Università
degli Studi
di Ferrara

Université
de Limoges

ISTITUTO
ITALIANO DI
TECNOLOGIA

Doctoral Course in
Translational Neurosciences and Neurotechnologies

Co-tutelle thesis with Université de Limoges (France)

in agreement with
Istituto Italiano di Tecnologia - IIT

CYCLE XXXII

COORDINATOR Prof. FADIGA Luciano

Modern Approaches to Data Collections in Neuroepidemiology

Scientific/Disciplinary Sector (SDS) MED 26 / 06d6

Candidate
Dott. BALDIN Elisa

____
(signature)

Supervisors
Prof. PUGLIATTI Maura

____
(signature)

Prof. PREUX Pierre-Marie

____
(signature)

Year 2016/2019

TABLE OF CONTENTS

1. BACKGROUND

Epidemiology studies the frequency, distribution, and determinants of health- and disease-related states and outcomes on a population or group level (Rothman KJ, Greenland S et al. 2008)

1.1 From conventional survey methodology to medical records linkage

Clinical epidemiological research is greatly interested in the definition of etiology of complex diseases where susceptibility genes, environmental factors, constitutional and behavioral factors interact to determine the pathological outcome. Moreover, all these factors can unfold over the entire life course and their effect could have an impact in any phase (in utero, at birth -perinatal phase, in childhood-adolescence, in adult life or during aging).

Epidemiological studies could be divided in descriptive studies, observational and experimental studies.

Descriptive studies are applied to populations. The objective is mostly to determine the frequencies of factors and diseases in the population under study and characterize the distributions in time, place and person of these factors. The methodological approach that best represents this type of study is the population survey, collecting data from a sample of the population to measure prevalence, incidence, identify cases and exposures, study familial aggregations, screen candidates for preventions or early treatments.

Observational studies, cohort and case-control studies, are focused on individuals. The objective is to test causal hypothesis, which is the same aim of experimental studies, i.e. clinical trials. Even though the experimental designs are the ideal option to move as close as possible to the truth of nature, observational studies, and among these longitudinal studies, are the preferred design in order to have long follow-up from prospective or historical cohorts of individuals that were disease-free at recording of exposure information. (Rothman KJ, Greenland S et al. 2008)

The ideal observational study, considering the recruitment of sufficient number of subjects, their assessments, the follow-up long enough to collect all the relevant information, is usually costly, in terms of finances but also time and personnel involved. Therefore, the use of already existing collection of data is a fundamental cost-effective alternative.

The first element to consider for the quality of a study is defining the population, in terms of geographical limits, time periods and sampling method. The most desirable is the population-based sample.

Another fundamental aspect is the definition of the disease under study, describing it as precisely as possible. This can limit for instance the risk of definition changes over time (e.g. definition of dementia changed from DSM III in 1983 vs. DSM V in 2013).

The knowledge of the characteristic of the disease helps characterizing the different possible conditions of the diseased subject so to identify the desired population. In fact, the disease could be always diagnosed, thus recognized and recorded, but in some cases the diagnosis could have been missed, despite symptoms or with signs. Moreover some subjects could be diseased but asymptomatic.

Based on its specific characteristics a condition could be defined in relation to the different medical contacts required and therefore the information that can be collected. Some chronic diseases may require outpatient visits only, or drug prescriptions or periodical laboratory tests. In other situation the hospitalization might be mandatory. The disease could be identified as cause of death on death certificates, or it could need a pathological confirmation to have a diagnosis. (St Germaine-Smith, Metcalfe et al. 2012)

Therefore, based on the specific characteristics of the disease we are interested in, the method for ascertainment needs to be different.

Whenever the disease determines a contact with the health system, it will be recorded and that information can be used. This can be done though the access to the records of hospitals and other institutions, or combining information from hospitals with those from specialists, general practitioners and drug use records.

One important instrument in this case is the use of medical records-linkage system.

Record linkage systems match data across administrative health databases and other information-rich repositories thus supplying very large samples for long-term observational studies. (Jutte, Roos et al. 2011) In contexts where a universal healthcare coverage exists, they provide information on the entire population. (OECD 2013)

Examples of this valuable approach are the Rochester Epidemiology Project (REP) (St Sauver, Grossardt et al. 2012) or collaborations across Nordic countries (Denmark, Finland, Iceland, Norway and Sweden). (Kieler, Artama et al. 2012)

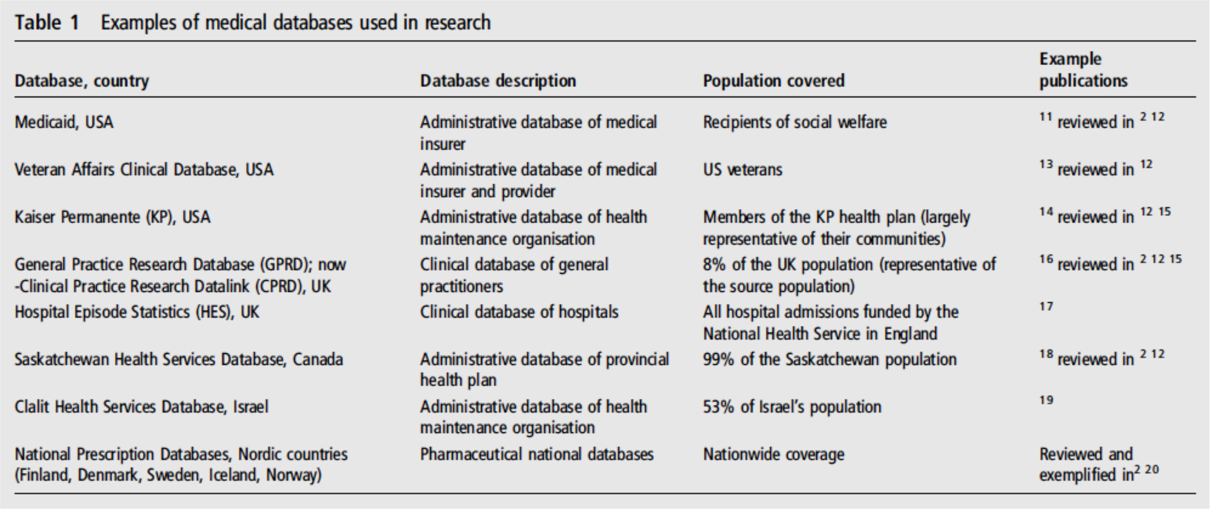


Table 1 Examples of medical databases used in research

Database, country	Database description	Population covered	Example publications
Medicaid, USA	Administrative database of medical insurer	Recipients of social welfare	[11] reviewed in [2,12]
Veteran Affairs Clinical Database, USA	Administrative database of medical insurer and provider	US veterans	[13] reviewed in [12]
Kaiser Permanente (KP), USA	Administrative database of health maintenance organisation	Members of the KP health plan (largely representative of their communities)	[14] reviewed in [12,15]
General Practice Research Database (GPRD); now -Clinical Practice Research Datalink (CPRD), UK	Clinical database of general practitioners	8% of the UK population (representative of the source population)	[16] reviewed in [2,12,15]
Hospital Episode Statistics (HES), UK	Clinical database of hospitals	All hospital admissions funded by the National Health Service in England	[17]
Saskatchewan Health Services Database, Canada	Administrative database of provincial health plan	99% of the Saskatchewan population	[18] reviewed in [2,12]
Clalit Health Services Database, Israel	Administrative database of health maintenance organisation	53% of Israel's population	[19]
National Prescription Databases, Nordic countries (Finland, Denmark, Sweden, Iceland, Norway)	Pharmaceutical national databases	Nationwide coverage	Reviewed and exemplified in[2,20]

From: Gavrielov-Yusim N, Friger M. J Epidemiol Community Health 2014;68:283-287.
With Permission from BMJ Publishing Group Ltd.

In the few last decades, the mining of these electronic healthcare databases or other medical records collections has been increasingly applied, through different systems, from the use of ICD codes to the implementation of artificial intelligence and machine learning systems. (Davenport and Kalakota 2019)

1.2 Healthcare databases

Administrative databases, possibly linked to other data collections, as disease registries and patients' surveys, can facilitate the epidemiological research on clinical questions, otherwise more complicated if not impossible sometimes to study with conventional techniques. (Johnson and Nelson 2013)

The use of administrative databases has been implemented in epidemiological fields, especially in pharmacoepidemiology, to study efficacy-safety of treatment interventions and detection of adverse drug events. The gold standard in this case would be randomized clinical trials (RTCs), however observational studies can be used as complementary. (Takahashi, Nishida et al. 2012) Moreover the availability of long follow-up, larger sample and the presence of subjects with detectable comorbidities in administrative databases, provide fundamental data for the evaluation of safety of new drugs. (Takahashi, Nishida et al. 2012, Gavrielov-Yusim and Friger 2014)

Another important application is in drug repurposing. The use of administrative data can be considered almost as ‘circular research’. It can be used as validation substrate for laboratory findings (Mittal, Bjornevik et al. 2017), thus allowing for other more targeted clinical studies, or, based on new and specific associations between exposures and outcomes (Xu, Aldrich et al. 2014, Peter, Dubinsky et al. 2018, Cai, Zhang et al. 2019, Foltynie 2019), it can suggest further evaluations of the mechanism of action and pathophysiological paths, generating hypothesis and thus new research paths.

Administrative databases include data on hospitalization, surgical or other procedures, outpatient clinic billing data; vital statistics data (e.g. births, deaths) (Jette, Atwood et al. 2013), drug prescriptions, long term care services and admissions; other health support services or billing exemption and other data collected for administrative purposes.

Almost every contact of the inhabitants with the health system of their country is collected in these administrative databases.(Gavrielov-Yusim and Friger 2014) The data are generally electronically recorded prospectively as a routine procedure, thus large populations of individuals (with and without a disease) are followed for long periods.

In countries where universal health coverage is available, these healthcare databases represent a valuable source for population-based epidemiological studies (Mazzali and Duca 2015), also in case of rare events and disease.

Some characteristics of neurological diseases can add challenges to the epidemiological study in this field. Many neurological diseases are relatively rare, the diagnosis of the disease may need an expert, or definite diagnosis may require postmortem examination. In many cases the precise onset can be uncertain since symptoms or signs can develop subtly and therefore often there is a long latency between onset and the diagnosis.

The different databases, administrative or disease specific registries and surveys, need to be linked in order to be analyze. Most linkage methods are based on the use of a unique identifier across databases, through a deterministic method, and the unique identifier varies across countries.

The availability of great amount of healthcare data is not sufficient for answering a research question. The knowledge of expert in the field for understanding the processes in data collection and the use of accurate measurements are fundamental aspects to consider in order to obtain reliable findings. (Ehrenstein, Nielsen et al. 2017)

1.3 Mandatory National registries

A patient registry is defined as "an organized system that uses observational study methods to collect uniform data (clinical and other) to evaluate specified outcomes for a population defined by a particular disease, condition, or exposure, and that serves a predetermined scientific, clinical, or policy purpose(s)" (2010).

Patient disease registries collect and organize information on a defined disease for a population of patients. They can be clinical-based or population-based and can thanks to the origin of the data from clinical practice, they can add important information, especially in rare diseases. These registries can be completed with additional data through the linkage with other databases. (Registries and EMA 2018)

The origin of registries can be traced back to 1856, in Norway, where the first national Leprosy Registry was established (IRGENS and BJERKEDAL 1973, Irgens 2012). In 1943 the Danish Cancer Registry was established as well (Gjerstorff 2011) and other registries in Nordic countries were started before 2000.

In Nordic countries the potential of data collection has been used by epidemiologist through the linkage of registries data with administrative and health related databases. (Sorensen 1997)

Several networks collaborations have so far been established, across Nordic countries (Kieler, Artama et al. 2012), but also across different European countries or multinational, thus providing large data collection (big data) with the advantage of improving precision of estimates, allowing analysis in subgroups of subjects usually less represented.

2. AIMS OF THE WORK

Big data from routinely recorded health care information is increasingly used as powerful instrument in clinical epidemiology research, with linkage of different datasets within a single country or on a multinational level. (Ehrenstein, Nielsen et al. 2017)

The present project is based on the need for the application of modern methodological approaches to the study of epidemiology of neurological diseases.

Most neurological diseases are rare (i.e. multiple sclerosis, amyotrophic lateral sclerosis) and it is often difficult to identify a clear onset.

The study of risk and protective factors requires sufficient sample size and thus the use of administrative data, registries and already collected data is of great importance.

However, the accuracy of the data used need to be validated in order to ensure the quality of the findings.

The general objective is to describe and discuss different approaches applied to population-based health data collections for epidemiological studies of neurological diseases. Specifically, to highlight the potential of the different types of study in answering to clinical questions, as well as the limits encountered and how to overcome them.

These issues have been the focus of specific articles:

I. To investigate the association between exposure to breastfeeding and the occurrence of MS in adulthood using prospectively collected community-based data on the exposure of interest.

II. To assess the effect of antibiotic exposure on the risk of developing MS in the Emilia-Romagna region (RER), Italy.

III. To assess the accuracy of Death Certificates in the identification of subjects with Amyotrophic Lateral Sclerosis in the Limousin region, France, through the validation of these administrative data using the FRALim registry as gold standard. Moreover, to identify whether specific factors are associated with the accuracy of the Death Certificates.

3. MATERIALS AND METHODS

The specific methodology of each study is presented.

3.1 Study I

To investigate the association between exposure to breastfeeding and the occurrence of MS in adulthood using prospectively collected community-based data on the exposure of interest.

Previously collected data from a set of Norwegian community-based surveys were accessed. These surveys are part of the Cohorts of Norway consortium (CONOR) (Naess, Sogaard et al. 2008) and were carried out during the period 1994-2002.

CONOR allowed us to identify mothers and link them to their offspring recorded in the Medical Birth Registry of Norway (MBRN) at the Norwegian Institute of Public Health.

Moreover offspring information was further linked with the Norwegian MS Registry and Biobank, a systematic collection of clinical and epidemiological data, as well as biological samples of subjects with a diagnosis of MS. (Myhr, Grytten et al. 2015)

3.1.1 Surveys- Cohort of Norway (CONOR)

The CONOR cohort has been established 1994 as a multipurpose study. (Naess, Sogaard et al. 2008, Aamodt, Søgaard et al. 2010) The first survey to contribute with data and blood samples to the CONOR was the Tromsø Study in 1994–95. In total 10 different health surveys were carried out in the period 1994-2003, in several areas of Norway.

Overall 309,742 Norwegians were invited to participate to the surveys, and 173,236 individuals gave their written consent to be included.

The CONOR collaboration is currently a research collaboration between the Norwegian Institute of Public Health and the Universities of Bergen, Oslo, Tromsø and Trondheim.

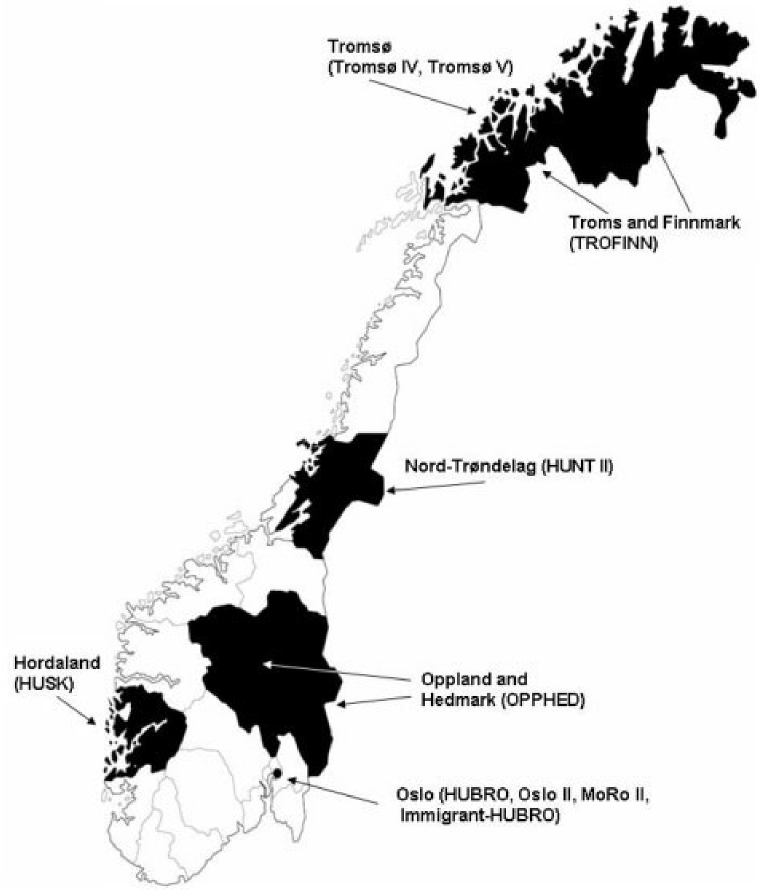


Figure 1 Map of Norwegian counties with location of each sub-study included in cohort of Norway (CONOR)

From: Næss O. et al. Cohort Profile: Cohort of Norway (CONOR), International Journal of Epidemiology; 2008: 37(3), 481–485. With permission from Oxford University press

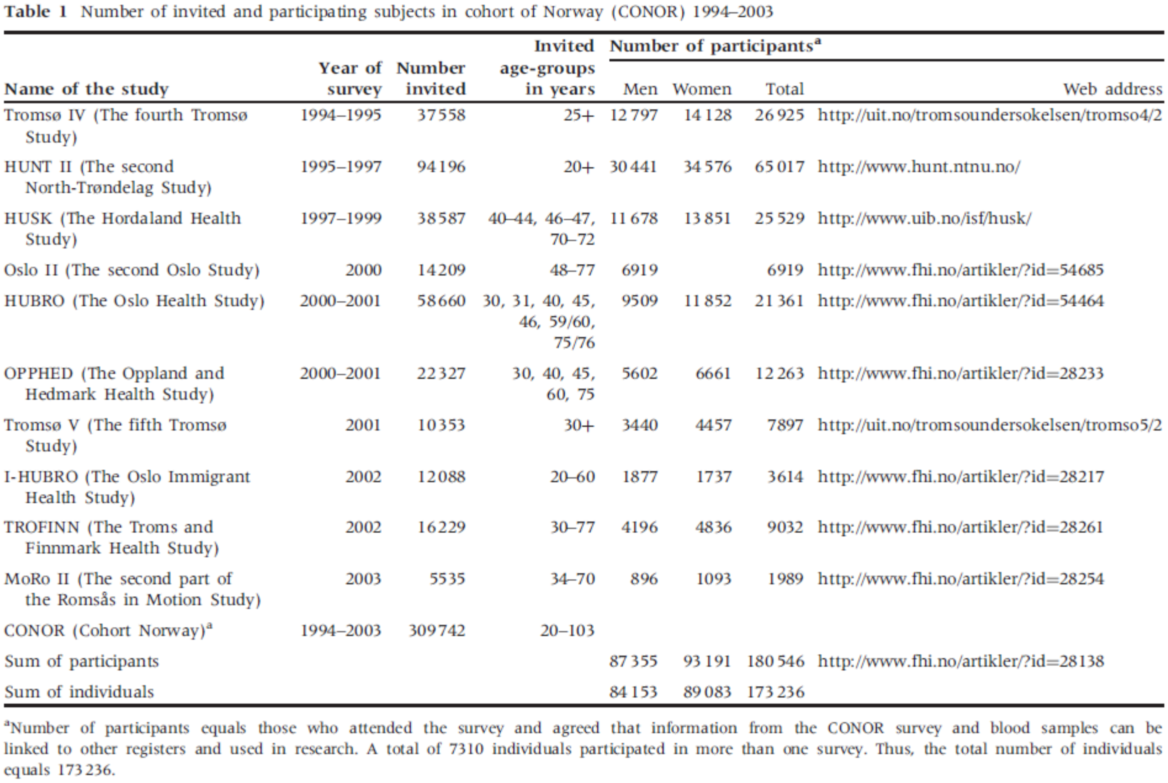


Table 1 Number of invited and participating subjects in cohort of Norway (CONOR) 1994–2003

Name of the study	Year of survey	Number invited	Invited age-groups in years	Number of participants[a]			
				Men	Women	Total	Web address
Tromsø IV (The fourth Tromsø Study)	1994–1995	37 558	25+	12 797	14 128	26 925	http://uit.no/tromsoundersokelsen/tromso4/2
HUNT II (The second North-Trøndelag Study)	1995–1997	94 196	20+	30 441	34 576	65 017	http://www.hunt.ntnu.no/
HUSK (The Hordaland Health Study)	1997–1999	38 587	40–44, 46–47, 70–72	11 678	13 851	25 529	http://www.uib.no/isf/husk/
Oslo II (The second Oslo Study)	2000	14 209	48–77	6919		6919	http://www.fhi.no/artikler/?id=54685
HUBRO (The Oslo Health Study)	2000–2001	58 660	30, 31, 40, 45, 46, 59/60, 75/76	9509	11 852	21 361	http://www.fhi.no/artikler/?id=54464
OPPHED (The Oppland and Hedmark Health Study)	2000–2001	22 327	30, 40, 45, 60, 75	5602	6661	12 263	http://www.fhi.no/artikler/?id=28233
Tromsø V (The fifth Tromsø Study)	2001	10 353	30+	3440	4457	7897	http://uit.no/tromsoundersokelsen/tromso5/2
I-HUBRO (The Oslo Immigrant Health Study)	2002	12 088	20–60	1877	1737	3614	http://www.fhi.no/artikler/?id=28217
TROFINN (The Troms and Finnmark Health Study)	2002	16 229	30–77	4196	4836	9032	http://www.fhi.no/artikler/?id=28261
MoRo II (The second part of the Romsås in Motion Study)	2003	5535	34–70	896	1093	1989	http://www.fhi.no/artikler/?id=28254
CONOR (Cohort Norway)[a]	1994–2003	309 742	20–103				
Sum of participants				87 355	93 191	180 546	http://www.fhi.no/artikler/?id=28138
Sum of individuals				84 153	89 083	173 236	

[a] Number of participants equals those who attended the survey and agreed that information from the CONOR survey and blood samples can be linked to other registers and used in research. A total of 7310 individuals participated in more than one survey. Thus, the total number of individuals equals 173 236.

