# A new way to estimate neurologic disease prevalence in the United States

Illustrated with MS

Lorene M. Nelson, PhD, MS, Mitchell T. Wallin, MD, MPH, Ruth Ann Marrie, MD, PhD, W.J. Culpepper, PhD, Annette Langer-Gould, MD, PhD, Jon Campbell, PhD, Stephen Buka, ScD, Helen Tremlett, PhD, Gary Cutter, PhD, Wendy Kaye, PhD, Laurie Wagner, MPH, and Nicholas G. Larocca, PhD, for the United States Multiple Sclerosis Prevalence Workgroup

**Correspondence**
Dr. Nelson
lnelson@stanford.edu

## Abstract

### Objective

Considerable gaps exist in knowledge regarding the prevalence of neurologic diseases, such as multiple sclerosis (MS), in the United States. Therefore, the MS Prevalence Working Group sought to review and evaluate alternative methods for obtaining a scientifically valid estimate of national MS prevalence in the current health care era.

### Methods

We carried out a strengths, weaknesses, opportunities, and threats (SWOT) analysis for 3 approaches to estimate MS prevalence: population-based MS registries, national probability health surveys, and analysis of administrative health claims databases. We reviewed MS prevalence studies conducted in the United States and critically examined possible methods for estimating national MS prevalence.

### Results

We developed a new 4-step approach for estimating MS prevalence in the United States. First, identify administrative health claim databases covering publicly and privately insured populations in the United States. Second, develop and validate a highly accurate MS case-finding algorithm that can be standardly applied in all databases. Third, apply a case definition algorithm to estimate MS prevalence in each population. Fourth, combine MS prevalence estimates into a single estimate of US prevalence, weighted according to the number of insured persons in each health insurance segment.

### Conclusions

By addressing methodologic challenges and proposing a new approach for measuring the prevalence of MS in the United States, we hope that our work will benefit scientists who study neurologic and other chronic conditions for which national prevalence estimates do not exist.

From the Division of Epidemiology, Department of Health Research and Policy (L.M.N.), Stanford University School of Medicine, Stanford, CA; Department of Veterans Affairs Multiple Sclerosis Center of Excellence (VA MSCoE) and Georgetown University School of Medicine (M.T.W.), Washington, DC; Department of Internal Medicine (R.A.M.), Max Rady College of Medicine, Rady Faculty of Health Sciences, University of Manitoba, Winnipeg, Canada; VA MS Center of Excellence and University of Maryland (W.J.C.), Baltimore; Neurology Department (A.L.-G.), Kaiser Permanente Southern California, Los Angeles; University of Colorado (J.C.), Denver; Brown University (S.B.), Providence, RI; University of British Columbia (H.T.), Vancouver, Canada; University of Alabama at Birmingham (G.C.); McKing Consulting Corporation (W.K., L.W.), Atlanta, GA; and National Multiple Sclerosis Society (N.G.L.), New York, NY.

# Glossary

**AHC** = administrative health claims; **BRFSS** = Behavioral Risk Factor Surveillance Survey; **CI** = confidence interval; **DMT** = disease-modifying treatment; **DOD** = Department of Defense; **ICD-9** = International Classification of Diseases–9; **MEPS-HC** = Household Component of the Medical Expenditure Panel Survey; **MS** = multiple sclerosis; **NHIS** = National Health Interview Survey; **SWOT** = strengths, weaknesses, opportunities, and threats; **VHA** = Veterans Health Administration.

Multiple sclerosis (MS) is the most common chronic progressive neurologic disease of young adults,[1,2] affecting individuals in their most productive years, and placing a heavy burden on affected persons, their family members, and the health care system. Unlike some other world regions, considerable gaps exist in knowledge regarding the incidence and prevalence of MS in the United States because there is no robust method for estimating either epidemiologic statistic on a national basis. The only population-based estimates of MS incidence (the number of new cases of MS in a population) have been obtained in the population of Rochester, Minnesota, where the presence of the Mayo Clinic enables rigorous periodic studies of MS incidence and prevalence.[3] The absence of a unified single health care system in the US does not enable robust estimates of MS incidence on a national level; therefore, our focus is on estimating national MS prevalence in the United States.

Obtaining an accurate estimate of MS prevalence in the United States is critical because prevalence supports comparisons of MS burden across demographic subgroups, geographic regions, and time periods. Health care providers and policymakers rely on prevalence estimates to make decisions about future health service needs. Researchers studying MS need information on disease prevalence for evaluating changes in disease frequency in relation to changing population demographics.

MS is not a reportable condition, and there is no national population-based registry for MS. Although many studies examining the prevalence of MS in select US communities have been conducted in the past 25 years,[4] the variable methodology and quality of these studies prevent the data from being aggregated into a single robust national estimate of MS in the United States.

In 2014, the National Multiple Sclerosis Society established the MS Prevalence Working Group, a group composed of scientists (neurologists, epidemiologists, statisticians) and policy advocates from the United States and Canada, to obtain a valid estimate of MS prevalence in the United States. We sought to evaluate possible methods for estimating MS prevalence in the current health care era with the ultimate goal of providing a scientifically sound estimate of the overall prevalence of MS in the United States.

We review MS prevalence studies conducted in the United States, critically examine possible methods for estimating national MS prevalence, and use this information to develop a rationale for a new 4-step paradigm for estimating MS prevalence in the United States.

# Previous estimates of MS prevalence in the United States

## Regional MS prevalence studies

A recent systematic review[4] identified 9 population-based MS prevalence studies conducted in small regions of the United States between 1985 and 2011 (table e-1; https://doi.org/10.5061/dryad.t1k42p8). MS point prevalence was lowest in the 19 counties surrounding Lubbock, Texas (age-standardized prevalence 39.9/100,000; 95% confidence interval [CI] 34.0–45.7)[5] and highest in Olmstead County, Minnesota (age-standardized prevalence 191.2/100,000; 95% CI 165.6–216.8).[3] A noteworthy advantage of such regional studies is that the relatively small population size enables investigators to use many overlapping methods to ascertain possible MS cases in the study population (case ascertainment) and apply standardized criteria to determine which cases have MS (case definition). Regional MS prevalence studies have used a variety of methods to maximize the completeness of case ascertainment, with the most common sources being specialty clinics or academic MS centers, community neurologists or other physicians, patient service associations, nursing homes or long-term care facilities, and death certificates.[4] Case definitions are usually established using medical chart reviews abstracted to apply standard diagnostic criteria.[6] However, because MS prevalence can have considerable geographic variation,[4] projecting estimates from one or a few regions to the entire United States is of uncertain validity.

