohort but was lower for patients with other conditions (53.7% [259/482] in the RIS subgroup and 45.3% [678/1497] in the NMOSD/MOGAD subgroup).

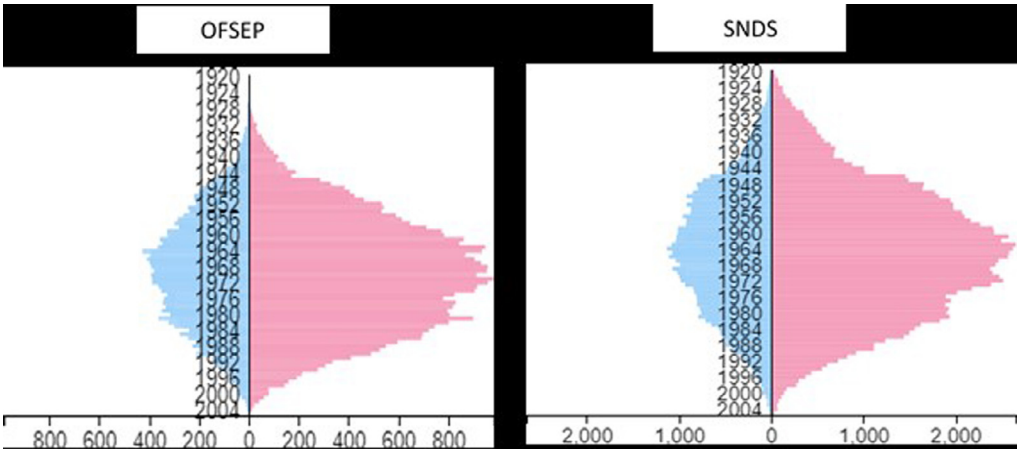

Fig. 1 – Age pyramid of the two populations to be linked. Footnote: The figure shows the number of people with MS according to age and sex in the two datasets (OFSEP and SNDS). There are more elderly people in the SNDS, which is not surprising since older people with MS are probably seen less in OFSEP centres.

Table 2 – The most frequent signatures leading to linked or unlinked records and associated indicators.

Signature	Number of OFSEP patients	Robustness (R)	Number of SNDS candidates	Number of variables in the signature	Number of missing variables in the SNDS candidates	% of shared information
Linked records						
SSS.VVVIIII..MM	4767	5.37	1	13	1.5	88
SSS.VVVVIIII..M.	3013	5.17	1	12	1.6	86
SSS.VVVVIII. . .MM	1070	4.42	1	12	1.6	86
SSS.VVVII. . ..M.	1054	3.51	1	10	1.3	87
SSS.VVVV. . .. . .M.	1029	2.04	1	8	0.8	90
SSS.VVVVIIIIHHMM	1020	6.91	1	15	1.8	88
SSS.VVVVIII. . .M.	1011	4.34	1	11	1.4	87
SSS.VVVVI. . ...M.	992	2.61	1	9	1.2	86
SSS.VVVVII. . ..MM	956	3.51	1	11	1.5	86
SSS.VVVVIIII. . ..	943	4.44	1	11	2.0	81
Unlinked records						
SSS.V.	989	NA	4.3	4	0.8	80
SSS. . .. .M.	814	NA	7.2	4	0	100
SSS.	635	NA	8.0	3	3.0	0
SSS.V. . .. . .M.	498	NA	5.0	5	0.9	82
SSS. . ...I.	449	NA	4.1	4	0.8	80
SSS. . .. .MM	341	NA	5.5	5	0	100
SSS.V.	329	NA	2.9	5	1.7	66
SSS.VV.	254	NA	3.7	5	1.7	66
SSS. . ...II.	200	NA	2.9	5	1.6	68
SSS.VVVVIIII. . ..	191	NA	3.4	11	7.3	33

Each linkage variable was represented by a letter: S: demographic data (3 digits); D: death (1 digit); V: neurological visits (4 digits); I: MRI exams (4 digits); H: hospitalizations (2 digits); M: DMT (2 digits);.: if a variable was missing. OFSEP: Observatoire français de la sclérose en plaques (French Multiple Sclerosis registry); SNDS: *Système national des données de santé* (French health insurance database).

Table 2 shows the ten most frequent signatures that resulted in a linked or unlinked pair and also some descriptive statistics and quality indicators. For pairs that were successfully linked, the number of SNDS candidates for one OFSEP patient was always 1. The number of OFSEP variables was higher in the linked records (most often $\geq$ 10) as well as the R value. Conversely, the number of variables not found in the SNDS database was limited (1 or 2), leading to a high level of confidence in the linkage. For the unlinked records, the number of OFSEP variables was low, meaning that the amount of data that could be used for the linkage was limited. For some signatures, the percentage of common data was high (up to 100%), but they were not sufficiently discriminating, leading to several potential candidates and therefore to the linkage failure.