From: Næss O. et al. Cohort Profile: Cohort of Norway (CONOR), International Journal of Epidemiology; 2008: 37(3), 481–485. With permission from Oxford University press

All participants, regardless the specific survey, underwent a physical examination and blood sample withdrawal in addition to a questionnaire.

The 50 core questions present in each CONOR survey are on the following topics: self-reported health and diseases, family history of disease, risk factors and lifestyle, social network and social support, education, work and housing, occupation, use of medication and reproductive health (for women). Every survey has additional information collected.

These surveys include subjects from different areas of Norway, rural and urban; through the use of the 11-digit personal identification number they can be linked to other administrative databases or national health databases (e.g. Medical birth registry of Norway (MBRN), census data).

The overall participation however is low, reporting 58% of those contacted (lowest in Oslo and in subjects ≤30 years of age), thus the actual representativeness cannot be fully assumed, given the risk of bias.

3.1.2 The Medical Birth Registry of Norway

Scandinavian countries are among the few places in the world in which all births can be systematically linked. And Norway was the first Scandinavian country to apply this methodology.

The Medical Birth Registry of Norway (MBRN), initiated in 1967, is a national health registry with information on all births in Norway. (Irgens 2000)

The Medical Birth Registry of Norway provides linkages through the unique personal identification number assigned to every person at birth.

“All maternity units in Norway must notify births to the MBRN (midwives and physicians). The notification form includes the name and personal identity number of the child and of the parents, information about maternal health before and during the pregnancy, and any complications, labour interventions, drugs used, during pregnancy or birth.” The registry records also whether the baby is born alive, any diagnoses in the child or evidence of congenital abnormalities.(Irgens 2000, NIPH 2019) All pregnancies ending after week 12 can be notified to the MBRN, including terminations after week 12. (NIPH 2019)

Upon mother’s consent, additional information about mother's occupation, smoking and alcohol habits and assisted conception can be registered.

The management of the MBRN is operated by The Norwegian Institute of Public Health and to ensure data quality, the MBRN is routinely linked with the Central Population Register.

The information is anonymized using a code system.

Researchers can apply for data of the MBRN (Norwegian Institute of Public Health's health registries and health studies) in order to conduct research on them. (NIPH 2019)

3.1.3 The Norwegian Multiple Sclerosis Registry

The Norwegian MS Registry has been approved by the Regional Committee for Research Ethics and the Norwegian Data Protection Authority in 1998.

The actual enrollment of patients began in 2001. The local neurologist, responsible for the diagnosis and treatment of the patients, has the responsibility of recording the patient in the registry, after obtaining the patient's written informed consent. (Myhr, Grytten et al. 2015)

The MS registry is located at the Norwegian MS Competence Centre at Haukeland University Hospital in Bergen.

The registry aims at providing quality control of diagnosis and treatment of patients with MS but data and biological samples for MS research can be requested and used. Indeed all researchers, through specific applications, can have access to these data.

In addition to clinical and demographic information of each patient, among which: year of birth, year of onset and of diagnosis, symptoms at onset, and disease course, information on diagnostic procedures, possible risk factors, comorbidity, specific and concomitant treatments are collected in the registry, as well as patient-reported outcomes.

3.1.4 Study procedures

This study is a community-based cohort study.

The MBRN provides, among others, information related to birth such as birth weight, gestational age at birth, mode of delivery, and pregnancy complications such as preeclampsia, for all individuals born in 1967 or later in Norway. Permission to use information from the National registry to link mothers with children born before 1967 was obtained from the Norwegian Tax Administration. The linkage was performed by the MBRN. The study population was characterized by children with information on breastfeeding and born between 1922 and 1986, in order to have at least 30 years of follow-up. In order to link the different datasets, we used the national identification number unique to every Norwegian citizen and resident (Figure 1).

Outcome: Offspring who developed MS were identified by linking the cohort dataset with the National MS Registry of Norway (Norwegian Competence Center for Multiple Sclerosis, Bergen), where MS patients are recorded up to 2016. The registry contains information on age at diagnosis and clinical course and covers more than 70% (biobank 2017) of the cases of MS in the country.

This is estimated based on the Norwegian National Patient Registry (NPR), which contains MS codes according to the International Statistical Classification of Diseases, 10th edition (ICD-10) given to every inpatient or outpatient, and partly to patients treated at private practices.

Exposure and covariates: The primary exposure of interest was 'breastfeeding'. In the CONOR surveys, mothers were asked how many children they gave birth to and for how many months they breastfed each of them. Duration of breastfeeding was considered as a categorical variable according to different cutoffs: at least 1 month vs. less than 1 month; at least 3 months vs. less than 3 months; at least 4 months vs. less than 4 months; at least 6 months vs. less than 6 months.

Possible confounders and/or effect modifiers of the association between breastfeeding and risk of MS were offspring sex, year of birth by 5-year categories, offspring birth order, level of education of the mother (mandatory, high school, college/university, missing), cigarette smoking habits of the mother (ever vs. never smoked), and mother's age at delivery. Among the offspring born in 1967 or later, additional factors were evaluated as possible confounders: preeclampsia during the pregnancy, mode of delivery, offspring birth weight (low: <2500g, medium: 2500-3499 g, high: ≥3500 g) and gestational age (preterm: < 37 weeks vs. full term: ≥37 weeks).

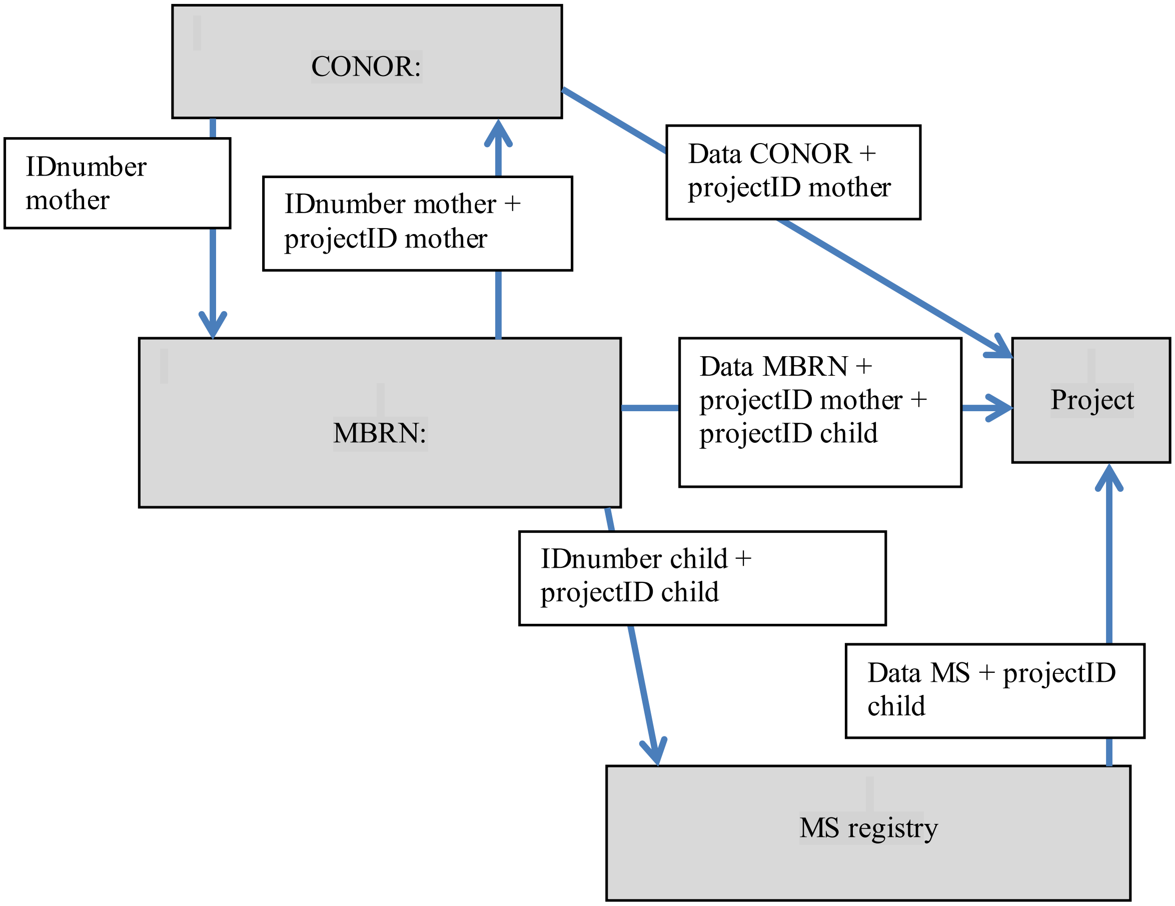

CONOR:
IDnumber mother
IDnumber mother + projectID mother
Data CONOR + projectID mother
MBRN:
Data MBRN + projectID mother + projectID child
Project
IDnumber child + projectID child
Data MS + projectID child
MS registry

Figure 1. Flowchart of the data process

3.2 Study II

To assess the effect of antibiotic exposure on the risk of developing MS in the Emilia-Romagna region (RER), Italy.

3.2.1 RER Pharmaceutical reimbursement database for outpatient and hospitalized subjects (AFT and AFO)

The two systems were activated on January 1st 2002, in Emilia Romagna region. They include complete and analytic data on outpatients and in hospital drug prescriptions, with the aim of monitorizing the characteristics of pharmaceutical consumptions, with analytics, quantitative and qualitative information on drugs prescribed and used in hospital institutions. (SISEPS 2019) The data are updated on a monthly basis.

3.2.2 RER Hospital discharges database (SDO)

The hospital discharge record (SDO) has been established by the Italian Ministry of Health in 1991 as part of the clinical record and as a routine instrument for collecting information on patients admitted in any public and private hospital of the country. In 2000 the ICD 9 CM classification system was introduced to codify all diagnoses and interventions. (SISEPS 2019)

3.2.3 Study procedures

This is a population-based nested case-control study.

Emilia Romagna is a region located in North-East Italy, and on 31/12/2017 the inhabitants were 4,452,629. (Emilia-Romagna.)

Multiple sclerosis care is provided by 12 MS dedicated units across the region, authorized to prescribe disease modifying drugs and other specific drugs to people with MS, according to the National Health System (NHS) regulations. Generally, no other private or public service provides care to MS patients. All but one regional MS units agreed to participate to the present study.

Subjects: All patients seen at a MS center between 01/02/2015 and 31/12/2017 were invited to participate to the study. They were asked, after signing an informed consent, to investigate their exposure to antibiotics as obtained from of the drug prescriptions reimbursement database of RER. All patients with a MS diagnosis, relapsing or progressive, regardless the diagnostic criteria (Poser, Paty et al. 1983, McDonald, Compston et al. 2001, Polman, Reingold et al. 2005, Polman, Reingold et al. 2011, Thompson, Banwell et al. 2018), with MS onset after January 1St 2005 and aged ≥18 years at inclusion were eligible. For each patient the participating neurologist collected the following information: fiscal code, year of birth, place of birth, place of residence, sex, year of MS onset (medical records), year of diagnosis and initial clinical course (relapsing-remitting or primary progressive). The fiscal code was then transformed by the Emilia Romagna Regional Health Service in an anonymous code that allow the anonymized link to the information on the database of drugs prescription. The index year is the year of MS onset.

Exclusion criteria was having had onset of MS before January 1st 2005.

For each MS patient five controls were identified among RER residents, matched on age (± 1 year), sex, place of residence and time in the cohort before the index year. Controls had to be alive at the time of the analysis.

Individuals who, based on drug prescription record, had received disease-modifying drugs for MS before or after the index date, were excluded from the eligible controls. Eligible controls were subjects resident in the RER in the year of onset of the corresponding case (index year) and should have not been prescribed with any MS disease modifying drugs.

Exposure: Antibiotics are provided to all Italian citizens by the National Health System with no or minimal expense for the subject. Information on antibiotic prescription was obtained through the linkage with the RER regional health information system, specifically the drug prescription database available with complete data since 2002. The data included all antibiotics prescribed by the general practitioners and those administered during hospital admissions. All prescriptions reporting the Anatomical Therapeutic Chemical (ATC) code "J01", corresponding to "Antibacterial for Systemic Use", were selected. For each antibiotic prescription a personal identification number, date of dispensation, type of drug (ATC-code), defined daily dose (DDD= the assumed average maintenance dose per day for a drug used for its main indication in adults) and estimated duration of use (days = number of packs * DDD) were obtained. Antibiotics were classified as "all antibiotics" (ATC code J01), tetracycline (ATC code J01A), penicillin (ATC code J01C), penicillin with extended spectrum (ATC code J01CA), cephalosporin (ATC code J01D), sulfonamides and trimethoprim (ATC code J01EE), macrolides (ATC code J01FA), quinolone antibacterial (ATC code J01M) and "other antibiotics" (ATC code J01XX).(WHO 2019)

Exposure to antibiotics was defined as having had at least one prescription of antibiotics in the 3 years before the index year.

Additional analyses were performed using as exposure period the 8 and 13 years preceding the index year.

Data Linkage process: The SISEP ("Servizio Informativo Politiche per la Salute e Politiche Sociali E-R") (SISEPS 2019) assigns a unique and anonymous identification number (PROG_PAZ) to all residents with at least one contact with Emilia Romagna Regional Health Service.

This identification number is present in the health regional administrative data -bases and allow to linkage between all databases and to keep anonymity.

The fiscal code obtained from each case was thus transformed by the Emilia Romagna Regional Health Service in an anonymous code (PROG_PAZ) that allow the anonymized link to the information on the database of drugs prescription.

A deterministic linkage between regional residents, clinical database and drug prescription database was then performed.

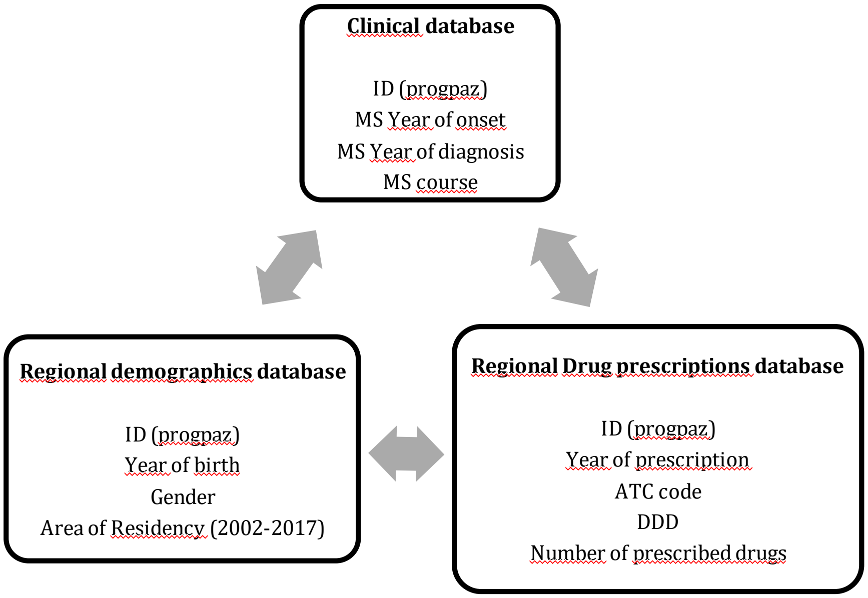


Figure 2. Deterministic linkage between regional residents, clinical database and drug prescription of RER.

3.3 Study III

To assess the accuracy of Death Certificates in the identification of subjects with Amyotrophic Lateral Sclerosis in the Limousin region, France, through the validation of these administrative data using the FRALim registry as gold standard. Moreover, to identify whether specific factors are associated with the accuracy of the Death Certificates.

3.3.1 Amyotrophic lateral Sclerosis

Amyotrophic lateral sclerosis (ALS) is an adult-onset, rare neurodegenerative disorder characterized by progressive degeneration of upper and lower motor neurons. (Logroscino, Traynor et al. 2008) It causes the paralysis of all voluntary muscles leading to death or tracheostomy in 2 to 4 years from symptoms onset, usually due to respiratory failure. (Logroscino, Traynor et al. 2008, Irgens 2012) The diagnosis is based on careful clinical examination, appropriate neurophysiologic studies, and exclusion criteria. The follow-up is usually organized in referral centers.

The incidence of ALS worldwide shows an heterogeneous distribution (Marin, Boumédiene et al. 2016) and the reported standardized incidence in Europe 1.81/100,000 (95% CI 1.66-1.97) persons-year of follow up (PYFU), and 2.58/100,000 (95% CI 2.27-2.89) PYFU (Chiò, Logroscino et al. 2013, Marin, Hamidou et al. 2014) standardized on European population, in the Limousin region, France.

In up to 10% of people a family history of ALS in a first-degree relative is identified (Al-Chalabi, Fang et al. 2010), but the etiology of ALS, still largely unknown, involves the concomitant effects of genetic and environmental factors.(Al-Chalabi and Hardiman 2013, Ingre, Roos et al. 2015)

3.3.2 The French register of ALS in Limousin - FRALIm registry

The French register of ALS in Limousin (FRALim) is the first ALS register in France to be established, with the aim of assessing the incidence of ALS in Limousin region, characterized by an ageing population (28.8% of inhabitants older than 60 years).(Marin, Hamidou et al. 2014)

Patients are included if they lived in the Limousin region at the time of the diagnosis of ALS and if they were identified by at least one source of ascertainment (see below).

To identify incident cases the register uses multiple sources, using the same criteria for case identification to ensure uniformity across sources. All information is reviewed by one neurologist, who then defines the ALS case according to El Escorial revised criteria (EERC). (Brooks, Miller et al. 2000, Marin, Hamidou et al. 2014)

Patients diagnosed with ALS can be referred to an ALS referral center by their general practitioner (GP), neurologist or another practitioner. All referral centers are coordinated by the French national body coordinating ALS referral centers (source 1), which also collects information on all patients recorded in a database, for daily routine practice. All ALS patients living in the Limousin region and diagnosed in one of the 18 ALS tertiary centers in France are included.

Of the 25 public and private hospitals in the region, 15 participate (source 2), and identify inpatients with a G12.2 code according to the International Classification of Disease (ICD) 10th version at any place in their diagnosis-related group. These kinds of data are generally generated by hospitals for administrative and reimbursement purposes within the national health system. These medical administrative data are used to identify only ALS patients and the diagnosis (according to EERC) and important dates are verified from medical records, allowing also the collection of other clinical data.