## National MS prevalence studies in the United States

For a country the size of the United States (~310 million population in 2010), it is not feasible to use the intensive case ascertainment methods that are routine in smaller geographic regions. In the last 4 decades, 4 studies have estimated the prevalence of MS in the United States by using methods intended to provide representative samples of the US population (table 1). The first study was conducted in 1976 using a probability sample of 725 acute short-term general hospitals and 8,800 physicians in the contiguous United States and surveying them to estimate the number of persons with MS with whom they had contact.[7] MS prevalence was estimated to be 58 per 100,000 population, corresponding to 123,000

**Table 1** Epidemiologic studies of multiple sclerosis (MS) prevalence in the United States using probability sample surveys

| Reference: Authors (year) | Study years | Case definition | Population | National MS prevalence per 100,000 population | Number of persons with MS in the United States (95% CI) |
|---|---|---|---|---|---|
| Baum and Rothschild (1981)[7] | 1976 | Survey of physicians and hospitals, asked to report patients meeting uniform criteria for probable and possible MS | National sample of 8,800 physicians and 725 hospitals, a probability sample of the 1976 US population | 58 per 100,000 (all ages) | 123,000 |
| Collins (1997)[10] | 1990–1992 | Patients or family members reporting physician-diagnosed MS | NHIS, a household probability sample of US population | 70 per 100,000 (all ages) | 180,000 |
| Noonan et al. (2002)[11] | 1982–1996 | Patients or family members reporting physician-diagnosed MS | NHIS, a household probability sample of US population | 85 per 100,000 (all ages) | 211,000 (191,000–231,000) |
| Campbell et al. (2014)[12] | 2008–2009 | Patients having 1 or more medical claims with MS diagnostic code 340 during a single year | MEPS-HC, a household probability sample of US population | 191 per 100,000 (≥18 years) | 572,312 (397,004–747,619) |

Abbreviations: CI = confidence interval; MEPS-HC = Household Component of the Medical Expenditure Panel Survey; NHIS = National Health Interview Survey.

cases in the United States, but Anderson et al.[8] believed that the prevalence of MS was underestimated due to several methodologic factors. A primary source of underascertainment was relying on physician recall to identify MS cases, as at the time these studies were conducted, computerized medical records systems were rare. Another source of underascertainment was the single-year surveillance window, which would have missed patients with MS who did not seek medical attention for MS during that year. Evidence for a third source of underascertainment came from a prevalence survey conducted in the Weld and Larimer Counties of northern Colorado,[9] which used multiple case ascertainment methods and found that 21% of all prevalent cases identified came exclusively from the records of 3 MS service organizations in Colorado, a source that was not used in the 1976 study.[7] Anderson et al.[8] used several adjustments for these sources of underascertainment to provide an updated estimate of MS prevalence in the United States in 1990, ranging from 250,000 to 350,000 persons.

The next 2 national MS prevalence studies were conducted using the National Health Interview Survey (NHIS), an annual probability sample of the noninstitutionalized US population. During 1982–1996, one-sixth of the NHIS sample was surveyed about MS and other health conditions resulting in limitation of activity. Collins[10] used NHIS data for a 3-year period (1990–1992) to estimate national MS prevalence because a single year of data did not yield enough persons with MS for stable estimation. MS prevalence was estimated to be 70 per 100,000, corresponding to an estimated count of 180,000 cases. A second study also used NHIS data, over 1982–1996.[11] MS prevalence was estimated to be 85 per 100,000 population, corresponding to an estimated count of 211,000 cases in the United States (95% CI 191,000–231,000). A major weakness of the NHIS studies is that survey data were collected from one family member within each sampled household who reported the medical conditions for the entire household. This method

would underestimate MS prevalence if household members underreport the condition but could also overestimate MS prevalence if self-reported data resulted in the inclusion of provisional cases (or noncases) who would not meet formal case definition criteria.

More recently, a national MS prevalence study was carried out using the Household Component of the Medical Expenditure Panel Survey (MEPS-HC), a nationally representative survey of the US civilian noninstitutionalized population drawn from NHIS that collects data on health status, health utilization, health costs, and health insurance coverage.[12] MS cases were identified on the basis of having one or more medically coded diagnoses of MS (ICD-9 code 340). Between 1998 and 2009, 526 MEPS-HC participants met this criterion, and the annual average number of persons with MS was estimated to be 572,312 (95% CI 397,004–747,619). Requiring only 1 ICD-9 code during the study year could have overestimated MS prevalence if false-positives were included because individuals who had a single ICD-9 340 code were ultimately not diagnosed with MS or the code was entered erroneously. The MEPS-HC probability sample only captured noninstitutionalized persons with MS, which could have resulted in underascertainment because it excludes the estimated 5%–10% of persons with MS residing in nursing homes.[13,14]

Two studies have used administrative health claims (AHC) data to estimate the number of persons with MS. Both studies assessed subpopulations of the United States and then used this information to estimate the total number of persons with MS in the United States as a whole (table 2). Pope et al.[15] obtained administrative claims data from privately insured individuals, those aged into Medicare, and Medicaid blind or disabled populations to determine the prevalence of MS in 6 states for 1991–1997. MS prevalence (per 100,000 population) was estimated to be 240 in the privately insured, 180 among Medicare age 65 and above, and 710 among Medicaid

**Table 2** Epidemiologic studies of multiple sclerosis (MS) prevalence in large regions of the United States using administrative health claims databases

| Reference: Authors (year) | Study years | Case definition | Population | National MS prevalence per 100,000 population (95% CI) | Number of persons with MS in the United States (95% CI) |
|---|---|---|---|---|---|
| **Pope et al. (2002)**[15] | 1994–1997 | One or more ICD-9 codes for MS (diagnosis code 340) | Nationally representative samples of administrative claims from 6 states (California, Colorado, Georgia, Michigan, Missouri, Tennessee) | | |
| | | | Privately insured (Medstat MarketScan database) | 240 per 100,000 (95% CI NR) (all ages) | NR |
| | | | Medicare, age 65 and above | 180 per 100,000 (95% CI NR) (age ≥65) | NR |
| | | | Medicaid disabled population, all ages | 710 per 100,00 (95% CI NR) (all ages) | NR |
| **Dilokthornsakul et al. (2016)**[16] | 2012 | At least 2 inpatient claims with ICD-9 code 340, at least 3 ICD-9 code 340 outpatient claims, or at least 1 MS-indicated DMT | Commercially insured population (Truven Market-Scan) | 149 per 100,000 (147.5–150.8) (all ages; age-adjusted to US 2012 population) | 403,630 (387,445–419,833) |

Abbreviations: CI = confidence interval; DMT = disease-modifying treatment; ICD-9 = International Classification of Diseases–9; NR = not reported.

disabled populations. These estimates are not strictly comparable due to the different age distributions and the expectation that MS prevalence would be much higher in the Medicaid disabled population. In addition, no algorithm was developed or used for MS cases. Dilokthornsakul et al.[16] estimated MS prevalence for 2008–2012 in the privately insured US population by applying a case definition algorithm using the PharMetrics Plus Health Plan Claims data, a large US private insurance claims database that contained claims data for approximately 42 million enrollees in 2011. The case definition algorithm required at least 2 inpatient claims with ICD-9 code 340, at least 3 ICD-9 code 340 outpatient claims, or at least one MS-specific disease-modifying treatment (DMT) claim. The estimated MS prevalence for 2012 was 149 per 100,000 (95% CI 147.6–150.9), and when extrapolated to the United States as a whole, yielded an estimate of 403,630 individual cases (95% CI 387,445–419,833).