Age and sex were not significantly different between linked and unlinked records (Table 3).

5. Discussion

The OFSEP-SNDS data linkage was successful for 81.9% of patients with MS or related diseases. This means that the clinical and administrative data of 42,603 patients with MS or MS-related conditions for the period 2009–2019 are now available for research. Due to the lack of a common unique identifier between the OFSEP and SNDS databases, an indirect record linkage was performed using a set of variables that were available in both data sources. The absence of a common identifier is linked to regulatory aspects. When it was set up, the OFSEP team did not use or ask for the NIR to be registered, as this would have involved more complex, cumbersome and restrictive procedures, and would not necessarily have been approved by the regulatory body. However, this affects the linkage possibilities and performance, unlike in other countries where the same unique number is available in all

Table 3 – Comparison of the demographic characteristics between linked and unlinked patients.

	Linked records n = 42,603	Unlinked records n = 9431
Women, n (%)	30,096 (70.6%)	6,613 (70.1%)
Men, n (%)	12,507 (29.4%)	2,818 (29.9%)
Age (years), mean ± SD (min–max)		
Women	49 ± 13.7 (12–96)	49 ± 18.1 (5–98)
Men	50 ± 13.8 (12–95)	50 ± 15.5 (2–93)

OFSEP: Observatoire français de la sclérose en plaques (French Multiple Sclerosis registry); SNDS: *Système national des données de santé* (French health insurance database)

databases and registers (for instance, Denmark, Sweden, Canada). For this reason, a set of 16 linkage variables was defined, and as expected, the linkage was more often successful when a high number of variables was available. Sensitivity analyses were carried out by increasing the number of linkage variables, for example by increasing the number of visits from 4 to 5 or 6. This did not increase the linkage rate, but only the confidence level of pairs that were already linked with a high degree of confidence. For some patients, very little information was available, besides sex and birth date, and the linkage was difficult. Indeed, the linkage completeness rate is influenced by the data accuracy and completeness in each data source [2]. In agreement, the linkage rate of the OFSEP-HD cohort was 96.9%. This high success rate was related to the fact that this sub-cohort includes patients with MS who are regularly followed at MS expert centers and 80.5% of them were treated at baseline. Now, 2554 of these patients have been linked to the SNDS and their detailed data will give the opportunity to examine (or adjust for) lifestyle factors, comorbidity, quality of life, individual and neighborhood socioeconomic indicators, healthcare utilization, in addition to clinical, MRI and laboratory data, and to determine their contribution to the overall disease prognosis.

Conversely, the linkage rates of the RIS and NMOSD/MOGAD subgroups were only 53.7% and 45.3%, respectively. These low rates could be explained by a ''bias'' in the SNDS target population. Indeed, record extraction focused on classical MS; however, patients with RIS are at risk of future demyelinating events, but will not necessarily receive a diagnosis of MS, and patients with NMOSD/MOGAD have a disease close to MS but that may be recorded using other ICD-10 codes (for instance G36 Other acute disseminated demyelination).

Before the linkage described in this study, in June 2021, a first data linkage process was performed using 47,256 patients from the OFSEP cohort with better linkage performances: 87.9% of success (i.e., 41,545 linked records and 5711 unlinked records), 89.1% of common information and R of 3.24. The decreased linkage rate (81.9%) was due to the inclusion of recent patients with fewer available data due to the shorter follow-up duration. However, the choice was made to keep the most recent and the larger dataset, although the results were slightly poorer. This once again underlines the need of reliable data in sufficient quantity for indirect matching. The CUBRLINK tool has been used in other contexts (stroke, cancer) and has provided excellent results based upon rich and well-defined datasets, with linkage rates of over 95% [11].

Although the overall linkage rate was good (81.9%; median among centers 83.5%), discrepancies were observed among geographical regions. Several hypotheses can be proposed. This may reflect inconsistent coding practices or varying degree of accuracy and completeness in the two data sources, although there are guidelines on OFSEP data collection [4] and SNDS data entry [5]. This may also indicate different recruitment practices among OFSEP centers, such as size or width of the geographic catchment area, level of MS diagnosis certainty, time since center opening.

The main limitation of our approach is the absence of a gold standard to evaluate the quality of the linkage results [13]. Indeed, we defined rules in the process and computed some indicators, but we were not able to measure the linkage error. Such error corresponds to either false matches (when records from different individuals link erroneously) or missed matches (where records from the same individual fail to link). However, the comparison of the linked and unlinked records also contributes to the assessment of the quality, and we did not identify major selection bias regarding age and sex distributions.