Patients with a declared long-term disorder [Affection de Longue Durée (ALD) no. 9 related to ALS] are reported to health insurance bodies (source 3) by GPs in order to obtain exemptions from health fees for their patients. The main sources among these are (i) the ‘Regime General’ or general scheme for employees and the unemployed covering 75% of the French population, (ii) the ‘Regime Agricole, Mutualite Sociale Agricole’ covering people involved in agriculture, (iii) the ‘Regime Social des Independants’, dealing with artisans, traders, self-employed professionals, and (iv) the ‘Caisse Nationale Militaire de Securite Sociale’ for military employees.

The correct diagnosis of ALS and the date of the diagnosis itself for these subjects are verified by contacting the GP who declared the ALD no. 9 and the neurologist of the patient.

For each patient several data are collected in the register: demographic and clinical characteristics, including date of birth, gender, date of onset and of diagnosis, type of onset, manual muscular testing (MMT) and ALS Functional Rating Scale_Revised (ALS-FRS-R) evaluated at diagnosis, and EERC assessed at diagnosis and at latest follow-up. Diagnosis of ALS for subjects with possible ALS according to EERC was revised after 4 years of stable possible ALS and amended to progressive lateral sclerosis (excluded from the register).(Marin, Hamidou et al. 2014)

3.3.3 Validation of medico-administrative data – ALS

The vast amount of information recorded for administrative purposes needs to be validated to assess their accuracy, fundamental for their use in epidemiological research. (van Walraven, Bennett et al. 2011)

Based on the specific characteristics of the disease that has to be studied, different combinations of data should be selected as the most appropriate for the validation process.

The assessment of accuracy has been performed on different kind of administrative data.

Among neurological diseases one example is Guillain-Barrè syndrome (GBS) - acute inflammatory polyradiculoneuropathy, a rare neurological disorder with low incidence across continents (crude incidence 0.81- 1.89 cases per 100,000 person-years).(Sejvar, Baughman et al. 2011) GBS patients need hospital admission, thus this disease represents a good target to assess the validity of hospital discharge data. An Italian study (Lombardy, northern Italy) assessed the validity of the hospital discharges from neurological wards, as an identifier of incident patients with GBS. (Bogliun and Beghi 2002) This study confirmed previous data of accuracy of hospital discharge data (HDD) from neurology departments as fairly valid surrogate of GBS incidence. (Jiang, de Pedro-Cuesta et al. 1995)

Among the available processes for the validation of medico-administrative data (Nissen, Quint et al. 2019), the use of gold standard information for case identification has been applied to some neurological diseases (eg., Creutzfeldt-Jakob Disease (Barash, West et al. 2014), Guillain-Barré Syndrome (Bogliun and Beghi 2002)), including amyotrophic lateral sclerosis (ALS). (Chancellor, Swingler et al. 1993, Beghi, Logroscino et al. 2001, Chio, Ciccone et al. 2002)

The accuracy of death certificates has been assessed in the diagnosis of Creutzfeldt-Jakob disease (Conti, Masocco et al. 2005, Brandel, Welaratne et al. 2011, Barash, West et al. 2014), mostly showing the need to consider multiple sources of data for the identification.

In case of chronic diseases the use of administrative data is more complex and need a combination of information in algorithms that need to be validated as well.(Tirschwell and Longstreth 2002, Muggah, Graves et al. 2013)

3.3.4 Mortality data use to estimate incidence of ALS

For those diseases with relatively short course, with mandatory hospitalizations or fatal outcome, the administrative data could be a source of identification of incident cases.

ALS is rapidly and invariably fatal (Chiò, Magnani et al. 1995) and only in about 10% of patients the survival is longer than 5-8 years. (Forbes, Colville et al. 2004, Zoccolella, Beghi et al. 2008, Pupillo, Messina et al. 2014)

It is therefore assumed that all patients with the disease will eventually be identified through death certificates (DCs) and this would reflect the incidence pattern, with a 2-4 years delay. (Marin, Couratier et al. 2011)

DCs are therefore another important source of information in the ascertainment of the disease, due to the information recorded and the availability of the causes of death. (Jougla, Pavillon et al. 1998, Marin, Couratier et al. 2011)
Moreover, it is reported that incorrectly classifying ALS as cause of death occurs less frequently compared to other neurodegenerative disorders as Parkinson's disease or dementia. (Johansen and Olsen 1998, Marin, Couratier et al. 2011, Hobson and Meara 2018)
However, given the administrative nature of the data on death certificates, their accuracy needs to be evaluated and confirmed, so as to ensure a high consistency with incident data.(Marin, Couratier et al. 2011)
A systematic review on the methodological quality of mortality studies on ALS (Marin, Couratier et al. 2011) used 'six good epidemiological practice criteria' to evaluate each study.

The good practice epidemiological criteria for ALS mortality studies reported were: i) definition of the population at risk; ii) accuracy of DC; iii) mortality data based on 'underlying' and 'contributory' causes; iv) examination of ALS rate time trends; v) comparability and high quality of health care and DC system of the geographical region considered; vi) consider ethnic factors or access to health care.

Out of 29 articles in which ALS mortality calculation was based on DC data, only a minority was compliant with all criteria. When all criteria were applied, this ensured a high consistency between the mortality and incidence rates.(Marin, Hamidou et al. 2014)

Fundamental, for the use of administrative data or other data collection is to assess the accuracy in identifying the disease.

Previous studies have been published on the validation of administrative data and specifically death certificate in the identification of ALS diagnosis, few recently. (Chiò, Magnani et al. 1992, Ragonese, Filippini et al. 2004, Marin, Couratier et al. 2011, Kioumourtzoglou, Seals et al. 2015) The results vary and in most cases it is suggested to use death certificates data in combination with other sources of information for a better performance.

The accuracy of death certificate showed to be of satisfactory quality in most of the countries where a validation study was performed.(Marin, Couratier et al. 2011) However certain variability was reported across countries or regions due to the validation methodology applied. (Ragonese, Filippini et al. 2004, Marin, Couratier et al. 2011)

Different approaches could be applied. National DC databases could be search in order to identify and select those DCs where ALS was indicated as underlying or contributory cause of death. The gold standard to validate this kind of data would be hospital discharges databases or clinical registries. In these studies positive predictive values are usually reported.

Another possibility would be to base the selection on already collected morbidity data (from registries of ALS) and then select the DCs of only those subjects with a known diagnosis of ALS that have died. This approach presents the advantage of allowing the revision also of the 'alternative diagnoses' reported on DCs when ALS was not recorded. In this case the sensitivity of DC data can be obtained.(Marin, Couratier et al. 2011)

The variability of methodologies and results emerging from the evaluations reinforces the need for an assessment of the DC validity for each population considered.

3.3.5 Study procedures

This is a validation study prospective population-based.

The study was set in Limousin, a French region with a population of 741,100 inhabitants in 2011 (Institut National de la Statistique et deas Etudes Economiques). (Insee 2013)

This retrospective population-based validation study used the *FRALim register* (Vasta, Boumediene et al. 2017) as the gold standard for ALS case identification.

A detailed *FRALim register* methodology description has been previously published (Marin, Hamidou et al. 2014), and described above.

The register was initiated in 2011, collecting ALS cases diagnosed since year 2000 and it is based on multiple sources for case ascertainment.

ALS cases that received a diagnosis between 2000 and 2011 were included in this study, to avoid inclusion of prevalent cases.

Death Certificate information was provided by the CEPIDC (Centre d'épidémiologie sur les causes médicales de décès). DC specific data were provided, following a specific anonymizing procedure and only for those patients with diagnosis of ALS included in the *FRALim register*.

The DCs data are related to those patients recorded in the *FRALim register* deceased between 2000 and 2011.

Each DC reported information on demographics, date and place of death. The clinical section provides information on the primary cause of death (immediate cause- i.e. the final disease or condition resulting in death) and secondary cause/causes (initial cause- i.e. list of conditions leading to the immediate cause) and the underlying cause as last, which was the disease that initiated the events resulting in death. In addition, a list of those comorbidities which could be related to the initial cause, are listed as well.

DCs are filled by medical doctors and the ICD10 classification codes are used to codify the cause of death.

The DC diagnosis was considered correct if ICD10 code G12.2 was recorded at any position of the listed causes of death.

Aggregated data on all deaths, and those deaths with ICD10 code corresponding to G12.2, were available for the Limousin region for the same period (2000-2011).

ICD10 code G12.2 includes all motor neuron disease (subtypes of ALS, progressive muscular atrophy, progressive bulbar palsy, primary lateral sclerosis, others) but excludes the spinal type of muscular atrophy

3.4 Statistical analysis

In **study I t**he frequency of breastfeeding varied according to other potential risk factors; we tested this difference using the $\chi 2$ test or Fisher's exact test for categorical variables (offspring sex, maternal smoking) and Wilcoxon rank-sum test for continuous variables (mother's education and offspring birth order). Cox proportional hazard regression models were used to estimate the association between the exposures and the risk of MS onset during the observation period given as hazard ratios (HRs). Follow-up started at birth and subjects were censored at death or the end of the study period, whichever came first. Different models were used considering different cutoffs for duration of breastfeeding. Each factor was examined in univariate models. All multivariable Cox models were adjusted for infant sex, birth order, and year of birth by 5-year categories, as well as mother's age at delivery, mother's cigarette smoking habit (ever vs. never smoker), and mother's education level (mandatory, high school, college/university, missing data). These covariates were those considered clinically relevant and those that showed an association with MS risk with a p-value < 0.05 in the univariate analyses. We considered as potential effect modifiers the factors sex, age at MS onset (<30 years vs. ≥ 30 years), and year of birth (born before 1971 vs. born in 1971 or after), and estimated their effects in regression models with different cutoffs for breastfeeding duration (test of interaction term). Moreover, to check for any possible recall bias, we repeated all analyses including only cases whose mothers reported about breastfeeding prior to disease onset in their offspring. We further conducted a sub-analysis for individuals born after 1967 including also the following perinatal factors into the multivariable models: mode of delivery, birth weight (low, medium, high), and gestational age (preterm vs. full term birth). We used a two-sided significance level of 0.05 for all analyses.

All analyses were conducted using SPSS ® software, v.25.

In **study II** descriptive statistical analyses were performed to analyze demographic and clinical characteristics, using frequencies and percentages to summarize categorical variables and median and Interquartile range (IQR) to summarize continuous variables.

Conditional logistic regression models were applied to estimate the odds ratios (OR) of the association between antibiotic use before the index date and MS onset. Models considered matched factors (age, sex, place of residence and time in the cohort before the index year). Any antibiotic prescription and each single antibiotic category (ATC code) were examined as independent variable. The 0.05 level of significance and two-sided tests were used for all analyses.

The information on pharmaceutical prescriptions was available since 2002, thus we performed separate analyses considering three time windows for the exposure: use of antibiotics during the three years before index, during 8 years and 13 years before index.

To evaluate possible differences among time windows of exposure, the use of any antibiotics was analyzed separately for the following lags: from 9 to 6 years before the index, from 6 to 3 years before and during the three years preceding the index.

In order to evaluate the possible dose-response effect the association between the number of accumulated weeks of treatment both consecutive and not consecutive, and risk of MS was studied. The number of days of treatment has been calculated based on the DDD.

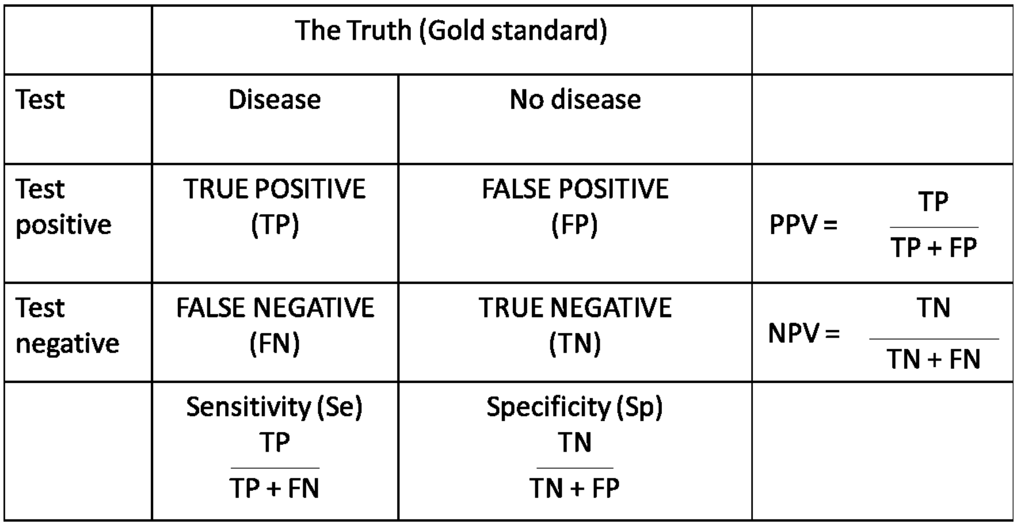
All analyses were conducted using STATA® software, v 14.0.

For study **III s**ensitivity, positive predictive value (PPV), given true-positive, true-negative, false-positive subjects and false negative were calculated.

Calculations were performed overall, and stratified by age groups (55-64, 65–75, ≥75 years), sex, and time-period of death ([2000–2003]; [2004–2007]; [2008– 2011]). Ninety- five percent confidence intervals (CI) were calculated.

In order to identify possible characteristics of subjects, independently associated with a correct ALS diagnosis on the DC (true positive), univariate logistic regression models were applied.

Factors possibly involved that were evaluated: sex, age at diagnosis (continuous and categorical), age at death (continuous and categorical), time period at death (categorized), source of ascertainment (ALS referral Centre, Hospital or Health insurance), number of sources of ascertainment (1, 2 or 3), diagnosis delay, site of first symptoms (bulbar vs. respiratory or spinal), Airlie House criteria at diagnosis (possible, probable+ laboratory, probable or definite), Manual Muscular testing at diagnosis, familial ALS, treatment with riluzole, gastrostomy placement, non-invasive ventilation placement, ALS Functional Rating Scale-Revised (ALSFRSR) score, variation of usual weight at diagnosis (%), Body mass index (BMI).



	The Truth (Gold standard)		
Test	Disease	No disease	
Test positive	TRUE POSITIVE (TP)	FALSE POSITIVE (FP)	PPV = $\frac{TP}{TP + FP}$
Test negative	FALSE NEGATIVE (FN)	TRUE NEGATIVE (TN)	NPV = $\frac{TN}{TN + FN}$
	Sensitivity (Se) $\frac{TP}{TP + FN}$	Specificity (Sp) $\frac{TN}{TN + FP}$	

Figure 3. Calculation of sensitivity, specificity, and positive and negative predictive values

3.5 Ethics – Protection of privacy

Permission to use information for **study I** from the National registry to link mothers with children born before 1967 was obtained from The Norwegian Tax Administration. The linkage was performed by the MBRN. The study population was characterized by children with information on breastfeeding and born between 1922 and 1986, in order to have at least 30 years of follow up. In order to link the different datasets, we used the national identification number unique to every Norwegian citizen and resident. The study was approved by the responsible regional Ethics committee (2017/298/REK vest).

Study II with data collection from people with MS and access to their pharmacological history was approved by the Bologna Health Trust Ethics Committee (CE 15069).

For **study III**, the FRALim registry has been approved by the CCTIRS ('Comité consultatif sur le traitement de l'information en matière de recherche dans le domaine de la santé', authorization no. 10487bis), by the CNIL ('Commission Nationale de l'Informatique et des Libertés', authorization no. 911038) and by an ethics review board ('Comité de Protection des Personnes Sud-Ouest Outre Mer').

4. SUMMARY OF RESULTS

4.1 Risk factors for multiple sclerosis (studies I-II)

4.1.1 Study I

The study included 50,069 mothers who participated in at least one CONOR survey between 1994 and 2002 and for whom information on breastfeeding duration was available. If the mother participated in more than one survey, we used information from the first survey. Based on this cohort of mothers, 122,286 children were identified who had been born between 1922 and 1999. The main analysis was performed among 95,891 children (78.4%) with information on breastfeeding and born between 1922 and 1986. Of these 6,231 (6.5%) had never been breastfed, whereas 89,660 (93.5%) had been breastfed for at least 1 month.

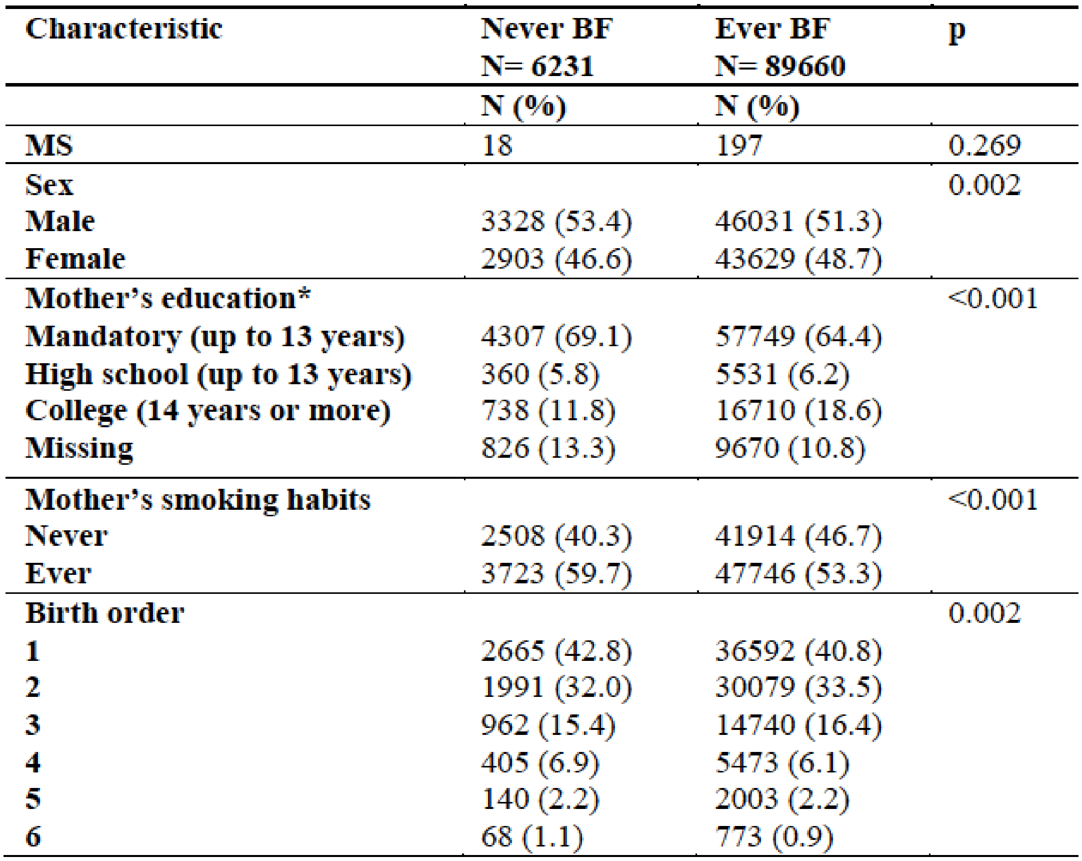
Of the 95,891 children, 58,803 (61.3%) were born in 1967 or later and for these, additional information on pre- and perinatal factors was available from the MBRN.

During follow-up, 215 offspring had a clinical onset of MS during adulthood. The mean age at onset was 31.2 years (standard deviation (SD) 8.9), and the mean age at diagnosis 35.2 years (SD 9.2). The majority of the cases were women (73%).

There were marked differences in exposure to breastfeeding according to maternal education and smoking, as well as offspring birth order and sex (Table 2a, 2b, 2c).