## Methods for conducting a national study of MS prevalence in the current era

Before designing our study to estimate the US prevalence of MS, the working group members carried out a strengths, weaknesses, opportunities, and threats (SWOT) analysis for 3 approaches to estimate MS prevalence: (1) population-based MS registries (table 3); (2) national probability health surveys (table 4); and (3) analysis of AHC databases (table 5). Studies of each type were evaluated as potential sources for estimating the national prevalence of MS based on the populations studied.

### Population-based MS registries
Longstanding MS registries have been successfully used to estimate MS prevalence in 4 countries: Denmark,[17,18] Sweden,[19] Norway,[20] and Poland.[21] However, a similar US population-based MS registry does not exist. In the United States, there are several voluntary MS registries for conducting outcomes research; however, none of these registries attempts the near-complete case ascertainment in a defined geographic region that would enable the estimation of MS prevalence. The benefits and pitfalls of population-based disease registries were recently reviewed[22] and are summarized in table 3. A nationwide population-based registry would be costly to implement; however, it is possible that a geographically representative system of regional registries could be implemented, similar to the approach used for cancer surveillance.[23] In the United States, only one neurologic disease registry has been implemented. The National Amyotrophic Lateral Sclerosis Registry was established by a Congressional mandate and began estimating national amyotrophic lateral sclerosis prevalence in 2010.[24,25]

### National probability health surveys
The primary method used to estimate national MS prevalence in the past has been probability health surveys. However, almost none of the national probability surveys conducted by the US government are sufficiently large enough to contain enough people with MS to enable precise estimation of MS prevalence, much less prevalence within demographic or regional subgroups. Moreover, the national survey that once allowed an overall estimate of MS prevalence (NHIS) no longer includes questions about MS and other disabling conditions. National probability survey data rely on patient or

**Table 3** Strengths, weaknesses, opportunities, and threats (SWOT) analysis using population-based multiple sclerosis (MS) registries to estimate national MS prevalence in the US

| Strengths | Weaknesses | Opportunities | Threats |
|---|---|---|---|
| Multiple overlapping case ascertainment methods are used to yield a high degree of case ascertainment | Organizational demands make a nationwide registry infeasible in the United States, with rare exceptions (National Amyotrophic Lateral Sclerosis Registry[24,25]) | A systematic approach could be developed to extrapolate results from multiple regional registries to national MS prevalence; however, a geographically valid sample of select states would need to be selected to represent the nation | If registry relies on patient volunteers or requires consent, a large degree of nonresponse can seriously affect the validity of the registry |
| Data are rigorously collected for research purposes according to standard protocols | Very high cost to implement in population-based regions; possibly infeasible for nationwide implementation in the United States | For regions where registry participation is high, can obtain valid and precise estimates of MS prevalence | Confirming clinician diagnoses is costly and impractical for large numbers of patients |
| Some US registries rely in part on case ascertainment through neurologists, primary care physicians, and hospitals | In the United States, registries often also rely on MS diagnosis reported by patient or family member, which might be misreported; medical charts may be requested for verification of case definition but may be difficult to obtain for all patients in a population-based registry | | Need to collect patient identifiers to identify duplicates identified by more than one case ascertainment method (risks HIPAA noncompliance) |
| Data can be collected on race/ethnicity and other sociodemographic characteristics | | | |
| Registries often contain more clinical details (e.g., initial symptomatology, clinical course, disability status) than do other sources | | | |

Abbreviation: HIPAA = Health Insurance Portability and Accountability Act.
With the exception of studies conducted by Mayo Clinic researchers in Rochester, Minnesota,[17] population-based MS registries do not exist in the United States. However, there are several notable population-based MS registries in Europe that enable ongoing estimation of national MS incidence and prevalence, including in Denmark,[17,18] Sweden,[19] Norway,[20] and Poland.[21] Such studies often rely on case ascertainment through multiple data sources including neurologists, primary care physicians, and hospitals.

family member reports of physician-diagnosed MS,[10,12] risking the inclusion of false-positive MS cases or provisional cases that would not meet formal case definition criteria. Only one national survey, the Behavioral Risk Factor Surveillance Survey (BRFSS), which surveys more than 500,000 individuals each year, has a sample size that would potentially yield precise estimates of MS prevalence. Feasibility work would be needed to determine if this BRFSS could be used for this purpose.

## Administrative health claims data

In other countries with universally funded health care systems such as Canada and many European nations, AHC databases have been used to provide a population-based alternative to the traditional intensive methods of case ascertainment used in regional studies.[1] In the United States, AHC data are maintained by government (public) insurance programs, including Medicare and Medicaid, and by private (commercial) insurance providers. AHC databases are a tremendous resource for studying neurologic disease incidence and prevalence because of their rich prospective data on patients from defined populations and their use of the standardized international disease-coding framework (using the ICD system).[26] However, AHC databases are not created for research purposes, making it

necessary to demonstrate the validity and the accuracy of case ascertainment algorithm prior to using them for estimating disease prevalence. These challenges exist for estimating prevalence of all other common chronic neurologic conditions.[26] AHC data sources have potential strengths and limitations when applied for the purpose of estimating disease prevalence (table 5). The information from AHC databases begins with enrollment into health insurance plans and provides a standardized source of longitudinal patient health care utilization. These data contain near complete capture of demographic information (sex, date of birth, race/ethnicity, zip code, or state of residence) and health care utilization data, including diagnostic codes and pharmacy records. They can be used for research without specific consent because the databases are de-identified. Identifying persons with MS through pharmacy records is an excellent case ascertainment strategy given that many of the DMTs are prescribed only for MS, thus, the persons with MS using such medications are highly likely to be true MS cases. However, over 40% of persons with MS in the United States do not take DMTs, necessitating additional case-finding methods. Because clinical details are often not available, such as whether a diagnostic code was entered by a neurologist, physician, or nonphysician, most claims database studies have relied on case-finding algorithms that have not been judged against