As mentioned in the Introduction, the OFSEP-SNDS data linkage had two main objectives: data enrichment and analysis of recruitment bias in the OFSEP cohort. Administrative data have several important strengths, particularly they are free from many of the measurement and loss-to-follow-up problems associated with longitudinal surveys [3]. Moreover, the SNDS database includes information widely and diligently collected for other purposes, with a quasi-exhaustive coverage of the French general population. The comparison of patients in the OFSEP cohort with patients with MS in the SNDS database will allow assessing the recruitment bias in OFSEP, and then better characterizing the scope of the results obtained using OFSEP data alone or linked OFSEP-SNDS data. The recruitment bias will be assessed by comparing patients seen and not seen within the OFSEP network, focusing particularly on the following variables from the SNDS database: age, sex, region, type of neurological follow-up (e.g., university hospital, general hospital, private practice, mixed, comorbidities, visit frequency, hospitalizations, MRI exams, treatments). We hypothesize that patients in the OFSEP cohort have more active and/or complex disease because recruitment in this cohort mainly relies on expert or tertiary centers, located in university hospitals.

Making the linked OFSEP-SNDS data available for academic research will broaden research opportunities. Indeed, the main strengths of the linked dataset are the combination of disease-specific clinical, imaging and treatment data, which rely on neurological expertise, with the large data collection related to use of healthcare services, the number of patients ($>$ 40,000), the length of the study period ($>$ 10 years) and the follow-up duration ($>$ 15 years). Data enrichment will facilitate the detection of comorbidities that may be insufficiently collected (incomplete, irregular, or neurologist-dependent) in MS registries. Using the linked OFSEP-SNDS dataset, a project will be launched soon to evaluate the impact of selected comorbidities (i.e., cancer, autoimmune, cardiovascular and psychiatric diseases) on MS prognosis and management (DMT use, effectiveness and safety, overall healthcare utilization) in France. Through this linkage, data on healthcare use for and outside MS are now available. It will be possible to describe the care pathways, overall and in the presence of (specific) comorbidities.

A study based on the linked dataset that focused on the sub-population of women with MS who underwent In Vitro Fertilization (IVF) was recently published [14]. Our group previously showed that the risk of relapse is not increased after IVF, based on the analysis of 225 women with IVF cycles identified in the SNDS database [8]. In this previous study, the relapse outcome was identified through healthcare utilization and may not be as accurate as through neurological expertise. Therefore, the question was then reassessed using the linked

OFSEP-SNDS dataset, with data on relapses and DMT coming from the OFSEP cohort and data on IVF, stimulation protocols and pregnancies coming from the SNDS database. All the results were consistent and reassuring as they confirmed the absence of risk of relapse after IVF.

The linked dataset was also used to examine healthcare usage in the three years before the first scan in RIS cases from 2010 to 2019 in comparison to the general population and to MS patients [15].

In accordance with the French legislation, the obtained approval and the GUILD recommendations [1], all studies using the linked dataset are (and will be) performed in a secure environment. Each study will have its own silo, and only the needed variables and records will be made available to the investigators. Specific anonymized identifiers will be created for each study. An application for a new record linkage using the same process has been submitted and recently approved, in order to update the data and study populations (e.g., longer period and longer follow-up up to 2023, inclusion of additional ICD-10 diagnoses for MS-related conditions). During the reviewing process of the present publication, we had the opportunity to carry out the linkage of the new dataset. Of the 66,312 eligible OFSEP patients over the period 2009–2023, 55,876 were successfully linked to the SNDS, representing a linkage rate of 84.3% (OFSEP-HD: 98.9%; RIS: 65.0%; NMOSD/MOGAD: 67.8%). The R robustness value was 3.62 and the percentage of common information in the two datasets was 91.6%. The indicators have all improved, demonstrating the relevance of the linkage tool and the feasibility of regular OFSEP-SNDS linkage processes in the future.

To conclude, the present study was made in accordance with the GUILD recommendations [1] to highlight the choices and decisions made during data processing that may affect linkage error and hence the results of analyses using these data. Sharing information along the data linkage pathway could improve the transparency and reproducibility of research, promote the use of improved methods to address linkage errors, and improve the interpretation of studies based on linked data. This manuscript offers the opportunity to increase transparency on the linkage process in order to facilitate communication among the research teams who are using or will use the linked OFSEP-SNDS dataset.

Funding

This work was conducted using data from the Observatoire français de la sclérose en plaques (OFSEP) that is supported by a grant by the French State and handled by the ''Agence nationale de la recherche'', within the framework of the ''Investments for the Future'' program, under the reference ANR-10-COHO-002, by the Eugène Devic EDMUS Foundation against multiple sclerosis, and by the ARSEP Foundation. The funding agencies had no role in the study design and execution, including the data collection, management, analysis and interpretation; the preparation, review or approval of the manuscript; or decision to submit the manuscript for publication.