Table 2. Characteristics of study participants and their mothers by different cutoffs of breastfeeding duration (2a. never vs. ever; 2b. <4 months vs. ≥ 4 months; 2c. < 6 months vs. ≥ 6 months)

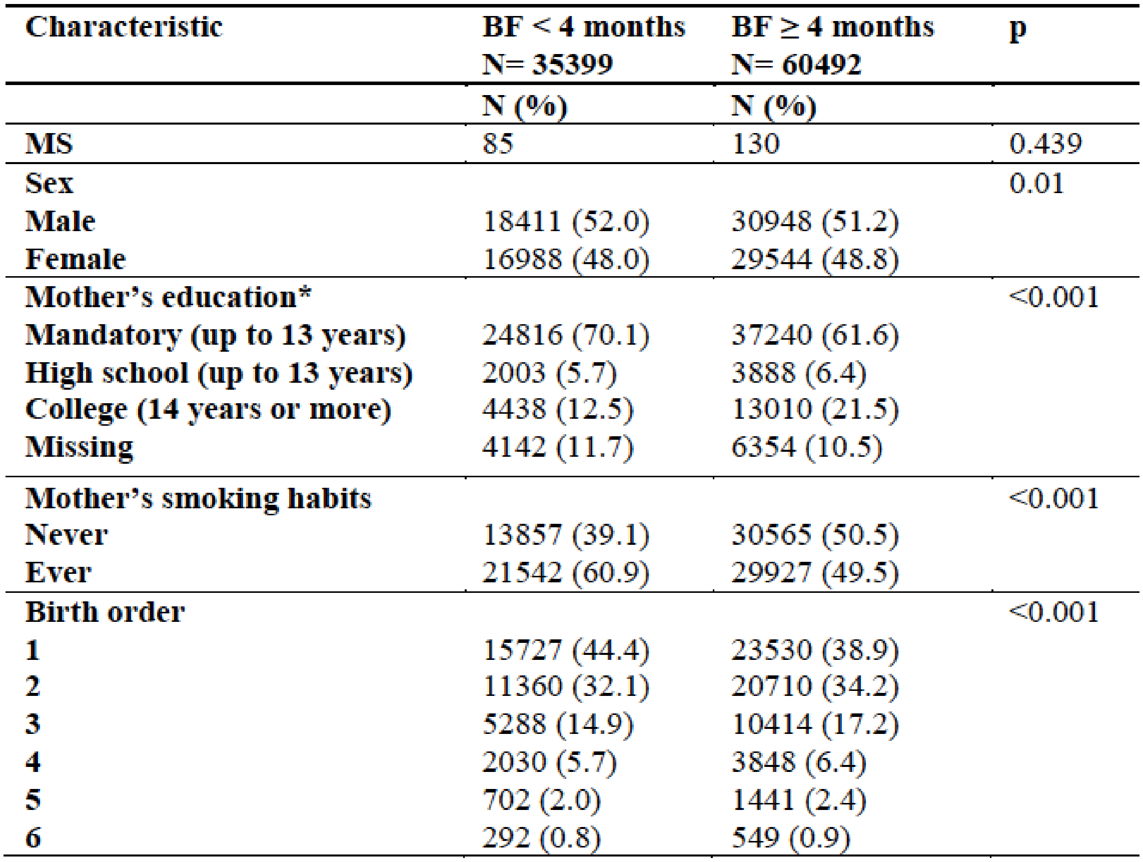
Table 2a. Never breastfed vs. breastfed



Characteristic	Never BF N= 6231	Ever BF N= 89660	p
	N (%)	N (%)	
MS	18	197	0.269
Sex			0.002
Male	3328 (53.4)	46031 (51.3)	
Female	2903 (46.6)	43629 (48.7)	
Mother's education*			<0.001
Mandatory (up to 13 years)	4307 (69.1)	57749 (64.4)	
High school (up to 13 years)	360 (5.8)	5531 (6.2)	
College (14 years or more)	738 (11.8)	16710 (18.6)	
Missing	826 (13.3)	9670 (10.8)	
Mother's smoking habits			<0.001
Never	2508 (40.3)	41914 (46.7)	
Ever	3723 (59.7)	47746 (53.3)	
Birth order			0.002
1	2665 (42.8)	36592 (40.8)	
2	1991 (32.0)	30079 (33.5)	
3	962 (15.4)	14740 (16.4)	
4	405 (6.9)	5473 (6.1)	
5	140 (2.2)	2003 (2.2)	
6	68 (1.1)	773 (0.9)	

BF= Breastfeeding; * N=58722. p= p values corresponding to Chi2 tests

Table 2b. Breastfed for <4 months vs. ≥ 4 months



Characteristic	**BF < 4 months** **N= 35399**	**BF ≥ 4 months** **N= 60492**	**p**
	N (%)	**N (%)**	
MS	85	130	0.439
Sex			0.01
Male	18411 (52.0)	30948 (51.2)	
Female	16988 (48.0)	29544 (48.8)	
Mother's education*			<0.001
Mandatory (up to 13 years)	24816 (70.1)	37240 (61.6)	
High school (up to 13 years)	2003 (5.7)	3888 (6.4)	
College (14 years or more)	4438 (12.5)	13010 (21.5)	
Missing	4142 (11.7)	6354 (10.5)	
Mother's smoking habits			<0.001
Never	13857 (39.1)	30565 (50.5)	
Ever	21542 (60.9)	29927 (49.5)	
Birth order			<0.001
1	15727 (44.4)	23530 (38.9)	
2	11360 (32.1)	20710 (34.2)	
3	5288 (14.9)	10414 (17.2)	
4	2030 (5.7)	3848 (6.4)	
5	702 (2.0)	1441 (2.4)	
6	292 (0.8)	549 (0.9)	

BF= Breastfeeding; * N=58722. p= p values corresponding to Chi2 tests

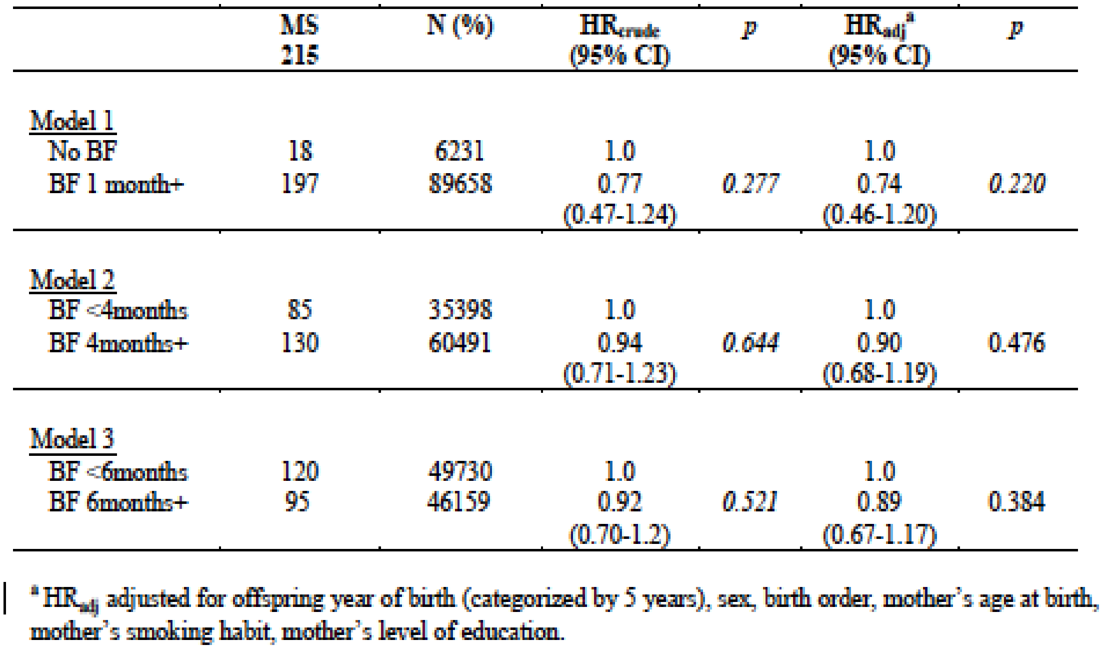
Table 2c. Breastfed for < 6 months vs. ≥ 6 months

Characteristic	p	BF < 6 months N= 49732	BF ≥ 6 months N= 46159	p
		N (%)	N (%)	
MS	0.439	120	95	0.221
Sex	0.01			0.04
Male		25760 (51.8)	23599 (51.1)	
Female		23972 (48.2)	22560 (48.9)	
Mother's education*	<0.001			<0.001
Mandatory (up to 13 years)		34568 (69.5)	27488 (59.6)	
High school (up to 13 years)		2810 (5.7)	3081 (6.7)	
College (14 years or more)		6582 (13.2)	10866 (23.5)	
Missing		5772 (11.6)	4724 (10.2)	
Mother's smoking habits	<0.001			<0.001
Never		20088 (40.04)	24334 (52.7)	
Ever		29644 (59.6)	21825 (47.3)	
Birth order	<0.001			<0.001
1		21096 (42.4)	18161 (39.3)	
2		16279 (32.7)	15791 (34.2)	
3		7810 (15.7)	7892 (17.1)	
4		3015 (6.1)	2863 (6.2)	
5		1115 (2.2)	1028 (2.2)	
6		417 (0.8)	424 (0.9)	

BF= Breastfeeding; * N=58722.   p= p values corresponding to Chi2 tests

We found no association between breastfeeding and MS onset in adulthood, also when considering different durations of breastfeeding (at least 1 month vs. less than 1 month, at least 4 months vs. less than 4 months, and at least 6 months vs. less than 6 months) (Table 2). Adjusting the models for offspring year of birth, sex, birth order, mother's age at delivery, level of education, and smoking habits did not modify the results (Table 3).

Table 3. Association between duration of breastfeeding (BF) and development of MS in 95,891 individuals. Crude and adjusted hazard ratios (HR) with 95% confidence intervals (CI).



	MS 215	N (%)	HR_{crude} (95% CI)	*p*	HR_{adj}[a] (95% CI)	*p*
Model 1						
No BF	18	6231	1.0		1.0	
BF 1 month+	197	89658	0.77 (0.47-1.24)	*0.277*	0.74 (0.46-1.20)	*0.220*
Model 2						
BF <4months	85	35398	1.0		1.0	
BF 4months+	130	60491	0.94 (0.71-1.23)	*0.644*	0.90 (0.68-1.19)	0.476
Model 3						
BF <6months	120	49730	1.0		1.0	
BF 6months+	95	46159	0.92 (0.70-1.2)	*0.521*	0.89 (0.67-1.17)	0.384

[a] HR_{adj} adjusted for offspring year of birth (categorized by 5 years), sex, birth order, mother's age at birth, mother's smoking habit, mother's level of education.

In analyses stratified by sex, median age at MS onset (<30 years vs. ≥30 years), or median year of birth (born before 1971 vs. born in 1971 or after), there was no evidence of an effect modification by these variables (test of interaction term, data not shown).

The majority of the mothers whose children developed MS (120 of 215) provided information on breastfeeding before knowledge of disease outcome. When we repeated the analysis including only cases whose mothers reported about breastfeeding prior to disease onset in their offspring, there were no major changes in the results (compare with Table 2, for model 1 (cutoff 1 month) the HR was 0.84 (95% CI: 0.43-1.67), for model 2 (cutoff of 4 months) the HR was 1.08 (0.73-1.57) and for the third model using cutoff 6 months, the HR was 1.05 (0.72-1.51)).

Nor did the results change in a sub-analysis, limited to the 58,803 subjects born after 1967 and therefore with available data from the MNBR, considering perinatal factors (mode of delivery, preeclampsia, gestational age, birth weight) (data not shown).

4.1.2 Study II

During the study period 877 subjects with MS were identified (Figure 4).

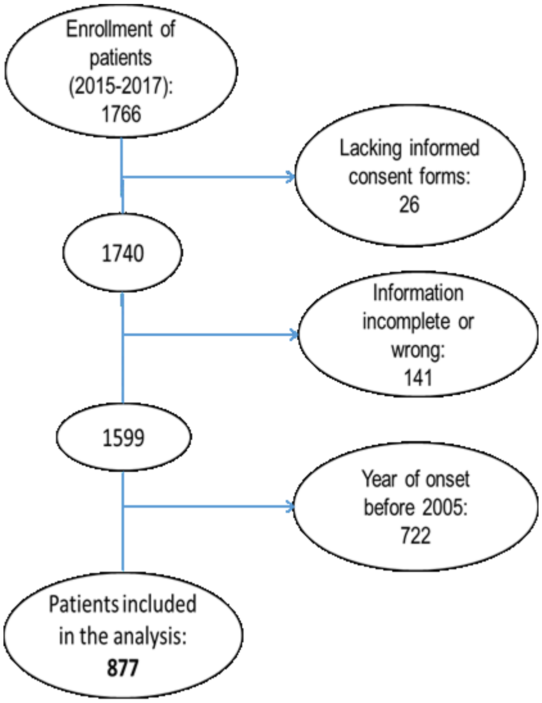
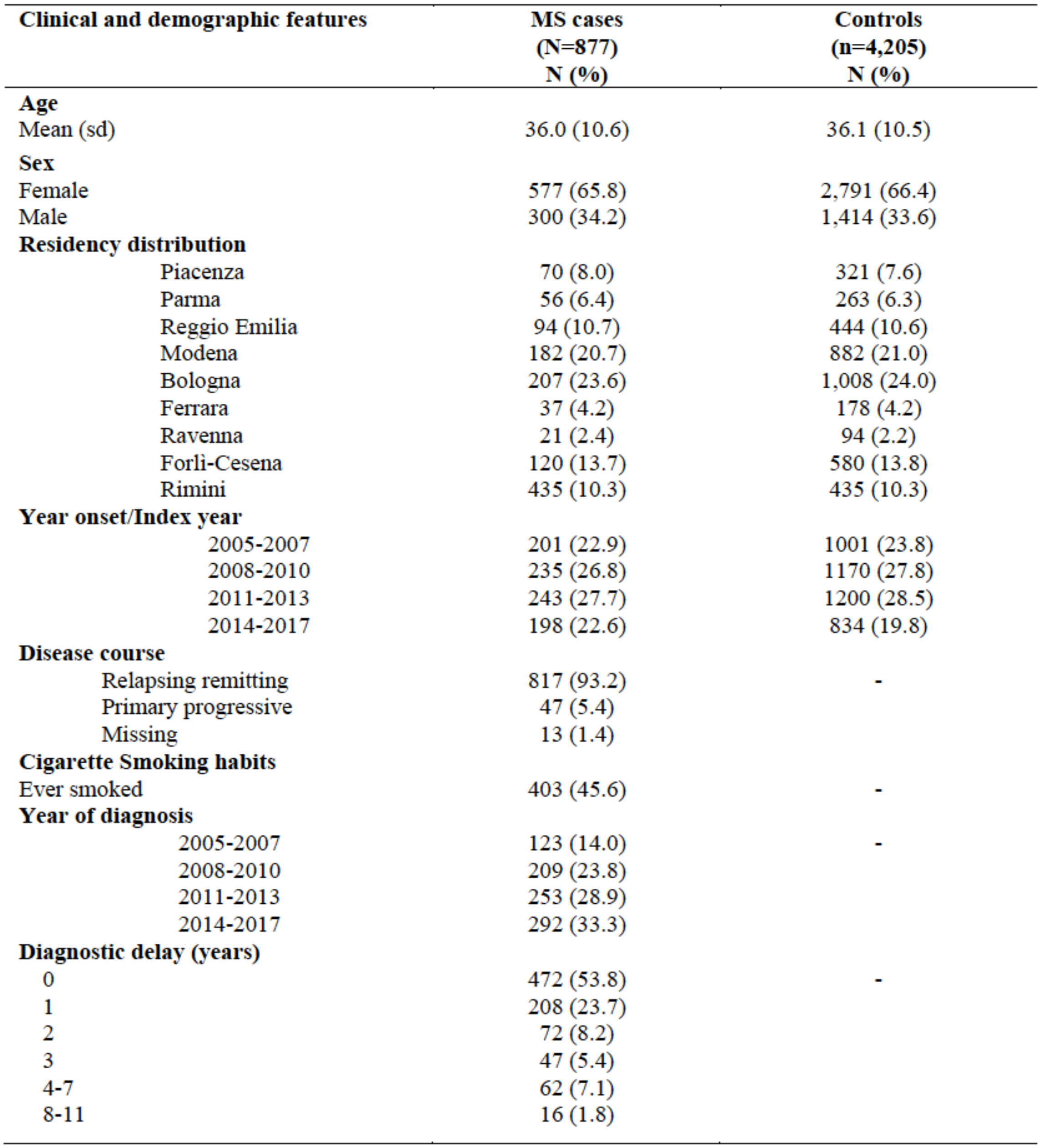

Enrollment of patients (2015-2017): 1766
Lacking informed consent forms: 26
1740
Information incomplete or wrong: 141
1599
Year of onset before 2005: 722
Patients included in the analysis: **877**

Figure 4. Flow-chart of study participants (patients with MS)

Mean age at onset was 36 years (sd 10.6), 65.8% were women and 93.2% of subjects had a relapsing remitting course of disease.

Cigarette smokers at any time were 45.6% of patients. (Table 4)

Table 4. Clinical and demographic characteristics distribution: 877 MS cases and 4,205 controls



Clinical and demographic features	MS cases (N=877) N (%)	Controls (n=4,205) N (%)
Age		
Mean (sd)	36.0 (10.6)	36.1 (10.5)
Sex		
Female	577 (65.8)	2,791 (66.4)
Male	300 (34.2)	1,414 (33.6)
Residency distribution		
Piacenza	70 (8.0)	321 (7.6)
Parma	56 (6.4)	263 (6.3)
Reggio Emilia	94 (10.7)	444 (10.6)
Modena	182 (20.7)	882 (21.0)
Bologna	207 (23.6)	1,008 (24.0)
Ferrara	37 (4.2)	178 (4.2)
Ravenna	21 (2.4)	94 (2.2)
Forlì-Cesena	120 (13.7)	580 (13.8)
Rimini	435 (10.3)	435 (10.3)
Year onset/Index year		
2005-2007	201 (22.9)	1001 (23.8)
2008-2010	235 (26.8)	1170 (27.8)
2011-2013	243 (27.7)	1200 (28.5)
2014-2017	198 (22.6)	834 (19.8)
Disease course		
Relapsing remitting	817 (93.2)	-
Primary progressive	47 (5.4)	
Missing	13 (1.4)	
Cigarette Smoking habits		
Ever smoked	403 (45.6)	-
Year of diagnosis		
2005-2007	123 (14.0)	-
2008-2010	209 (23.8)	
2011-2013	253 (28.9)	
2014-2017	292 (33.3)	
Diagnostic delay (years)		
0	472 (53.8)	-
1	208 (23.7)	
2	72 (8.2)	
3	47 (5.4)	
4-7	62 (7.1)	
8-11	16 (1.8)	

Exposure to any antibiotics prescribed in the three years before index was associated with MS: OR= 1.52 (CI 95%=1.29–1.79). Similar results were found for different antibiotics classes (Table 5) or considering only MS cases with relapsing remitting disease course.

Table 5. Exposure to antibiotics (at least one prescription in the 3 years preceding the index year) and risk of MS. Odds ratio (OR) and CI 95%

Antibiotics	**Cases** **N (%)**	**Controls** **N (%)**	**OR (95% CI)**
J01 – Antibacterials for systemic use	603 (68.8)	2,528 (60.1)	1.52 (1.29 -1.79)
J01C – Beta-lactam antibacterials (Penicillins)	378 (43.1)	1,621 (38.6)	1.22 (1.05-1.42)
J01D – Other beta-lactam antibacterials	202 (23.0)	872 (20.7)	1.19 (0.99-1.43)
J01FA – Macrolides	265 (30.2)	1,086 (25.8)	1.26 (1.07-1.48)
J01M – Quinolone antibacterials	165 (18.8)	680 (16.2)	1.22 (1.01-1.48)

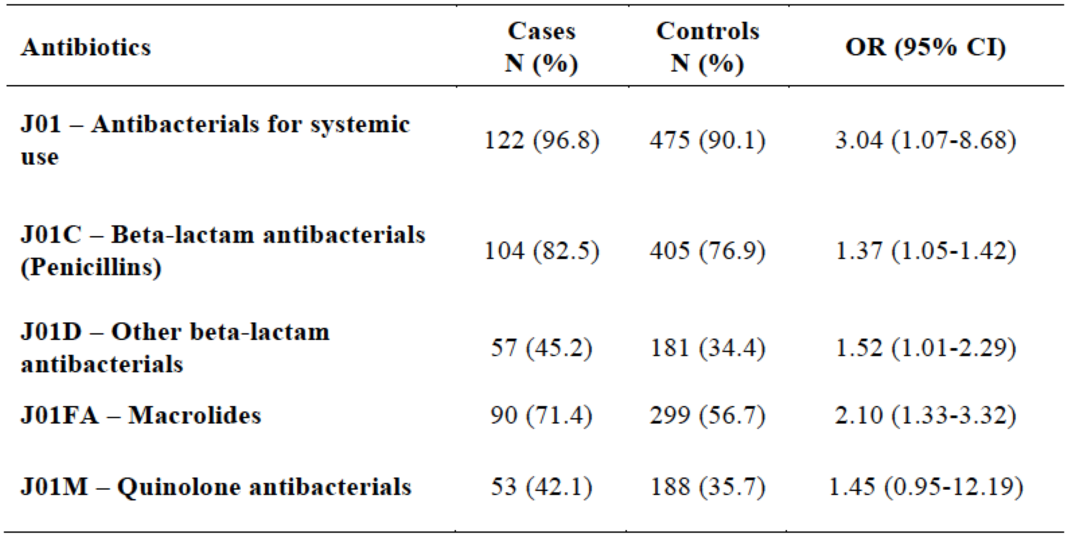
Moreover, considering exposure to antibiotics during the 8 and 13 years before the index date, the associations showed an increased effect size (OR=1.95, 1.44-2.63 and OR=3.04, 1.07-8.68, respectively) (Table 6a- 6b).