**Table 4** Strengths, weaknesses, opportunities and threats (SWOT) analysis using national probability survey samples to estimate national multiple sclerosis (MS) prevalence in the US

| Strengths | Weaknesses | Opportunities | Threats |
|---|---|---|---|
| Can be obtained from existing government surveys | Dedicated MS-specific patient surveys are costly to carry out and infeasible for national scope | This method can be feasible if multiple years of survey data can be aggregated | Existing national probability samples usually comprise only segments of the US population (i.e., noninstitutionalized) |
| Sampling designs are rigorous (probability samples) of the US population, rigorously collected for research purposes according to standard protocols | Survey participation rates are at historically low levels | Successful application would require access to a very large survey sample such as the BRFSS conducted annually by the CDC; however, a geographically valid sample of select states would need to be selected to represent the nation | Survey questions for patients and family members have not been validated for accuracy against a gold standard (neurologist diagnosis) |
| Data are accessible and either cost-free or low-cost | Surveys rely on MS diagnosis reported by patient or a family member, which are susceptible to both underreporting and false-positive reports | If survey questions were validated against a gold standard (neurologist diagnosis), this method could be feasibly implemented in a subset of US states | Restrictions on number of questions limits collection of clinical details (e.g., clinical course, disability status) |
| Data are available on age, sex, race/ethnicity, broad geographic location, and other sociodemographic characteristics | Count of MS patients in single years of data are often too small for precise estimation | | |
| Some surveys collect self-reported data that can inform other analyses, e.g., prevalence of comorbidity, health care utilization and cost, DMT use, lifestyle risk factors | Overall prevalence might be estimable, but data are often too sparse for getting breakdowns by sex, age, race, or geographic region | | |

Abbreviations: BRFSS = Behavioral Risk Factor Surveillance Survey; CDC = Centers for Disease Control and Prevention; DMT = disease-modifying treatment. In the United States, national probability survey samples include the National Health Interview Survey, Medical Expenditure Panel Survey: Household Component, National Nursing Home Survey, and the BRFSS (a probability survey completed by ~500,000 Americans per year, and for which disease-specific questions can be added on a state-by-state basis).

case definitions or even neurologist diagnoses for accuracy. Two studies used a case definition that required only a single health care encounter for MS,[12,15] which is likely to be highly nonspecific, risking an unknown number of false-positives. The study by Dilokthornsakul et al.[16] that used private insurance claims data was strengthened by conducting sensitivity analyses that examined the performance of 2 alternative algorithms, one of which produced estimates very similar to the primary algorithm.

## Developing a new approach for estimating MS prevalence in the United States

Based on the SWOT analysis, and through discussion and consensus, we concluded that AHC data would provide the most cost-effective method to develop a national prevalence estimate for MS. Collectively, government and private health claims databases represent a very large proportion of the US population (84% in 2010),[28] which together enables precise estimates of MS prevalence. The data are sufficiently distributed geographically and socioeconomically to be representative of the entire US population. Figure 1 illustrates a new 4-step approach we devised for estimating the prevalence of MS for the United States as a whole using a segmental approach for each insurance source and then combining them into a national estimate.

Step 1: Identify datasets covering US publicly and privately insured populations: figure 2 illustrates the fragmented nature of insurance coverage of the US population, which does not enable a single health care source for estimating MS prevalence as is often available in Scandinavian countries with national unified health care systems. Nevertheless, AHC data are available for almost all of the insurance segments of the US population, allowing the estimation of MS prevalence separately for each insurance segment, which can then be aggregated into an overall national estimate of MS prevalence. Figure 2 shows US Census Current Population Survey data for 2008–2010, presented separately for individuals under age 65 vs those ages 65 and older because the sources of insurance coverage vary considerably for these 2 age groups. In 2010, for individuals below age 65, private insurance comprised the largest segment of insurance coverage, 65%, when employer-purchased and self-purchased private insurance were considered together. The next largest segment of insurance coverage in 2010 for individuals under age 65 is Medicaid (17%), a government insurance program that provides free medical insurance to low-income people. Smaller percentages of individuals under age 65 in 2010 were covered by a military source such as Department of Defense (DOD) or Veterans

**Table 5** Strengths, weaknesses, opportunities, and threats (SWOT) analysis using administrative health claims databases to estimate national multiple sclerosis (MS) prevalence in the US

| Strengths | Weaknesses | Opportunities | Threats |
|---|---|---|---|
| Can develop systematic approach that could be repeatedly utilized over time to obtain estimates now and in the future | Lack of universal health care in the United States, therefore not all patients are identifiable through administrative data | Can develop a systematic approach that can be repeatedly utilized over time to obtain estimates now and in the future | Lack of certainty whether case definitions developed in one dataset will have same performance characteristics in other datasets |
| Data are accessible and relatively low-cost | Must access multiple administrative datasets | A case definition can be validated in 2 or 3 different administrative datasets to show consistency of findings across datasets, then applied even in datasets where validation is not possible[4] | No easily accessible national source of data for uninsured or self-pay populations |
| Consistent capture of sex and date of birth supports calculation of age- and sex-specific prevalence estimates | Data are collected for health system management; therefore their validity for research must be carefully assessed | If a careful analytic approach is used, can minimize overlap between datasets and avoid duplicate counting | No easily accessible data for incarcerated or aboriginal populations |
| Data often include zip code, which can be geocoded to determine socioeconomic status and establish region of residence | Lack of identified data to enable identification of duplicates creates difficulty identifying the same individuals in overlapping datasets | | Would ideally include collection of patient identifiers to identify duplicates present in more than one database; however, nongovernment researchers are not allowed access to such identifiers |
| These databases often have additional information that could inform other analyses, e.g., prevalence of comorbidity, health care utilization, and disease-modifying treatment use | Some datasets lack information regarding race/ethnicity or have poor geographic resolution | | Representativeness: cannot access all administrative health databases in United States; must select few to estimate MS prevalence and extrapolate to all |
| | Issue of misdiagnosis as a source of error; lack of clinical detail makes it difficult to distinguish inaccurately coded diagnoses from clinician misdiagnosis | | |