SNDS data were made available to OFSEP by CNAM (French National Health Insurance Fund), and the indirect matching between OFSEP and SNDS data was carried out by EHESP and CUBR.

Disclosure of interest

EL received personal compensation for consulting, serving on a scientific advisory board, speaking, or other activities with Alexion, Biogen, Merck, Novartis, Roche, Sanofi-Genzyme; nothing related to the contents of the present work.

E.D. and R.C. declare that they have no competing interest.

SV received non-personal consulting and lecturing fees, travel grants and unconditional research support from Biogen, Janssen, Merck, Novartis, Roche, Sandoz and Sanofi.

REFERENCES

[1] Gilbert R, Lafferty R, Hagger-Johnson G, Harron K, Zhang L-C, Smith P, et al. GUILD: guidance for information about linking data sets. J Public Health (Oxf) 2018;40:191.
[2] Bohensky MA, Jolley D, Sundararajan V, Evans S, Pilcher DV, Scott I, et al. Data linkage: a powerful research tool with potential problems. BMC Health Serv Res 2010;10:346. http://dx.doi.org/10.1186/1472-6963-10-346.
[3] Jutte DP, Roos LL, Brownell MD. Administrative record linkage as a tool for public Health Research. Annu Rev Public Health 2011;32:91–108.
[4] Vukusic S, Casey R, Rollot F, Brochet B, Pelletier J, Laplaud DA, et al. Observatoire français de la sclérose en plaques (OFSEP): a unique multimodal nationwide MS registry in France. Mult Scler 2020;26(1):118–22. http://dx.doi.org/10.1177/1352458518815602.
[5] Tuppin P, Rudant J, Constantinou P, Gastaldi-Ménager C, Rachas A, de Roquefeuil L, et al. Value of a national administrative database to guide public decisions: from the système national d'information interrégimes de l'Assurance Maladie (SNIIRAM) to the système national des données de santé (SNDS) in France. Rev Epidemiol Sante Publique 2017;65:S149–67.
[6] Pierret C, Mainguy M, Leray E. Prevalence of multiple sclerosis in France in 2021: data from the French health insurance database. Rev Neurol (Paris) 2024;180(5):429–37. http://dx.doi.org/10.1016/j.neurol.2023.12.007 [S0035-3787(24)00369-2].
[7] Moisset X, Leray E, Chenaf C, Taithe F, Vukusic S, Mulliez A, et al. Risk of relapse after COVID-19 vaccination among patients with multiple sclerosis in France: a self-controlled case series. Neurology 2024;103(5):e209662. http://dx.doi.org/10.1212/WNL.0000000000209662.
[8] Mainguy M, Tillaut H, Degremont A, Le Page E, Mainguy C, Duros S, et al. Assessing the risk of relapse requiring corticosteroids after in vitro fertilization in women with multiple sclerosis. Neurology 2022;99(17):e1916–25. http://dx.doi.org/10.1212/WNL.0000000000201027.
[9] Roux J, Guilleux A, Lefort M, Leray E. Use of health care services from patients with multiple sclerosis in France over 2010–2015: a nationwide population-based study using health administrative data. Mult Scler J Exp Transl Clin 2019;5(4). http://dx.doi.org/10.1177/2055217319896090 [2055217319896090].
[10] Confavreux C, Compston DA, Hommes OR, et al. EDMUS, a European database for multiple sclerosis. J Neurol Neurosurg Psychiatry 1992;55:671–6.

[11] Drezen E, Happe A, Kerbrat S, Balusson F, Oger E. New metrics for assessing linkage quality in deterministic record linkage of health databases. HAL 2022, https://hal.science/hal-03601245.
[12] Scailteux LM, Droitcourt C, Balusson F, Nowak E, Kerbrat S, et al. French administrative health care database (SNDS). The value of its enrichment. Therapie 2019;74(2):215–23. http://dx.doi.org/10.1016/j.therap.2018.09.072.
[13] Harron K, Dibben C, Boyd J, Hjern A, Azimaee M, Barreto ML, et al. Challenges in administrative data linkage for research. Big Data Soc 2017;4 [2053951717745678].
[14] Mainguy M, Casey R, Vukusic S, Lebrun-Frenay C, Berger E, Kerbrat A, et al. Assessing the risk of relapse after in vitro fertilization in women with multiple sclerosis. Neurol Neuroimmunol Neuroinflamm 2025;12(2):e200371. http://dx.doi.org/10.1212/NXI.0000000000200371.
[15] Lebrun-Frenay C, Kerbrat S, Okuda DT, Landes-Chateau C, Kantarci OH, Pierret C, et al. Analysis of healthcare utilization before the diagnosis of radiologically isolated syndrome. Mult Scler 2025;31(2):184–96. http://dx.doi.org/10.1177/13524585241291471.