Table 6. Exposure to antibiotics (at least one prescription in the 8 (a) and 13 (b) years preceding the index year) and risk of MS

6a. 529 cases and 2,474 controls

Antibiotics	**Cases** **N (%)**	**Controls** **N (%)**	**OR (95% CI)**
J01 – Antibacterials for systemic use	468 (88.5)	2,000 (80.8)	1.95 (1.44-2.63)
J01C – Beta-lactam antibacterials (Penicillins)	374 (70.7)	1,559 (63.0)	1.44 (1.17-1.76)
J01D – Other beta-lactam antibacterials	156 (29.5)	669 (27.0)	1.16 (0.94-1.42)
J01FA – Macrolides	291 (55.0)	1,159 (46.9)	1.44 (1.19-1.76)
J01M – Quinolone antibacterials	179 (33.8)	718 (29.0)	1.31 (1.06 -1.61)

6b. 126 cases and 527 controls



Antibiotics	Cases N (%)	Controls N (%)	OR (95% CI)
J01 – Antibacterials for systemic use	122 (96.8)	475 (90.1)	3.04 (1.07-8.68)
J01C – Beta-lactam antibacterials (Penicillins)	104 (82.5)	405 (76.9)	1.37 (1.05-1.42)
J01D – Other beta-lactam antibacterials	57 (45.2)	181 (34.4)	1.52 (1.01-2.29)
J01FA – Macrolides	90 (71.4)	299 (56.7)	2.10 (1.33-3.32)
J01M – Quinolone antibacterials	53 (42.1)	188 (35.7)	1.45 (0.95-12.19)

In order to evaluate possible differences among time windows, the use of any antibiotics was analyzed separately for the periods from 9 to 6 years before the index (6 years lag-time), from 6 to 3 years before (3 years lag-time) and from 3 years before until the index year (no lag-time). The associations found for each period were similar (Figure 5). Similar results were identified for different antibiotics categories (data not shown).

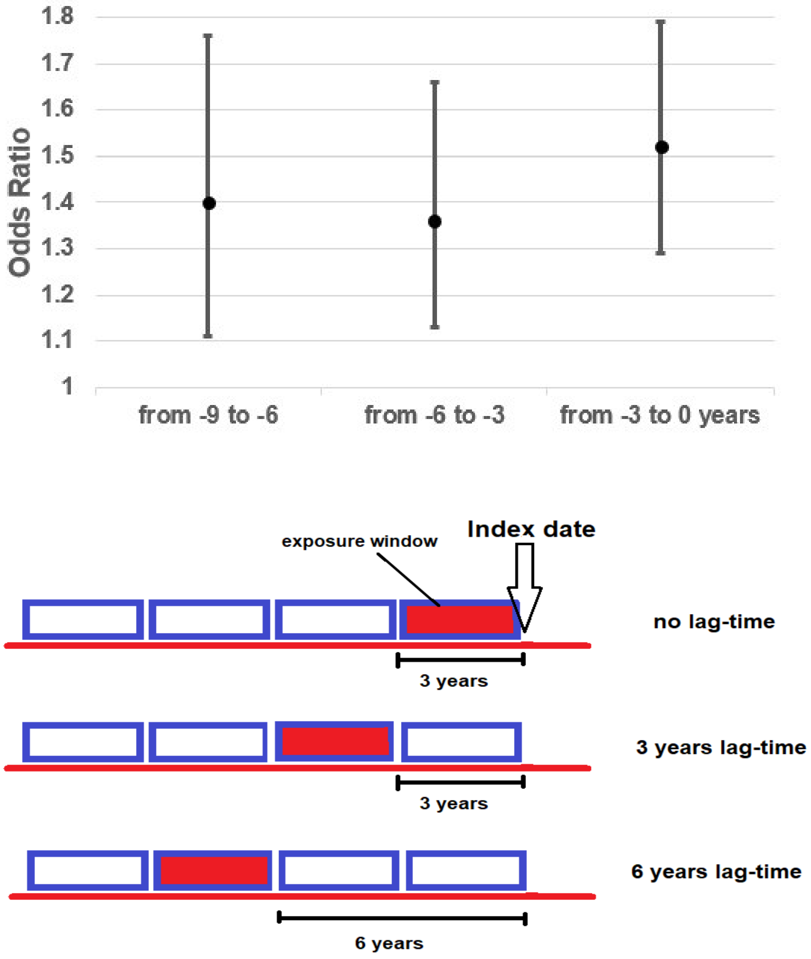


Figure 5. Lag-time exposure to antibiotics before index year and risk of MS

When considering the use of antibiotics as number of accumulated weeks of treatment both consecutive and not consecutive, during the 3 years before the index year, there was an association with increased risk of MS for treatment prescribed for 1 or 2 consecutive weeks. When the accumulated weeks of treatment were not consecutive the association was found across different durations of therapy (data not shown).

4.2 Death certificate validation

4.2.1 Study III

The *FRALim register* recorded 279 patients with a diagnosis of ALS between 2000 and 2011, of whom 224 died between 2000 and 2011. Among these deceased patients, 197 (87.9%) had a DC and 185 (93.9%) subjects were identified as having a diagnosis of ALS, corresponding to an ICD 10 code G12.2, among the causes of death in death certificates. In 12 (6.1%) DCs corresponding to patients recorded in the register there was no ALS diagnosis in any position of the document. The alternative causes of death recorded for these patients were respiratory arrest, malignant neoplasia and symptoms associated, degenerative disease of CNS, acute myocardial infarction/ischemic heart disease, polyneuropathy, hereditary sensory and motor neuropathy, suicide, nontraumatic intracerebral hemorrhage (Figure 6).

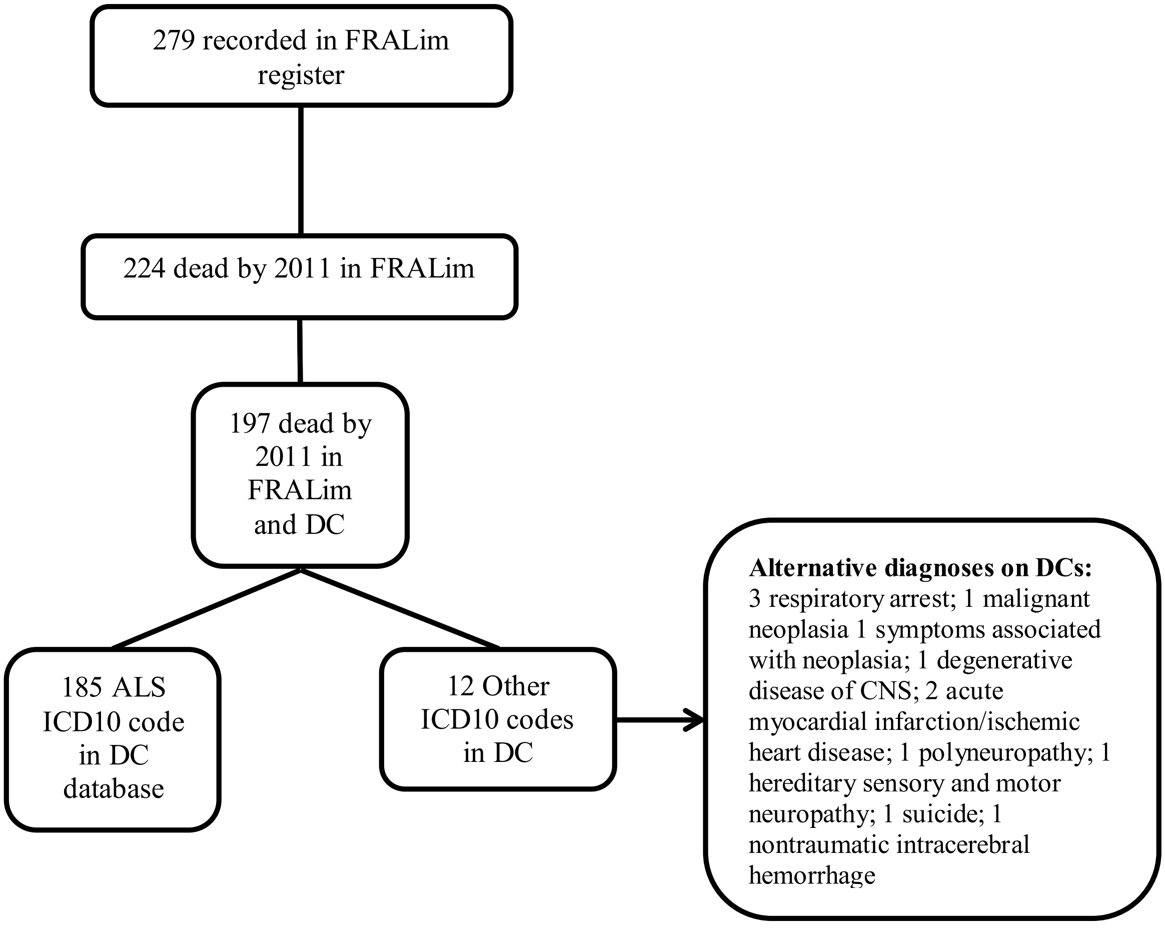

279 recorded in FRALim register
224 dead by 2011 in FRALim
197 dead by 2011 in FRALim and DC
185 ALS ICD10 code in DC database
12 Other ICD10 codes in DC
Alternative diagnoses on DCs:
3 respiratory arrest; 1 malignant neoplasia 1 symptoms associated with neoplasia; 1 degenerative disease of CNS; 2 acute myocardial infarction/ischemic heart disease; 1 polyneuropathy; 1 hereditary sensory and motor neuropathy; 1 suicide; 1 nontraumatic intracerebral hemorrhage

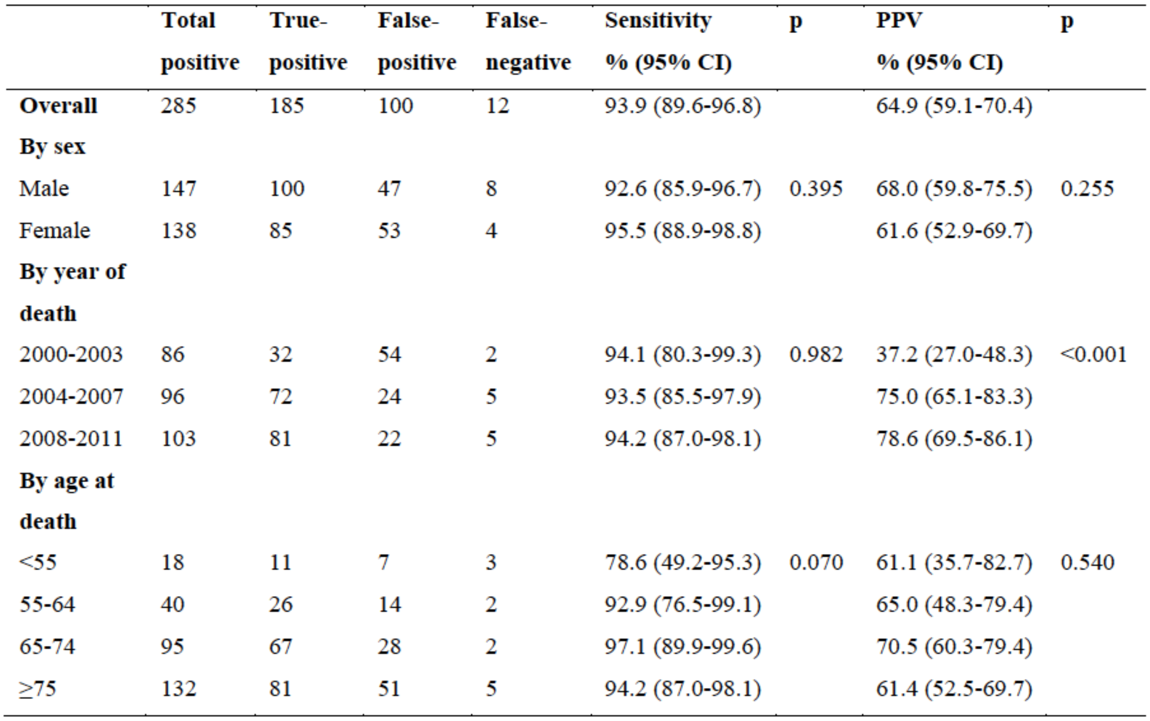
Figure 6. Flow chart of data collection

DC accuracy

Overall, sensitivity was 93.9% (95% CI 89.6-96.8) and PPV was 64.9% (59.1-70.4) (Table 7). Slightly higher Se and lower PPV for female compared to males (p=0.395 and p=0.255 respectively) were reported. The period of death did not show a clear trend in Se, while there was a lower PPV for those patients who died during 2000-2003 compared to the other periods (p<0.001).

For individuals who died before 55 years of age, the sensitivity of DC was lower compared to older patients (Table 7).

Table 7. Sensitivity and PPV of G12.2 code in DCs



	Total positive	True-positive	False-positive	False-negative	Sensitivity % (95% CI)	p	PPV % (95% CI)	p
Overall	285	185	100	12	93.9 (89.6-96.8)		64.9 (59.1-70.4)	
By sex								
Male	147	100	47	8	92.6 (85.9-96.7)	0.395	68.0 (59.8-75.5)	0.255
Female	138	85	53	4	95.5 (88.9-98.8)		61.6 (52.9-69.7)	
By year of death								
2000-2003	86	32	54	2	94.1 (80.3-99.3)	0.982	37.2 (27.0-48.3)	<0.001
2004-2007	96	72	24	5	93.5 (85.5-97.9)		75.0 (65.1-83.3)	
2008-2011	103	81	22	5	94.2 (87.0-98.1)		78.6 (69.5-86.1)	
By age at death								
<55	18	11	7	3	78.6 (49.2-95.3)	0.070	61.1 (35.7-82.7)	0.540
55-64	40	26	14	2	92.9 (76.5-99.1)		65.0 (48.3-79.4)	
65-74	95	67	28	2	97.1 (89.9-99.6)		70.5 (60.3-79.4)	
≥75	132	81	51	5	94.2 (87.0-98.1)		61.4 (52.5-69.7)	

PPV= positive predictive value

Characteristics associated with correct ALS diagnosis on DC

Considering 185 subjects true positive and 12 false negative few factors showed an association with a correct identification of ALS diagnosis in DCs (Table 8). Having an age of 55-64 years at diagnosis or at death increased the probability of being correctly identified in the DC diagnosis compared to other age categories.

Having been ascertained through hospital data, the use of riluzole, and a bulbar onset were associated with higher probability as well (OR: 4.94, 95%CI: 1.35-18.13; 5.30, 1.43-19.55; 6.45, 0.81-51.04, respectively).

Having a familial or unknown form of ALS showed an inverse association with a correct diagnosis on the DC (0.17, 0.03-0.97; 0.22, 0.05-0.92, respectively).

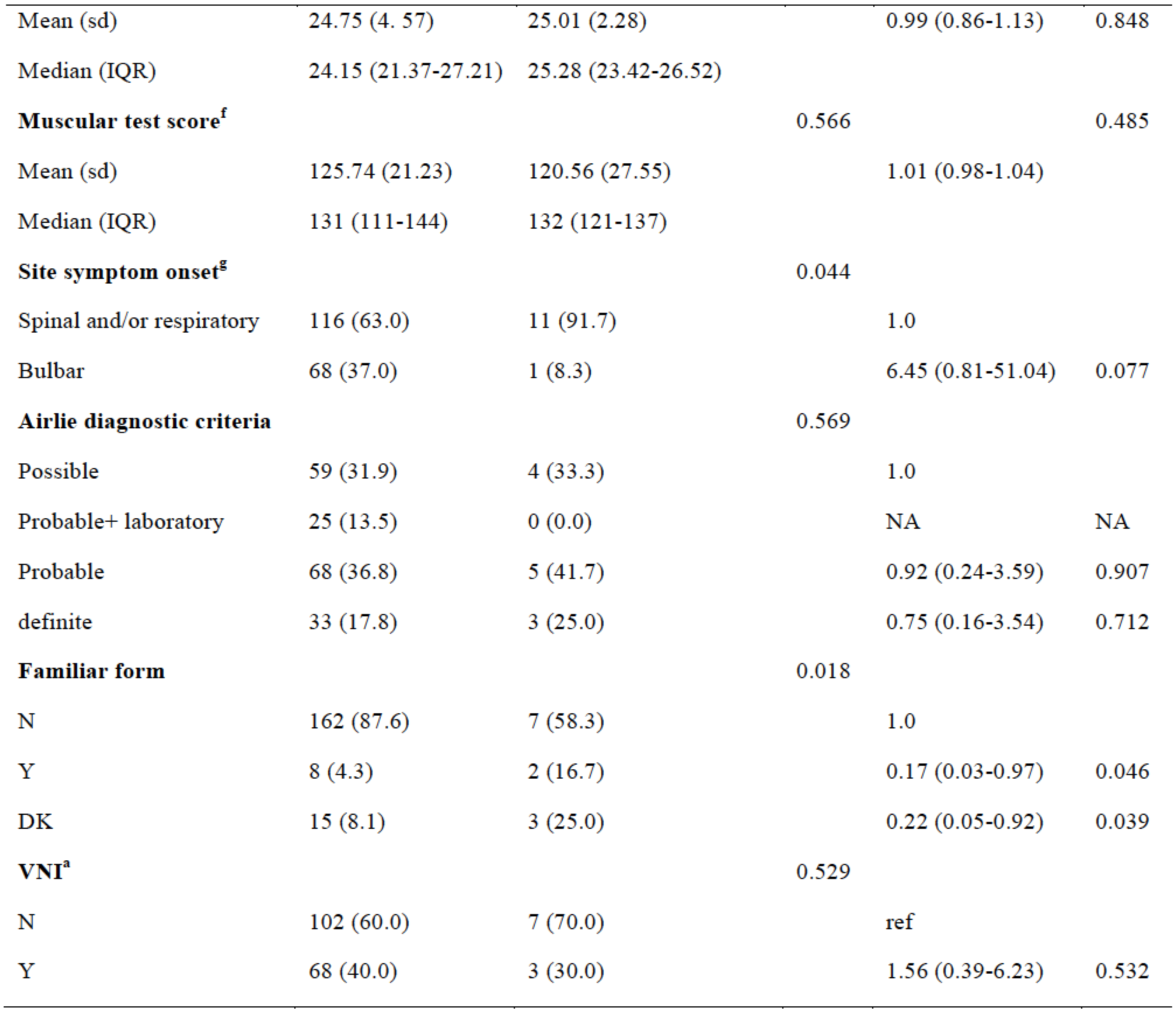
Table 8. Association between demographic and clinical factors and correct ALS diagnosis on DC in 197 individuals. Crude odds ratios (OR) with 95% confidence intervals (CI).