In the United States, administrative health claims databases include Medicare (a government insurance program covering 93% of US residents age 65 and above), Medicaid (a government insurance program for low-income parents and children and disabled persons in the United States), the Minimum Data Set (administrative data on Medicare- or Medicaid-certified nursing facilities), the Veterans Health Administration (health care program for veterans of US wars), commercial health insurance claims data (for employer-paid or self-paid insurance), and health maintenance organization health records.
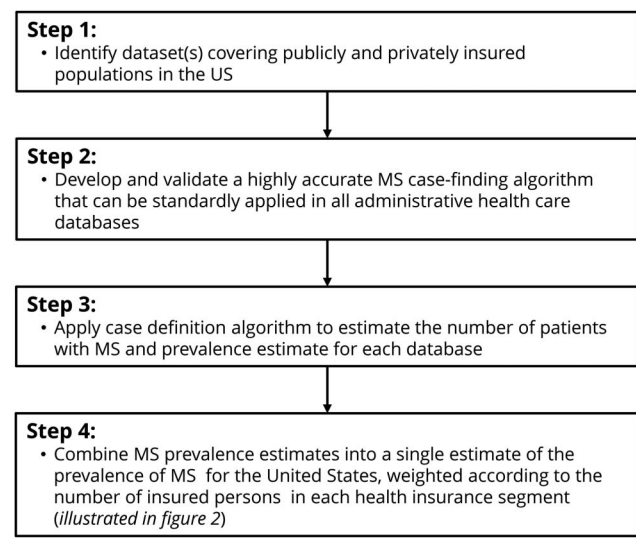
Health Administration (VHA) (4%), or by another government-provided insurance plan (Medicare, 3%). In 2010, 18% of individuals under age 65 had no insurance coverage. In 2010, for individuals ages 65 and above, government insurance provided by Medicare covered 93% of the population and was often supplemented by private coinsurance (33% employer-provided, 29% self-purchased). Medicaid, either alone or in combination with Medicare, covered 9% of those age 65 and above. In 2010, 8% of individuals age 65 and above were covered by a military source (DOD or VHA), and only 2% of the elderly were uninsured. Thus, despite the fragmentation of the US health care system, AHC databases for all of the insured segments of the US population are available and can be used to estimate MS prevalence in each dataset separately before combining the estimates into a whole (step 4).

Step 2: Develop and validate a highly accurate MS case-finding algorithm that can be standardly applied in all AHC databases. Our approach requires a highly accurate case definition to identify MS cases that performs consistently across AHC databases with differing characteristics so that it can be applied even in datasets where formal validation is not possible. Notably, no MS prevalence studies in the United States have used validated algorithms for identifying persons with MS, yet having a validated algorithm for this purpose is a necessary step for carrying out valid prevalence estimation.[26] Culpepper et al.[27] recently conducted a multi-site validation study to develop a valid and reliable algorithm for identifying MS cases using AHC data, with excellent sensitivity, specificity, and positive predictive value. For the first time ever, this provides a high-performing algorithm that performs consistently well across datasets and can be applied across all AHC databases. Similar methodology could be employed for other chronic neurologic diseases, using existing literature to provide a starting point for the development of such case definitions.[26]

Step 3: Apply the case definition algorithm to estimate the number of MS cases and prevalence estimates separately for each database: The third step of this approach is simply to use the validated algorithm (step 2) within each of the

**Figure 1** Stepwise process for determining the prevalence of multiple sclerosis (MS) in the United States

**Step 1:**
- Identify dataset(s) covering publicly and privately insured populations in the US

↓

**Step 2:**
- Develop and validate a highly accurate MS case-finding algorithm that can be standardly applied in all administrative health care databases

↓

**Step 3:**
- Apply case definition algorithm to estimate the number of patients with MS and prevalence estimate for each database

↓

**Step 4:**
- Combine MS prevalence estimates into a single estimate of the prevalence of MS for the United States, weighted according to the number of insured persons in each health insurance segment *(illustrated in figure 2)*

databases (step 1) to obtain valid MS prevalence estimates (overall and specific with respect to age, sex, race/ethnicity, and geographic location).

Step 4: Combine MS prevalence estimates into a single estimate of the prevalence of MS for the United States, weighted according to the number of insured persons in each health insurance segment: after obtaining MS prevalence estimates separately for each of the US health insurance segments, they can be combined into a single national prevalence estimate. Information from the Current Population Survey can be used to obtain the necessary weights by determining the number of US residents with each type of insurance coverage, stratified by sex, age group, race, and geographic region. The total estimated number of MS cases for each age, sex, race, and geographic region–specific category can then be summed to estimate the total number of MS cases in the United States. For example, MS prevalence for the private insurance segment can be estimated by accessing AHC data for Optum (Eden Prairie, MN) and Truven Health (Ann Arbor, MI) which, in 2010, together comprise approximately 68 million individuals, or 35% of the 2010 privately insured US population. MS prevalence in the government-insured segment of the population can be directly and completely estimated for 100% of people covered by using administrative claims data for Medicaid and Medicare. MS prevalence in the military population, which is predominantly male, can be estimated used the Veterans Health Administration database. A limitation of this method is that no claims data exist for persons without health insurance, estimated at 16.3% in 2010[28]; however, for this uninsured group, MS prevalence estimates from the Medicaid health care claims data could be applied because the 2 groups have similar sociodemographic characteristics.

Our new approach cannot be applied to estimate the incidence of MS in the United States. Incidence—the number

of new MS cases per year in a population—is an important measure of disease frequency. However, it can only be derived for countries that have single uniform health care systems where longitudinal AHC records for each person enable relatively precise indicators of the onset of new MS diagnosis and the exclusion of existing prevalent cases. The lack of a single payer health care system or national tracking system in the United States does not allow reliable estimates of MS incidence because some individuals may change health care coverage over time, and it is not possible to link patient records across insurance sources to provide a single longitudinal record within which newly diagnosed persons with MS may be ascertained. This creates short periods of observation, which do not allow accurate estimates of incidence for conditions such as MS and cancer due to the inability to definitively determine which individuals are truly newly diagnosed.[29]

## Discussion

We have proposed a new approach for applying validated case-finding algorithms to AHC data from both government and private sources to produce robust estimates of MS prevalence in the United States. In a companion article, we applied this approach to estimate the total number of MS cases in the United States.[30] This approach could be applied to estimate prevalence of other neurologic disorders and chronic conditions that face similar methodologic challenges such as Parkinson disease, dementia, and epilepsy.