Factor	**DC+** **N= 185 (%)**	**DC- N= 12 (%)**	**p***	**Univariate OR (95% CI)**	**p**
Sex			0.395		
Male	100 (54.1)	8 (66.7)		0.59 (0.17-2.02)	0.400
Female	85 (45.9)	4 (33.3)		ref	
Year at diagnosis			0.590		
2000-2003	52 (28.1)	2 (16.7)		1.0	
2004-2007	82 (44.3)	7 (58.3)		0.45 (0.09-2.25)	0.332
2008-2011	51 (27.6)	3 (25.0)		0.65 (0.10-4.08)	0.649
Age at diagnosis			0.726		
Mean (SD)	70.70 (9.89)	66.71 (16.31)		1.03 (0.98-1.09)	0.200
Median (IQR)	71.65 (65.65-78.35)	72.17 (52.73-80.72)			
Age at diagnosis			0.033		
<55	16 (8.6)	4 (33.3)		1.0	
55-64	28 (15.1)	1 (8.3)		7.00 (0.72-68.15)	0.094
65-74	73 (39.5)	2 (16.7)		9.13 (1.54-54.19)	0.015
≥75	68 (36.8)	5 (41.7)		3.4 (0.82-14.11)	0.092
Age at death			0.668		
Mean (SD)	72.06 (9.55)	68.18 (15.39)		1.04 (0.98-1.09)	0.196
Median (IQR)	72.90 (67.06-79.35)	73.53 (55.42-81.61)			
Age at death			0.070		
<55	11 (5.9)	3 (25.0)		1.0	
55-64	26 (14.1)	2 (16.7)		3.55 (0.52-24.26)	0.197
65-74	67 (36.2)	2 (16.7)		9.14 (1.37-61.05)	0.022
≥75	81 (43.8)	5 (41.6)		4.42 (0.92-21.11)	0.063
Source of ascertainment: ALS center			0.758		
Y	160 (86.5)	10 (83.3)		1.28 (0.26-6.19)	0.759
N	25 (13.5)	2 (16.7)		ref	

Table 8. Continued

Source of ascertainment: hospital			0.009		
Y	168 (90.8)	8 (66.7)		4.94 (1.35-18.13)	0.016
N	17 (9.2)	4 (33.3)		ref.	
Source of ascertainment: health insurance			0.944		
Y	106 (57.3)	7 (58.3)		0.96 (0.29-3.13)	0.944
N	79 (42.7)	5 (41.7)		ref	
Number of source of ascertainment			0.429		
1	24 (13.0)	3 (25.0)		1.0	
2	73 (39.4)	5 (41.7)		1.83 (0.41-8.21)	0.433
3	88 (47.6)	4 (33.3)		2.75 (0.58-13.13)	0.205
Use of riluzole[a]			0.006		
Y	143 (84.1)	5 (50)		5.30 (1.43-19.55)	0.012
N	27 (15.9)	5 (50)		ref	
Gastrostomy[a]			0.130		
Y	56 (32.9)	1 (10.0)		4.42 (0.55-35.77)	0.163
N	114 (67.1)	9 (90.0)		ref	
Diagnostic delay[b]			0.903		
Mean (sd)	10.13 (11.24)	8.74 (5.96)		1.02 (0.94-1.10)	0.669
Median (IQR)	7.5 (4.97-11.68)	7.30 (4.11-12.60)			
Score FRSR[c]			0.860		
Mean (sd)	35.70 (7.84)	35.45 (7.37)		1.00 (0.92-1.10)	0.929
Median (IQR)	37.2 (32-42)	36.6 (30.8-40.8)			
Weight variation at diagnosis (%)[d]			0.830		
Mean (sd)	-8.19 (9.00)	-7.42 (8.33)		0.99 (0.92-1.07)	0.803
Median (IQR)	-6.82 (-13.79; 0)	-5.45 (-11.84; 0)			
BMI[e]			0.523		

Table 8. Continued



Mean (sd)	24.75 (4. 57)	25.01 (2.28)		0.99 (0.86-1.13)	0.848
Median (IQR)	24.15 (21.37-27.21)	25.28 (23.42-26.52)			
Muscular test score[f]			0.566		0.485
Mean (sd)	125.74 (21.23)	120.56 (27.55)		1.01 (0.98-1.04)	
Median (IQR)	131 (111-144)	132 (121-137)			
Site symptom onset[g]			0.044		
Spinal and/or respiratory	116 (63.0)	11 (91.7)		1.0	
Bulbar	68 (37.0)	1 (8.3)		6.45 (0.81-51.04)	0.077
Airlie diagnostic criteria			0.569		
Possible	59 (31.9)	4 (33.3)		1.0	
Probable+ laboratory	25 (13.5)	0 (0.0)		NA	NA
Probable	68 (36.8)	5 (41.7)		0.92 (0.24-3.59)	0.907
definite	33 (17.8)	3 (25.0)		0.75 (0.16-3.54)	0.712
Familiar form			0.018		
N	162 (87.6)	7 (58.3)		1.0	
Y	8 (4.3)	2 (16.7)		0.17 (0.03-0.97)	0.046
DK	15 (8.1)	3 (25.0)		0.22 (0.05-0.92)	0.039
VNI[a]			0.529		
N	102 (60.0)	7 (70.0)		ref	
Y	68 (40.0)	3 (30.0)		1.56 (0.39-6.23)	0.532

*Fisher's exact test; [a]180 subjects; [b] 191 subjects; [c] 163 subjects; [d] 152 subjects; [e] 167 subjects; [f] 160 subjects; g 196 subjects

DC+ = correctly identified ALS diagnosis on death certificate

DC- = not identified ALS diagnosis on death certificate

DK= Don't know**; VNI**= Non-Invasive Ventilation placement

5. DISCUSSION

The experience based on the studies presented showed the values and issues in applying epidemiological methods to large collections of data.

5.1 Interpretation and contribution of the findings

5.1.1 MS environmental risk factors

Multiple sclerosis (MS) is an inflammatory and neurodegenerative demyelinating disease of the central nervous system (CNS) with peak of onset in young adulthood. The origin is likely autoimmune (Hemmer, Kerschensteiner et al. 2015) but the etiology is multifactorial and complex, involving both genetic and environmental factors.(Compston and Coles 2008)

Genetics likely contributes to the overall population susceptibility to MS. The HLA DRB1*1501 -DRB5*0101 allele is the most strongly associated in Caucasian (Dyment, Ebers et al. 2004), but over 100 gene variants are reported in genome-wide association studies. (Beecham, Patsopoulos et al. 2013, Sawcer, Franklin et al. 2014)

However the incomplete concordance among identical twins (Mumford, Wood et al. 1994, Willer, Dyment et al. 2003) and specific epidemiological patterns of MS, such as time trends in the incidence documented during the last decades (Pugliatti, Harbo et al. 2008), changes in risk among immigrants (Hammond, English et al. 2000), suggest an important role of lifestyle factors in disease initiation and modulation. (Compston and Coles 2008, Ascherio and Munger 2016)

The environmental factors mostly reported as associated with the risk of developing multiple sclerosis (MS) Epstein-Barr virus (EBV) infection, low vitamin D levels, cigarette smoking, obesity could play an important role during early life (from perinatal phase to adolescence) in predisposing to MS.(Pugliatti, Harbo et al. 2008, McLeod, Hammond et al. 2011, Ascherio 2013) The timing of the effect for these factors could be different and should be considered.(Handel, Giovannoni et al. 2010)

Results from migration studies suggest that an individual's risk of MS is determined during the first two decades of life (Gale and Martyn 1995), and space-time cluster studies provided further suggestions of early life as critical age for the influence of environmental factors. (Riise, Grønning et al. 1991, Pugliatti, Riise et al. 2006)

EBV infection is a strong risk factor even though the mechanism by which this infection increases the risk of MS is still not clearly defined. EBV infection later in life is associated with an increase of risk of MS. (Ascherio and Munger 2016) According to the hygiene hypothesis, multiple infectious exposures in early childhood could reduce the risk of MS.(Bach 2002) Thus other viral infections have been evaluated and some bacterial infections as well; in particular Chlamydophila Pneumoniae has been considered as potential risk factors. (Bagos, Nikolopoulos et al. 2006)

Another factor evaluated for long time is the vitamin D deficiency, due to the link with latitude, sunlight exposure, diet. (Bjornevik, Riise et al. 2014, Munger, Hongell et al. 2017) High levels of vitamin D are associated with decreased risk of MS, but it is not completely clear in which phase of life this effect could be stronger. (Mirzaei, Michels et al. 2011, Munger, Chitnis et al. 2011, Ascherio and Munger 2016, Munger, Aivo et al. 2016, Berezowska, Coe et al. 2019)

Partially related with low levels of vitamin D is also obesity (high body max index - BMI), which increases the risk of MS by two- three-fold. This factor as well had a stronger effect in children and adolescents. (Pereira-Santos, Costa et al. 2015, Amato, Derfuss et al. 2018)

Robust evidence supports the role of cigarette smoking as risk factor in MS onset with a dose-response effect. The findings are however not concordant on the most susceptible age period to the exposure, since part of the literature suggest adolescence and young adulthood as the most at risk for the negative effect on the development of MS. (Salzer and Sundström 2013, Ascherio and Munger 2016, Amato, Derfuss et al. 2018)

Other environmental, modifiable factors have been proposed, e.g. sodium, caffeine, alcohol, reproductive factors, stress and the gut microbiome, but evidence is still limited. (Ascherio and Munger 2016, Amato, Derfuss et al. 2018)

Among early life exposures, also breastfeeding and early nutrition deserve particular attention for MS risk.

Due to the health benefits of human breast milk, the World Health Organization currently recommends that infants should be exclusively breastfed for up to 6 months, with continued breastfeeding along with appropriate complementary foods up to two years of age or beyond. (WHO 2019) Breastfeeding has been of interest as a potential protective factor against MS. A beneficial effect would be in line with evidence suggesting a protective effect of breastfeeding against other autoimmune diseases (i.e. inflammatory bowel disease, celiac disease, type-1 diabetes).(Klement, Cohen et al. 2004, Alves, Figueiroa et al. 2012, Patelarou, Girvalaki et al. 2012, Pozo-Rubio, Olivares et al. 2012) However, the association between having been breastfed and the risk of developing adult-onset MS is not fully elucidated. Case-control studies have been performed in different populations with inconsistent findings. (Spencely and Dick 1982, Pisacane, Impagliazzo et al. 1994, Conradi, Malzahn et al. 2013, Ragnedda, Leoni et al. 2015, Dalla Costa, Romeo et al. 2019)

To our knowledge, **study I** was the first study of breastfeeding as a potential protective factor for the development of MS using data from community-based health and disease registries with ascertainment of breastfeeding exposure before disease occurrence in most cases. In this cohort of individuals followed from the perinatal phase, there was no association between breastfeeding and development of MS, even after considering different cutoffs for duration.

In **study II** we found an association between the prescription of antibiotics and increased risk of subsequent onset of MS, defined by neurologists, among inhabitants of the Emilia Romagna during the period 2002 e 2017. There was no evidence of a dose-effect relationship and the association showed similar effect estimates considering specific categories of antibiotics in addition to the whole antibiotics group. The result was confirmed also exploring different exposure time-lags for the whole class and for the subclasses of antibiotics.

5.1.2 Validity of DC in ALS

Study III showed that death certificates identifying subjects with a diagnosis of ALS in the French Limousin region had an overall good sensitivity. However the PPV for DC was low-moderate, and it increased during the time frame evaluated. There was no difference of the accuracy related to sex and age at death. A bulbar onset of symptoms, the use of riluzole, and the case ascertainment by a hospital decreased the probability of misdiagnosis in the DC, for

the specificity of the signs and symptoms recorded and the treatment prescribed. The increased accuracy for the age group of 65-75 years was likely related to the incidence peak of the disease for subjects around 70 years of age.
The possibility of absence of ALS diagnosis as the main and underlying cause of death on DCs highlights the need to use DC in combination with other administrative data to create algorithms with higher accuracy performances.

5.2 Strengths and challenges of observational studies using administrative data

Data-linkage studies (**studies I-III**) are included among observational designs, since the data collected in administrative databases (including exposure information) are prospectively recorded, independently from later disease occurrence.

Since the primary functions of administrative databases are different from research, it is fundamental for researchers to determine whether the information available from these databases is appropriate for answering their specific research questions. Thus some issues normally assessed in epidemiological studies also apply when the source of data is an administrative dataset.

Information bias

This error occurs during data collection. In administrative databases the incorrect recording of diagnoses, procedures and other information is a likely event, intrinsic to the process. There are indeed several steps between the event and its correct record on the dataset.

One of the most frequent is the misclassification of exposure or outcome.

When the misclassification is differently distributed between the groups (exposed/unexposed or cases/controls) this could lead to differential misclassification, affecting the estimates either away or towards the null value.

When the misclassification is the same across the groups that are studied and the estimates will be biased towards the null.(COPELAND, CHECKOWAY et al. 1977)

Misclassification of exposure

In **study II** the risk of misclassification could be present in considering the exposure to the prescription of antibiotics as proxy for a bacterial infectious disease.

Misclassification of this exposure could have acted on different levels. While considering the prescription of antibiotics as proxy of disease we are assuming that antibiotics are prescribed, with a defined dosage, to treat an infectious disease. However we don't have this information and thus we have to assume that the presence of an infection justified the antibiotics prescription. Most prescription claim databases do not have records of the reason for prescription (one exception is Norway), nor the potential adverse events or outcomes (e.g. adverse drug reactions voluntary database).

Moreover the records refer to the prescription of the drug. It is not possible to determine whether the patient actually took the drug during this period and whether he or she followed the indication (based on DDD). A blood test might be needed to prove that, which is impossible in this context.

Another situation to be considered is that subjects might have bought antibiotics without any prescription, thus the antibiotics exposure won't be recorded. This event is however limited.

These possible misclassifications for exposure however likely happened both in cases and controls, given that the recording of drug prescriptions is mandatory and automatic, and not related to the other comorbidities of the subject. This is the case of a non-differential misclassification and our findings would therefore be conservative due to a bias towards the null.

In **study I,** one of the possible issues was recall bias.

In most retrospective studies the information on breastfeeding is collected from the mothers only years later and after disease onset in their offspring, thus increasing the risk of differential misclassification of the exposure. Moreover, when the exposure information is not assessed directly from the mothers, but for example from the offspring, the risk of misclassification is higher.

In our prospective study, breastfeeding information was collected directly from the mothers, during surveys unrelated to the present research, and in most cases before any knowledge of an MS diagnosis of their offspring. This method limits the risk of bias, but in case there have been a misclassification it would have occurred equally in cases and controls, thus non-differential. An analysis including only mothers not knowledgeable about the disease status of the offspring at exposure assessment showed indeed no major changes in the results, thus excluding a bias in our study.

Misclassification of outcome

This issue could derive from the use of incorrect or unclear clinical documentation or a misdiagnosis at each level of the diagnostic process and recording. It depends on the modality the data are collected, e.g. more expensive medical procedures could be better documented since they would require a reimbursement.

Therefore it is important to confirm that the condition recorded is indeed a correct proxy or representative of the disease under investigation.

In **study II**, a proxy for MS diagnosis was used to identify subjects within the general population. The prescription of disease modifying drugs was considered as indicator of the diagnosis of MS. Differently from antibiotics, in this case the drug prescribed has more likely been actually taken, following the prescribed dosage, due to the specific modalities of prescription (special prescription from MS specialists only). Moreover, the drugs considered were specific for MS, thus it was unlikely that the prescription was done for a different disease. Those subjects identified through this method were excluded from the source of controls.

Change in definition (exposure and outcome)

The availability of long observational periods is one the strengths of the use of healthcare data, and it is often sought in studies where the outcomes are rare, or in prognostic studies. However, changes in clinical practices, cultural and social habits could occur.

Disease treatment or management guidelines, clinical definitions, the development and implementation of new diagnostic criteria, and the clinical coding system might change over time. Since this issue cannot be avoided, it is important to specify it in the research protocol.

In **study I**, breastfeeding habits were derived over a long period (1922-1986). During this time cultural habits, historical context, and political and societal rules, involved in orienting towards a suggested duration of breastfeeding, might have changed. In order to control for time trends, all analyses were adjusted for offspring year of birth. Moreover, breastfeeding information in our study did not refer to exclusive breastfeeding and therefore includes any combination of breastfeeding with other feeding habits. We thus considered different cutoffs for duration of breastfeeding, according to the literature and current suggestions from WHO.(WHO 2019) Norway is among the countries with highest rates of breastfeeding (recently reported: 71% at 6 months and 46% at 12 months) thanks to longer and remunerated maternity and paternity leave, and a generalization of the findings could be difficult.(Heiberg Endresen and Helsing 1995)

The diagnosis of MS in **study II** was done in different periods, since both incident and prevalent cases were included, and different diagnostic criteria were used at the time of the diagnosis. However, the inclusion of each patient was done by a neurologist, expert in MS care, thus with clinical confirmation.

Another aspect is the attitude in prescription of antibiotics, which might be different in different areas and historical periods. In this case antibiotics might be prescribed more often than in other countries (Vaccheri, Bjerrum et al. 2002), but this was true for both cases and controls.

In **study III**, in order to avoid the inclusion of different ICD coding revisions or diagnostic criteria, a limited time of observation was selected and the diagnostic criteria for the inclusion were specified in the study protocol, applied by the neurologist who evaluated all medical records.

Moreover, when comparing the findings with other literature, this variability should always be considered.

<u>Accuracy of measurement –validity</u>

To ensure the correct attribution of exposure and outcomes is mandatory to assess the accuracy and completeness of databases used.

Data quality and specific information recorded in administrative datasets may vary across different countries.(Sorensen, Sabroe et al. 1996, Iezzoni 1997, Takahashi, Nishida et al. 2012) Determining the accuracy of the data used is fundamental in order to allow the comparisons.

Different medical conditions require specific approaches, as described above.

In **study II** two different algorithms have been applied in order to identify persons with MS among the population in the administrative databases. A first approach considered subjects with the prescription of MS specific disease-modifying treatments as criteria to identify cases on the entire drug prescription RER database. It showed that 77% of cases in our study population were identified by the drug-based algorithm. When including in the algorithm also the ICD9 code for MS (ICD9 code: 340) in the diagnosis upon hospital discharge, the percentage of cases included that were identified was 91%.

In **study III** the main aim was the validation of death certificates as alternative source of identification of incident ALS cases in the Limousin region. The accuracy was evaluated with sensitivity and PPV, thus limiting the effect of low incidence of the disease. (BRENNER and GEFELLER 1997) The evaluation indicated that the DC data were not accurate enough to be used for case identification, but rather they should be used in combination with other administrative data to create more accurate algorithms.

Selection bias

Selection bias is an error due to the selection of a study population that does not represent the target population. (Ellenberg 1994) It is an issue when the comparison groups are identified using differing criteria (non-random) from the general source population and these criteria are related to the exposure status or disease of interest. (Delgado-Rodriguez and Llorca 2004)

This condition happens more likely in case-control studies that are retrospective and thus the recruitment is performed after the disease onset and could affect the probability of participating.

In **study II,** the case-control design is mitigated since it is nested in a general population. MS cases were identified from the MS clinics of the RER, but controls were selected and matched to the cases, from the same population.

The enrollment rate in this study was about 50% of the estimated cases in the region. There was different participation rate in different centers but no patient-related factors that could lead us to suspect a selection of cases, different depending on the center. Nonetheless more severe patients that were not able to go to the MS clinics might have not been seen and included in the study. Similarly, less severe cases might have not gone to the clinics for a follow up visit during the study period and therefore they were not invited to participate to the project. Thus less severe and more severe cases could be underrepresented in the study population.

Population-based cohort studies, considering the entire population, are however only partially protected from the risk of selection bias.

In **study I,** which is a community-based study, the original surveys (CONOR) did not include the entire Norwegian population and not all mothers in the survey reported information on breastfeeding. Consequently their offspring were excluded from our study.

However, the question on breastfeeding was included in each survey conducted in different years across different regions in the country. The surveys' participation could not be related to the exposure or the outcome in study. Moreover, the MS cases were those identified by the Norwegian MS registry, which at the time of linkage for this study had a coverage of about 70%.