There are several reasons why our new approach is a valuable methodologic development to estimate the national MS prevalence. First, this paradigm offers the potential to use existing AHC data for case ascertainment across the United States without requiring the cost and effort involved in regional MS prevalence studies, which typically employ intensive case ascertainment methods. The data are sufficiently distributed, both geographically and socioeconomically, to be representative of the entire US population, and do not require consent because the databases are de-identified. Second, our method is efficient and cost-effective, and can be applied repeatedly with datasets that represent a large proportion of the country, enabling updates to population-based estimates of MS prevalence over time. Third, applying this approach with data that represent the entire United States allows us to obtain more stable and precise prevalence estimates, and to fill in knowledge gaps regarding the prevalence of MS in many regions for which we have never had estimates. Once AHC data are able to accurately and completely capture race and ethnicity, this approach could be applied to racial and ethnic minorities that have received little study in the past. Fourth, the use of a highly accurate case-finding algorithm[4] ensures that there will be greater comparability between estimates derived from these sources in the future. By adopting a standard paradigm for ascertaining MS cases and applying validated case definition

**Figure 2** Source of health insurance by age (estimates from the current population survey, 2008–2010)
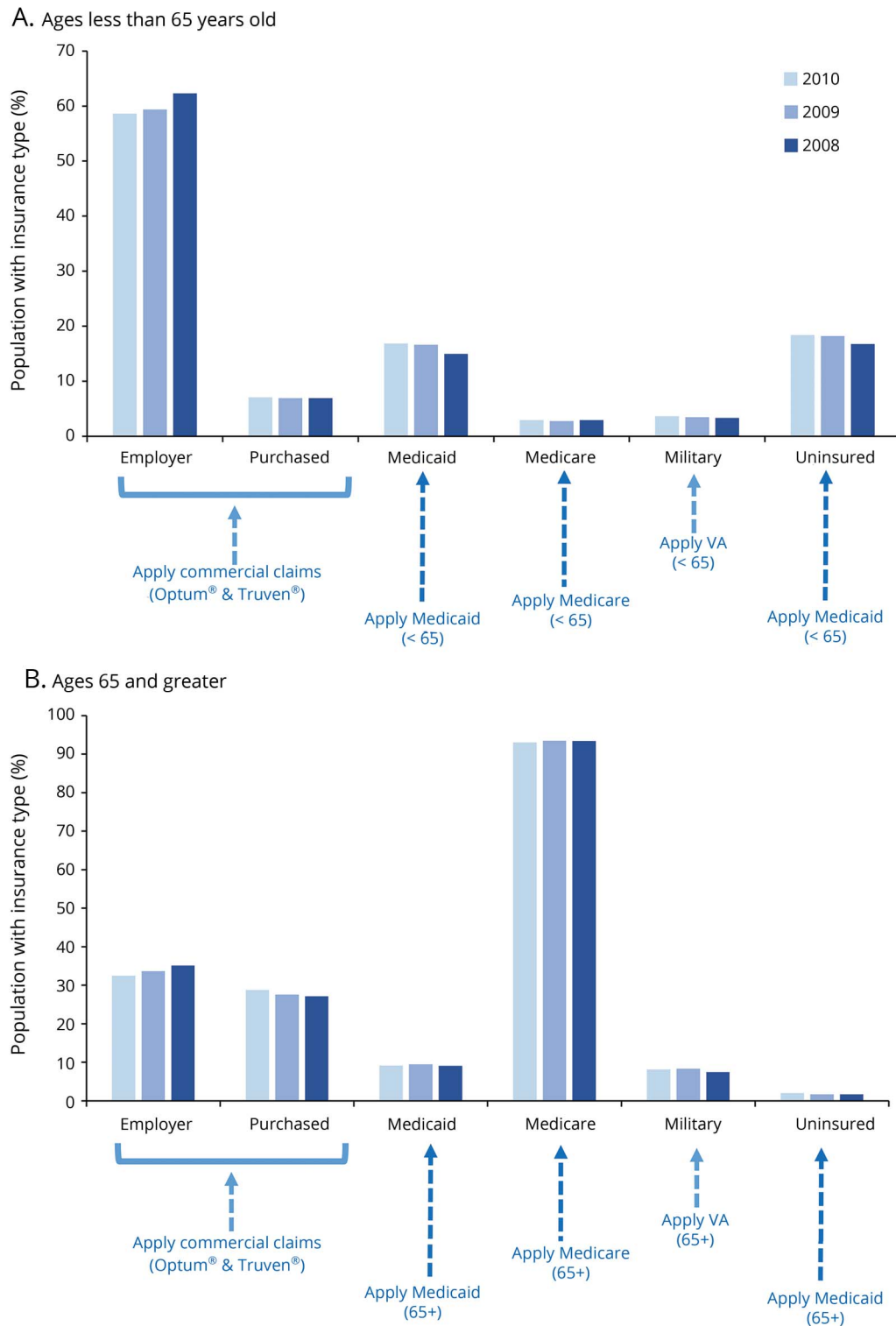


(A) Population younger than 65 years. (B) Population 65 years of age and older.

criteria, we will be better able to monitor and compare spatial differences in MS prevalence and to examine changes in disease prevalence over time without the variation introduced by differing methodologic approaches.

Despite many strengths, our new approach is limited in that the case ascertainment methods do not cover all segments of the US population. It excludes native American residents who choose to receive care through the Indian Health Service

(0.6% of the US population in 2010)[31] and US residents who are incarcerated (approximately 0.5% of the US population in 2010).[32] If health care databases existed for these populations, we could similarly apply an MS algorithm to estimate MS prevalence and incorporate them into the overall national estimate. Our approach also cannot distinguish between different racial and ethnic groups; thus we cannot address to what extent variations in prevalence by geography or other factors may be better explained by racial/ethnic differences in prevalence.[33]

Our work has important implications for the new era of MS epidemiologic research that will be possible through the 21st Century Cures Act legislation, enacted in December 2016, which authorized the establishment of a federally funded neurologic disorders surveillance system.[34] A national registry for MS could be implemented similar to the National Amyotrophic Lateral Sclerosis Registry, with AHC data as a primary method for case ascertainment and a web portal available for all persons with MS to register.[24,25,35] Without access to a national registry in the intermediate term, we propose that our approach will deliver the most current and scientifically sound US estimate of the number of persons with MS. By addressing methodologic challenges and proposing a new paradigm for measuring the prevalence of MS in the United States, we hope that our work will benefit scientists who study all chronic conditions for which national prevalence estimates do not currently exist.

### Disclosure
L. Nelson receives grants from the Centers for Disease Control and Prevention, NIH, and National Multiple Sclerosis Society, and contracts from the Agency for Toxic Substances and Diseases Registry. She receives compensation for serving as a consultant to Acumen, Inc. and is on the Data Monitoring Committee of Neuropace. M. Wallin has served on data safety monitoring boards for the National Institutes of Neurological Disease and Stroke–NIH, has been an associate editor for the *Encyclopedia of the Neurological Sciences*, and received funding support from the National Multiple Sclerosis Society and the Department of Veterans Affairs Merit Review Research Program. R. Marrie is supported by the Waugh Family Chair in Multiple Sclerosis, receives research funding from Canadian Institutes of Health Research, Research Manitoba, Multiple Sclerosis Society of Canada, Multiple Sclerosis Scientific Foundation, National Multiple Sclerosis Society, and Rx & D Health Research Foundation, and has conducted clinical trials funded by Sanofi-Aventis. She also serves on the Editorial Board of *Neurology*®. W. Culpepper has additional research funding from the National Multiple Sclerosis Society, receives support from the VHA MS Center of Excellence, and is a member of the National Multiple Sclerosis Society Health Care Delivery and Policy Research study section. A. Langer-Gould is site principal investigator for 2 industry-sponsored phase 3 clinical trials (Biogen Idec; Hoffman-LaRoche) and was site PI for one industry-sponsored observation study (Biogen Idec). She receives grant support from the NIH, National Institute of Neurological Disorders and Stroke, Patient-Centered Outcomes Research Institute, and the National Multiple Sclerosis Society. J. Campbell has consultancy or research grants over the last 5 years with Agency for Healthcare Research and Quality, ALSAM Foundation, Amgen, AstraZeneca, Bayer, Biogen Idec, Boehringer Ingelheim, Centers for Disease Control and Prevention, Colorado Medicaid, Enterprise Community Partners Inc., Institute for Clinical and Economic Review, Mallinckrodt, NIH, National Multiple Sclerosis Society, Kaiser Permanente, PhRMA Foundation, Teva, Research in Real Life Ltd., Respiratory Effectiveness Group, and Zogenix Inc. S. Buka receives research funding from the NIH and the National Multiple Sclerosis Society. H. Tremlett is the Canada Research Chair for Neuroepidemiology and Multiple Sclerosis. She receives research support from the National Multiple Sclerosis Society, the Canadian Institutes of Health Research, the Multiple Sclerosis Society of Canada, and the Multiple Sclerosis Scientific Research Foundation. In addition, in the last 5 years she has received research support from the Multiple Sclerosis Society of Canada, the Michael Smith Foundation for Health Research, and the UK MS Trust, and speaker honoraria and/or travel expenses to attend conferences. G. Cutter is a member of Data and Safety Monitoring Boards for AMO Pharmaceuticals, Apotek, Horizon Pharmaceuticals, Modigenetech/Prolor, Merck, Merck/Pfizer, Opko Biologics, Neurim, Sanofi-Aventis, Reata Pharmaceuticals, Receptos/Celgene, Teva pharmaceuticals, NHLBI (Protocol Review Committee), and NICHD (OPRU oversight committee), and is on the Consulting or Advisory Boards for Atara Biotherapeutics, Argenix, Bioeq GmbH, Consortium of MS Centers (grant), Genzyme, Genentech, Innate Therapeutics, Klein-Buendel Incorporated, Medimmune, Medday, Novartis, Opexa Therapeutics, Roche, Savara Inc., Somahlution, Teva Pharmaceuticals, Transparency Life Sciences, and TG Therapeutics. W. Kaye receives funding from the Agency for Toxic Substances and Disease Registry, the National Multiple Sclerosis Society, and the Association for the Accreditation of Human Research Protection Programs. L. Wagner receives funding from the

## Appendix 1 Authors

| Name | Location | Role | Contribution |
|---|---|---|---|
| Lorene M. Nelson, MS, PhD | Stanford University, CA | Author | Study concept and design, preparation of draft manuscript, revision of the manuscript for content, approval of final manuscript |
| Mitchell T. Wallin, MD, MPH | Veterans Health Administration; Georgetown University, Washington, DC | Author | Study concept and design, critical review, revision of the manuscript for content, approval of final manuscript |
| Ruth Ann Marrie, MD, PhD | University of Manitoba, Winnipeg, Canada | Author | Study concept and design, critical review, revision of the manuscript for content, approval of final manuscript |
| William J. Culpepper, PhD, MA | Veterans Health Administration, Washington, DC; University of Maryland, Baltimore | Author | Study concept and design, critical review, revision of the manuscript for content, approval of final manuscript |
| Annette Langer-Gould, MD, PhD | Kaiser Permanente Southern California, Los Angeles | Author | Study concept and design, critical review, revision of the manuscript for content, approval of final manuscript |
| Jonathan Campbell, PhD | University of Denver, CO | Author | Study concept and design, critical review of the manuscript for content, approval of final manuscript |
| Stephen Buka, PhD | Brown University, Providence, RI | Author | Study concept and design, critical review of the manuscript for content, approval of final manuscript |
| Helen Tremlett, PhD | University of British Columbia, Vancouver, Canada | Author | Study concept and design, critical review of the manuscript for content, approval of final manuscript |
| Gary Cutter, PhD | University of Alabama, Birmingham | Author | Study concept and design, critical review of the manuscript for content |

## Appendix 1 (continued)

| Name | Location | Role | Contribution |
|---|---|---|---|
| Wendy Kaye, PhD | McKing Consulting Corp., Atlanta, GA | Author | Study concept and design, critical review of the manuscript for content, approval of final manuscript |
| Laurie Wagner, MPH | McKing Consulting Corp., Atlanta, GA | Author | Study concept and design, critical review of the manuscript for content, approval of final manuscript |
| Nicholas G. LaRocca, PhD | National Multiple Sclerosis Society, New York, NY | Author | Study concept and design, critical review of the manuscript for content, approval of final manuscript |

## Appendix 2 Coinvestigators and members of the US Multiple Sclerosis Prevalence Workgroup

| Name | Location | Role | Contribution |
|---|---|---|---|
| Albert Lo, MD, PhD | Brown University, Providence, RI | Coinvestigator | Study concept and design |
| Robert McBurney, PhD | Accelerated Cure Project, Boston, MA | Coinvestigator | Study concept and design |
| Oleg Muravov, PhD | Agency for Toxic Substances and Disease Registry Division of Health Studies, Atlanta, GA | Coinvestigator | Study concept and design |
| Bari Talente, Esq | National Multiple Sclerosis Society, Washington, DC | Coinvestigator | Study concept and design |
| Leslie Ritter | National Multiple Sclerosis Society, Washington, DC | Coinvestigator | Study concept and design |