Lack of covariates (confounders, effect modifiers)

The potential for confounding effects is intrinsic in the observational study design, since comparison groups might not be exchangeable with regard to other characteristics except for the exposure we are interested to study, as in experimental trials. (Rothman KJ, Greenland S et al. 2008) Thus exposure and outcome might have a common cause that alters the association.

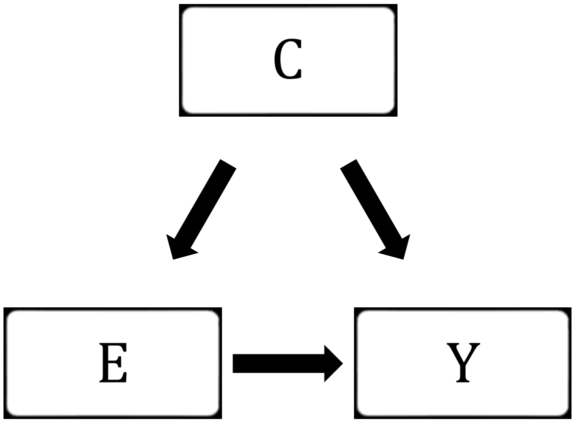


Figure 7. Causal Diagram of a Single Confounder: one outcome (Y), one exposure variable (E), and one confounder (C).

In administrative databases, information on possible confounders is usually incomplete, thus limiting the appropriate adjustments in the analyses (e.g. lifestyle factors, demographic features, comorbidities).

The cause of the missing data should also be defined, distinguishing between lack of data and absence of an event. (Lanes, Brown et al. 2015) The linkage of several databases and demographic data can help following subjects' medical history and integrate missing information.

Without this important information we might obtain precise estimates from the analysis of the large database, but confounded. Therefore we would give an altered interpretation of the truth.

The use of appropriate surrogate variables (proxy) can reduce this risk.

In **study I**, given the use of previously recorded data, we were not able to evaluate all potentially important confounding factors (i.e. socioeconomic status of the offspring in adulthood). However, we used maternal education as a proxy for socioeconomic status, and mother's smoking habits as proxy for offspring exposure to smoke, even though we were not able to define whether the individual was exposed prenatally or through passive smoking during infancy/childhood/adolescence.

Moreover applying linkage of records across different databases we had information on a series of other potential maternal and perinatal confounders similarly recorded.
Known environmental factors associated with MS should be considered as possible confounders but in **study II,** data on confounders were not available from the datasets used. Additional information could only be collected for MS cases that were questioned about their cigarette smoking habits. This feature was however not available for controls.

Privacy protection

Record linkage, fundamental for the correct connection between different sources of data, must be anonymous and individual, through the use of personal identifier (specific for each country or area) that should be the same in each database to allow the linkage process. In order to maintain the anonymity, the identifier will be removed after the linkage. (Gavrielov-Yusim and Friger 2014)

However the use of a unique identifier is not implemented in every country and, due to privacy reasons on personal health data, in some cases sensitive data might be provided only aggregated. (OECD 2013)

When individual level data are lacking, aggregated data can be used as proxy as well, or as such to give a different level of information (e.g. socio-demographic context of a neighborhood; air/water pollution) which could be useful in ecological studies.(van Walraven and Austin 2012, Gavrielov-Yusim and Friger 2014)

Temporality

The use of administrative data and registries can help to assess the correct temporality of the association.

Events and procedures in administrative databases are usually recorded with the specific dates. This information can confirm the adherence to the temporality criteria (Hill 1965) whenever the exposure is recorded before the outcome we are interested in.

However when neurological diseases are characterized by a subtle onset course, a defined onset might be difficult to identify.

In administrative databases, moreover, the disease onset is usually not recorded.

As a proxy for the diagnosis, the use of algorithm with hospital discharge codes, drug prescription, and payment exemption can be used. However it cannot be used as an indication of date of onset or diagnosis.

In the three studies (**I-II-III**) the information on outcome of interest (MS onset in **study I** and **II** and death for ALS) was recorded actively by the neurologists and collected in specific data collection (i.e. Norwegian MS registry (**I**), ad hoc data collection (**II**), FRALim registry (**III**)).

In **study I** breastfeeding could only happen during the first months of life, therefore many years before a potential onset of MS.

The exposure to antibiotics in **study II** could be traced back up to 13 years before the MS onset through the drug prescription data available, given the risk for protopathic bias, considering only a short exposure-lag.

This bias is related to the fact that the disease could be already present even though not yet apparent (before the clinical onset). This possibility is indeed relevant in multiple sclerosis, since at onset most cases have neuroradiological elements supporting an ongoing activity which began in an unknown time. The health status of the cases will be different from the general, healthy population. Thus subjects that will be diagnosed with MS might actually be more susceptible to infectious episodes and to use more antibiotics.

The observation of different lag-time of exposure can help in this situation, showing a constant effect regardless the lag period considered

Clinical significance versus statistical significance

Clinical significance and statistical significance are not always concordant.

This aspect is common to any epidemiological study. However, it is more noteworthy in the use of population-based healthcare data, given the often very large sample size. (van Walraven and Austin 2012, Gavrielov-Yusim and Friger 2014) This characteristic, which is an advantage in the study of rare diseases, should be kept in mind when interpreting the findings. Minimal effect sizes indeed can yield a statistical significance in large numbers samples, but clinical relevance must be demonstrated with, or sometimes more, than the statistical one. (van Walraven and Austin 2012, Gavrielov-Yusim and Friger 2014)

6. CONCLUSIONS AND PERSPECTIVES

The use of administrative data, linked with disease-specific data collections, offers an important opportunity for epidemiological research. The availability of great amount of healthcare data is not sufficient to give a better answer to a research question due to the risk of bias.

We have shown the application of different methodological approaches in population-based settings to healthcare databases. We assessed the association of possible risk factors as use of antibiotics or protective factors as having been breastfed, with the development of MS. The use of prerecorded data allowed the evaluation of the effect of these factors in different life periods, from perinatal phases to adulthood.

We also assessed the validity of death certificate in identifying the diagnosis of ALS.

In order to overcome intrinsic limitations of administrative databases on lack of important information, it could be useful to consider the application of more complete, disease-specific databases or registries, including all fundamental covariates (e.g. Vitamin D levels, or socio-economic status for MS), that could be considered as confounders or effect modifiers.

The application of neuroepidemiology to healthcare databases should therefore be an effort to maximize the signal over the noise. The large amount of data, indeed, may amplify the possibility of biases and thus, the use of the information needs to be considered carefully and their validity to be assessed in order to correctly interpret the findings.

7. REFERENCES

(2010). AHRQ Methods for Effective Health Care. Registries for Evaluating Patient Outcomes: A User's Guide. nd, R. E. Gliklich and N. A. Dreyer. Rockville (MD), Agency for Healthcare Research and Quality (US).

Aamodt, G., A. J. Søgaard, Ø. Næss, A. C. Beckstrøm and S. O. Samuelsen (2010). CONOR-databasen - et lite stykke Norge.

Al-Chalabi, A., F. Fang, M. F. Hanby, P. N. Leigh, C. E. Shaw, W. Ye and F. Rijsdijk (2010). "An estimate of amyotrophic lateral sclerosis heritability using twin data." J Neurol Neurosurg Psychiatry **81**(12): 1324-1326.

Al-Chalabi, A. and O. Hardiman (2013). "The epidemiology of ALS: a conspiracy of genes, environment and time." Nat Rev Neurol **9**(11): 617-628.

Alves, J. G. B., J. N. Figueiroa, J. Meneses and G. V. Alves (2012). "Breastfeeding protects against type 1 diabetes mellitus: a case–sibling study." Breastfeeding Medicine **7**(1): 25-28.

Amato, M. P., T. Derfuss, B. Hemmer, R. Liblau, X. Montalban, P. Soelberg Sørensen and D. H. Miller (2018). "Environmental modifiable risk factors for multiple sclerosis: Report from the 2016 ECTRIMS focused workshop." Multiple Sclerosis Journal **24**(5): 590-603.

Ascherio, A. (2013). "Environmental factors in multiple sclerosis." Expert Review of Neurotherapeutics **13**(sup2): 3-9.

Ascherio, A. and K. L. Munger (2016). "Epidemiology of Multiple Sclerosis: From Risk Factors to Prevention-An Update." Semin Neurol **36**(2): 103-114.

Bach, J. F. (2002). "The effect of infections on susceptibility to autoimmune and allergic diseases." N Engl J Med **347**(12): 911-920.

Barash, J. A., J. K. West and A. DeMaria, Jr. (2014). "Accuracy of administrative diagnostic data for pathologically confirmed cases of Creutzfeldt-Jakob disease in Massachusetts, 2000-2008." Am J Infect Control **42**(6): 659-664.

Beecham, A. H., N. A. Patsopoulos, D. K. Xifara, M. F. Davis, A. Kemppinen, C. Cotsapas, T. S. Shah, C. Spencer, D. Booth, A. Goris, A. Oturai, J. Saarela, B. Fontaine, B. Hemmer, C. Martin, F. Zipp, S. D'Alfonso, F. Martinelli-Boneschi, B. Taylor, H. F. Harbo, I. Kockum, J. Hillert, T. Olsson, M. Ban, J. R. Oksenberg, R. Hintzen, L. F. Barcellos, C. Agliardi, L. Alfredsson, M. Alizadeh, C. Anderson, R. Andrews, H. B. Sondergaard, A. Baker, G. Band, S.

E. Baranzini, N. Barizzone, J. Barrett, C. Bellenguez, L. Bergamaschi, L. Bernardinelli, A. Berthele, V. Biberacher, T. M. Binder, H. Blackburn, I. L. Bomfim, P. Brambilla, S. Broadley, B. Brochet, L. Brundin, D. Buck, H. Butzkueven, S. J. Caillier, W. Camu, W. Carpentier, P. Cavalla, E. G. Celius, I. Coman, G. Comi, L. Corrado, L. Cosemans, I. Cournu-Rebeix, B. A. Cree, D. Cusi, V. Damotte, G. Defer, S. R. Delgado, P. Deloukas, A. di Sapio, A. T. Dilthey, P. Donnelly, B. Dubois, M. Duddy, S. Edkins, I. Elovaara, F. Esposito, N. Evangelou, B. Fiddes, J. Field, A. Franke, C. Freeman, I. Y. Frohlich, D. Galimberti, C. Gieger, P. A. Gourraud, C. Graetz, A. Graham, V. Grummel, C. Guaschino, A. Hadjixenofontos, H. Hakonarson, C. Halfpenny, G. Hall, P. Hall, A. Hamsten, J. Harley, T. Harrower, C. Hawkins, G. Hellenthal, C. Hillier, J. Hobart, M. Hoshi, S. E. Hunt, M. Jagodic, I. Jelcic, A. Jochim, B. Kendall, A. Kermode, T. Kilpatrick, K. Koivisto, I. Konidari, T. Korn, H. Kronsbein, C. Langford, M. Larsson, M. Lathrop, C. Lebrun-Frenay, J. Lechner-Scott, M. H. Lee, M. A. Leone, V. Leppa, G. Liberatore, B. A. Lie, C. M. Lill, M. Linden, J. Link, F. Luessi, J. Lycke, F. Macciardi, S. Mannisto, C. P. Manrique, R. Martin, V. Martinelli, D. Mason, G. Mazibrada, C. McCabe, I. L. Mero, J. Mescheriakova, L. Moutsianas, K. M. Myhr, G. Nagels, R. Nicholas, P. Nilsson, F. Piehl, M. Pirinen, S. E. Price, H. Quach, M. Reunanen, W. Robberecht, N. P. Robertson, M. Rodegher, D. Rog, M. Salvetti, N. C. Schnetz-Boutaud, F. Sellebjerg, R. C. Selter, C. Schaefer, S. Shaunak, L. Shen, S. Shields, V. Siffrin, M. Slee, P. S. Sorensen, M. Sorosina, M. Sospedra, A. Spurkland, A. Strange, E. Sundqvist, V. Thijs, J. Thorpe, A. Ticca, P. Tienari, C. van Duijn, E. M. Visser, S. Vucic, H. Westerlind, J. S. Wiley, A. Wilkins, J. F. Wilson, J. Winkelmann, J. Zajicek, E. Zindler, J. L. Haines, M. A. Pericak-Vance, A. J. Ivinson, G. Stewart, D. Hafler, S. L. Hauser, A. Compston, G. McVean, P. De Jager, S. J. Sawcer and J. L. McCauley (2013). "Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis." Nat Genet **45**(11): 1353-1360.

Beghi, E., G. Logroscino, A. Micheli, A. Millul, M. Perini, R. Riva, F. Salmoiraghi and E. Vitelli (2001). "Validity of hospital discharge diagnoses for the assessment of the prevalence and incidence of amyotrophic lateral sclerosis." Amyotroph Lateral Scler Other Motor Neuron Disord **2**(2): 99-104.

Berezowska, M., S. Coe and H. Dawes (2019). "Effectiveness of Vitamin D Supplementation in the Management of Multiple Sclerosis: A Systematic Review." Int J Mol Sci **20**(6).

biobank, N. M.-r. o. (2017). "https://http://www.kvalitetsregistre.no/registers/505/resultater." Retrieved october 29th 2019.

Bjornevik, K., T. Riise, I. Casetta, J. Drulovic, E. Granieri, T. Holmoy, M. T. Kampman, A. M. Landtblom, K. Lauer, A. Lossius, S. Magalhaes, K. M. Myhr, T. Pekmezovic, K. Wesnes, C. Wolfson and M. Pugliatti (2014). "Sun exposure and multiple sclerosis risk in Norway and Italy: The EnvIMS study." Mult Scler **20**(8): 1042-1049.

Bogliun, G. and E. Beghi (2002). "Validity of hospital discharge diagnoses for public health surveillance of the Guillain-Barre syndrome." Neurol Sci **23**(3): 113-117.

Brandel, J. P., A. Welaratne, D. Salomon, I. Capek, V. Vaillant, A. Aouba, S. Haik and A. Alperovitch (2011). "Can mortality data provide reliable indicators for Creutzfeldt-Jakob disease surveillance? A study in France from 2000 to 2008." Neuroepidemiology **37**(3-4): 188-192.

BRENNER, H. and O. GEFELLER (1997). "VARIATION OF SENSITIVITY, SPECIFICITY, LIKELIHOOD RATIOS AND PREDICTIVE VALUES WITH DISEASE PREVALENCE." Statistics in Medicine **16**(9): 981-991.

Brooks, B. R., R. G. Miller, M. Swash and T. L. Munsat (2000). "El Escorial revisited: revised criteria for the diagnosis of amyotrophic lateral sclerosis." Amyotroph Lateral Scler Other Motor Neuron Disord **1**(5): 293-299.

Cai, R., Y. Zhang, J. E. Simmering, J. L. Schultz, Y. Li, I. Fernandez-Carasa, A. Consiglio, A. Raya, P. M. Polgreen, N. S. Narayanan, Y. Yuan, Z. Chen, W. Su, Y. Han, C. Zhao, L. Gao, X. Ji, M. J. Welsh and L. Liu (2019). "Enhancing glycolysis attenuates Parkinson's disease progression in models and clinical databases." J Clin Invest **129**(10): 4539-4549.

Chancellor, A. M., R. J. Swingler, H. Fraser, J. A. Clarke and C. P. Warlow (1993). "Utility of Scottish morbidity and mortality data for epidemiological studies of motor neuron disease." J Epidemiol Community Health **47**(2): 116-120.

Chio, A., G. Ciccone, A. Calvo, M. Vercellino, N. Di Vito, P. Ghiglione and R. Mutani (2002). "Validity of hospital morbidity records for amyotrophic lateral sclerosis. A population-based study." J Clin Epidemiol **55**(7): 723-727.

Chiò, A., G. Logroscino, B. J. Traynor, J. Collins, J. C. Simeone, L. A. Goldstein and L. A. White (2013). "Global epidemiology of amyotrophic lateral sclerosis: a systematic review of the published literature." Neuroepidemiology **41**(2): 118-130.

Chiò, A., C. Magnani, E. Oddenino, G. Tolardo and D. Schiffer (1992). "Accuracy of death certificate diagnosis of amyotrophic lateral sclerosis." Journal of Epidemiology and Community Health **46**(5): 517-518.

Chiò, A., C. Magnani and D. Schiffer (1995). "Gompertzian Analysis of Amyotrophic Lateral Sclerosis Mortality in Italy, 1957–1987; Application to Birth Cohorts." Neuroepidemiology **14**(6): 269-277.

Compston, A. and A. Coles (2008). "Multiple sclerosis." Lancet **372**(9648): 1502-1517.

Conradi, S., U. Malzahn, F. Paul, S. Quill, L. Harms, F. Then Bergh, A. Ditzenbach, T. Georgi, P. Heuschmann and B. Rosche (2013). "Breastfeeding is associated with lower risk for multiple sclerosis." Mult Scler **19**(5): 553-558.

Conti, S., M. Masocco, V. Toccaceli, M. Vichi, A. Ladogana, S. Almonti, M. Puopolo and M. Pocchiari (2005). "Mortality from human transmissible spongiform encephalopathies: a record linkage study." Neuroepidemiology **24**(4): 214-220.

COPELAND, K. T., H. CHECKOWAY, A. J. McMICHAEL and R. H. HOLBROOK (1977). "BIAS DUE TO MISCLASSIFICATION IN THE ESTIMATION OF RELATIVE RISK." American Journal of Epidemiology **105**(5): 488-495.

Dalla Costa, G., M. Romeo, F. Esposito, F. Sangalli, B. Colombo, M. Radaelli, L. Moiola, G. Comi and V. Martinelli (2019). "Caesarean section and infant formula feeding are associated with an earlier age of onset of multiple sclerosis." Multiple Sclerosis and Related Disorders **33**: 75-77.

Davenport , T. and R. Kalakota (2019). "The potential for artificial intelligence in healthcare. ." Future Healthc J **6**(2): 94–98.

Delgado-Rodriguez, M. and J. Llorca (2004). "Bias." J Epidemiol Community Health **58**(8): 635-641.

Dyment, D. A., G. C. Ebers and A. D. Sadovnick (2004). "Genetics of multiple sclerosis." Lancet Neurol **3**(2): 104-110.

Ehrenstein, V., H. Nielsen, A. B. Pedersen, S. P. Johnsen and L. Pedersen (2017). "Clinical epidemiology in the era of big data: new opportunities, familiar challenges." Clin Epidemiol **9**: 245-250.

Ellenberg, J. H. (1994). "Selection bias in observational and experimental studies." Stat Med **13**(5-7): 557-567.

Emilia-Romagna., R.-R. 2019, from https://statistica.regione.emilia-romagna.it/factbook/fb/popolazione/pop_res.

Foltynie, T. (2019). "Glycolysis as a therapeutic target for Parkinson's disease." Lancet Neurol **18**(12): 1072-1074.

Forbes, R. B., S. Colville, G. W. Cran and R. J. Swingler (2004). "Unexpected decline in survival from amyotrophic lateral sclerosis/motor neurone disease." Journal of Neurology, Neurosurgery &amp; Psychiatry **75**(12): 1753-1755.

Gale, C. R. and C. N. Martyn (1995). "Migrant studies in multiple sclerosis." Prog Neurobiol **47**(4-5): 425-448.

Gavrielov-Yusim, N. and M. Friger (2014). "Use of administrative medical databases in population-based research." Journal of Epidemiology and Community Health **68**(3): 283-287.

Gjerstorff, M. L. (2011). "The Danish Cancer Registry." Scand J Public Health **39**(7 Suppl): 42-45.

Hammond, S. R., D. R. English and J. G. McLeod (2000). "The age-range of risk of developing multiple sclerosis: evidence from a migrant population in Australia." Brain **123 (Pt 5)**: 968-974.

Handel, A. E., G. Giovannoni, G. C. Ebers and S. V. Ramagopalan (2010). "Environmental factors and their timing in adult-onset multiple sclerosis." Nat Rev Neurol **6**(3): 156-166.

Heiberg Endresen, E. and E. Helsing (1995). "Changes in breastfeeding practices in Norwegian maternity wards: national surveys 1973, 1982 and 1991." Acta paediatrica **84**: 719-724.

Hemmer, B., M. Kerschensteiner and T. Korn (2015). "Role of the innate and adaptive immune responses in the course of multiple sclerosis." Lancet Neurol **14**(4): 406-419.