## References

1. Wallin MT, Kurtzke JF. Neuroepidemiology. In: Bradley WG, Daroff RB, Fenichel GM, Jankovic J, editors. Neurology in Clinical Practice. Philadelphia: Butterworth-Heinemann; 2008:762–780.
2. Browne P, Chandraratna D, Angood C, et al. Atlas of multiple sclerosis 2013: a growing global problem with widespread inequity. Neurology 2014;83:1022–1024.
3. Mayr WT, Pittock SJ, McClelland RL, Jorgensen NW, Noseworthy JH, Rodriguez M. Incidence and prevalence of multiple sclerosis in Olmsted County, Minnesota, 1985-2000. Neurology 2003;61:1373–1377.
4. Evans C, Beland SG, Kulaga S, et al. Incidence and prevalence of multiple sclerosis in the Americas: a systematic review. Neuroepidemiology 2013;40:195–210.
5. Noonan CW, Williamson DM, Henry JP, et al. The prevalence of multiple sclerosis in 3 US communities. Prev Chronic Dis 2010;7:A1.
6. Polman CH, Reingold SC, Banwell B, et al. Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria. Ann Neurol 2011;69:292–302.
7. Baum HM, Rothschild BB. The incidence and prevalence of reported multiple sclerosis. Ann Neurol 1981;10:420–428.
8. Anderson DW, Ellenberg JH, Leventhal CM, Reingold SC, Rodriguez M, Silberberg DH. Revised estimate of the prevalence of multiple sclerosis in the United States. Ann Neurol 1992;31:333–336.
9. Nelson LM, Hamman RF, Thompson DS, et al. Higher than expected prevalence of multiple sclerosis in Northern Colorado: dependence on methodological issues. Neuroepidemiology 1986;5:17–28.
10. Collins JG. Prevalence of selected chronic conditions: United States, 1990-1992. Vital Health Stat 10 1997;194:1–89.

11. Noonan CW, Kathman SJ, White MC. Prevalence estimates for MS in the US and evidence of an increasing trend in women. Neurology 2002;58:136–138.

12. Campbell JD, Ghushchyan V, Brett McQueen R, et al. Burden of multiple sclerosis on direct, indirect costs and quality of life: national US estimates. Mult Scler Relat Disord 2014;3:227–236.

13. Buchanan RJ, Radin D, Huang C, Zhu L. Caregiver perceptions associated with risk of nursing home admission for people with multiple sclerosis. Disabil Health J 2010;3: 117–124.

14. Northrop DE, Frankel D, editors. Nursing Home Care of Individuals with Multiple Sclerosis: Guidelines and Recommendations for Quality Care: Monograph. New York: National Multiple Sclerosis Society Professional Resource Center. Available at: nationalmssociety.org/NationalMSSociety/media/MSNationalFiles/Brochures/Brochure-Nursing-Home-Care-of-Individuals-with-MS.pdf. Accessed July 1, 2018.

15. Pope GC, Urato CJ, Kulas ED, Kronick R, Gilmer T. Prevalence, expenditures, utilization, and payment for persons with MS in insured populations. Neurology 2002;58:37–43.

16. Dilokthornsakul P, Valuck RJ, Nair KV, Corboy JR, Allen RR, Campbell JD. Multiple sclerosis prevalence in the United States commercially insured population. Neurology 2016;86:1014–1021.

17. Koch-Henriksen N, Magyari M, Laursen B. Registers of multiple sclerosis in Denmark. Acta Neurol Scand 2015;132:4–10.

18. Koch-Henriksen N, Sorensen PS. The changing demographic pattern of multiple sclerosis epidemiology. Lancet Neurol 2010;9:520–532.

19. Hillert J, Stawiarz L. The Swedish MS registry: clinical support tool and scientific resource. Acta Neurol Scand 2015;132:11–19.

20. Myhr KM, Grytten N, Aarseth JH, Nyland H. The Norwegian multiple sclerosis National Competence Centre and National Multiple Sclerosis Registry: a resource for clinical practice and research. Acta Neurol Scand Suppl 2006;183:37–40.

21. Brola W, Sobolewski P, Fudala M, et al. Self-reported quality of life in multiple sclerosis patients: preliminary results based on the Polish MS Registry. Patient Prefer Adherence 2016;10:1647–1656.

22. Rooney JPK, Brayne C, Tobin K, Logroscino G, Glymour MM, Hardiman O. Benefits, pitfalls, and future design of population-based registers in neurodegenerative disease. Neurology 2017;88:2321–2329.

23. National Cancer Institute (NCI), National Institutes of Health, US Department of Health and Human Services. Surveillance, Epidemiology and End Results (SEER) Program. Available at: seer.cancer.gov/registries/. Accessed July 1, 2018.

24. Mehta P, Kaye W, Bryan L, et al. Prevalence of amyotrophic lateral sclerosis: United States, 2012-2013. MMWR Surveill Summ 2016;65:1–12.

25. Mehta P, Antao V, Kaye W, et al. Prevalence of amyotrophic lateral sclerosis: United States, 2010-2011. MMWR Suppl 2014;63:1–14.

26. St Germaine-Smith C, Metcalfe A, Pringsheim T, et al. Recommendations for optimal ICD codes to study neurologic conditions: a systematic review. Neurology 2012;79:1049–1055.

27. Culpepper JW, Marrie RA, Langer-Gould A, et al. Validation of an algorithm for identifying MS cases across administrative health claims datasets. Neurology (in press 2019).

28. US Census Bureau. Income, Poverty, and Health Insurance Coverage in the United States: 2010. Washington, DC: US Department of Commerce, Economics and Statistics Administration; 2011.

29. Griffiths RI, O'Malley CD, Herbert RJ, Danese MD. Misclassification of incident conditions using claims data: impact of varying the period used to exclude pre-existing disease. BMC Med Res Methodol 2013;13:32.

30. Wallin MT, Culpepper JW, Campbell J, et al. The prevalence of MS in the United States: a population-based estimate using health claims data. Neurology (in press 2019).

31. US Department of Health and Human Services Indian Health Service. Trends in Indian Health, 2014 edition. Office of Public Health Support, Division of Program Statistics. Available at: ihs.gov/dps/includes/themes/newihstheme/display_objects/documents/Trends2014Book508.pdf. Accessed July 1, 2018.

32. US Department of Justice, Bureau of Justice Statistics. Prisoners in 2010. 2011. Available at: bjs.gov/content/pub/pdf/p10.pdf. Accessed July 1, 2018.

33. Langer-Gould A, Brara SM, Beaber BE, Zhang JL. Incidence of multiple sclerosis in multiple racial and ethnic groups. Neurology 2013;80:1734–1739.

34. 21st Century Cures Act, Enacted by the 114th United States Congress in December 2016. An Act to Accelerate the Discovery, Development, and Delivery of 21st Century Cures, and for Other Purposes. Public Law 114–255. Available at: congress.gov/bill/114th-congress/house-bill/6. Accessed July 1, 2018.

35. Horton DK, Mehta P, Antao VC. Quantifying a nonnotifiable disease in the United States: the National Amyotrophic Lateral Sclerosis Registry model. JAMA 2014;312:1097–1098.