Hill, A. B. (1965). "THE ENVIRONMENT AND DISEASE: ASSOCIATION OR CAUSATION?" Proceedings of the Royal Society of Medicine **58**(5): 295-300.

Hobson, P. and J. Meara (2018). "Mortality and quality of death certification in a cohort of patients with Parkinson's disease and matched controls in North Wales, UK at 18 years: a community-based cohort study." BMJ open **8**(2): e018969-e018969.

Iezzoni, L. I. (1997). "Assessing quality using administrative data." Ann Intern Med **127**(8 Pt 2): 666-674.

Ingre, C., P. M. Roos, F. Piehl, F. Kamel and F. Fang (2015). "Risk factors for amyotrophic lateral sclerosis." Clinical epidemiology **7**: 181-193.

Insee. (2013). "Institut national de la statistique et des études économiques. Estimation de la population au 1er Janvier par région, dèpartement (1975-2012), gender and age. ." 2019, from https://http://www.insee.fr/fr/statistiques.

Irgens, L. M. (2000). "The Medical Birth Registry of Norway. Epidemiological research and surveillance throughout 30 years." Acta Obstetricia et Gynecologica Scandinavica **79**(6): 435-439.

Irgens, L. M. (2012). "The origin of registry-based medical research and care." Acta Neurol Scand Suppl(195): 4-6.

IRGENS, L. M. and T. BJERKEDAL (1973). "Epidemiology of Leprosy in Norway: the History of The National Leprosy Registry of Norway from 1856 until today." International Journal of Epidemiology **2**(1): 81-89.

Jette, N., K. Atwood, M. Hamilton, R. Hayward, L. Day, T. Mobach, C. Maxwell, C. M. Fortin, G. Fiebelkorn, K. Barlow, M. Shevell, M. K. Kapral, S. Casha, M. Johnston, S. Wiebe, L. Korngut and T. Pringsheim (2013). "Linkage between neurological registry data and administrative data." Can J Neurol Sci **40**(4 Suppl 2): S32-34.

Jiang, G. X., J. de Pedro-Cuesta and S. Fredrikson (1995). "Guillain-Barré syndrome in South-West Stockholm, 1973–1991, 1. Quality of registered hospital diagnoses and incidence." Acta Neurologica Scandinavica **91**(2): 109-117.

Johansen, C. and J. H. Olsen (1998). "Mortality from amyotrophic lateral sclerosis, other chronic disorders, and electric shocks among utility workers." Am J Epidemiol **148**(4): 362-368.

Johnson, E. K. and C. P. Nelson (2013). "Values and pitfalls of the use of administrative databases for outcomes assessment." The Journal of urology **190**(1): 17-18.

Jougla, E., G. Pavillon, F. Rossollin, M. De Smedt and J. Bonte (1998). "Improvement of the quality and comparability of causes-of-death statistics inside the European Community. EUROSTAT Task Force on "causes of death statistics"." Rev Epidemiol Sante Publique **46**(6): 447-456.

Jutte, D., L. Roos and M. Brownell (2011). "Administrative record linkage as a tool for public health research." Annu Rev Public Health **32**: 91-108.

Kieler, H., M. Artama, A. Engeland, O. Ericsson, K. Furu, M. Gissler, R. Nielsen, M. Nørgaard, O. Stephansson, U. Valdimarsdottir, H. Zoega and B. Haglund (2012). "Selective serotonin reuptake inhibitors during pregnancy and risk of persistent pulmonary hypertension in the newborn: population based cohort study from the five Nordic countries. ." BMJ **344**: d8012.

Kioumourtzoglou, M.-A., R. M. Seals, L. Himmerslev, O. Gredal, J. Hansen and M. G. Weisskopf (2015). "Comparison of diagnoses of amyotrophic lateral sclerosis by use of death certificates and hospital discharge data in the Danish population." Amyotrophic lateral sclerosis & frontotemporal degeneration **16**(3-4): 224-229.

Klement, E., R. V. Cohen, J. Boxman, A. Joseph and S. Reif (2004). "Breastfeeding and risk of inflammatory bowel disease: a systematic review with meta-analysis." The American journal of clinical nutrition **80**(5): 1342-1352.

Lanes, S., J. S. Brown, K. Haynes, M. F. Pollack and A. M. Walker (2015). "Identifying health outcomes in healthcare databases." Pharmacoepidemiology and Drug Safety **24**(10): 1009-1016.

Logroscino, G., B. J. Traynor, O. Hardiman, A. Chio, P. Couratier, J. D. Mitchell, R. J. Swingler and E. Beghi (2008). "Descriptive epidemiology of amyotrophic lateral sclerosis: new evidence and unsolved issues." J Neurol Neurosurg Psychiatry **79**(1): 6-11.

Marin, B., F. Boumédiene, G. Logroscino, P. Couratier, M.-C. Babron, A. L. Leutenegger, M. Copetti, P.-M. Preux and E. Beghi (2016). "Variation in worldwide incidence of amyotrophic lateral sclerosis: a meta-analysis." International Journal of Epidemiology **46**(1): 57-74.

Marin, B., P. Couratier, P. M. Preux and G. Logroscino (2011). "Can Mortality Data Be Used to Estimate Amyotrophic Lateral Sclerosis Incidence?" Neuroepidemiology **36**(1): 29-38.

Marin, B., B. Hamidou, P. Couratier, M. Nicol, A. Delzor, M. Raymondeau, M. Druet-Cabanac, G. Lautrette, F. Boumediene and P. M. Preux (2014). "Population-based epidemiology of amyotrophic lateral sclerosis (ALS) in an ageing Europe – the French register of ALS in Limousin (FRALim register)." European Journal of Neurology **21**(10): 1292-e1279.

Marin, B., B. Hamidou, P. Couratier, M. Nicol, A. Delzor, M. Raymondeau, M. Druet-Cabanac, G. Lautrette, F. Boumediene and P. M. Preux (2014). "Population-based epidemiology of amyotrophic lateral sclerosis (ALS) in an ageing Europe--the French register of ALS in Limousin (FRALim register)." Eur J Neurol **21**(10): 1292-1300, e1278-1299.

Mazzali, C. and P. Duca (2015). "Use of administrative data in healthcare research." Intern Emerg Med **10**(4): 517-524.

McDonald, W. I., A. Compston, G. Edan, D. Goodkin, H. P. Hartung, F. D. Lublin, H. F. McFarland, D. W. Paty, C. H. Polman, S. C. Reingold, M. Sandberg-Wollheim, W. Sibley, A. Thompson, S. van den Noort, B. Y. Weinshenker and J. S. Wolinsky (2001). "Recommended diagnostic criteria for multiple sclerosis: guidelines from the International Panel on the diagnosis of multiple sclerosis." Ann Neurol **50**(1): 121-127.

McLeod, J. G., S. R. Hammond and J. F. Kurtzke (2011). "Migration and multiple sclerosis in immigrants to Australia from United Kingdom and Ireland: a reassessment. I. Risk of MS by age at immigration." Journal of Neurology **258**(6): 1140-1149.

Mirzaei, F., K. B. Michels, K. Munger, E. O'Reilly, T. Chitnis, M. R. Forman, E. Giovannucci, B. Rosner and A. Ascherio (2011). "Gestational vitamin D and the risk of multiple sclerosis in offspring." Ann Neurol **70**(1): 30-40.

Mittal, S., K. Bjornevik, D. S. Im, A. Flierl, X. Dong, J. J. Locascio, K. M. Abo, E. Long, M. Jin, B. Xu, Y. K. Xiang, J. C. Rochet, A. Engeland, P. Rizzu, P. Heutink, T. Bartels, D. J. Selkoe, B. J. Caldarone, M. A. Glicksman, V. Khurana, B. Schule, D. S. Park, T. Riise and C. R. Scherzer (2017). "beta2-Adrenoreceptor is a regulator of the alpha-synuclein gene driving risk of Parkinson's disease." Science **357**(6354): 891-898.

Muggah, E., E. Graves, C. Bennett and D. G. Manuel (2013). "Ascertainment of chronic diseases using population health data: a comparison of health administrative data and patient self-report." BMC Public Health **13**(1): 16.

Mumford, C. J., N. W. Wood, H. Kellar-Wood, J. W. Thorpe, D. H. Miller and D. A. Compston (1994). "The British Isles survey of multiple sclerosis in twins." Neurology **44**(1): 11-15.

Munger, K. L., J. Aivo, K. Hongell, M. Soilu-Hanninen, H. M. Surcel and A. Ascherio (2016). "Vitamin D Status During Pregnancy and Risk of Multiple Sclerosis in Offspring of Women in the Finnish Maternity Cohort." JAMA Neurol **73**(5): 515-519.

Munger, K. L., T. Chitnis, A. L. Frazier, E. Giovannucci, D. Spiegelman and A. Ascherio (2011). "Dietary intake of vitamin D during adolescence and risk of multiple sclerosis." J Neurol **258**(3): 479-485.

Munger, K. L., K. Hongell, J. Aivo, M. Soilu-Hanninen, H. M. Surcel and A. Ascherio (2017). "25-Hydroxyvitamin D deficiency and risk of MS among women in the Finnish Maternity Cohort." Neurology **89**(15): 1578-1583.

Myhr, K.-M., N. Grytten, Ø. Torkildsen, S. Wergeland, L. Bø and J. H. Aarseth (2015). "The Norwegian Multiple Sclerosis Registry and Biobank." Acta Neurologica Scandinavica **132**(S199): 24-28.

Naess, O., A. J. Sogaard, E. Arnesen, A. C. Beckstrom, E. Bjertness, A. Engeland, P. F. Hjort, J. Holmen, P. Magnus, I. Njolstad, G. S. Tell, L. Vatten, S. E. Vollset and G. Aamodt (2008). "Cohort profile: cohort of Norway (CONOR)." Int J Epidemiol **37**(3): 481-485.

NIPH. (2019). "Medical Birth Registry of Norway." Retrieved 13.11.2019, 2019, from https://http://www.fhi.no/en/hn/health-registries/medical-birth-registry-of-norway/.

Nissen, F., J. K. Quint, D. R. Morales and I. J. Douglas (2019). "How to validate a diagnosis recorded in electronic health records." Breathe **15**(1): 64-68.

OECD (2013). Strengthening Health Information Infrastructure for HealthCare Quality Governance: Good Practices, New Opportunities and Data Privacy Protection Challenges , OECD Publishing.

Patelarou, E., C. Girvalaki, H. Brokalaki, A. Patelarou, Z. Androulaki and C. Vardavas (2012). "Current evidence on the associations of breastfeeding, infant formula, and cow's milk introduction with type 1 diabetes mellitus: a systematic review." Nutrition Reviews **70**(9): 509-519.

Pereira-Santos, M., P. R. Costa, A. M. Assis, C. A. Santos and D. B. Santos (2015). "Obesity and vitamin D deficiency: a systematic review and meta-analysis." Obes Rev **16**(4): 341-349.

Peter, I., M. Dubinsky, S. Bressman, A. Park, C. Lu, N. Chen and A. Wang (2018). "Anti-Tumor Necrosis Factor Therapy and Incidence of Parkinson Disease Among Patients With Inflammatory Bowel Disease." JAMA Neurol **75**(8): 939-946.

Pisacane, A., N. Impagliazzo, M. Russon, R. Valiani, A. Mandarini, C. Florio and P. Vivo (1994). "Breast feeding and multiple sclerosis." BMJ **308**(6941): 1411-1412.

Polman, C. H., S. C. Reingold, B. Banwell, M. Clanet, J. A. Cohen, M. Filippi, K. Fujihara, E. Havrdova, M. Hutchinson, L. Kappos, F. D. Lublin, X. Montalban, P. O'Connor, M. Sandberg-Wollheim, A. J. Thompson, E. Waubant, B. Weinshenker and J. S. Wolinsky (2011). "Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria." Ann Neurol **69**(2): 292-302.

Polman, C. H., S. C. Reingold, G. Edan, M. Filippi, H. P. Hartung, L. Kappos, F. D. Lublin, L. M. Metz, H. F. McFarland, P. W. O'Connor, M. Sandberg-Wollheim, A. J. Thompson, B. G. Weinshenker and J. S. Wolinsky (2005). "Diagnostic criteria for multiple sclerosis: 2005 revisions to the "McDonald Criteria"." Ann Neurol **58**(6): 840-846.

Poser, C. M., D. W. Paty, L. Scheinberg, W. I. McDonald, F. A. Davis, G. C. Ebers, K. P. Johnson, W. A. Sibley, D. H. Silberberg and W. W. Tourtellotte (1983). "New diagnostic criteria for multiple sclerosis: guidelines for research protocols." Ann Neurol **13**(3): 227-231.

Pozo-Rubio, T., M. Olivares, E. Nova, G. De Palma, J. R. Mujico, M. D. Ferrer, A. Marcos and Y. Sanz (2012). "Immune development and intestinal microbiota in celiac disease." Clin Dev Immunol **2012**: 654143.

Pugliatti, M., H. F. Harbo, T. Holmoy, M. T. Kampman, K. M. Myhr, T. Riise and C. Wolfson (2008). "Environmental risk factors in multiple sclerosis." Acta Neurol Scand Suppl **188**: 34-40.

Pugliatti, M., T. Riise, M. A. Sotgiu, W. M. Satta, S. Sotgiu, M. I. Pirastru and G. Rosati (2006). "Evidence of early childhood as the susceptibility period in multiple sclerosis: space-time cluster analysis in a Sardinian population." Am J Epidemiol **164**(4): 326-333.

Pupillo, E., P. Messina, G. Logroscino and E. Beghi (2014). "Long-term survival in amyotrophic lateral sclerosis: a population-based study." Ann Neurol **75**(2): 287-297.

Ragnedda, G., S. Leoni, M. Parpinel, I. Casetta, T. Riise, K. M. Myhr, C. Wolfson and M. Pugliatti (2015). "Reduced duration of breastfeeding is associated with a higher risk of multiple sclerosis in both Italian and Norwegian adult males: the EnvIMS study." J Neurol **262**(5): 1271-1277.

Ragonese, P., G. Filippini, G. Salemi, E. Beghi, A. Citterio, R. D'Alessandro, C. Marini, M. R. Monsurrò, I. Aiello, L. Morgante, A. Tempestini, C. Fratti, M. Ragno, M. Pugliatti, A. Epifanio, D. Testa and G. Savettieri (2004). "Accuracy of Death Certificates for Amyotrophic Lateral Sclerosis Varies Significantly from North to South of Italy: Implications for Mortality Studies." Neuroepidemiology **23**(1-2): 73-77.

Registries, T. C.-C. T. F. o. P. and EMA. (2018). "Discussion paper:
Use of patient disease registries for regulatory purposes – methodological and operational considerations." 2019, from https://http://www.ema.europa.eu/en/human-regulatory/post-authorisation/patient-registries.

Riise, T., M. Grønning, M. R. Klauber, E. Barrett-Connor, H. Nyland and G. Albrektsen (1991). "Clustering of Residence of Multiple Sclerosis Patients at Age 13 to 20 Years in Hordaland, Norway." American Journal of Epidemiology **133**(9): 932-939.

Rothman KJ, Greenland S and L. TL (2008). Modern Epidemiology. Philadelphia, PA.

Salzer, J. and P. Sundström (2013). "Timing of cigarette smoking as a risk factor for multiple sclerosis." Therapeutic advances in neurological disorders **6**(3): 205-205.

Sawcer, S., R. J. Franklin and M. Ban (2014). "Multiple sclerosis genetics." Lancet Neurol **13**(7): 700-709.

Sejvar, J., A. Baughman, M. Wise and O. Morgan (2011). "Population Incidence of Guillain-Barré Syndrome: A Systematic Review and Meta-Analysis." Neuroepidemiology **36**: 123-133.

SISEPS. (2019). "Assistenza Farmaceutica." from http://salute.regione.emilia-romagna.it/siseps/sanita/assistenza-farmaceutica.

SISEPS. (2019). "Schede di dimissione ospedaliera." from http://salute.regione.emilia-romagna.it/siseps/sanita/sdo.

Sorensen, H. T. (1997). "Regional administrative health registries as a resource in clinical epidemiologyA study of options, strengths, limitations and data quality provided with examples of use." Int J Risk Saf Med **10**(1): 1-22.

Sorensen, H. T., S. Sabroe and J. Olsen (1996). "A framework for evaluation of secondary data sources for epidemiological research." Int J Epidemiol **25**(2): 435-442.

Spencely, M. and G. Dick (1982). "Breast-Feeding and Multiple Sclerosis." Neuroepidemiology **1**(4): 216-222.

St Germaine-Smith, C., A. Metcalfe, T. Pringsheim , J. Roberts, C. Beck , B. Hemmelgarn, J. McChesney, H. Quan and N. Jette (2012). "Recommendations for optimal ICD codes to study neurologic conditions: a systematic review." Neurology **79**(10): 1049-1055.

St Sauver, J., B. Grossardt, B. Yawn, L. r. Melton, J. Pankratz, S. Brue and W. Rocca (2012). "Data resource profile: the Rochester Epidemiology Project (REP) medical records-linkage system." Int J Epidemiol **41**(6): 1614-1624.

Takahashi, Y., Y. Nishida and S. Asai (2012). "Utilization of health care databases for pharmacoepidemiology." European Journal of Clinical Pharmacology **68**(2): 123-129.

Thompson, A. J., B. L. Banwell, F. Barkhof, W. M. Carroll, T. Coetzee, G. Comi, J. Correale, F. Fazekas, M. Filippi, M. S. Freedman, K. Fujihara, S. L. Galetta, H. P. Hartung, L. Kappos, F. D. Lublin, R. A. Marrie, A. E. Miller, D. H. Miller, X. Montalban, E. M. Mowry, P. S. Sorensen, M. Tintore, A. L. Traboulsee, M. Trojano, B. M. J. Uitdehaag, S. Vukusic, E. Waubant, B. G. Weinshenker, S. C. Reingold and J. A. Cohen (2018). "Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria." Lancet Neurol **17**(2): 162-173.

Tirschwell, D. L. and W. T. Longstreth, Jr. (2002). "Validating administrative data in stroke research." Stroke **33**(10): 2465-2470.

Vaccheri, A., L. Bjerrum, D. Resi, U. Bergman and N. Montanaro (2002). "Antibiotic prescribing in general practice: striking differences between Italy (Ravenna) and Denmark (Funen)." J Antimicrob Chemother. **50**(6): 989-997.

van Walraven, C. and P. Austin (2012). "Administrative database research has unique characteristics that can risk biased results." Journal of Clinical Epidemiology **65**(2): 126-131.

van Walraven, C., C. Bennett and A. J. Forster (2011). "Administrative database research infrequently used validated diagnostic or procedural codes." J Clin Epidemiol **64**(10): 1054-1059.

Vasta, R., F. Boumediene, P. Couratier, M. Nicol, A. Nicoletti, P. M. Preux and B. Marin (2017). "Validity of medico-administrative data related to amyotrophic lateral sclerosis in France: A population-based study." Amyotroph Lateral Scler Frontotemporal Degener **18**(1-2): 24-31.

WHO. (2019). "https://http://www.whocc.no/atc_ddd_index." 2019.

WHO, W. H. O. (2019). Breastfeeding.

Willer, C. J., D. A. Dyment, N. J. Risch, A. D. Sadovnick and G. C. Ebers (2003). "Twin concordance and sibling recurrence rates in multiple sclerosis." Proc Natl Acad Sci U S A **100**(22): 12877-12882.

Xu, H., M. C. Aldrich, Q. Chen, H. Liu, N. B. Peterson, Q. Dai, M. Levy, A. Shah, X. Han, X. Ruan, M. Jiang, Y. Li, J. S. Julien, J. Warner, C. Friedman, D. M. Roden and J. C. Denny (2014). "Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality." Journal of the American Medical Informatics Association **22**(1): 179-191.

Zoccolella, S., E. Beghi, G. Palagano, A. Fraddosio, V. Guerra, V. Samarelli, V. Lepore, I. L. Simone, P. Lamberti, L. Serlenga and G. Logroscino (2008). "Predictors of long survival in amyotrophic lateral sclerosis: a population-based study." J Neurol Sci **268**(1-2): 28-32.